Author:
Juvinao-Quintero, Diana

Title:
Appraising the causal relationship between DNA methylation and type 2 diabetes

Appraising the causal relationship between DNA methylation and type 2 diabetes

Diana Lizeth Juvinao-Quintero

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Bristol Medical School

MRC Integrative Epidemiology Unit
Department of Population Health Sciences
Bristol Medical School
University of Bristol
UK

May 2019

Words: 103,149

Abstract

Type 2 diabetes (T2D) is a multifactorial disease with a range of genetic and environmental risk factors. Increasing evidence shows that DNA methylation is likely to play a role in the aetiology of T2D. In this thesis, I investigated the association of DNA methylation with T2D focusing on middle-age adults of European background, and ascertained causality of this association by using Mendelian randomization.

I have used genome-wide DNA methylation data to identify markers associated with prevalent T2D in four European cohorts, combining results across studies via meta-analysis to increase power. This approach allowed the confirmation of well-known markers mapping to the genes *TXNIP* and *ABCG1*, and the identification of novel markers which are good candidates for validation. I also investigated DNA methylation in relation to different glycaemic traits using disease-free participants. Results suggested that variation in methylation preceding disease occurrence can indicate early stage of disease. Various *in silico* explorations allowed me to explore the potential mediating role of DNA methylation in the genetics of T2D and gene expression, biological pathways enriched for differentially-methylated sites, and the relevance of blood as a source of methylation markers for T2D. Genetic variants associated with T2D were used to determine causality of signals detected observationally using a bidirectional Mendelian randomization. Results revealed that methylation can be both, a consequence and a cause of T2D.

The findings of this thesis suggest that DNA methylation might be an important factor in the aetiology of T2D. However, further studies are needed to confirm novel signals, and to assess their generalizability in other populations. Results of the causal analysis should be reinforced by including larger datasets, especially when DNA methylation is regarded as the outcome.

To my grandma Hilda for being my inspiration in life

Acknowledgements

The completion of this work was made possible due to the following people and organizations, to whom I express my most sincere gratitude.

I want to thank the Government of Colombia and the Colciencias Scholarship for supporting the completion of my PhD.

I am profoundly grateful to my supervisors Prof. Caroline Relton, Dr Hannah Elliott and Dr Gemma Sharp for their constant support and patience in discussing aspects of my research. I also want to thank them, particularly Caroline, for advancing my professional development.

Especial thanks to the participants of the studies included in my work, not least the Avon Longitudinal Study of Parents and Children, Lothian Birth Cohort (1936), the Rotterdam Studies, the Cooperative Health Research in the Region of Augsburg and the Southall & Brent Revisited study. I want to thank the external cohorts invited to this study directly and the researchers Riccardo Marioni, Carolina Ochoa-Rosales, Stefan Brandmaier and Therese Tilling who conducted important analyses and provided the crucial access to data.

My thanks to Dr Santiago Rodriguez for his insightful feedback during my annual reviews and to Dr Josine Min, Dr Matthew Suderman and Dr Gibran Hemani whose advice was so valuable when conducting the genetic, epigenetic, and causal analyses aspects of my work.

I also want to thank my PhD colleagues and post-docs who participated in the weekly meetings in epigenetics for their sharing of research and the resulting discussions that provided many new ideas. To the Bristol Medical School for fostering a leading centre for research through seminars, special lectures and intensive courses that enriched my knowledge throughout the course of my PhD.

In addition, I want to thank the administrative department of the Bristol Medical School, particularly Sharen O'Keefe and Anne Walsh for assisting me with important paperwork and the financial aspects of my PhD.

I am especially grateful to my friends and colleagues Dr Carolina Bonilla, Dr Cilia Mejia, Dr Cynthia Ochieng, Yenenesh Kelile, Dr Ana Goncalves and Dr Carolina Borges, whose emotional and

professional support have helped me to complete my thesis particularly during the last months of the writing process.

Outside the work environment, I want to give immense thanks to my family for their persistent emotional support and guidance from afar, which allowed me to accomplish my goals even though any adversity. I will be forever thankful to my parents Maria and Blasco for being important role models in my life and true examples of perseverance, and to my siblings Maria and Felipe who fill my life with joy.

I also owe huge thanks to Chris, who has been my company for the last three years, during which time he has encouraged me to give my best and to dedicate to my work, thank you for your patience and support.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ……. DATE:……

Table of Contents

List of Tables

List of Figures

Abbreviations

ADA: American Diabetes Association

ALSPAC: Avon Longitudinal Study of Parents and Children

ARIES: Accessible Resource for Integrated Epigenomic Studies

CGI: CpG island

CRP: C-reactive Protein

CVD: Cardiovascular Disease

DIAGRAM: Diabetes Genetics Replication and Meta-Analysis

DMP: Differentially Methylated Positions

DMR: Differentially Methylated Region

EGP: Endogenous Glucose Production

eQTL: Expression Quantitative Trait Loci

eQTM: Expression Quantitative Trait Methylation

EWAS: Epigenome-Wide Association Studies

FDR: False Discovery Rate

FG: Fasting Glucose

GDM: Gestational Diabetes Mellitus

GO: Gene Ontology

GWAS: Genome-wide Association Study

HbA1c: Glycated Haemoglobin A1c

2-h Gluc/2-h PG/ OGTT: Two-Hours Post-Load Glucose Test or Oral Glucose Tolerance Test

HOMA-B: homoeostasis model assessment of β-cell disfunction

HOMA-IR: homoeostasis model assessment of insulin resistance

IFG: Impaired Fasting Glucose

IGT: Impaired Glucose Tolerance

IQR: Interquartile Range

IV: Instrumental Variable

KEGG: Kyoto Encyclopaedia of Genes and Genomes

KORA: Cooperative Health Research in the Region of Augsburg

LBC1936: Lothian Birth Cohort (1936)

LD: Linkage Disequilibrium

LRT: Likelihood-Ratio Test

MAF: Minor Allele Frequency

MAGIC: Meta-Analyses of Glucose and Insulin-related traits Consortium

meQTL: Methylation Quantitative Trait Loci

MR: Mendelian randomization

OR: Odds Ratio

PRS: polygenic risk score

QC: Quality Control

RSIII-1/RS-Bios: first follow-up of the Rotterdam Study Cohort-III and Rotterdam-Bios.

SABRE: Southall and Brent Revisited study

SAT: Subcutaneous Adipose Tissue

SES: Socioeconomic status

2SMR: Two-Sample Mendelian Randomization

SNP: Single Nucleotide Polymorphisms

SSMR: Single Sample Mendelian Randomization

SV(A): Surrogate Variable (Analysis)

T2D: Type 2 Diabetes

TSS: Transcription Start Site

UTR: Untranslated Region

VAT: Visceral Adipose Tissue

WHO: World Health Organization

Chapter 1 Introduction

1.1 Overview

1.1.1 What is Type 2 Diabetes?

Type 2 Diabetes (T2D) is a metabolic disorder characterized by hyperglycaemia, β-cell disfunction, and the abnormal metabolism of lipids in response to different levels of overnutrition, inactivity, obesity, and insulin resistance[1, 2]. In addition, T2D is considered an inflammatory condition since the metabolic stress caused by overnutrition can lead to a low-grade inflammatory response[3]. There are multiple genetic and environmental risk factors playing a role in disease onset, and to date, growing evidence indicates that epigenetic factors are also implicated in the aetiology of T2D, acting at the interface between the environment and co-ordinated transcriptional control[4].

T2D is a phenotypically heterogeneous disease, and even though obesity is regarded as the strongest risk factor for T2D, not all overweight (BMI >30 kg/m^2) people undergoes T2D, and there are also circumstances in which lean people (BMI <25 kg/m^2) are at risk of the disease. For instance, Perry *et al.*[5] demonstrated that by stratifying T2D cases between lean and obese participants, lean individuals had a stronger genetic predisposition to T2D reflected in larger odds ratios when measuring the effect size of 36 well-established genetic variants for T2D individually, and in a genetic risk score. The strongest variant in association with T2D in lean participants of European descent was reported in the *LAMA1* locus, while in obese cases the strongest variant was identified in the *HMG20A* locus[5]. In conclusion, Perry and colleagues demonstrated that lean cases of T2D were more enriched in T2D risk loci compared to obese cases, in whom the development of T2D was favoured by the physiological strains imposed by obesity and insulin resistance[5].

Non-insulin dependent diabetes mellitus can also develop in response to autosomal dominant mutations in the absence of obesity in young adults[6], while normoglycemic obese people can have normal insulin secretion even though obesity is a major risk factor for insulin resistance[6]. Since multiple risk factors can trigger the onset of T2D (see section 1.2.4), it is important to recognize that T2D is essentially a failure of pancreatic β-cells to compensate for the increasing degree of insulin resistance (i.e. abnormal biological action of insulin). Failure of the β-cell to secrete high levels of insulin to cope with the effects of reduced insulin sensitivity, can lead over time to impaired glucose tolerance with mild increase in postprandial glucose levels, and to diabetes with persistent hyperglycaemia[6]. Factors associated with insulin resistance are obesity, hypertension,

hyperlipidaemia, or molecular defects that affect the expression of the receptor of insulin, and the underlying signal-transduction pathways[6].

1.1.2 Why is it important to study T2D?

T2D is a chronic condition affecting 8.3% of the adult population worldwide, and it is one of the most common non-communicable diseases of current times[3, 7], accounting for approximately 90% of the cases of diabetes[1]. The importance of studying T2D is due to its high prevalence, the systemic damage that it causes to many organs in the body, the economic burden impinged on the health system through direct medical costs[7, 8], and the indirect economic impact to the society through loss of workforce[1, 7]. Only in the UK, it was estimated that costs associated directly or indirectly with T2D were up to £21 billion between 2010-2011[9], and they are expected to rise to £35.6 billion by 2035-2036[9].

1.1.3 Pathophysiology of T2D

Normal glucose homeostasis

In the fasting state, a balance is maintained between the endogenous glucose production (EGP) in the liver, and the use of EGP by glucose-independent organs as the brain[1]. EGP avoids hypoglycaemia during the fasting state, and its usage is prioritized in the brain relative to insulin-dependent organs, which are in turn fed by non-glucose nutrients[1]. Also, in the fasting state, levels of glucagon are higher than those of insulin in pancreatic islets. By contrast, after the intake of nutrients, levels of glucose in blood rise after being absorbed in the gut, stimulating the production of insulin in β-cells, and downregulating the production of glucagon in α-cells of the pancreas. In addition, EGP in the liver is inhibited by insulin in response to increasing levels of glucose in blood, and the supply of glucose to glucose-dependent tissues such as the skeletal muscle, adipose tissue and the heart, is consequently activated[1]. Furthermore, neurohormonal response to the intake of nutrients include the production of GLP-1 (glucagon-like peptide 1), which activates glucose-stimulated secretion of insulin, and the inhibition of glucagon[1].

Abnormal glucose homeostasis

In the presence of chronic overnutrition, people with a genetic predisposition to T2D experience an abnormal storage of fat in the form of visceral fat and ectopic fat in important organs such as the liver, pancreas, heart, and skeletal muscle, leading over time to tissue damage[1] (Figure 1-1). Furthermore, the adverse metabolic conditions in T2D trigger the production of cytokines in plasma, which contributes to the aggravation of insulin-dependent tissues[1, 2]. Even though obesity is a major risk factor for T2D, not all overweight people develop T2D, and this is due to compensatory

mechanisms that allow overnourished diabetes-resistant people to maintain relatively normal levels of nutrients in blood[1]. Some of these adaptive mechanisms include the expansion of subcutaneous adipose tissue (SAT) over the expansion of visceral adipose tissue (VAT) and ectopic fat in important organs, and the maintenance of adequate levels of insulin secretion in β-cells[1] (Figure 1-1). These adaptive mechanisms prevent diabetes-resistant people from experiencing tissue damage in the long-term, and from developing insulin resistance.

1.1.4 T2D aetiology

Overt T2D can arise in response to genetic predisposition or environmental triggers (see section 1.2.4 "Major risk factors"), or after gestational diabetes, or in response to treatment with certain medications (i.e. steroids or thiazides) that impair insulin secretion or action, and most rarely, due to damage of the pancreas (i.e. infection, pancreatitis, trauma, pancreatic cancer, cystic fibrosis, haemochromatosis, exposure to toxins, etc) and endocrinopathies (i.e. excess of hormones that are antagonists to insulin action)[10]. T2D can also arise in association with autoimmune disorders (i.e. stiff-man syndrome, systemic lupus erythematosus), and various genetic syndromes (i.e. Down syndrome, Klinefelter, Turner and Wolfram's syndrome)[10]. Thus, it is important that the clinician and the patient understand the pathogenesis leading to hyperglycaemia and T2D to treat it accordingly[10].

Overall, mechanisms leading to T2D in diabetes-susceptible people include: impaired compensatory response of β-cells to overnutrition, increased secretion of glucagon and endogenous glucose production even in the nourished state, gradual loss of β-cell mass, hyperglycaemia, disposal of fat in important insulin-dependent tissues leading to tissue damage, production of cytokines in response to metabolic stress, inflammation of adipose tissue, insulin resistance in peripheral tissues and β-cell disfunction[1, 2].

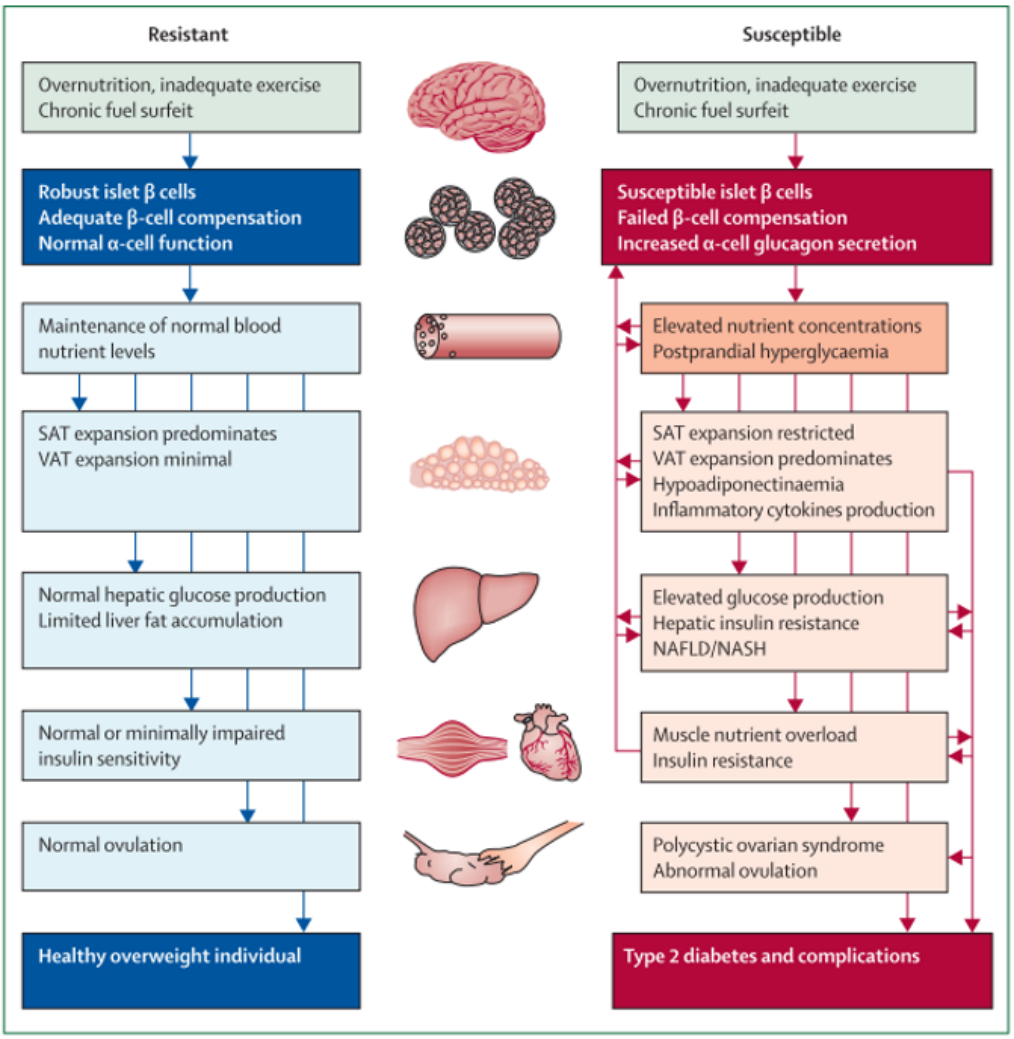


Figure 1-1 Pathways to T2D. The flow-diagram on the left describes the compensatory mechanisms developed by diabetes-resistant obese people to cope with the excess of nutrients in the body, and to the right, are the pathological conditions characteristic of diabetes-susceptible people. Diagram taken with permission from Nolan et al. 2011. Type 2 diabetes across generations: from pathophysiology to prevention and management. Lancet; 378:169-61.[1]

1.1.5 Clinical diagnosis

Symptoms characteristic of T2D are less pronounced than in other forms of diabetes, and they tend to be detected at a later stage regularly accompanied by secondary complications[2]. Symptoms of T2D include polyuria, polydipsia, weight loss with polyphagia, and blurred vision[10]. Other symptoms related to chronic hyperglycaemia are growth impairment and susceptibility to infections, while more acute and life-threatening symptoms include ketoacidosis and nonketotic hyperosmolar syndrome[10]. Long-term conditions associated with T2D are retinopathy leading to loss of vision, nephropathy leading to renal failure, peripheral neuropathy leading to foot ulcers and amputations, and autonomic neuropathy leading to genitourinary, gastrointestinal, and cardiovascular complications, also to sexual disfunction[10] (see section 1.1.6 "T2D complications"). Hypertension and abnormal metabolism of lipoproteins, are two other traits characteristic of people with diabetis[10].

Normally, people with T2D retain some function of the pancreas, without needing exogenous insulin. However, as the disease progresses, the degree of β-cell destruction can reach a point where exogenous insulin is required for survival[10].

Even though T2D is recognized as a disease of adults, recent changes in lifestyle factors towards a more sedentary life, and the incorporation of a westernised diet (i.e. highly caloric), has resulted in a decrease in the age of diagnosis of T2D, with increasing rates observed in adolescents and young adults (<19 years)[9]. Early diagnosis of T2D is a phenomenon likely to become more common in minority ethnic groups, and in socially deprived communities[9]. Because incidence of T2D at young age is still rare[1, 2, 10], this work is largely focused on T2D in adulthood.

1.1.5.1 Detection of prediabetes

Because of the long time elapsed between the start of the metabolic disturbance until progression towards hyperglycaemia and the adverse conditions of T2D, it is important to identify participants at higher risk of developing T2D. Prediabetes is a term used to define individuals with impaired fasting glucose (IFG) and/or impaired glucose tolerance (IGT), where levels of glucose in blood are higher than the normal range, without reaching hyperglycaemia to formally diagnose T2D[10]. Similar disease processes occur in the prediabetic state, and participants with IFG and IGT are commonly obese, with dyslipidaemia (i.e. high triglycerides and/or low HDL cholesterol levels) and hypertension[10]. Prevalence of IFG is more common among men and associated to higher internal glucose production and impaired insulin secretion, whereas IGT is more common in women and is associated with peripheral insulin resistance[11]. According to the World Health Organization (WHO), IGT is a better predictor of future risk of T2D and CVD among women.

In asymptomatic adults of any age, criteria considered for testing for prediabetes are BMI > 25kg/m^2, obesity-related insulin resistance, dyslipidaemia, family history of diabetes, ethnic background (higher risk if African-American, Latino, Native-American, Asian-American or Pacific Islander), presence of hypertension, physical inactivity, women with polycystic ovaries syndrome, women with GDM or history of GDM, child born to a women with GDM, and obese children[2]. Formal testing for prediabetes should start after 45 years of age in asymptomatic adults[2].

Biomarkers used for the diagnosis of prediabetes are similar to those used for T2D, and criteria suggested for assessing IFG and IGT are described below in Table 1-1. When using HbA1c, the range suggested for the identification of prediabetes is between 5.7-6.4%, which corresponds to the range where higher liability of future T2D was identified according to results from multiple prospective

cohort studies[2]. In addition, people identified with prediabetes by HbA1c should be informed of their higher cardiovascular risk and counselled about mechanisms to lowering future risk of T2D and CVD[2, 10] (see section 1.1.8 "Prevention and management of T2D").

1.1.5.2 Biomarkers used in the diagnosis of T2D

Markers in blood used for the diagnosis of T2D are fasting plasma glucose (FG, mmol/L), with fasting defined as no caloric intake for at least 8 hours; 2-hours post-load glucose using 75g of anhydrous glucose dissolved in water (OGTT or 2-h PG, mmol/L), and glycated haemoglobin A (HbA1c, % or mmol/L), with a test performed using a method that is certified by the National Glycohemoglobin Standardization Program (NGSP), and standardised to the Diabetes Control and Complications Trial (DCCT) reference assay[10]. Proinsulin (pmol/L or µIU/ml) is another blood-borne marker used to estimate insulin resistance[12, 13].

Important international organizations providing guidelines for the detection of diabetes and prediabetes (i.e. IFG, IGT), are the WHO and the American Diabetes Association (ADA), and for gestational diabetes, it is the International Association of Diabetes in Pregnancy Study Groups (IADPSG). Selection of glucose cut points of FG and 2-h PG for the diagnosis of T2D, were derived from cross-sectional studies looking at the prevalence of retinopathy across different glucose concentrations in plasma[10]. Similarly, the cut point for HbA1c coincided with an inflection point for the prevalence of retinopathy[10]. Different from FG, HbA1c is a marker that reflects long-term (2-3 months) concentration of glucose in blood, and it is widely used for the management of diabetes and adequate glycaemic control, as levels of HbA1c associate positively with microvascular complications[10]. Criteria suggested by the WHO and ADA, and by the IADPSG, for the diagnosis of T2D and GDM, respectively, are described below in Table 1-1.

Table 1-1 Criteria for the diagnosis of T2D and other related phenotypes

	Diabetes (WHO and ADA) *	**IFG and IGT (WHO)***	**Prediabetes (ADA)***	**GDM and ODP (IADPSG)**
HbA1c	≥6.5%	NA	≥5.7% and <6.5%	ODP, ≥6.5%
FG (mmol/L)	≥7.0	IFG ≥6.1 and <7.0	≥5.6 and <7.0	GDM, ≥5.1, ODP ≥7.0
75g OGTT (mmol/L)	2h ≥11.1	IGT 2h, ≥7.8 and <11.1	2h, ≥7.8 and <11.1	GDM, 1h ≥10.1 and 2h, ≥8.5
Random glucose (mmol/L)	≥11.1	NA	NA	ODP ≥11.1

HbA1c=glycated haemoglobin A_{1c}, FG=Fasting glucose, OGTT=post-load plasma glucose, IFG=impaired fasting glucose, IGT=impaired glucose tolerance, GDM=gestational diabetes mellitus, ODP=overt diabetes in pregnancy, NA=not applicable. *Even though ADA and WHO have similar criteria for the diagnosis of T2D, they differ in the classification of intermediate hyperglycaemia and prediabetes.

1.1.5.3 Comparison among T2D tests

FG, 2-h PG and HbA1c are all used in practice to diagnose T2D[2]. FG and 2-h PG tests are more readily available than HbA1c, particularly in developing countries[10]. Compared to FG, HbA1c is a more practical test as it does not require fasting for its assessment, and the glucose measured is more stable and less susceptible to day-to-day variability due to illness or stress, allowing HbA1c results to be used in the management of glycaemia[10].

Despite the advantages of HbA1c over glucose-based tests, HbA1c is only a proxy of the average level of glucose in blood, and results of this test can be affected by factors including age, ethnicity (i.e. increased HbA1c in African Americans vs non-Hispanic whites), hemoglobinopathies (i.e. haemoglobin variants, sickle cell disease), and other conditions that alter red-blood cell turn over (i.e. pregnancy, haemodialysis, blood loss or transfusion, etc)[2, 10]. In cases where HbA1c is not adequate, it is recommended to use only glucose-based tests. In terms of specificity in capturing cases, it is known that FG and 2-h PG perform better at identifying cases than HbA1c at the specific glucose cut points, but this limitation can be offset by the greater practicality of HbA1c[10].

As with T2D, FG, 2-h PG and HbA1c can be implemented to detect prediabetes, but higher impact in primary prevention of T2D has been demonstrated by identifying patients with IGT rather than IFG or prediabetes measured by HbA1c[2]. Thus the advantage of 2-h PG over other tests for the early detection and prevention of T2D[2]. Overall, it is important to recognise that, regardless of the test applied, diabetes should be diagnosed within a continuum clinical spectrum that includes non-symptomatic individuals where a glucose test is randomly performed, individuals with known T2D risk, and finally, symptomatic patients[2].

1.1.6 T2D complications

Uncontrolled T2D generates a series of systemic complications that decrease quality of life of patients, and increases their risk of premature mortality[7]. Affected tissues and organs include blood vessels, heart, kidneys, nerves and eyes[7]. T2D is one of the leading causes of CVD, blindness, lower-limb amputations and kidney failure in adults between 20-74 years[7, 14]. Only in the US, 44% cases of end-stage renal failure and 60% of non-traumatic lower-limb amputations, are attributed to T2D[1]. T2D complications can be divided into acute and chronic complications. Acute complications include hypoglycaemia, diabetic ketoacidosis, hyperglycaemic hyperosmolar state, hyperglycaemic diabetic coma, seizures or loss of consciousness and infections[14]. Chronic complications can be divided into microvascular and macrovascular complications. Chronic microvascular complications are retinopathy, nephropathy and neuropathy, while chronic macrovascular complications are CVD, diabetic foot, and diabetic encephalopathy[14].

1.1.7 T2D comorbidities

Comorbidities are conditions that affect people with diabetes more frequently than age-matched healthy people, and these conditions can complicate the management of diabetes[15]. The type of comorbidity accompanying diabetes has changed over time due to an increase in the longevity of the population[7]. Some comorbidities associated with T2D are specific types of cancer, different levels of physical and cognitive disability, kidney failure, non-alcoholic fatty liver disease, pancreatitis, fractures, deaf, HIV, low testosterone in men, obstructive sleep apnoea, periodontal disease, polycystic ovarian syndrome, tuberculosis and depression[1, 7, 14-16]. Additionally, people with diabetes are more susceptible to infection by hepatitis B, and more likely to develop complications from influenza and pneumococcal disease[15]. The relationship between some of these comorbidities and type 2 diabetes can be found in the corresponding ADA report[15].

1.1.8 Prevention and management of T2D

Due to the severity of T2D symptoms and complications, and the great costs involved in diabetes care, prevention of T2D should follow a life course approach, from early life in the gestational and neonatal period, through childhood and adulthood. In addition, since different risk factor are involved in disease onset and progression, it is necessary that management of established T2D includes a multifactorial approach.

Prevention in early life

Adverse metabolic conditions during early life, such as malnutrition and hyperglycaemia, are likely to impact on developmental programming, increasing the child's susceptibility to obesity, insulin resistance, β-cell disfunction and T2D later in life[1]. However, an alternative hypothesis to the developmental programming hypothesis, known as the "fetal insulin hypothesis", suggests that the genetic component of insulin resistance, together with fetal environmental factors, are responsible for the observed association between low birthweight, and adverse cardiometabolic outcomes later in life[17]. In other words, the fetal insulin hypothesis postulates that low birthweight and future risk of insulin resistance, T2D, and adverse cardiovascular outcomes, are all phenotypes of the same insulin-resistant genotype[17]. Thus, insulin-mediated fetal growth is not only determined by the mother's glucose levels, but also by fetal genetic factors that control insulin secretion and insulin sensitivity in peripheral tissues of the foetus[17]. Evidence for this hypothesis comes from family studies of monogenic forms of T2D, where autosomal dominant mutations that affect pancreatic glucose sensing, insulin secretion, and insulin resistance, have been associated with impaired fetal growth[17, 18]. Different effects in birthweight can be observed depending on the mutation and its origin (if in the mother, the foetus or both).

Further studies are required to determine if common genetic variants (i.e. SNPs) associated with increased insulin resistance in the general population, can also influence birthweight and the risk of cardiometabolic outcomes in adulthood. Primary evidence supporting this concept was reported by Horikoshi *et al*., who conducted a multi-ethnic GWAS meta-analysis of birthweight, identifying 60 loci that explained 15% of the variance in birthweight based on fetal genotype from 153,781 unrelated participants[19]. Genetic variants associated with birthweight were inversely correlated with the genetics of systolic blood pressure, T2D and coronary artery disease in adulthood. Pathway analysis further indicated that genes mapping to birthweight-associated variants, were related to glucose homeostasis, insulin signalling, glycogen biosynthesis and chromatin remodelling[19]. In conclusion, Horokoshi *et al.* provided evidence of potential shared genetics between fetal growth phenotypes and adulthood cardiometabolic outcomes, which supports the fetal insulin hypothesis.

Despite contrasting views between the developmental programming hypothesis and the fetal insulin hypothesis on the influence of the intrauterine environment in latter cardiometabolic outcomes, this factor continues to be an important mediator of health outcomes in the index child. Thus, pregnant women are required to adopt a healthy lifestyle during and after pregnancy to improve the wellbeing of the child[1]. A healthy lifestyle includes an appropriate diet, exercise, good quality of

obstetric, neonatal and paediatric care, breastfeeding, adequate control of glucose, blood pressure, and appropriate intake of folic acid[1, 9]. In pregnant women with GDM, it is often likely that it develops onto overt T2D after pregnancy[1, 9]. Thus, untreated or lately diagnosed GDM imposes a risk for the mother and her offspring, in both the index and (potentially) future pregnancies[1, 20, 21].

Prevention in childhood, early adulthood and adulthood

Similar lifestyle changes to prevent T2D apply for the child, young adult and adult, even though the focus for the prevention of early onset T2D is on improving diet and physical activity to reduce obesity[9]. In adults, lifestyle intervention remains the most cost-effective strategy to prevent T2D[22], and this includes structured diabetes education, exercise, reduction of fat intake while increasing that of fibre, moderate weight reduction, smoking cessation, control of blood pressure, among others[1, 7, 22]. Two follow-up studies have demonstrated the long-term effect of lifestyle changes on reducing the incidence of T2D by 43% over 20 years[23], and on reducing mortality associated to CVD and other causes over 23 years of follow-up[24].

In addition to lifestyle intervention, multiple drugs are available to prevent or treat T2D and obesity, including metformin, thiazolidinedione compounds, acarbose, orlistat and insulin[22], with accompanying clinical trials supporting their effectiveness[22]. From these drugs, metformin has the largest body of evidence in preventing diabetes, and a meta-analysis showed a reduction in approximately 40% in the risk of T2D by using metformin[25]. Some antihypertensive drugs are also recommended to reduce the incidence of T2D in people at higher risk who are hypertensive, although results suggest that these drugs are more effective on maintaining normoglycemia rather than in reducing diabetes incidence[22]. A more extreme preventive intervention in morbid obese people is the use of bariatric surgery, with reported benefits on weight loss, glycaemic control, and remission of T2D in those who previously had diabetes[9, 22].

Management of established T2D

Management of diabetes is conducted at the primary care level by a team of specialized health-care providers who deliver basic interventions including medication, self-care education and follow-up[7]. Thus, the aim of primary care intervention is to promote heathy lifestyles (exercise, balanced diet) to control for risk factors that lead to diabetes complications and premature death[7]. Risk factors to be controlled are levels of blood glucose, lipid concentrations, blood pressure, smoking and bodyweight[1, 7]. In addition, it is necessary to maintain periodic referrals of T2D patients to specialists for the screening and management of microvascular and macrovascular complications [1, 7, 22]. New

pharmacological approaches for the treatment of diabetes will be based on a better knowledge of its pathophysiology, with the aim of using more personalized therapies that reduce unwanted effects, mortality associated with diabetes complications, and improve the quality of life of patients[1, 7].

1.2 Epidemiology of Type 2 Diabetes

1.2.1 Global prevalence of T2D

According to the International Diabetes Federation (IDF), global prevalence of diabetes is projected to increase from 415 million people in 2015, to approximately 642 million people by 2040[14], with almost 87% to 91% of the cases corresponding to T2D based on data from high-income countries[14]. Rate of increase in prevalence of T2D is expected to be higher in adults between 40-60 years of age from low- and middle-income countries, compared to adults from high-income countries[7, 14], where T2D is more common in older age (>60 years)[1, 7, 22].

Based on 2017 estimates by the IDF, age distribution of diabetes was almost three times higher among the working-age population (20-64 years, 326.5 million people affected) compared to the elder population (65-99 years, 122.8 million people affected)[14], and this proportion is expected to be maintained by 2045 figures[14]. In terms of gender distribution of diabetes, it was estimated to be 8.4% for women between 20-79 years of age, and 9.1% for men of similar age-range[14]. Comparing between rural and urban areas, prevalence of diabetes was estimated to be higher in urban compared to rural areas (10.2% versus 6.9%, respectively)[14].

Looking at regional differences in diabetes, age-standardised prevalence of diabetes in 2014 was the highest in the Western Pacific and South-East Asia regions (WHO regions), accounting these for almost half of the cases of diabetes in the world [7, 26]. For the same year, prevalence of diabetes was the lowest in the African and European regions[7]. A more recent map of global prevalence rates of T2D is presented in Figure 1-2 according to IDF estimates for 2017. Based on these estimates, the highest prevalence of T2D in 2017 was reported in North America and the Caribbean regions, while the lowest prevalence was reported in the African region[14]. Lower prevalence of T2D in the African region can be related to more rural than urban areas, under-nutrition, higher rates of communicable diseases, and lower rates of obesity[14]. Countries with the highest number of people leaving with diabetes according to IDF records for 2017, were China, India and the United States, the growing economies, and these rates are projected to remain for the year 2045[14].

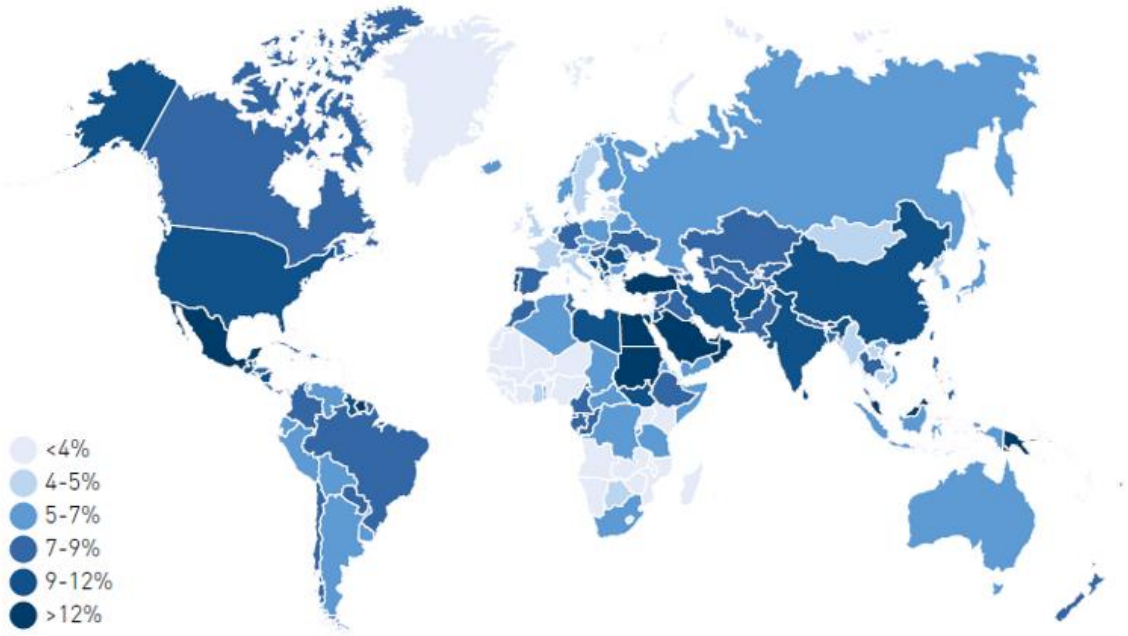


Figure 1-2 Age-adjusted global prevalence of Diabetes in adults (20-79 years) by 2017. Figured taken with permission from: IDF Diabetes Atlas. Eight edition, 2017 [14].

Reasons leading to an increase in prevalence of diabetes are population growth and aging, longevity of people with diabetes, change in the ratio of diagnosed to undiagnosed cases, and increase in age-specific prevalence of diabetes[7, 22], with each factor having a different contribution depending on the region studied[22]. According to global trends from 1980 to 2014, diabetes increased almost four times during this period[7], with 40% of this increase due to population growth and aging, 28% to a raise in age-specific diabetes, and the remaining 32% to an interaction between both factors[26].

Compared to studies of prevalence of T2D, those of incidence of T2D are more complicated and scarce due to the high proportion of undiagnosed cases[7], which may reach global proportions of 30% to 50% [7, 14]. As a result, there is almost no data on true incidence of T2D[7]. Undiagnosed diabetes is a problem affecting global economies and the healthcare system. Since approximately 50% of the people living with diabetes are undiagnosed[14], with 84.5% of these cases in low- and middle-income countries[14], the current challenge is to improve screening and diagnosis of this group. Providing early treatment for diabetes in undiagnosed participants, will help to avoid the future economic burden caused by diabetic complications[14].

1.2.2 Mortality rates

In the year 2000, 2.9 million deaths were attributed to diabetes[1], with most cases corresponding to T2D. For the year 2012, mortality rates went down to 1.5 million deaths directly caused by diabetes,

positioning diabetes as the eighth leading cause of deaths in men and women worldwide, and the fifth leading cause of death in women alone[7].

The latest index reported by the IDF for the year 2017, indicated that global diabetes mortality ascended to 4.0 million (3.2-5.0 million) deaths in people between 20-79 years[14], and almost 46.1% of these deaths occurred before the age of 60[14]. Considering all global causes of mortality, diabetes accounted for 10.7% of the deaths, and this proportion was higher than that of deaths attributed to infectious diseases[14]. Globally, deaths for diabetes were higher in women (2.1%) than in men (1.8%), except for the North America and Caribbean region, where mortality rates were higher in men[14]. Across regions, the highest rate of premature deaths (under 60 years) for diabetes was observed in the African region (77%), while the lowest rate was observed in the European region (32.9%)[14].

1.2.3 Prevalence of T2D and mortality rates in Europe

Evidence presented throughout this thesis corresponds to prevalence of T2D in the European region alone based on two population studies from England, and three other studies from Scotland, Germany and the Netherlands. In the global context, prevalence of T2D in Europe is the second lowest (8.8%) among people between 20-79 years[22], but there are contrasts in rates of T2D between countries due to differences in population composition, availability of recent data, and economic disparities[22]. One of the main risk factors for T2D in Europe is aging, knowing that 45.1% of the population is between 50-99 years[14].

In the UK, prevalence of T2D was one of the lowest (6.2%) according to IDF figures in 2015[22]. Some of the factors leading to T2D in inner-cities in the UK were poverty and social deprivation, obesity, physical inactivity and smoking, which are factors that tend to co-segregate[22]. Another important factor when considering prevalence of T2D in the UK is multi-ethnicity of the population. Several studies have shown that burden of T2D is higher among immigrant groups, especially those from the Indian subcontinent, compared to white Europeans[22]. In the Netherlands, T2D affects 8% of the elderly white Dutch population[22], and a prospective study showed that elderly over 70 years represented 50% of the population affected with T2D[22]. Prevalence of T2D in Germany is the second highest in Europe, ascending to 8-9% according to 2017 records[14]. In a prospective population-based study conducted in middle-aged adults in Germany, incidence rates of T2D were 10.5% among men and women, with main risk factors attributed to age, family history, BMI, uric acid, current smoking, and high HbA1c, 2-h PG and FG[27]. The same study concluded that risk of T2D among German adults was higher in people with isolated IGT, compared to those with isolated IFG[8].

Mortality rates of diabetes in Europe ascend to 9%, with more than 60% of these deaths occurring in people over 60 years[14]. Lower rates of premature mortality in Europe can be attributed to the age distribution of the population, early detection of diabetes, and improved healthcare[14]. Number of deaths for diabetes in Europe are higher in women (413,807 deaths) compared to men (279,543 deaths) due to higher prevalence of diabetes in women, and the relatively higher proportion of women to men in this population.

1.2.4 Major risk factors

Multiple genetic and environmental risk factors are involved in the pathophysiology of T2D. Similar risk factors are observed across different populations, but each risk factor has specific characteristics depending on the population under study. Further details of well-established genetic and non-genetic risk factors for T2D are described in this section.

1.2.4.1 Genetic risk factors

Extensive research has been conducted to investigate the genetics of T2D, including family, twin, and population studies. Family and twin studies have determined the large heritability component of T2D (estimated to be >50%), supporting the genetic origin of the disease[28]. In fact, people with first degree relatives with T2D have approximately 5-10 times higher risk of developing T2D in their life-time, compared to people without a family history of T2D, but with other risk factors including obesity and low physical activity[28]. In addition, first-degree relatives of T2D patients are more likely to manifest the pathological conditions leading to T2D (i.e. insulin resistance, β-cell disfunction) long before the onset of the disease[28]. Likewise, evidence from twin studies suggests that among monozygotic twins discordant for T2D, the unaffected twin has a 90% chance of developing T2D later in life[28].

To date, the implementation of high-throughput genotyping technologies in population studies has facilitated the meta-analyses of large-scale genome-wide association studies (GWAS), which in combination with candidate loci studies, have allowed the identification of over 60 common genetic variants associated with T2D[1, 8, 28, 29]. These genes encode for proteins that can alter different pathways leading to disease, including pancreatic development, insulin synthesis, processing and secretion, amyloid deposition in β-cells, cellular insulin resistance, and impaired regulation of gluconeogenesis[8, 28]. Most of the genes identified in T2D are related with β-cell disfunction (*KCNJ11, TCF7L2, WFS1, HNF1B, SLC30A8, CDKAL1, IGF2BP2, CDKN2A, CDKN2B, NOTCH2, CAMK1D, THADA, KCNQ1, MTNR1B, GCKR, GCK, PROX1, SLC2A2, G6PC2, GLIS3, ADRA2A,* and *GIPR*), and to a less

extent, with impaired insulin sensitivity (*PPARG, IRS1, IGF1, FTO,* and *KLF14*) and obesity (*FTO*)[1]. In comparison, there is no representation of genes related with the control of glucose tolerance among reported risk variants for T2D[28].

Findings from GWAS demonstrate that most of the genetic predisposition for T2D is related to the biology of the pancreas, β-cell function and insulin secretion[8]. This means that the genetic make-up of β-cells is likely to determine if the requirement for more insulin under specific environmental conditions or stressors, is fulfilled by increasing insulin secretion (and subjects remain healthy), or not (and subjects develop T2D)[8]. In general, risk variants for T2D are widely distributed around the genome and have a small effect on disease risk[1, 8]. Currently, the common variant with the strongest association in T2D GWAS has been identified at the *TCF7L2* gene in relation with β-cell disfunction[1, 30].

Genetic studies on T2D have made a large contribution in understanding the biology of β-cells, the mechanisms controlling glucose tolerance, and the response to anti-diabetic drugs, but they have also indicated that insulin resistance alone cannot explain the entire risk of T2D[28]. In fact, common genetic variants in T2D only explain 10%-15% of the heritability of the disease[1, 8, 31]. The remaining heritability of T2D can potentially be explained by rare genetic variants (allele frequency < 5%), which are more difficult to assess with current GWAS technologies[1].

All in all, evidence suggests that it is unlikely that the genetic make-up for T2D alone has a big impact in the current epidemic of T2D[8, 28]. This contrasts with the well-established impact of environmental risk factors in T2D incidence and progression[28]. Some environmental risk factors for T2D (i.e. diet, physical activity, SES) are likely to be shared among family groups[28].

1.2.4.2 Non-genetic risk factors

1.2.4.2.1 Age

Risk of T2D increases with age. Thus, it is recommended to begin a screening for hyperglycaemia at 45 years of age in asymptomatic adults, or earlier if additional risk factors are present[2]. T2D can also be observed among young adults, normally in association with obesity. In this group, treatment of diabetes requires of a stronger polypharmacy component, with little advantages seen in implementing lifestyle changes to lowering the risk of cardiovascular complications, which manifests more aggressively in young patients[9].

1.2.4.2.2 Ethnicity

Higher risk of T2D has been identified among Pima Indians, Hispanics, Native Americans and Afro-Caribbean compared to Caucasians[1, 32]. According to several cross-sectional and prospective studies of insulin sensitivity in ethnic groups in the U.S., African-Americans, Latino Hispanics and Native Americans, have a 2.0, 2.5 and 5.0 increased risk of developing T2D compared to Caucasians, respectively [32]. In developed countries, the increased prevalence of T2D among minority groups, compared to the rest of the population, is commonly attributed to SES disparities[32].

1.2.4.2.3 Sex

T2D is a clear example of a non-communicable diseases with sex differences in disease prevalence. Global records show that there are more men affected with T2D than women[11], as it was reflected by figures in the year 2013, where there were 14 million men more affected than women[11]. Despite this difference, prevalence of IGT is higher among women, as well as prevalence of obesity, especially in women over 45 years of age[11]. Sex difference in the risk of T2D is independent of age, as both sexes have similar patterns of age-dependency in the onset of T2D[11]. In addition, there is data available of regional differences in the proportion of men and women affected by T2D. For example, in Oceania, equal rates of T2D are observed in both sexes, while in South and Central Asia, the Middle East, and North Africa, T2D growth is higher among women. For the high-income Asia-Pacific and Western region, men are more affected than women[11].

1.2.4.2.4 Insulin secretion and glucose tolerance

Two major predictors of T2D are increased fasting insulin (proxy for insulin resistance), and low insulin sensitivity, measured as the change in insulin concentration divided by the change in glucose concentration after 30 min of an oral glucose tolerance test[32]. Increased insulin secretion at baseline was correlated with higher incidence of T2D among Mexican-Americans[32]. However, there was an inverse relationship between insulin sensitivity and the risk of T2D[32]. A combination of insulin secretion and insulin sensitivity was demonstrated to have an additive effect in predicting T2D. Thus, participants with one of these two factors under normal functioning and the other impaired or low, translated into five times higher risk of T2D compared to subjects with normal values of insulin secretion and insulin sensitivity[32]. If both factors were impaired, then the risk of T2D was 14 times higher[32].

1.2.4.2.5 Imbalance of sex hormones

Balance of sex hormones is important to maintain energy metabolism, body composition and sexual function[11]. In women, elevated levels of androgens over oestrogens is associated with increased

body weight and VAT [11]. In addition, risk of polycystic ovary syndrome (PCOS) in women was associated with higher levels of androgens and hyperinsulinemia related to obesity, T2D, and higher cardiometabolic risk[11]. In males, lower levels of androgens were observed in obese and diabetic subjects, in addition to increased VAT[11].

1.2.4.2.6 Incretin imbalance

Incretins are positive regulators of insulin secretion in the nourished state. The action of this hormone is partly regulated by the GLP-1 hormone secreted in the gut. A 25% lower response of GLP-1 to an oral glucose challenge was observed in women with IGT and T2D in a large cohort study, and this association was independent of age and BMI[11]. Lower response of GLP-1 was correlated with lower insulin secretion. Thus, it was hypothesized that lower incretin levels could partially explain the reduced β-cell function observed in women with IGT and T2D[11].

1.2.4.2.7 Inflammatory markers

The observed association between T2D and higher concentrations of inflammatory markers including CRP, interleukin-6 (IL-6) and interleukin-1 (IL-1), has led to consider inflammation as an important risk factor for T2D[33]. However, this association can be confounded by increased BMI, which is associated with both, inflammation and T2D. Alternatively, inflammation can be a consequence of T2D through reverse causation[33].

1.2.4.2.8 Obesity

Obesity accounts for 80-85% of the overall risk of T2D, and underlies the current pandemic proportions of the disease[22]. Prevalence of obesity has increased worldwide despite the intention to halt it by 2025. According to WHO estimates for 2014, one in three adults aged over 18 years were overweight, and more than one in ten adults were obese[7]. Among sexes, women tend to be more obese than men[7], and across regions, the region of the Americas had the highest proportion of overweight and obese people[7]. Comparing across income levels, high- and middle-income countries almost doubled the proportion of obese people in low-income countries[7].

Between general and central obesity, this latter type is a stronger predictor of T2D[22], and it is associated with insulin resistance and β-cell disfunction, partly through increased levels of fatty acids and lipotoxicity[22]. Anthropometric measures including waist circumference, waist-hip ratio and BMI, are important risk factors for T2D[1, 11], with variations in their effect population-wise[7] and sex-wise[11]. For instance, in Mexican-Americans, waist circumference was a better predictor of T2D in men and women compared to waist-hip ratio, and in men BMI was a better predictor of T2D compared to

waist-hip ratio[32]. Similar findings have been observed in European populations[32]. With respect to sex differences, men are diagnosed with T2D at a BMI 1-3 kg/m^2 lower than females, who tend to be more obese[11]. According to WHO and ADA criteria, a BMI ≥ 25kg/m^2 should be considered a risk factor for T2D, independently of sex[2]. However, in Asian Americans and other populations, this threshold should be lowered to 23kg/m^2, which is the cut point for BMI where a higher prevalence of T2D is observed[2].

Another marker of obesity is body fat composition, which is measured using VAT and SAT distribution. In healthy men, deposition of fat is higher in the trunk, VAT and upper extremities, compared to women of same BMI and age[11]. However, no difference in body fat composition was observed between sexes in adults with similar levels of insulin resistance[11]. Levels of leptin and adiponectin, which are both fat biomarkers, can also serve to determine risk of T2D[11]. Leptin helps in the regulation of food intake, satiety and energy expenditure[11], while adiponectin helps in the metabolism of glucose and lipids, and it improves insulin sensitivity in insulin-dependent organs[11]. In obese and diabetic subjects, an inverse correlation is seen between adiponectin levels in plasma and insulin sensitivity, and this inverse correlation is more pronounced in women[11]. In contrast, higher levels of leptin, which associate positively with SAT, are related with higher risk of diabetes in men[11].

1.2.4.2.9 Gestational Diabetes Mellitus (GDM)

Recognized as any type of glucose intolerance developed during the gestational period, identified first during the second or third trimester of pregnancy, which is clearly not pre-existing type 1 or type 2 diabetes[2]. Nowadays, the rate of undiagnosed GDM has increased in line with an increase in the prevalence of T2D, especially among women of childbearing age[10]. According to the ADA, for the year 2010, a cross-population study revealed that around 7% of all pregnancies were complicated by GDM, which translated into more than 200,000 cases of GDM every year[10]. Generally, GDM is screened in the first prenatal visit in women at higher risk of T2D, and 24-28 weeks of gestation in women with lower risk of diabetes[2].

1.2.4.2.10 Intrauterine environment

Research of the effect of elevated glucose during the gestational period has shown that the offspring of mothers with overt T2D are more likely to develop T2D earlier in life compared to those born to mothers without T2D[1]. Also, among siblings, the risk of obesity and T2D is higher for those born after the mother developed T2D[1]. With regards to the time in pregnancy when the risk is increased, a Danish study found that hyperglycaemia during the third trimester impacts more in the risk of prediabetes or T2D in the offspring[34]. Research in animal models has shown that early life

programming can influence neurohormonal networks of weight control and development of pancreatic islets[35, 36].

1.2.4.2.11 Diet

Poor nutritional habits, with higher intake of saturated fatty acids, processed food and sweetened drinks, leads to unhealthy gain of bodyweight and higher risk of T2D[7]. Also, adoption of a westernised diet, characterized by high caloric intake and low physical activity, is a strong contributor to the current pandemics of obesity and T2D[1]. Transitional generations who move from a rural environment characterized by good dietary practices, with foods rich in fibre and low in sugar, to an urban environment with abundance of processed food high in energy, are also at higher risk of developing T2D due to a discordance between the two life-styles[1]. Furthermore, deficiency in micronutrients such as vitamin B12 and vitamin D in the context of high folic acid and iron storage, have been associated with the pathogenesis of T2D[1]. Alteration in the gut microbiota by events in early life (i.e. mode of delivery, type of infant feeding, hospitalization, and prematurity), or later in life (i.e. antibiotics and diet), is also implicated in the risk of T2D[37], with promising beneficial effects in lowering T2D risk by using probiotics.

1.2.4.2.12 Physical inactivity

Physical inactivity is a well-known risk factor for T2D[1, 22, 27], with increased physical inactivity leading to obesity, higher insulin resistance and β-cell disfunction[1]. Reduction in the duration and intensity of physical activity patrons, is a major concern worldwide. Across regions and income groups, reports by the WHO in 2010 showed that the global percentage of physical inactivity in women was equal to 27%, and that of men was 20%[7]. Looking across countries, prevalence of physical inactivity was almost double in high-income countries compared to low-income countries, and among regions, the Eastern Mediterranean region showed the highest prevalence of inactivity in adults and adolescents[7]. As with diet, higher index of physical inactivity can be attributed to an increased Westernized life-style, accompanied by urbanization and mechanization[22].

1.2.4.2.13 Socioeconomic status (SES)

SES, evaluated in terms of educational level, position and income, is inversely associated with risk of T2D worldwide[1, 11, 32, 38]. Inverse association between T2D and SES was stronger in women from specific communities compared to men, even after adjustment for obesity and physical activity[11]. Even though more studies are needed in this aspect, evidence suggests that adverse measures of socioeconomic predictors of T2D, can affect more women than men in their future risk of T2D[11].

1.2.4.2.14 Smoking

Evidence suggests that risk of T2D increases with smoking, being this either passive or active smoking, and that small differences are observed by sex[11]. Heavy smokers tend to have higher risk of T2D compared to casual smokers, and this increased risk persists even after years of smoking cessation[39]. When comparing among sexes and smoking behaviours, a recent meta-analysis of prospective studies indicated that if smoking was a causal factor for T2D, 11.7% of T2D cases in men and 2.4% in women would be attributed to current smoking[40].

1.3 Epigenetics

1.3.1 Types of Epigenetic Modifications

Epigenetic modifications can be found in the form of mitotically stable DNA methylation (DNAm), post-translational modification of histone proteins and non-coding RNAs (ncRNAs)[41]. All these marks reside in the epigenome, which contains the entire epigenetic composition of a cell[42], is highly dynamic and influenced by the interplay between genetic and environmental factors[41]. The predominant form of DNAm are methyl groups added to carbon five of cytosines in the context of cytosine-guanine (CpG) dinucleotides[4, 41, 42], a mark commonly associated with transcriptional repression[4]. Other forms of DNAm including CpH (H=A/C/T) and 5-hidroximethylation of cytosines (5hmC), are less common but have a potential role in gene regulation and differentiation[41]. Histone modifications can result from different levels of methylation, acetylation and citrullination of amino acids located in the amino-terminal tails of proteins forming the nucleosome (i.e. core histones)[41]. Non-coding RNAs, on the other hand, can be transmitted independently of the DNA sequence and are represented by microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), large-intergenic non-coding RNAs (lincRNAs), among others[41]. All these epigenetic marks are tissue-specific and vary with age, and in response to environmental exposures[4]. Figure 1-3 illustrates the different epigenetic marks described.

From the available epigenetic markers, the most commonly used in epigenetic epidemiological studies is DNAm[41-43]. There are various reasons why DNAm is the marker of preference in epigenetic epidemiological studies. The foremost is that of the stability of the methyl-cytosine bond, which allows to capture DNAm using routine DNA extraction protocols[43]. Also, due to the long-term stability of the methylated cytosine, DNAm can be obtained from newly collected samples, or from appropriately processed and stored samples[43]. This contrasts with the more analytically demanding and costly protocols required to analyse histone modifications and miRNA from large samples[4]. Another advantage of DNAm is that it is relatively easy to measure as it exists as a continuous signal

between the methylated and unmethylated state[43], in comparison to the more complex post-translational modifications observed in histones[43].

Independent of the mark selected, useful epigenetic markers in epidemiological studies must show higher interindividual rather than intraindividual variation, and the variation captured should be systematic rather than stochastic[43]. Systematic epigenetic variation allows individuals to be distinguished by differences in lifestyle, exposure to an environmental factor, and susceptibility to disease[43].

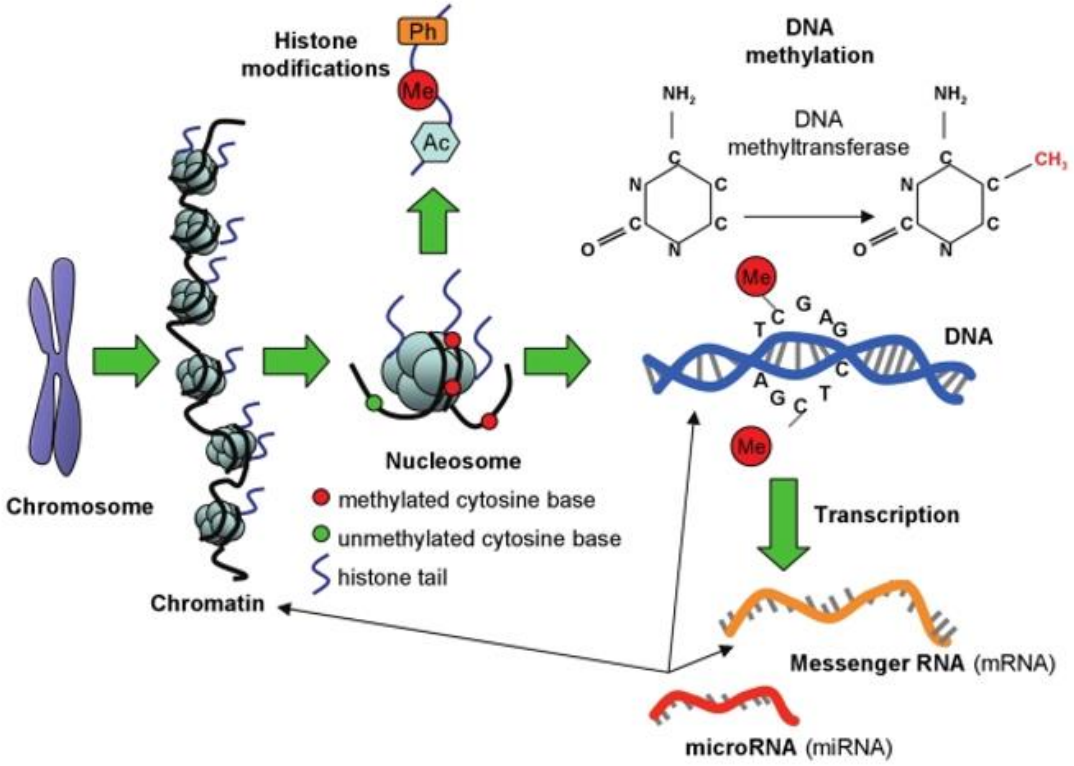


Figure 1-3 Epigenetic modifications of interest in epidemiological studies. From left to right, histone modifications in the protruding aminoacid tail of core histone proteins surrounding the DNA. Histone modifications appear as phosphorilation (Ph), methylation (Me) and acetylation (Ac). DNA methylation (DNAm) at the carbon 5 of a cytosine located next to a guanine base. Methylation is enabled by DNA methyltransferase enzymes, and it is a sign of transcriptional repression. Messenger RNA (mRNA) is the product of transcriptionally active genes, and it is translated into active proteins by the ribosome. However, micro RNAs (miRNAs) can repress mRNA translation, and influence DNAm and chromatine structure through the regulation of histone modifiers. Figure taken with permission from Relton & Davey Smith. 2010.[4]

When using DNAm as the marker of interest, variation between samples can be identified at the individual CpG site also known as a differentially methylated site, or it can be identified in groups of adjacent CpG sites within a differentially methylated region or DMR[41, 42]. In general, CpG sites are

underrepresented in the genome, around 70% of them are methylated, and the remainder are unmethylated and usually located within CpG islands[44], which can be found in the promoter or in the body of the gene[44]. CpG density can be as high as 60% in the promoter of human genes[44]. The influence of methylation on gene expression depends on the genomic context. Generally, unmethylated CpG islands within promoters are positively associated with gene expression, as is methylation within the gene body[44].

Association of variation in DNAm with disease susceptibility has generally been restricted to core promoters and CpG islands near candidate genes[41, 42]. However, current evidence also highlights the importance of more distant variation in DNAm occurring in CpG island shores and shelves, in the study of complex diseases[41]. Furthermore, variation in DNAm can be investigated at specific loci for genes known to play a role in disease processes[4, 43], or they can be investigated genome-wide to identify changes in DNAm and discover new associations[4, 43]. Assays used to measure DNAm using a candidate gene approach commonly provide higher coverage, than assays used for genome-wide methylation (i.e. Infinium arrays, ChIP-on-chip), which are highly reliant on bioinformatic methods to interpret results[43]. The approach selected to study DNAm will depend on the research question, costs, and on the availability of adequate tools to assess methylation at the level of CpG site density required.

1.3.2 Importance of epigenetics in the study of complex diseases

The purpose of including epigenetic markers in the study of complex diseases, is to contribute to understanding disease aetiology and to improve clinical detection and disease subtyping[42], with the ultimate goal of developing interventions to prevent and treat disease[43]. The epigenome mediates the interaction between the environment and the genotype, and it provides a mechanism to link environmental risk factors with disease risk in the general population[43]. Thus, the incorporation of epigenetic measures in epidemiological studies has the potential to identify interindividual epigenetic variation associated with susceptibility to disease, understanding how environmental exposures related with disease can influence the epigenome throughout life, and explaining sex differences in disease susceptibility[42]. Challenges common to epigenetic epidemiological studies are confounding, effect modification, and power to identify epigenetic marks with sufficient interindividual variation[43] (see section 1.5.5 for more on methodological considerations).

One method widely used to identify epigenetic variation in the population has been epigenome-wide association studies or EWAS, which include DNA methylation as the primary marker to identify

epigenetic differences associated with future risk of disease, disease state or progression[42]. EWAS have been conducted in multiple complex phenotypes, using cross-sectional or longitudinal studies, and comparing across accessible tissues and internal primary target tissues.

Epigenetic changes measured in an EWAS can be deterministic and occur in response to environmental exposures (i.e. nutritional, chemical, physical, psychosocial), or they can be stochastic and occur during development[42]. In both cases, epigenetic changes are transmitted stably across cell generations through mitosis[42], which has led to consider the importance that early life epigenetic modifications have in later susceptibility to disease. For instance, it has been demonstrated that exposure to adverse conditions during early life, such as prenatal famine, determines epigenetic changes that can be detectable years after the time of exposure[42]. Similarly, early life hypermethylation in association with obesity at a CpG island in the proopiomelanocortin (*POMC*) gene, can potentially persist across tissues and across the life course[42].

1.3.3 Epigenetic technologies in the study of DNA methylation: methylation arrays

Methods for epigenomic profiling are nowadays more affordable and allow investigators to conduct large-scale EWAS. The marker selected for large-scale EWAS should be stable, convenient for high-throughput analysis, and easily accessible from routine clinical samples[41]. All these characteristics are fulfilled by DNAm, as it was described before (see section 1.3.1). Technologies available for genome-wide methylation profiling in EWAS are array- and sequencing-based technologies[41], and the final choice between them will rely on the balance between resolution, coverage, accuracy, specificity, throughput and cost[41]. Even though sequencing-based technologies are more sensitive and provide higher resolution of the epigenome, they tend to be analytically more demanding and costlier compared to array-based technologies, which offer a good balance of genome coverage, resolution, throughput and cost[41]. These characteristics make array-based technologies more suitable for large-scale EWAS.

Further detail of the different sequencing- and array-based technologies available to date for EWAS, can be found elsewhere[31, 41, 45]. In principle, the common procedure used in array- and sequencing-based technologies is the treatment of DNA with sodium bisulphite. This chemical treatment converts unmethylated cytosines into uracil in single stranded DNA, while methylated cytosines remain unchanged, allowing for differences in the base sequence to be identified in later parts of the protocol[45]. In addition, bisulphite treatment generates a DNA sequence of less complexity, with

single-stranded DNA that is more fragile and requires of adequate storage to avoid DNA damage[45]. Bisulphite treated DNA is then amplified and analysed using the method of choice.

Examples of sequencing-based technologies are whole-genome bisulphite sequencing (WGBS) and reduced representation bisulphite sequencing (RRBS). The WGBS covers almost all the 28 million CpG sites that compose the human methylome, provides direct readout of non-CpG methylation, is less biased towards CpG enriched regions, and it is considered the gold standard method for DNA methylation typing[41, 42]. The RRBS method uses digestion of the DNA with methylation sensitive restriction enzymes, followed by purification of digested fragments, bisulphite conversion of the DNA, amplification and sequencing. This genome-wide technique allows the measurement of methylation at a single base-pair resolution, with a CpG coverage between 5-10%, reducing sequencing costs considerably[41, 45].

Examples of array-based technologies are the comprehensive high-throughput relative methylation (CHARM), and the Infinium methylation arrays. The first technique uses methylation-sensitive restriction enzymes, while the second one uses two different bead types to detect CpG methylation from bisulphite treated DNA[41, 42]. From the Infinium array-based methods available to date (Illumina Golden Gate, Infinium 27K and Infinium 450K), the 450K array has been the most commonly used in large-scale EWAS because of its quantitative accuracy, single base-pair resolution, relatively high coverage (>450K probes), high throughput (12 samples per chip and up to 96 samples per run) and lower cost compared to the WGBS[41].

Despite providing a relatively high coverage compared to other array-based methods, the 450K array presents drawbacks that need to be considered when interpreting results. This array covers <2% of all the CpG sites in the genome, with preference for sites located in CpG islands and gene promoters, and it is unable to discriminate 5-methyl-cytosine from 5hmC due to the bisulphite treatment of the DNA[42, 45]. Furthermore, because probes in the array are limited to those defined by the company, it is difficult to estimate the functional relevance of epigenetic variation detected by the 450K in the epigenetic study of complex phenotypes[42, 45]. A recent improvement from the 450K array has been the development of the Infinium MethylationEPIC (EPIC) array. This new array has >90% concordance with sites profiled in the 450K array, in addition to 350,000 CpG sites more included in enhancers, providing in total >850,000 methylation sites to be analysed[44]. As technologies move forward and become more affordable, the goal is to assess the methylome using whole-genome

methods that provide direct (i.e. without bisulphite conversion) and real-time determination of DNA methylation, hydroxy-methylation and DNA sequence, all within a single assay[41].

1.4 Epigenetics in Type 2 Diabetes

The growing field of epigenetics in T2D encompasses studies using different epigenetic marks, tissues, different study designs, and hypothesis driven (candidate gene assessment) or not (genome-wide assessment) approaches, to investigate markers associated with T2D aetiology that could be used in the prevention, detection, or treatment of T2D. For the purpose of this work, and in line with current evidence, studies reviewed were those that used DNA methylation as the marker of interest. Main findings from epigenetic studies in T2D are described in this section considering the tissue where associations were detected, the study design used, and the level of coverage of the genome considered. Figure 1-4 illustrates how DNA methylation can lead to T2D in response to environmental triggers common to the disease.

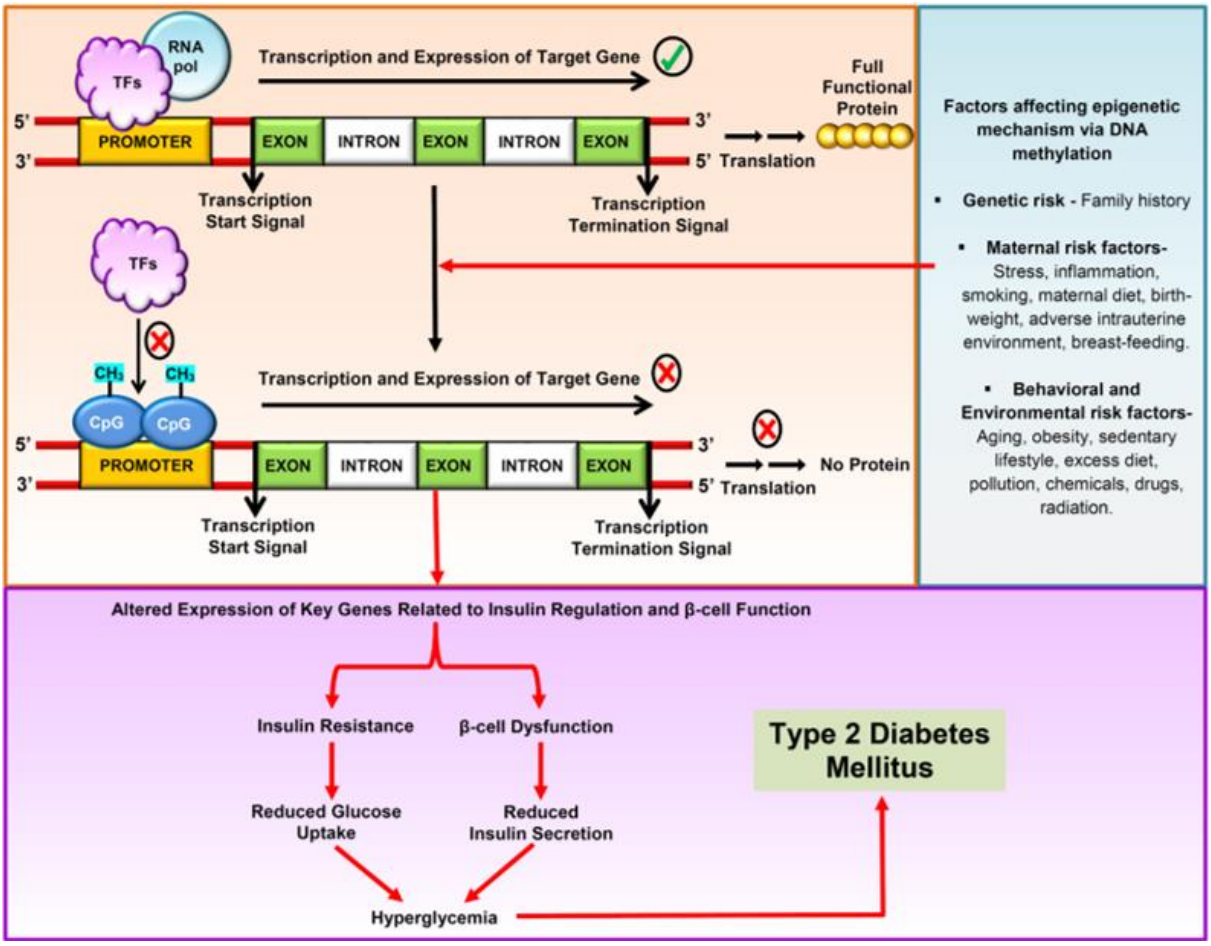


Figure 1-4 Example of mechanisms leading to T2D that can be influenced by changes in DNAm. Figure taken with permission from Alam et al. 2016.[31].

1.4.1 Previous studies using candidate loci in T2D

Epigenetic studies based on candidate loci rely on previous knowledge of the biology and functional implications of the gene(s) in T2D[31]. Genes taken forward for analysis are generally extracted from GWAS evidence, or from findings of previous epigenetic studies. At first sight, the disadvantage of this type of studies is that they overlook the importance that other unknown methylation variation might have on determining disease risk, assuming that variation in methylation relevant for T2D is constrained to the candidate locus or loci studied. This assumption is risky, considering that our knowledge of the epigenome is still limited[42]. In addition, because of biological differences between genetic and epigenetic markers, it cannot be expected a complete overlap between epigenetic and genetic signals of relevance in T2D[28]. DNA methylation signals detected in candidate gene studies of T2D are summarized below.

1.4.1.1 Studies in peripheral blood

In comparison to internal tissues of relevance in T2D, peripheral blood offers the possibility of analysing larger samples and identifying biomarkers of disease in a readily accessible tissue[4, 42], facilitating the use of DNA methylation as a diagnostic tool for T2D[46]. Most of the candidate gene studies in blood have been conducted using a retrospective case-control study design.

One of the first candidate gene studies in blood was reported by Zou *et al.*[47], who used bisulphite sequencing PCR to analyse methylation signals detected at the promoter region of the PRKCZ gene using a MeDIP-Chip array. The authors identified that hypermethylation of PRKCZ was associated with T2D[47], and it was inversely correlated with serum levels of PRKCZ transcripts[47]. PRKCZ encodes for a regulatory molecule in the insulin signalling pathway, and it has been recognized as an important contributor to the pathogenesis of T2D[47]. Evidence of differential methylation in PRKCZ in response to T2D, furtherly supports this concept.

Another candidate gene study in blood was conducted by Gu *et al.*[48] using bisulphite pyrosequencing to investigate difference in methylation at the *IGFBP-1* gene between T2D cases and controls. The *IGFBP-1* gene encodes for the insulin-like growth factor binding protein-1, a protein produced in the liver which downregulation has been associated with increased risk of insulin resistance and T2D[48]. In this study, it was demonstrated that methylation of *IGFBP-1* was higher in established cases of T2D and newly diagnosed participants, compared to age-matched control men[48]. In addition, hypermethylation was related to lower expression of *IGFBP-1* in T2D cases versus controls[48]. Thus, methylation of *IGFBP-1* provided further support of the involvement of this gene in the pathogenesis

of T2D. In a second study by Gu *et al*.[49], the authors investigated difference in methylation in relation to T2D at the promoter of the *IGFBP-7* gene, which has been linked to insulin resistance and T2D[49]. Results showed that newly diagnosed T2D cases were hypermethylated at specific CpG sites in *IGFBP-7* compared to controls, but there was no correlation between methylation and serum levels of *IGFBP-7* [49]. Thus, DNA methylation in peripheral blood was able to capture differential risk of T2D associated with *IGFBP-7*.

The study by Canivell *et al*.[50] provided another example of a well-known candidate gene in T2D with underlying variation in methylation between newly diagnosed T2D cases (without pharmacological treatment) and controls. In this study, differences in methylation at 22 sites located in the promoter region of the *TCF7L2* gene, a protein implicated in β-cell function, were analysed using the EpiTYPER system[50]. Difference in methylation was identified in 13 out of 22 CpG sites analysed after adjustment for covariates and correction for multiple testing[50]. Furthermore, some of these CpG sites were associated with metabolic traits including FG, total cholesterol and LDL-cholesterol[50]. Thus, it was demonstrated that differences in methylation of *TCF7L2*, an important gene in the pathogenesis of T2D, could be reflected in peripheral blood DNA[50].

Mitochondrial disfunction has a role in impaired insulin secretion from β-cells, insulin resistance, and its associated complications (i.e. obesity, T2D)[51, 52]. In line with this concept, Gemma and colleagues undertook a candidate gene study to determine the association between difference in methylation at the promoter region of the *TFAM* gene, and the metabolic syndrome in a sample of young adults [52]. *TFAM* encodes for the mitochondrial transcription factor A, a protein that regulates mitochondrial DNA replication and transcription[52]. Results showed that the promoter region of *TFAM* was hypomethylated, and that there was an inverse correlation between levels of methylation in *TFAM*, and insulin resistance status measured by fasting insulin, FG levels, or by HOMA-IR index[52].

1.4.1.2 Studies in pancreatic islets

Most of these studies were conducted using a MassARRAY EpiTYPER assay or sequencing of bisulphite treated DNA[53], and a retrospective case-control study design. From the list of candidate genes in T2D with influence in β-cell function and cellular metabolism, the promoters of *INS*[54] (encoding insulin), *PDX1*[55] (encoding the transcription factor important for the development of the pancreas and function of mature β-cells), *PPARGC1A*[56] (encoding the mitochondrial regulator PGC1α) and *GLP1R*[57] (encoding the GLP1 receptor which enhances insulin secretion and protects β-cell mass), were found hypermethylated in islets of T2D donors compared to controls. Hypermethylation

of these loci was also associated with reduced expression of the same genes in islets of T2D donors, and with high HbA1c levels, indicating some level of β-cell disfunction in T2D[53].

1.4.1.3 Studies in skeletal muscle

In the study by Kulkarni *et al*.[58], bisulphite sequencing was used to type DNA methylation of genes related with mitochondrial enzymes, in skeletal muscle biopsies from 33 T2D cases and 79 controls[58]. The authors identified that T2D cases were hypomethylated at the promoter region of *PDK4* compared to controls, and that hypomethylation was correlated with increased gene expression of this gene. In addition, increased expression of *PDK4* correlated positively with some metabolic traits including BMI, HbA1c, C peptide, FG and insulin[58]. *PDK4* encodes for a kinase involved in the metabolism of glucose and fatty acids[58], and differential regulation of this locus in association with T2D, indicates some level of impairment in the utilization of these substrates in diabetes patients[58].

1.4.2 Previous epigenome-wide association studies in T2D

Epigenome-wide association studies (EWAS) are analogous to GWAS, but in the former the systematic analysis is done over the epigenome, commonly using DNA methylation as the epigenetic marker to identify differences in disease predisposition and progression among individuals from large population-based studies[41, 42]. The increase in the number of EWAS in T2D has been aided by the availability of array technologies, which nowadays incorporate similar resolution and coverage as the latest genotyping arrays from GWAS (>500K SNPs)[41]. Despite technological advances, sample size is still a constraint in epigenome-wide studies because DNA methylation is a tissue-specific marker, with dynamic changes over time[42]. The above means that samples included in epigenome-wide studies need to come from the same tissue, with cases and controls matched temporally to avoid spurious associations[41, 42]. One approach to surpass sample size limitations and improve the strength of EWAS findings, is to replicate the EWAS in comparable samples, and to summarize evidence via well-conducted meta-analyses[42]. This section provides a description of the different EWAS in T2D conducted to date, using different tissues and study designs.

1.4.2.1 EWAS in peripheral blood

As mentioned before, peripheral blood is a more accessible tissue than internal primary target tissues in T2D [4, 42], and it represents the most widely used source of DNA in large scale epidemiological studies[59]. Thus, blood is an important tissue for identifying predictive and/or diagnostic biomarkers of T2D[51]. One important factor to be considered in blood-based EWAS is cellular composition, as it has been shown that top signals in these EWASs can be related to cellular heterogeneity induced by inflammation and immune response, rather than to T2D itself [46].

Therefore, it is necessary to account for differential cellular composition between samples to avoid spurious associations[46]. Various blood-based EWAS in T2D have been conducted to date in the context of retrospective case-control studies or longitudinal studies, and from them some of the most common differentially methylated sites have been identified at the *TXNIP, ABCG1, CPT1A* and *SREBF1* loci[46].

One of the first EWAS in T2D was conducted by Toperoff *et al*.[60] using unrelated age-matched T2D cases (n=710) and controls (n=459) from an Ashkenazi Jewish population. Methylation signals were measured using an array-based method, followed by sequencing of bisulphite converted DNA pools[60]. As a result, Toperoff *et al*.[60] identified 6 DMRs associated with T2D, which were annotated to well-known candidate genetic loci for T2D (*CENT2, FTO, KCNJ11, TCF7L2* and *WFS1*). CpG sites identified within significant DMRs were then replicated, resulting in 13 CpG sites with strong methylation difference between pooled T2D cases and controls[60]. These 13 CpG sites mapped within introns of the *THADA, JAZF1, TCF7L2, KCNQ1* and *FTO* genes, and within the 3'-UTR of *SLC30A8*[60]. From these signals, the strongest association was detected at the CpG site in *FTO* (position 52,366,689bp), where hypomethylation was associated with 6.1% higher risk of T2D[60]. In addition, Toperoff *et al*.[60] utilized an independent prospective cohort study to determine if replicated methylation signals can be used as predictors of T2D. From this second analysis, they established that hypomethylation of analysed sites was an early marker of T2D[60], identifying for the first time differentially methylated CpG sites that predicted T2D and that were outside the promoter region of genes[60].

Another regional analysis of differential methylation in T2D was conducted by Yuan *et al*.[61], who identified DMRs associated with T2D using a twin study, and DNA methylation measured by immunoprecipitation sequencing (MeDIP-seq). DMRs were initially identified in 27 T2D-discordant and concordant twin pairs, followed by an analysis of top T2D-associated DMRs only in the subset of 17 T2D-discordant twin pairs to obtain genetic-independent DMRs[61]. The top T2D-associated DMRs were mostly hypermethylated, and some of them were annotated to T2D loci from GWAS (*THADA, FTO, IRS1, ADAMTS9, SLC30A8* and *KCNJ11*)[61]. The strongest genetic-independent T2D-associated DMRs were then replicated in a second study using 263 unrelated T2D cases and controls[61]. The successfully replicated DMR was annotated to the *MALT1* gene, which is involved in insulin and glycaemic pathways. Other DMR's that were specific to T2D-discordant twins were in the *GPR61* and *PRKCB* genes[61].

Looking at identifying predictive DNA methylation markers for T2D, a recent nested case-control study conducted by Chambers *et al*.[62] reported five CpG sites in peripheral blood that predicted future risk of T2D among Indian Asians. In comparison to previous blood based EWAS in T2D, this study had a large sample size (n=3,805) and included incident cases of T2D. Furthermore, by using a multi-ethnic study, Chambers *et al*. were able to compare methylation risk factors between Indian Asians and Europeans, and to try to explain observed ethnic differences in the risk of T2D. CpG sites identified in association with incident T2D were detected at the *TXNIP* gene (cg19693031); *SREBF1* gene (cg11024682); *PHOSPHO1* gene (cg02650017); *SOCS3* gene (cg18181703) and *ABCG1* gene (cg06500161)[62]. Relative to a 1% increase in methylation, these markers were associated with an increased (*SREBF1* and *ABCG1*) or a decreased (*TXNIP, PHOSPHO1* and *SOCS3*) relative risk of incident T2D[62]. Methylation markers detected in Indian Asians were successfully replicated in Europeans of the same prospective study[62]. Replicated CpG sites were then combined in a score, which was associated with higher relative risk of incident T2D in Indian Asians after adjustment for established risk factors[62]. A higher magnitude of the score in Indian Asians compared to Europeans, was indicative of the higher susceptibility of this population to T2D [62]. Overall, the study by Chambers *et al*.[62] was able to provide predictive DNA methylation markers for T2D that helped to understand pathways to disease, and to explain underlying differences in the susceptibility to T2D between Indian Asians and Europeans.

One potential weakness of the study by Chambers *et al.* was the advanced age at which participants were recruited at baseline (48-59 years), and the short follow-up period (8.5 years). This meant that at baseline, some participants who were classified as normoglycemic and non-diabetic, could have had prediabetes and been already manifesting changes in methylation characteristic of the disease. Therefore, it is likely that some of the methylation markers detected by Chambers as predictive or causal of T2D, coincide with CpG sites identified in other studies in association with prevalent T2D.

To determine if predictive DNA methylation markers for T2D detected by Chambers *et al*. could be replicated in an independent prospective study, Dayeh *et al*.[63] conducted a replication analysis of CpG sites previously associated with incident T2D using European samples. DNA methylation was measured at baseline in normoglycemic middle-age participants from the Botnia prospective family-based study, achieving replication for associations reported at the *ABCG1* and *PHOSPHO1* loci[63]. No replication was detected for associations at the *SOCS3, TXNIP* and *SREBF1* loci[63].

Dayeh *et al*. were able to confirm that increased methylation in *ABCG1* was associated with higher future risk of T2D, while increased methylation in *PHOSPHO1* was associated with a protective effect against future risk of T2D[63]. In addition, Dayeh *et al*. measured the correlation between DNA methylation at the replicated CpG sites and metabolic factors. They found that methylation at *ABCG1* was positively correlated with BMI, triglycerides, HbA1c and fasting insulin[63], while methylation at *PHOSPHO1* was positively correlated with HDL levels[63]. Going one step further, this study evaluated if there was consistency between DNA methylation measured in blood, and DNA methylation measured in primary target tissues for T2D at the replicated loci. Consistent with findings in blood, there was increased methylation of *ABCG1* in adipose tissue, and decreased methylation of *PHOSPHO1* in skeletal muscle of T2D-discordant MZ twins[63]. Lastly, it was evaluated if methylation at the CpG sites in *ABCG1* and *PHOSPHO1* correlated with differential gene expression between T2D cases and controls. As a result, it was identified that methylation of *ABCG1* in muscle was inversely correlated with gene expression in the same tissue[63], but no correlation was identified in blood. For *PHOSPHO1*, no correlation was identified between methylation and gene expression in any of the tissues evaluated[63].

In contrast to Chambers *et al*. and Dayeh *et al*. who investigated methylation markers associated with incident T2D, Kulkarni *et al*.[64] used a cross-sectional family-based study with participants from Mexican-American origin, to identify methylation sites associated with prevalent T2D and some glycaemic traits[64]. In total, 51 CpG sites were detected in strong association with T2D, another 19 were associated with FG, and 24 CpG sites more were associated with HOMA-IR[64]. From the list of 51 CpG sites strongly associated with T2D, five of them were able to explain 7.8% of the heritability of T2D[64]. These five CpG sites were annotated to three well-known epigenetic loci for T2D: *ABCG1* (cg06500161), *TXNIP* (cg19693031) and *SAMD12* (cg07960624), while the remaining two sites (cg25217710 and cg08309687) were in intergenic regions. As expected, some of the methylation loci for incident T2D reported by Chambers *et al*. (i.e. *ABCG1* and *TXNIP*), were also captured in the cross-sectional study by Kulkarni *et al*.[64]. Additional CpG sites associated with T2D were identified at the *SREBF1, LOXL2* (cg24531955), *CPT1A* (cg17058475), *SOCS3* (cg10508317), *CALHM1* (cg26712428), *ICA1* (cg26804423), *ZBTB7A* (cg04727071), *CUX1*, *NFE2L3* (cg21699330) and *LDLRAP1* (cg04344749) genes[64]. Pathway analysis suggested that these genes were related with insulin signalling and lipid transport pathways, which are processes generally dysregulated in participants with T2D[64]. All in all, Kulkarni *et al*. were able to provide novel epigenetic indicators of T2D in a population at high risk of the disease as it is Mexican Americans, aiming to utilize these markers in the detection and treatment of T2D.

Following studies by Chambers *et al*. and Kulkarni *et al*., numerous exploratory EWAS in T2D have continued to be conducted to date, some of them using European samples, and others using different populations at high risk of T2D. Soriano-Tárraga *et al*.[65] provided evidence of one of the first EWAS of prevalent T2D using Europeans in the discovery sample. Participants included in this study were older adults with prevalent T2D and a history of ischaemic stroke[65]. Results showed a strong association between hypomethylation at the CpG site cg19693031 in *TXNIP* and T2D[65], confirming the importance of this marker in the pathogenesis of T2D not only in Europeans, but also in Indian Asians and Mexican-Americans. Additional CpG sites were detected at the *POR* (cg01676795) and *PFKFB3* (cg26262157) genes, and at the intergenic CpG site cg07805383 (chromosome 2)[65]. Except for the association at *TXNIP*, none of the remaining CpG sites were replicated in two additional independent European cohorts from Spain[65]. Furthermore, the authors identified a strong inverse correlation between methylation at *TXNIP* and levels of HbA1c, suggesting that sustained hyperglycaemia may be one of the factors driving hypomethylation of the CpG in *TXNIP* [65]. Because there are reports of the sensitivity of *TXNIP* expression to glucose concentrations, it was suggested that DNA methylation was the mediating mechanism[65]. In conclusion, Soriano-Tárraga *et al*. proposed that methylation at *TXNIP* was a potential early biomarker of impaired glucose homeostasis[65].

Work by Florath *et al.*[66] furtherly supported the role of DNA methylation at the *TXNIP* gene in T2D by using a population-based cohort of older adults from Germany (ESTHER study). These participants were analysed for genome-wide differences in DNA methylation using the 450k array, identifying 39 differentially methylated CpG sites associated with prevalent T2D at FDR of 5% [66]. Replication of strong signals detected in the discovery sample was attempted in a second sub-cohort from the same study, identifying strong replication only for the association detected at the CpG in *TXNIP* [66]. Similar to what was found before by Soriano-Tárraga *et al*.[65], it was demonstrated that the level of methylation at the CpG in *TXNIP* was inversely associated with the concentration of fasting glucose and HbA1c based on a dose-response analysis[66]. As an additional finding, Florath *et al*. determined that reduction in median values of methylation at *TXNIP* was around 5% lower in T2D patients with poor glycaemic control (HbA1c ≥ 7%), compared to participants free from T2D[66].

Exploring epigenetic risk factors for T2D in other populations different from Europeans is important because DNA methylation is strongly influenced by the environment[67], meaning that generalisability of top associations between populations is not always true[67]. In line with this concept, Al Muftah *et al*.[67] performed the first EWAS in T2D and BMI in an Arab population using a cross-sectional study in

samples from 15 families of Qatari descent (n=123). In this study, 30 T2D cases and 93 controls were analysed for peripheral blood DNA methylation using the 450k array, identifying a novel CpG site associated with prevalent T2D at the *DQX1* (cg06721411) gene[67]. This association was furtherly replicated in an independent sample of 810 female twins discordant for T2D from the TwinsUK study[67]. In addition to conducting these EWASs, Al Muftah *et al*. aimed to replicate in the Qatari sample 47 CpG sites previously identified in association with T2D and BMI in European samples. Successful results were obtained at replicating the CpG site cg19693031 in *TXNIP* associated with T2D, and at replicating other seven CpG sites associated with BMI[67]. Replicated epigenetic markers in the Qatari sample were then investigated in TwinsUK, and results were combined across studies via meta-analysis obtaining stronger associations but higher heterogeneity[67]. Higher heterogeneity across studies was attributed to differences in the underlying mechanisms, which might depend on genetic background and environmental pressure[67]. In addition, higher heterogeneity might result from differences in sample-size and power between the two studies. In this sense, results from the meta-analysis could have been more influenced by effect estimates detected in TwinsUK compared to the Qatari sample, which was smaller and probably more underpowered.

Another example of an EWAS of prevalent T2D conducted in a non-European population was recently published by Meeks *et al*.[68]. In this study, global DNA methylation from peripheral blood was measured in a subset of participants from the RODAM study, a cross-sectional population-based cohort from Ghana[68]. The aim of this study was to identify novel differentially methylated sites associated with T2D and with HbA1c among Ghanaians, and to replicate in this population some of the top signals previously identified in European samples[68]. After adjustment for common covariates and correction for inflation and multiple testing, Meeks and colleagues identified associations between DNA methylation and T2D at four loci: *TXNIP* (cg19693031), *C7orf50* (cg04816311), *CPT1A* (cg00574958) and *TPM4* (cg07988171), which together explained 25% of the variance in T2D[68]. The strongest association with T2D was reported at the *TXNIP* locus, and associations at the *TXNIP, TPM4* and *C7orf50* loci surpassed further correction for BMI[68].

Methylation at the CpG site in *TPM4* was reported as a novel epigenetic marker for T2D in the Ghanaian sample[68], whereas the remaining epigenetic associations were ubiquitous across different populations. *TXNIP* was also strongly associated with HbA1c, showing that a 0.1µl/mol increase in HbA1c was associated with an average 0.15% lower methylation of the CpG in *TXNIP* [68]. Furthermore, Meeks *et al*. performed a DMR analysis identifying a strong association between a DMR in chromosome 2, which was annotated to the *GDF7* gene, and T2D[68]. For this DMR, T2D cases

were hypomethylated compared to controls. Nine CpG sites composed the DMR in *GDF7*, and using average methylation beta-values from these CpG sites, it was estimated that an average of 1% increase in methylation was associated with 4.37 higher risk of T2D[68]. Total variance in T2D explained by the DMR in *GDF7* was 1.2%[68].

Considering the large number of epigenetic markers so far detected in association with incident and prevalent T2D across tissues and populations, Walaszczyk *et al*. aimed to identify the most significant associations from these EWAS, and to evaluate their replicability in an independent sample[46]. To accomplish this, Walaszczyk and colleagues performed a systematic literature search of all the EWAS in T2D and glycaemic traits published to date across populations and tissues, and selected 100 CpG sites with study-specific Bonferroni significance[46]. These 100 CpG sites were then tested for replication in a case-control subsample from the LIFELINES prospective population-based study from the Netherlands. This subsample was composed by 100 T2D cases and 100 age- and sex-matched controls, all unrelated and European ancestry samples[46]. Within the diabetic group, half of the participants had CVD complications. Based on peripheral blood DNA methylation measures, Walaszczyk and colleagues were able to replicate, with stringent multiple-testing correction, five (*ABCG1, LOXL2, TXNIP, SLC1A5* and *SREBF1*) out of 52 CpG sites previously reported in association with T2D in blood[46]. Two CpG sites associated with fasting glucose (*ABCG1* and *CCDC57*) were replicated at nominal significance ($p<0.05$) in a subset of healthy control individuals, but they were unable to replicate CpG sites associated with HbA1c[46]. It was of particular interest that none of the signals detected in internal primary tissues for T2D (i.e. pancreas, adipose tissue, liver), could be replicated in blood, meaning that these markers were tissue specific. Considering that some of the blood-based CpG sites associated with T2D were successfully replicated, independent of the population where they were originally identified, the authors concluded that there was promising evidence of the clinical use of these CpG sites as biomarkers of T2D[46].

So far, epigenetic studies reviewed in peripheral blood have been based on the use of DNA methylation from mixed white blood cells, without considering cell-type specific variations in DNA methylation. Because immune cells participate in the low-grade inflammation processes observed in obesity, insulin resistance and T2D[69], it is likely that cell-type specific differences in DNA methylation show association with these metabolic disturbances. Evidence of this was provided by Simar and colleagues, who identified hypermethylation of B-cells and natural killer cells in T2D cases compared to controls[69], and this difference in methylation correlated positively with insulin resistance[69]. This study concluded that DNA methylation provided a link between immune function and metabolic

disorders, and emphasized the need for additional cell type specific studies of DNA methylation in T2D[69].

1.4.2.2 EWAS in adipose tissue

Adipose tissue is an important regulator of energy balance in the body due to its capacity to accumulate excess of energy in the form of fat, and to release lipids[53]. It also works as a secretory organ of adipokines such as adiponectin, leptin and tumour necrosis factor alpha (TNFα), which control food intake, insulin sensitivity and metabolism[53]. When there is excess of fuel, the accumulation of fat will be directed towards important insulin-sensitive organs such as the liver, muscle and pancreas, causing in the long-term, insulin resistance and disfunction[53]. Thus, it is important to study how obesity, other risk factors for T2D, and T2D itself, can disturb the methylation profile of the adipose tissue[53].

Most EWAS in adipose tissue have followed a retrospective case-control study design. In studies using monozygotic (MZ) twin pairs discordant for T2D[70, 71], a different methylation pattern was observed in twins with T2D in various genes (*ZNF668, HSPA2, C8orf31, CD320, SFT2D3, TWIST1* and *MYO5A*)[70]. However, observed methylation differences were modest, suggesting that there was still a high genetic component in the heritability of methylation in adipose tissue in relation to T2D[70, 71]. When the same analysis was done in unrelated participants with and without T2D, 15,627 CpG sites were found differentially methylated in T2D participants, and they were annotated to 7,046 genes[71], some of them matching with GWAS loci for T2D (*PPARG*, *KCNQ1*, *TCF7L2*, and *IRS1*)[71]. In agreement with findings in blood, in adipose tissue of MZ twins discordant for T2D it was identified that established T2D was associated with hypermethylation of the widely reported CpG site cg06500161 in the *ABCG1* gene[63]. This gene encodes for a transporter protein of importance in the metabolism of lipids[53], and differential methylation at this site has been regarded as a predictor of T2D according to studies in peripheral blood[62, 63].

Other studies have implicated DNA methyltransferase enzymes (DNA methyl group donors) in the development of insulin resistance in adipose tissue. An example is *DNMT3A*, a *de novo* methyltransferase that has been demonstrated to target methylation of the *fgf21* gene based on *in vitro* experiments in adipocytes in mice[72]. In addition, hypermethylation of *FGF21* was identified in individuals with T2D, and this was negatively correlated with *FGF21* expression[72]. Because increased expression of *fgf21* in cultured adipose cells reversed insulin resistance[72], *dnmt3a* was described as a novel epigenetic mediator of insulin resistance in adipose tissue[72].

Obesity, which is an established risk factor for T2D, also influences differences in methylation in adipose tissue. From EWAS using obesity markers (i.e. BMI, fat distribution) as the exposure[73-75], changes in methylation in adipose tissue have been commonly identified at *HIF3A*, a transcription factor that regulates different adaptive responses to low oxygen tension[53]. In addition, differences in methylation in response to obesity have been identified in target genes for T2D, including *FTO*, *TCF7L2* and *IRS1* based on work conducted by Rönn *et al*.[73]. These same T2D loci were identified in an EWAS of T2D in adipose tissue conducted by Nilsson *et al*.[71]. Similarity in the signals detected between obesity and T2D, suggests that obesity can determine changes in methylation before the onset of T2D, perhaps affecting mechanisms that in turn lead to the disease[53].
Exercise is also capable of inducing changes in methylation in adipose tissue, as was demonstrated by an intervention study conducted by Rönn *et al*.[76]. In this study, healthy participants who were initially sedentary, with or without a family history of T2D, were subjected to some level of exercise for six months. Methylation was then compared before and after the exercise intervention, identifying differences in methylation at 17,975 CpG sites including 7,663 genes, and one third of these genes also showed differences in gene expression (i.e. *RALPB1*, *HDAC4* and *NCOR2*)[76]. Furthermore, among differentially methylated loci, Rönn *et al*. identified 18 obesity loci and 21 T2D loci, including *TCF7L2* and *KCNQ1*[76].

Diet also affects DNA methylation in adipose tissue. Gillberg *et al*. showed differences in methylation in 652 CpG sites before and after exposing healthy men to a high fat diet for 5 days[77]. Some of the CpG sites were annotated to the *CDK5*, *CIDEA*, *IGFBP5* and *SLC2A4* genes, which are related with adipose tissue metabolism or differentiation[77]. For genes with difference in methylation and gene expression after the dietary intervention, most were enriched in pathways related to oxidative phosphorylation and insulin signalling[77]. Moreover, diet content (saturated versus polyunsaturated fatty acids), fasting and weight loss (through diet or gastric bypass surgery), are all factors that influence differences in DNA methylation in adipose tissue[53].

1.4.2.3 EWAS in pancreatic islets

Pancreatic islets are one of the major regulators of the levels of glucose in the body through the secretion of insulin and glucagon[53]. T2D develops when there is lower secretion of insulin from β-cells, and an enhanced secretion of glucagon from α-cells during the fed state, leading to hyperglycaemia[53]. All the EWAS in T2D conducted in pancreatic islets have used a retrospective case-control study design.

Volkmar *et al.* implemented the Infinium HumanMethylation27 BeadChip array (27k array, 27,578 probes) to determine differences in methylation between pancreatic islet samples from five T2D donors and eleven controls[78]. The authors detected methylation differences in 276 CpG sites (96% hypomethylated) and 254 genes related with altered β-cell function and survival under normal and stress conditions[78]. Expanding on the number of methylation sites previously enquired by Volkmar and colleagues, Dayeh *et al*.[79] used the Infinium 450k array to compare methylation of islets between 12 T2D donors and 34 controls. In this study, Dayeh *et al*. identified over 3,000 differentially methylated sites located near or in 853 genes. Almost half of the CpG sites identified showed methylation differences >5% [79]. In addition, some of the genes mapping to differentially methylated sites were newly identified in T2D (*CDKN1A, PDE7B, EXOC3L2* and *HDAC7*), while others coincided with already known GWAS loci for T2D, including *ADAMTS9, ADCY5, FTO, HHEX, HNF1B, IRS1, JAZF1, KCNQ1, TCF7L2, THADA, VEGFA* and *EGF*[79]. A gene enrichment analysis revealed that genes annotated to T2D-associated CpG sites, were related with impaired β-cell function, while expression data showed that some of the changes in methylation were correlated with altered mRNA expression[79]. In the case of the *HDAC7* gene, encoding for a histone deacetylase, this was hypomethylated and overexpressed in islets from T2D donors[79]. Overexpression of Hdac7 in animal models and clonal β-cells, resulted in impaired mitochondrial function and insulin secretion[80]. This finding indicated that changes in methylation at *HDAC7* were involved in β-cell disfunction, a characteristic of T2D[53].

The most recent study evaluating difference in DNA methylation in pancreatic islets was conducted by Volkov *et al*.[81]. In this study, whole-genome bisulphite sequencing was used to identify DMRs associated with T2D. Around 26,000 DMRs were identified in association with T2D, the average size for a DMR was 414bp (6bp-3411bp), they were located in important genes related with β-cell function, and within known GWAS loci for T2D including *TCF7L2*, *THADA* and *KCNQ1*[81]. DMRs were also enriched in binding sites for transcription factors, indicating their role in regulating islet function[81]. Some of the genes annotated to the DMRs showed altered methylation and gene expression in T2D islets, including among these *NR4A3*, *PARK2*, *PID1*, and *SOCS2* loci[81]. When changes in gene expression of the above genes were induced in clonal β-cells, the result was an impairment in glucose-induced insulin secretion[81], providing further support that changes in DNA methylation underlie important mechanisms for the onset of T2D.

1.4.2.4 EWAS in skeletal muscle

Skeletal muscle is an important tissue to understand disease mechanisms in T2D because it is responsible for the majority of the insulin-induced uptake of glucose in the body, and for the ability

to perform exercise[53]. Epigenetic studies in T2D in skeletal muscle have been undertaken using biopsies and myocyte cell cultures from T2D donors and controls, and implementing various technologies including MeDIP-Chip arrays (targeting promoter regions)[82, 83] and the 27k array[70].

Studies using MZ twins discordant for T2D have identified differences in methylation in the *IL8* and *PPARGC1A* genes[70], *LINE1* elements, and in *HOX* genes[82], based on methylation measured in skeletal muscle samples. As part of an exercise intervention study conducted by Nitert *et al*.[82], the authors identified variation in methylation in the promoter of 65 genes when comparing sedentary and unrelated healthy participants with and without a first-degree relative with T2D[82]. Some of the genes identified with differential methylation were *MAPK1*, *MYO18B*, *HOXC6*, and *PRKAB1*[82], and most of these genes were hypomethylated (i.e. 60/65 genes) in people with FH of T2D. In total, 40% of all differentially methylated genes were validated in a study of MZ twins discordant for T2D[82], indicating that previously identified genes by Nitert *et al*. play an important role in the development of T2D[82]. Differentially methylated loci were related to the mitogen-activated protein kinase (MAPK) signalling pathway, and to the insulin, calcium, sphingolipid and adipocytokine pathways[82].

Another study using DNA methylation from skeletal muscle in unrelated participants with and without T2D, showed a strong association between T2D and hypermethylation of a CpG site in the promoter of the *PPARGC1A* gene[83]. Hypermethylation of *PPARGC1A* was accompanied by lower transcription of *PPARGC1A* and reduced mitochondrial DNA in T2D cases compared to controls[83]. Because *PPARGC1A* is a transcriptional coactivator of genes involved in energy metabolism[31], lower activity of *PPARGC1A* in T2D cases could be associated with decreased mitochondrial content and mitochondrial disfunction[31].

1.4.2.5 EWAS in the liver

The liver is an important organ that regulates levels of glucose in the body during the fed and the fasting states[53]. The liver responds to insulin stimulation by accumulating glycogen during the fed state, while during the fasting state it responds to increasing levels of glucagon by stimulating the internal production of glucose via glycogenolysis and gluconeogenesis[53]. Two main EWAS in T2D have been conducted to date using liver samples. The study designed implemented has been a retrospective case-control study, with the 450k array as the primary method for methylation profiling.

In a study by Nilsson *et al*. [84], the authors reported 251 CpG sites differentially methylated in T2D cases compared to controls, and most of these sites were hypomethylated[84]. Hypomethylation in T2D cases was potentially the result of reduced levels of erythrocyte folate, a dietary donor of methyl groups, in these participants[84]. Some of the genes with differential methylation in the liver were also known previously from GWAS in T2D, including *GRB10*, *ABCC3*, *MOGAT1* and *PRDM16*[84]. Two of the genes with differential methylation and expression in the liver were *H19* and *RIPK4*, with *RIPK4* also related with decreased insulin sensitivity in the liver[85].

A study by Kirchner *et al*. further confirmed that hypomethylation of differentially methylated sites in the liver was a characteristic of participants with T2D[86]. Most of the genes with differential methylation were related to the metabolism of glucose and lipids (*PRKCE, ABR, PDGFA, ARHGEF16, ADCY6, RPS6KA1, CTBP1, CCND1* and *WNT11*)[86], and with genes of the ATF-motif regulatory site[86]. Furthermore, Kirchner *et al*. demonstrated that hypomethylation of *PRKCE* and increased expression of this transcript, might induce insulin resistance in the liver, possibly by interfering directly with the kinase activity of the insulin receptor[86].

1.4.2.6 EWAS in glycaemic traits

The interest in epigenetics for determining mechanisms in T2D development, has also expanded towards investigating differences in DNA methylation associated with glycaemic traits that are used in the diagnosis of diabetes. These traits include levels of fasting insulin, fasting glucose, HbA1c, and the HOMA-IR and HOMA-B indexes, which represent measures of insulin resistance and β-cell disfunction, respectively. In EWAS of glycaemic traits, the discovery sample was composed of normoglycemic non-diabetic participants to avoid capturing changes in methylation directly associated with T2D metabolic disturbances. The aim of pursuing these studies has been to contribute to identify risk loci for T2D, similar to what has been done in GWAS of glycaemic traits.

To date, EWAS in glycaemic traits have been conducted by Hidalgo *et al*.[87] (in fasting insulin, fasting glucose, and HOMA-IR), Kriebel *et al*.[88] (in fasting glucose, fasting insulin, 2-h PG, 2-h insulin, HbA1c and HOMA-IR), Kulkarni *et al*.[64] (in fasting glucose and HOMA-IR) and Rönn *et al*.[73] (in HbA1c). Most of these studies have been based on European samples, except for the study by Kulkarni *et al*.[64] that included samples of Mexican American origin. The target tissue in these studies has been peripheral blood, and different associations have been identified for each glycaemic trait, some of them overlapping across traits and with blood based T2D associations, including CpG sites in the *ABCG1, CPT1A, TXNIP* and *SREBF1* genes[46].

Summary of epigenetic studies in T2D

There is compelling evidence of widespread differences in DNA methylation associated with T2D and diabetes risk factors, and this evidence comes from accessible tissues such as peripheral blood, and from primary target tissues in T2D. Major genes affected by methylation changes are related with cellular metabolism, β-cell function, cellular signalling and inflammatory processes, among others. Epigenetic markers in T2D are more likely to be tissue specific as there is low overlap between markers across tissues. Even though blood does not represent a primary target tissue for T2D, its accessibility makes it one of the main sources of DNA in epidemiological studies, and a good candidate tissue for the identification of DNA methylation biomarkers in T2D.

In comparison to epigenetic studies in T2D using less accessible tissues, those conducted in peripheral blood are more common and have included larger samples, providing further support for the use of blood-based epigenetic markers in the detection and treatment of T2D. Despite all the progress achieved to date in detecting epigenetic biomarkers for T2D, further work is needed to demonstrate the replicability of these associations, to address their causality, and to identify novel associations.

At present, most of the EWAS in T2D have been conducted in retrospective case-control studies, which are more convenient than nested case-control studies as they provide larger samples and the possibility to adjust for existing confounders. However, a retrospective case-control study does not allow one to disentangle the cause-effect relationship between differences in DNA methylation and T2D. Thus, it is necessary to incorporate more prospective studies that help to determine the causal role of DNA methylation in the pathogenesis of T2D. Alternatively, evidence from the genetics of DNA methylation (meQTL studies) should be incorporated to reinforce causal inference analyses of epigenetics in T2D.

1.5 Important considerations in epigenetic epidemiology studies

1.5.1 Selection of Tissue

Epigenetic markers are tissue-specific signals, and variation of epigenetic patterns between tissues within the same individual are higher, than the variation between individuals[4, 42]. Ideally, markers associated with disease risk should be extracted from tissues directly influenced by the disease[42, 43]. However, it is difficult to collect enough samples from internal target tissues to identify disease markers with sufficient strength, or to obtain internal tissues from control samples for analytical

comparison with cases, or to collect internal tissues prior the onset of disease in healthy participants from a prospective epidemiological study[43]. Despite these limitations, there are other more accessible tissues like blood, saliva, buccal cells, urine, skin cells and faeces, which are easy to collect and process, and are convenient for large-scale population studies[42, 43]. Markers obtained from accessible tissues can be used as surrogate markers, but it is unlikely that similar associations are observed in target tissues for the disease [43], as the correlation between methylation patterns across tissues is complex and locus dependent[4].

When using blood as the source of epigenetic markers, it is important to consider cellular heterogeneity and to apply statistical correction for this to avoid capturing associations related with underlying immunological conditions, rather than with the disease outcome of interest[43]. Different algorithms exist to estimate cellular composition in peripheral blood using epigenomic data, or routine cell counts[42]. These strategies will help to adjust analyses and to obtain stronger disease-related DNAm associations[42].

Further knowledge of cross-tissue correlation is required to evaluate the extent to which accessible tissues can be used to obtain surrogate markers of disease from target internal tissues. To contribute to this knowledge, it is important to identify within the same individual, epigenetic marks at specific genomic regions for which measures in blood are informative of any other internal tissue[42]. In addition, it is necessary to distinguish intra-individual inter-tissue correlation due to true epigenetic variation, from inter-tissue correlation derived from genetic variants with influence on DNA methylation[42].

1.5.2 Selection of study design

The study design of choice to conduct an epigenetic study depends greatly on the research question, and on the availability of sufficient samples and phenotypic information to support the association analysis. There are different epigenetic study designs, including cross-sectional studies, retrospective case-control studies, cohort studies, nested case-control studies, intervention studies, family-based studies and birth cohorts[43]. From them, the most common are retrospective and nested case-control studies[43].

Nested case-control studies are appropriate to identify methylation biomarkers that predispose to disease, which can be used for early disease detection[43]. In this prospective study design, methylation is measured before the onset of disease in healthy participants from a cohort study, and

some of them are selected again at any time during the follow-up period in regards to disease onset[43]. For comparison, controls are appropriately selected from participants who remained free from disease in the cohort during the follow-up. Due to the temporal distance between sample collection and disease onset, the epigenetic state reflected in the biospecimen preceded disease and is not influenced by disease[43]. In comparison to nested case-control studies, in longitudinal studies the follow-up period includes different timepoints where frequent biospecimens and phenotype data are collected to allow comparison of lifecourse changes in methylation, with temporal variation of risk factors and disease prevalence[41-43]. Although nested case-control studies and longitudinal studies have the advantage of providing robust biomarkers of disease predisposition, and of establishing the temporal origin and stability of disease-associated epigenetic marks[41], respectively, these studies are normally expensive to be maintained[41, 43].

Most commonly, epigenetic studies include a retrospective case-control study design[41], where methylation is measured at the same time of disease onset or diagnosis in cases, and in controls appropriately selected from the same study population[43]. Epigenetic markers detected by comparing methylation patterns between cases and controls are indicators of disease prevalence, but it is not possible to confidently assert a causal relationship as they might be influenced by the disease itself (reverse causation), post-disease processes, or disease treatment[41-43]. Cross-sectional studies are also common in epigenetic-epidemiology, where the aim is to characterize a methylation mark in a specific population, without performing case-control comparisons[43]. For instance, comparing the level of DNA methylation between older and young Europeans, or between men and women age 60-65 years.

Less common studies include intervention studies where the effect of a treatment like folate supplementation on the methylome is measured in the same sample before and after the intervention, or in independent samples by using a randomized controlled trial[43]; birth cohort studies, where epigenetic changes in the offspring are identified in response to pre-conceptional and prenatal exposures[42, 43]; family-based studies, where the aim is to identify transgenerational inheritance of epigenetic traits[41, 43]; disease-discordant monozygotic twins, where the contribution of epigenetics to disease risk is controlled for the effect of germline genetic variation[41], but to provide cause-effect markers, twin studies need to be conducted longitudinally[41, 42].

1.5.3 Covariates in epigenetic studies

Epigenetic markers are dynamic factors that can vary in response to the environment, meaning that conventional rules in the design of observational studies should be followed in epigenetic studies to avoid spurious associations and to ensure the reproducibility of results[42, 43]. These rules include the use of a representative study population with sufficiently large sample size, selection of cases and controls from the same study, matching samples by age and sex, and appropriate selection of covariates[42]. Variables with known effects in the methylome that need to be included as covariates in epigenetic studies are age, sex, smoking, cellular heterogeneity and batch effects[42], which are introduced by technical variation. Another covariate of relevance in epigenetic studies of T2D is BMI. BMI can influence the methylome directly, as it has been demonstrated by recent EWAS on this trait[89, 90], or indirectly through nutrition and a high-fat diet[4].

Even after applying statistical correction for common covariates, results can be still confounded by unmeasured exposures, demographic factors, or by the effect of the disease itself (i.e. reverse causation) or disease treatment[42]. Thus, the importance of applying causal inference methods in epigenetic studies, or selecting appropriate study designs, to determine causality in the association between DNAm and the exposure and/or the outcome of interest[4, 42].

1.5.4 Single site versus regional DNAm analysis

As mentioned before, difference in DNAm can be interrogated at the single CpG site or at the regional level (DMR), with pros and cons associated with either of these approaches. Usually, a genome-wide analysis starts by interrogating DNAm differences at the CpG site level, and this analysis is then reinforced or complemented by using DMR analysis. The single CpG site analysis is straightforward (i.e. generalized linear models, logistic regressions models, mixed and random-effect models, etc)[44], and it has the ability to capture variation in methylation in critical CpG sites with functional implications in transcription[42]. For instance, methylation of specific cytosines in transcription factor binding sites can reduce binding affinity and gene transcription[42]. CpG site analysis can also give an overview of the global methylation status[42]. Some of the disadvantages of the single CpG site analysis are the multiple-testing burden, more likelihood of identifying false-positives by batch effects, and the difficulty to link single-site methylation across the genome with functional biological implications[42].

DMR analysis requires more specific statistical methodologies than single CpG site analyses. Methods currently available for DMR detection can be classified between those that (a) do the

analysis at the CpG site level first and then draw statistical inferences about the regions, and those that (b) cluster CpG sites first to reduce dimensionality of the data, and then establish putative regions[44]. Each DMR method has its advantages and disadvantages, and a list of currently available methods in DMR analysis has been described by Breton *et al.*[44]. Compared to the single CpG site approach, the DMR analysis can be less affected by false positives caused by technical artefacts, and it also reduces multiple-testing burden and gain power by grouping together CpG sites within a similar genomic region[42]. The size of a DMR can vary from a few hundred to a few thousand bases, and this range coincides with the size of gene regulatory regions[44]. Furthermore, DMRs are more likely to be implicated in chromatin remodelling and transcriptional regulation processes, compared to single CpG sites[42]. This concept is supported by the observation that differences in DNAm between cells, in malignant tissues, and in response to prenatal environmental exposures, occur within genomic regions and not at specific CpG sites[42].

Disadvantages of using a DMR analysis are the lower consensus on the definition of a DMR, which generally depends on the methodology applied[44]. Also, DMR methodologies need to be further refined in terms of their performance, sensitivity, multiple-testing correction, and ability to establish the functional relevance of the identified regions[44]. Furthermore, DMR detection using single CpG site data from methylation arrays is problematic because the probe positions are sparse and most of the relevant data is missing due to the low genome coverage of array-based technologies[44].

1.5.5 Other methodological considerations

Apart from determining the type of methylation marker to be used, the tissue, the method for methylation typing and the study design, some other considerations are necessary in the design of epigenetic studies. One of them is the study population, in which case it is important to select a representative sample, so that results obtained can be generalized to the population of interest[43]. Another common limitation of any study is having a small sample size. In principle, high quality epigenetic epidemiological studies need sufficiently large samples, which can be estimated *a priori* using power calculators based on the effect size of interest[43]. Larger samples guarantee the validity and precision of results, and the ability to capture associations with moderate to modest effect size[43].

Because current population studies were generally not designed with a focus on epigenetics, there is commonly an imbalance between availability of phenotyping data and epigenetic information from sufficiently large samples that can support the validity of epigenetic associations[42]. To surpass

sample size constraints, avoid false positives and increase power, it is necessary to incorporate replication and meta-analysis of data across epigenetic epidemiological studies[42, 44]. Ideally, replication of an exposure/outcome-DNAm association should be performed in an independent but comparable sample, with similar measures of the exposure or outcome, and similar statistical modelling of the data[44]. Replication can also be achieved by splitting the same study population into a discovery and a replication sample. Generally, the replication sample is smaller than the discovery sample, but it maintains an adequate proportion of cases versus controls[44].

Replication and meta-analysis are facilitated by large consortia, where similar profiling methods for DNAm, and standardised protocols for data pre-processing and analysis, are used across studies[41, 42, 44]. Results generated by each study are combined via meta-analysis, increasing the sample size to the thousands, and improving power[44]. Despite the caveats of the 450K array, this method has been widely used across studies, and it is the best candidate technology for the establishment of large consortia with a focus on epigenetics[42]. Because the 450K array and other available typing methods still render some imprecision in the estimation of methylation, it is important to apply technical validation of top findings obtained in the replication and/or meta-analysis[42]. Generally, technical validation is performed on similar samples, but using different methods to estimate correlation coefficients between measurements across methods[44]. To reduce costs, methods implemented for technical validation tend to be locus specific[91], and some examples include EpiTYPER, locus-specific bisulphite sequencing, MethyLight and methylation-specific PCR (MSP)[91].

Other important considerations in epigenetic epidemiological studies are effect modification, confounding and misclassification. Effect modification occurs when the association between DNAm and the exposure and/or outcome of interest is differentially influenced by levels of other factors including age, sex and ethnicity[43]. If there is effect modification, associations with DNAm need to be stratified by levels of the modifying factor[43]. Confounding is an important phenomenon affecting the validity of results in epidemiological studies. In principle, the confounder should be associated with the exposure and outcome of interest, but it is not part of the causal pathway between the exposure and the outcome[43]. The effect of a confounder can be identified if after adjustment for the confounder the association of interest changes in terms of effect size, direction of effect, or statistical significance[43]. Therefore, identification and adjustment for confounders is necessary in epigenetic epidemiological studies to avoid spurious results. Lastly, misclassification of the methylation state of an epigenetic signal can also hinder the interpretation of results[43]. However, current advances in typing technologies favour precision in the estimation of methylation, and

bioinformatic resources help to distinguish true signals from background noise, reducing the chances of misclassification[43].

1.5.6 Functional interpretation of main findings

Once results of the EWAS are obtained, the next step is to understand the biological meaning of these findings with respect to the outcome and/or exposure of interest. Functional interpretation is important because most epigenetic findings have small effect sizes, in the range of 2-10% or even smaller[44], meaning that understanding the contribution of this variation to the exposure or disease of interest is not straightforward. Different methods can be used for functional analysis, most of them rely on the use of existing bioinformatic platforms, while others include *in vitro* and/or *in vivo* assays in human cells and animal models, respectively, to build up evidence on causality of EWAS findings.

Generally, the first approach is to characterize the genomic context of the region surrounding the CpG site or the DMR of interest to understand possible functional mechanisms leading to disease[42, 91]. This genomic characterization is important because DNAm does not act in isolation, as it can be influenced by the underlying genotype, and it is interrelated with other *cis* epigenetic marks (i.e. histone modifications, miRNA)[42, 91]. An example of bioinformatic tools that allow the visualization of the genomic context of CpG sites and DMRs are the UCSC genome browser, Ensembl, GREAT Webserver, Galaxy, HyperBrowser and EpiExplorer[91]. Another approach to determine biologically meaningful trends in EWAS findings, is gene enrichment analysis or pathway analysis[91]. This analysis is based on a list of genes annotated to CpG sites and DMRs of interest or based directly on the genomic regions. Using genomic regions rather than gene lists is considered a more convenient approach to obtain valid interpretation of results, because DNAm and gene expression are not completely correlated phenotypes[42, 92]. Therefore, it is incorrect to assume that the CpG site or the DMR of interest exert their function directly on the nearby gene, which is the one commonly used for genomic annotation.

Despite some caveats, assessing the correlation between DNAm and gene expression still represents one of the main functional inspections of EWAS results[42, 44, 92]. Once the *in silico* functional analysis has confirmed the biological relevance of epigenetic findings, a more detailed and causally-driven inspection might include *in vitro* reporter gene assays, *in vitro* experimental assays on human cell lines to study developmental and differentiation processes under controlled conditions, or *in vivo* validation of *in vitro* experiments using animal models[42].

1.6 Using genotype to understand causal epigenetic pathways in T2D

1.6.1 Causal inference analyses

The double nature of the epigenome as a heritable mark and as a plastic cellular phenotype responsive to disease-related environmental factors, makes it difficult to establish the causal role of EWAS findings in their association with the outcome of interest [91]. Several mechanisms explain how variation in methylation occurs prior to disease onset, either through transgenerational inheritance, stochastic changes during development, or in response to environmental triggers in adulthood or during the gestational period[41]. However, variation in methylation detected prior to disease is not always an indicator of causality, as the observed association can be influenced by unmeasured environmental or genetic confounders[41]. Because observational studies do not allow us to distinguish between causal and consequential epigenetic variation, causal inference methods based on Mendelian Randomization (MR) have been proposed as a good strategy to assess causality of EWAS findings[4, 42].

MR is a method for causal inference that uses germline genetic variation to establish the causal relationship between a modifiable exposure and a health-related outcome in observational epidemiology[4, 93, 94]. MR studies resemble randomized control trials, but instead of randomly allocating individuals to a treatment to avoid confounding and determine causality, MR studies use the random allocation of an individual's genotype before conception to make causal inferences in aetiological epidemiology[94]. Because genetic variants are fixed at conception, they are not influenced by behavioural, socioeconomic or physiological factors commonly affecting observational studies, or by the disease through reverse causation[4, 93, 94]. Furthermore, because the association between genetic variants and the modifiable exposure remains consistent throughout life, their use in causal inference avoids attenuation by error (regression dilution bias)[94]. These characteristics allow genetic variants to be used as instrumental variables to proxy levels of a modifiable exposure in MR studies[93], and if incorporated in appropriately sized samples, it is possible to infer causality by synthesising a "causal estimate"[33].

As with any other analytical method, MR has specific assumptions with respect to the selected instruments. Valid instruments in MR should be strongly associated with the exposure, but not with potential confounders of the exposure-outcome association to rule out pleiotropy[93, 94]. Furthermore, instruments should not be directly associated with the outcome (i.e. independence assumption), only through the exposure of interest[94]. Sources of genetic variants to conduct MR studies are well-powered GWAS, methylation quantitative trait loci (meQTL) studies and expression quantitative trait

loci (eQTL) studies. These studies focus on identifying genetic variants strongly associated with metabolic, anthropometric, behavioural and disease phenotypes in GWAS, with DNAm in meQTL studies (see section 1.6.3 "Genetics of DNA methylation"), and with differential expression in eQTL studies. To date, the two largest consortia for the study of the genetics of T2D and related glycaemic traits, are the Diabetes Genetics Replication and Meta-analysis (DIAGRAM), and the Meta-analysis of Glucose and Insulin-related traits consortium (MAGIC).

Depending on the source of data used to derive estimates of the association between the genotype, the modifiable exposure and the outcome, the MR approach can be a single sample MR or a two sample MR[93]. In single sample MR, estimates are derived from a single sample where individual level data is required for the analyses. In comparison, in two sample MR the genotype-exposure and genotype-outcome associations are retrieved from two independent but comparable populations using summary data[93]. Generally, two sample MR has more power to infer causality compared to single sample MR, but if there is overlap between samples, results of two sample MR can be biased towards the observational exposure-outcome association[93]. Different methods continue to be developed in MR studies, some of them allow to test for genetic pleiotropy when using multiple instruments, including the median-based method and MR-Egger[93].

Another important aspect of an MR study is the number of variants used to instrument the modifiable exposure. Some studies include a single variant with the strongest effect on the exposure and well-known function, where core assumptions are supported by biological knowledge[93]. However, most of the times a single variant only explains a small proportion of the total variation in the phenotype (i.e. weak instrument), requiring multiple variants to increase the proportion of variation captured, and the statistical power to detect causality[93]. The effect of multiple variants can be studied individually using a statistical regression model in a single sample or a two sample MR analysis[93], or this effect can be combined in a genetic risk score[93]. The genetic score is used as a single instrument to predict the exposure, and it can be weighted or not by the effect of each one of the variants in the score[93].

Some of the challenges associated with MR studies are population stratification due to different genetic background between individuals in a population[93]; canalization or the attenuation of possible adverse phenotypes from genetic variations through compensatory developmental processes[93]; weak instruments or the use of variants that explain a small proportion of the variation in the phenotype and bias results of the MR towards the null or towards the observational

association[93]; and horizontal pleiotropy or the effect of genetic variants on multiple biological pathways[93]. All these factors can bias results in an MR analysis.
MR has been applied to study the causal role of multiple modifiable exposures (i.e. HDL, LDL, smoking, alcohol, BMI) in their association with different outcomes (i.e. CVD, T2D, obesity, cancer). Particularly for T2D, MR studies have investigated the causal role of adiposity, blood lipids and inflammation in the risk of disease[33]. MR can also be extended to study the causal role of DNAm as a mediator in the exposure-outcome association, or as the exposure of interest. In either case, causality needs to be supported by identifying SNPs in strong association with the CpG site(s) of interest, also known as meQTL[4]. In comparison to genetic association studies of complex traits which are well-powered, studies with availability of epigenome-wide DNAm, genome-wide genetic data and phenotype information, are generally modest in size[95]. A reduced sample size can limit the use of DNAm studies in the context of a single sample MR, but not in a two sample MR, where associations are retrieved from summary data using two independent samples[95]. Direction of the causal effect is another factor to address in MR studies of DNAm, in which case implementation of a bidirectional MR can help to solve directionality issues.

One example of applying MR to infer causality in epigenetic studies has been recently published by Richardson *et al*.[95]. In this study, a systematic investigation of putative causal relationships between DNAm and various complex traits, including T2D and some glycaemic traits, was performed using a two sample MR[95]. DNAm estimates were obtained from the Accessible Resource for Integrated Epigenomic Studies (ARIES)[96], estimates for T2D were obtained from a GWAS conducted by Mahajan *et al*.[30], and estimates for the glycaemic traits were retrieved from different studies reported in the MAGIC consortium.

Richardson *et al*. identified strong evidence of genetic liability in the association between DNAm (exposure) and T2D (outcome) at the CpG sites cg04198914 (*HNF1B*), cg03864215 (*KCNJ11*), cg23956648 (*IGF2BP2*), and cg25064352 (*WFS1*), regarding meQTL detected at birth and childhood[95]. Furthermore, strong evidence of causality was identified between DNAm and T2D at the CpG sites cg15453836 (*PEAK1*) and cg01883759 (*JAZF1*), based on meQTL detected later in life[95]. With respect to the glycaemic traits, DNAm was causally associated with fasting proinsulin at five CpG sites mapping to the *PDE2A*, *PTPMT1*, *STARD10* and *ARAP1* genes, and with HbA1c at seven CpG sites mapping to the *G6PC2*, *TBCD* and *FN3K* genes[95]. The use of colocalization methods further indicated that for the associations previously identified at the CpG sites cg03864215 (*KCNJ11*) and cg25064352 (*WFS1*), the causal variant was the same for DNAm and for T2D[95]. In contrast, the remaining

associations were explained by the effect of two different but correlated causal variants, one explaining variation in DNAm, and the other explaining variation in T2D or the glycaemic trait[95]. To assess direction of causality, a reverse MR was performed using T2D as the exposure and DNAm as the outcome based on associations where there was colocalization of the causal variant. None of the associations at the *KCNJ11* and *WFS1* loci, surpassed statistical significance in the reverse MR analysis[95].

Comparing DNAm loci associated with T2D in the causal analysis conducted by Richardson *et al*., with the strongest loci detected in blood based EWAS in T2D, there is no overlap between loci across analyses. Discordance among signals detected across analyses can be explained by multiple factors. One of them is that variation in methylation associated with T2D in the causal analysis is entirely genetically driven in comparison to that identified in the EWAS, which can be confounded by exogenous factors. Also, methylation loci detected by Richardson *et al*. are more related with future liability to T2D rather than with established disease, since most of these loci were detected regarding genetic variation in methylation in young and most likely unaffected participants. Furthermore, in the causal analysis, associations were identified using DNAm as the exposure, but this causal relationship might not be the same for associations detected observationally, where variation in DNAm can be a consequence of T2D.

Another study using causal inference in epigenetics of T2D was reported by Elliott *et al.*[97]. In this study, the principles of MR were used to explain the mediating role of DNAm in the association between well-established causal anchors (i.e. GWAS SNPs for T2D) and T2D, without applying a formal MR analysis[97]. The first stage of the analysis involved the use 62 previously reported T2D SNPs to determine their influence on variation in DNAm using blood samples from unaffected young participants[97]. From this analysis, it was identified that around half of the genetic variants in T2D were associated with variation in DNAm at 118 CpG sites, establishing that DNAm at these sites may be in the causal pathway to disease, or that it could be a non-causal biomarker[97]. A second analysis identified 226 meQTL associated with some of the T2D CpG sites detected in the previous analysis[97]. These meQTL were independent of T2D SNPs, and there was no evidence that they were associated with T2D based on summary data from DIAGRAM, except for one meQTL located in the *KCNQ1* locus[97]. Thus, for most of the associations detected in the second analysis, it was postulated that the CpG site could be a non-causal biomarker of future disease. In conclusion, the study by Elliott *et al*. showed that except for methylation at the CpG in *KCNQ1*, there was no evidence that DNAm could be in the causal pathway to future disease.

These previous studies serve as examples of the use of MR analysis, or its principles, to assess causality of DNAm as a mediator in associations where the effect of the genetic variants on T2D is already well-established. In addition, these studies disregard previous observational evidence of an association between DNAm and T2D. In contrast, other studies infer causality only after a methylation locus has been identified in the observational analysis. For instance, formal MR analysis has been applied to infer causality, and direction of causality, in methylation markers detected in EWASs of BMI[89, 90]. Furthermore, some of the BMI-associated methylation markers have been demonstrated to effectively predict incident T2D[89]. MR analysis has also been applied to infer causality in the association between BMI and *HIF3A* methylation[98]. To my knowledge, similar studies have not been conducted in T2D to infer causality, and direction of causality, of methylation markers detected in well-powered EWAS, and this constitutes the aim of this thesis. Overall, the importance of knowing the causal direction of disease-related variation in methylation, is because this can help to identify markers of disease predisposition or progression, and putative targets for intervention[41].

1.6.2 GWAS of T2D

As referred to previously in section 1.2.4.1 when describing genetic risk factors for T2D, multiple GWAS meta-analyses have been conducted to date in T2D using European[99-102] or mixed populations[30], identifying across them more than 60 risk loci for T2D[103]. Most of the genetic loci identified in GWAS of T2D correspond to common genetic variants (MAF > 5%) rather than rare or low-frequency variants (MAF <1% or between 1-5%)[103]. However, advances in genotyping technologies, variant calling, and association testing in large samples, are some of the factors contributing to reduce this gap to identify the effect of rare or low-frequency variants in T2D[103].

Strong statistical significance identified in GWAS signals in T2D, contrasts with their small effect sizes, small proportion of variation explained in T2D (10-15%), and scarce knowledge of the mechanisms in which these genetic variants influence the pathogenesis of T2D[103, 104]. Despite these limitations, genetic variants are fundamental to conduct causal inference analyses based on MR methods[4, 93, 94], where they are used to instrument the risk of T2D, and generate causal estimates of the association between T2D and DNAm. Thus, GWAS data in T2D is necessary to retrieve strongest variants to be used in a single sample MR, or in a two sample MR approach based on summary data. Furthermore, GWAS data can be used to identify risk variants for T2D that correlate with genetic variation in DNAm, establishing shared genetics between these two traits that support causality, and allow to identify direction of causality. Genetic variants associated with T2D can be retrieved from

large-scale GWAS meta-analyses reported in the DIAGRAM consortium (http://www.diagram-consortium.org/), the GWAS catalog (https://www.ebi.ac.uk/gwas/), or in the MR-Base platform for Mendelian randomization analyses (http://www.mrbase.org/).

1.6.3 Genetics of DNA methylation

The role of the epigenome in regulating gene function is not independent of the genotype, and results from twin-, family- and population-based studies, have demonstrated that common genetic variants (SNPs) can explain much of the interindividual variation in DNA methylation[4]. MeQTL are genetic variants associated with variation in methylation in a temporal and tissue-specific manner[4]. Depending on the distance separating the SNP from the CpG site, an meQTL can be a *cis* (≤1Mb) or a *trans* (>1Mb) meQTL[4]. In addition, the SNP can locate in the same position as the CpG site, establishing CpG-SNP pairs, where the SNP affects directly methylation by introducing or removing a CpG site[53]. Almost 25% of the SNPs in the genome are deemed to be CpG-SNPs[53].

A clear understanding of meQTL maps can help to explain how methylation establishes and persist[4], and how the genotype interacts with the methylome to influence risk of disease, since many meQTL have been identified in loci previously reported in GWAS of some disorders[42]. Furthermore, the availability of well-powered meQTL studies allows us to incorporate meQTL as causal anchors for variation in DNAm in the context of MR studies.

MeQTL have been identified in various tissues relevant for T2D[82, 105-108], but large-scale meQTL studies have been primarily conducted using blood samples. One of the most comprehensive meQTL studies in blood was reported by Gaunt *et al*.[109], where they characterized the genetic contribution to genome-wide variation in DNAm at five different timepoints across the lifecourse, and results of the meQTL analysis are available in the meQTL database (http://www.mqtldb.org/)[109]. Samples included in this study were 753 mothers and 814 children from the ARIES sub-study. Overall, Gaunt and colleagues demonstrated that the genetic effects on DNAm remain highly consistent across the lifecourse, which reflects the high heritable component of DNAm[109]. They also showed that most of the genetic variation in DNAm was captured as *cis*-effects compared to *trans*-effects (7%), although the *trans* component is thought to be larger but highly polygenic[109]. Furthermore, results suggested that meQTL are likely to contribute to the variation in complex traits[109]. Another important meQTL study using blood samples from 3841 Dutch individuals was reported by Bonder *et al*.[110] Results of this meQTL analysis are available in the BIOS QTL browser (https://genenetwork.nl/biosqtlbrowser/). Lastly, the consortium for the study of the genetics of DNAm (GoDMC) is to date the largest meta-

analysis of meQTL studies using blood samples from more than 20,000 middle-age participants, predominantly of European ancestry. Access to results of the meQTL analysis in GoDMC is available upon submission of a project proposal and approval by the executive committee.

1.7 This thesis

Currently, epigenetic studies of DNAm in T2D suggest that this molecular phenotype is associated with the establishment of T2D based on evidence from retrospective and nested case control studies. Compared to genetic risk factors for T2D, DNAm variation identified across tissues explains a higher proportion of the variation in T2D. However, further work is needed using a prospective study design to corroborate these studies.

Furthermore, there is little evidence to support causality in the association between DNAm and T2D, where the direction of the causal effect remains unknown for all the associations that have been detected. Because DNAm is a plastic phenotype known to be influenced by environmental factors related to disease, and by the disease itself through reverse causation, it is fundamental to apply causal inference analysis to disentangle the cause-effect relationship between DNAm and T2D. Establishing which methylation markers are causally associated with T2D, will help to prioritize them for early screening of T2D. Furthermore, it is crucial to identify novel associations with T2D, as well as to confirm existing associations, by using EWAS in sufficiently large samples from population-based studies. In addition, it is necessary to evaluate the relevance of blood as a source of methylation markers for T2D, by comparing levels of DNAm in blood, with DNAm in more relevant tissues for T2D at the target CpG sites. Such cross-tissue comparison of DNAm can be done using publicly available datasets.

1.7.1 Aims and overview of chapters

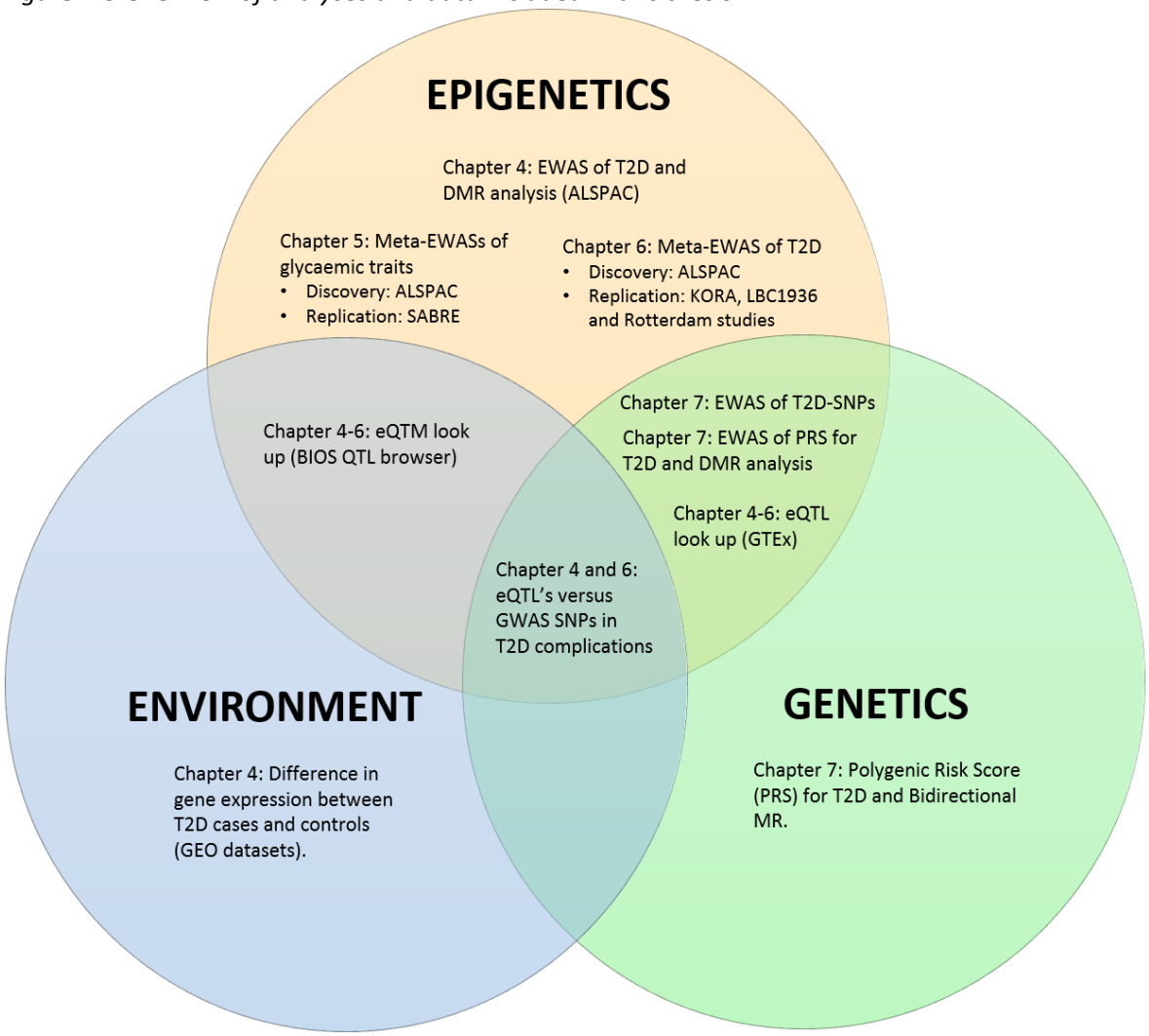
The aims of this thesis were to identify DNAm sites observationally associated with prevalent T2D and related glycaemic traits, and to infer causality and direction in these associations. Functional exploration of loci of interest showing variation in DNAm in T2D was also undertaken. These aims were accomplished by using different datasets and analyses, described throughout this thesis as follows:

1. Chapter 2: Description of cohorts and methods to conduct and interpret EWAS of prevalent T2D and glycaemic traits.

2. Chapter 3: Description of MR methods used to infer causality and direction in the association between DNAm and prevalent T2D.
3. Chapter 4: Results of the EWAS of prevalent T2D in middle-aged adults from the Avon Longitudinal Study of Parents and Children (ALSPAC). In addition, functional interpretation and replication of EWAS results in three independent European cohorts.
4. Chapter 5: Results of the EWAS of glycaemic traits using control samples from ALSPAC, with replication and meta-analysis of EWAS in an independent multi-ethnic cohort.
5. Chapter 6: Results of a meta-analysis of EWAS of prevalent T2D, using data from five cohorts.
6. Chapter 7: Results of a bidirectional MR study assessing the direction of causality of the top associations identified in the meta-analysis of EWAS of prevalent T2D.
7. Chapter 8: Discussion of overall findings, strengths and limitations of the present study, and future perspectives.

An overview of the analyses conducted throughout this thesis, and the data incorporated, are provided in Figure 1-5.

Figure 1-5 Overview of analyses and data included in this thesis.



Chapter 2 Methods: cohort descriptions and epigenome-wide association studies

The aim of this chapter is to outline studies and methods used throughout this work. The first part of the Chapter provides detail of the cohorts included in the epigenetic study of T2D and glycaemic traits, emphasizing the use of ALSPAC as the discovery study. The second part of the Chapter describes methods in epigenetic analysis, including the generation and processing of methylation data, EWAS, selection of covariates, region-based approaches, meta-analysis, and exploration of the biological function.

2.1 The Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC is a longitudinal birth cohort from the southwest of England that was established to study the influence of multiple behavioural, psychological, biological, genetic and epigenetic exposures, on different health, social, and developmental outcomes throughout the lifecourse[111, 112]. The study initially recruited 14,541 pregnant women with expected delivery date between 1991 and 1992, who attended any of three health districts within the old administrative county of Avon[111]. This initial sample was supplemented by including other participants at a follow-up period, who met the recruitment criteria, allowing the study to increase the sample to 15,247 pregnancies.

Data from participants in ALSPAC has been collected at multiple time-points for the children and their parents via clinics and self- and parent-reported questionnaires. Sampling of blood and buccal swabs provided DNA to conduct genetic and epigenetic studies. A complete list of variables available in the cohort can be searched in the variable search tool at http://www.bristol.ac.uk/alspac/researchers/our-data/. Ethical approval was obtained from the ALSPAC ethics and Law Committee and the Local Research Committee.

Selection of the relevant time-point for the study of Type 2 Diabetes

Since the purpose of this study was to identify epigenetic variation associated with T2D and related glycaemic traits, the focal time-point was adulthood, considering that this is the period where the disease reaches its highest prevalence (> 40y in developing countries, and >60y in developed countries)[1, 2]. However, there is growing evidence suggesting that this is no longer the case as T2D can also occur at younger age based on higher rates of obesity-related T2D in youth[1]. Despite this, it is difficult to determine the correct diagnostic criteria for diabetes in childhood[2].

Data collection for mothers and partners

Multiple questionnaires have been completed by parents in ALSPAC during the gestational period (4 questionnaires for mothers and 2 for the partner), and postnatally (16 questionnaires for the mother, and 14 for the partner). In addition, data collection via clinics has been conducted at four time-points for the mothers (Focus on Mothers 1-4 or FOM1-4), and at a single time-point for the partners (Focus on fathers 1 or FOF1). The first clinical assessment for the mothers was completed when children were approximately 19 years of age (FOM1), and the following clinics were conducted when children were in their early adulthood, approximately 22 to 24 years of age (FOM2-4). The only clinic available for the partners was completed when children were 22 years of age. Other opportunistic measures have been taken for the mothers when they accompanied the children at three focus clinics when children were 12-13, >13-14 and 15-16 years of age (Teen focus 1-3 or TF1-3). Further detail of variables collected for the mothers at the different points of assessment, can be found in one of the accompanying papers of the cohort[112].

2.1.1 Selection of samples

The subsample included in this retrospective case-control study of epigenetic variation in T2D, comprised a subset of participants from the core ALSPAC sample with availability of necessary data at the FOM1 and FOF1 clinics. Despite the presence of multiple clinics for the mothers, only data collected at the FOM1 was used based on the completeness of the metabolic records available for this clinic at the time of conducting this study. In total, 4,832 mothers and 2,001 partners attended the FOM1 and FOF1 clinics, respectively.

2.1.2 Selection of variables and covariates

Main variables considered throughout this study were fasting glucose (FG), 2-hours postload glucose (2-h PG), fasting insulin and proinsulin (see section 2.1.3), and HOMA scores (HOMA-B and HOMA-IR). HOMA scores, which are mathematical models used to explain the function of organs and tissues involved in the regulation of glucose during the homeostatic state[113], were calculated from fasting insulin (µIU/mL) and fasting glucose (mmol/l) levels using the formulae: HOMA-IR=(insulin x glucose)/22.5 and HOMA-B=(20 x insulin)/(glucose - 3.5)[113, 114]. HOMA scores were exclusively calculated for the mothers as records of fasting insulin were not available in the partners.

Several covariates were included to characterize the study population at baseline, and some of them were also included for the adjustment of main analyses. Covariates were extracted from clinic data, or from existent records in self-completed questionnaires. Commonly, the questionnaire used was the one closest in time to when the clinic visit was conducted. Sociodemographic covariates

considered were age, sex, ethnicity and socioeconomic status (SES). SES was proxied by educational level (low=no qualification, certificate of secondary education or vocational, middle=O level, and high=A level degree) as described by Bath *et al*.[115].

Anthropometric measures used were BMI, waist-circumference and waist-hip ratio. Weight was previously measured using the Tanita scales (TBF401-A) and recorded to the nearest 0.1kg, while standing height was measured using the Harpenden stadiometer and recorded to the nearest 1mm. Body mass index was derived from weight and height and calculated as: weight (kg)/ [height (m)]^2. Waist circumference was measured twice using the Seca 200 body tension tape and recorded to the nearest 1mm. Cardiovascular health covariates used were systolic and diastolic blood pressure, which were measured twice from the left arm using an Omron M6 upper arm BP/Pulse monitor, and the average between the two records was the value used for the analyses.

Metabolic variables of interest were different measures of lipids in blood (total cholesterol, HDL, LDL and triglycerides), and c-reactive protein (CRP) as an inflammatory marker. Further detail of the method used for the assessment of these variables can be found in section 2.1.3. Behavioural covariates included were smoking and physical activity. Smoking was extracted from questionnaires in the mothers, and from clinic data in the partners, and it was coded as never smokers, former smokers and current smokers. Due to the high proportion of missingness for smoking, this variable was latter predicted using a methylation score (see section 2.1.7). Physical activity was extracted from existing records, and it was defined based on the number of hours per week that participants dedicated to performing physical exercise (i.e. ≤4h/week or >4h/week).

Health history covariates were family history of T2D and coronary heart disease (CHD). Family history of T2D was derived from two records in the mothers based on self-completed questionnaires at 12 weeks of gestation and at 8 years after birth of the index child. Family history of T2D was extracted from an existing record in the partners based on a self-completed questionnaire at the time the index child was born. Family history of CHD was extracted from existing records based on a questionnaire completed by the mothers at 12 weeks of gestation, and by the partners at the time the index child was born.

2.1.3 Assessment of glycaemic traits and other metabolic variables

Analysis of glucose, lipids, and CRP at the FOM1 clinic for the mothers, was conducted by the staff of the Routine Lipids Section in the Biochemistry Department of the Glasgow Royal Infirmary using a Hitachi Modular P Analyser. Fasting insulin and proinsulin levels were analysed by the Metabolic Medicine Group in the Department of Vascular Biochemistry, Glasgow University, using the Tecan Sunrise plate reader and the Tecan Magellan v6.4 (2007 Tecan) software for calculation of results. Metabolic variables collected at the FOF1 clinic for the partners were analysed using in-house laboratory methods similar to those applied for the assessment of maternal metabolic variables. Description of specific laboratory methods used can be found in Table 2-1, while further documentation of laboratory protocols is available through formal request to the ALSPAC data access team (alspac-data@bristol.ac.uk).

Table 2-1 Description of laboratory methods used for the assessment of metabolic variables of interest in this study. Variables were collected at the FOM1 and FOF1 clinics for mothers and partners, respectively. All assays were conducted using EDTA plasma samples.

Measure	Units	Time-point	Method†
Fasting Glucose	mmol/l	FOM1/FOF1	Enzymatic colorimetric assay using a hexokinase enzyme. Gluco-quant Glucose/HK kit by Roche Diagnostics. Cat no. 1447513.
2-h PG*	mmol/l	FOM1	75-g of glucose load ‡ followed by hexokinase assessment of plasma glucose (Gluco-quant Glucose/HK kit, by Roche).
Fasting insulin	µIU/mL	FOM1	Enzyme immunoassay. Ultra-sensitive insulin ELISA kit by Mercodia. Cat no. 10-1132-01.
Fasting Proinsulin	pmol/l	FOM1	Enzyme immunoassay. Proinsulin ELISA kit by Mercodia. Cat no. 10-1118-01.
C-reactive protein	mg/L	FOM1/FOF1	Particle-enhanced immunoturbidimetric assay. Tina-quant CRPHS kit by Roche Diagnostics. Cat no. 11972855.
Total cholesterol	mmol/l	FOM1/FOF1	Enzymatic colorimetric test using cholesterol esterase and cholesterol oxidase enzymes. Cholesterol Kit (CHOD-PAP) by Roche Diagnostics. Cat no. 1491458.
Triglycerides	mmol/l	FOM1/FOF1	Enzymatic colorimetric test using a lipoprotein lipase. Triglyceride kit (GPO-PAP) by Roche Diagnostics. Cat no. 1730711.
HDL	mmol/l	FOM1/FOF1	Homogeneous enzymatic colorimetric test for the direct quantification of HDL. HDL-C plus (2nd Generation) kit by Roche Diagnostics. Cat no. 3045935.
LDL	mmol/l	FOM1/FOF1	Calculated using the Friedwald formulae: LDL= total Cholesterol-(HDL Cholesterol + Triglyceride/2.19).

*2-hours post-load glucose. †Laboratory method implemented for the analysis of metabolic variables in ALSPAC, including name of the kit used, provider company, and reference number of the catalogue utilized. ‡ glucose load equivalent to 75-g anhydrous glucose dissolved in water.

2.1.4 Baseline characteristics of the subsample of middle-age adults in ALSPAC

In total, there were 6,607 adults (mothers= 4,606 and partners= 2,001) eligible for this study after excluding samples with multiple-entry IDs based on pregnancy number, and samples without clinic data. Further detail of this subsample can be found in Table 2-2. Briefly, mean age of these participants was 49.53 years (SD=5.41), most of them were of European ancestry (only 8.45% were non-white), the proportion of females was twice as high as the proportion of males, the mean FG in the subsample indicated average normoglycemia, and family history of diabetes and CHD was reported by 9.41% and 13.91% participants from the total subsample. Only for the mothers, mean levels of fasting insulin, proinsulin, 2-h PG, HOMA-IR and HOMA-B, were 6.11μIU/ml (SD=6.31), 8.50pmol/l (SD=9.47), 4.48mmol/l (SD=0.88), 1.52 (SD=2.10) and 68.53 (SD=80.16), respectively. Similar glycaemic measures were not available in the partners dataset.

Table 2-2 Baseline characteristics of the subsample of middle-age adults in ALSPAC eligible for the epigenetic study of T2D.Continuous variables were described using the mean, standard deviation and the range (min, max), while categorical variables were described using the proportion and the number of samples per category. For each variable it is reported the proportion of missingness found.

N=6,607	**Mean (SD)**	**Range**	**% (number)**	**Missing (%)**
Age (yrs)	49.53 (5.41)	34-89	---	0.11
Ethnicity [% white]	---	---	91.55 (6,049)	6.49
Fasting Glucose (mmol/l)	5.42 (1.06)	2.6-23.14	---	9.04
Body mass index (kg/m^2)	26.94 (5.02)	14.58-55.04	---	0.71
Waist circumference (cms)	88.62 (13.42)	57.8-150	---	0.59
Waist-hip ratio (cms)	0.89 (0.11)	0.59-1.41	---	**0.59**
Systolic BP (mmHg)	122.79 (14.59)	88.5-207.5	---	1.44
Diastolic BP (mmHg)	74.14 (10.32)	47.5-155	---	1.44
Serum Total Cholesterol (mmol/L)	4.79 (0.91)	1.85-8.74	---	8.48
Triglycerides (mmol/l)	1.16 (0.69)	0.24-20.61	---	9.04
HDL cholesterol (mmol/l)	1.42 (0.36)	0.53-3.57	---	9.04
LDL cholesterol (mmol/l)	3.04 (0.82)	0.39-6.62	---	9.07
Fasting C-reactive protein (mg/L)	2.24 (5.43)	0.03-273	---	9.69
Sex [male %]	---	---	30.29 (2,001)	0.00
Medication for T2D [Yes %]	---	---	2.07 (120)	12.21
T2D [% cases]	---	---	5.28 (261)	**25.17**
Family History of Diabetes [Yes %]	---	---	9.41 (492)	**20.86**
Family History of CHD [Yes %]	---	---	13.91 (851)	7.43
Smoking [%]				**27.30**
Never smoker	---	---	53.53 (2,571)	
Former smoker	---	---	37.25 (1,789)	
Current smoker	---	---	9.22 (443)	
Physical activity [less than 4h/week %]	---	---	82.42 (3,202)	**41.20**
Socioeconomic status [%]				15.21
High income	---	---	49.96 (2,799)	
Middle income	---	---	33.70 (1,888)	
Low income	---	---	16.33 (915)	

2.1.5 Definition of T2D

A case of T2D was defined as an individual with at least one or more medical records of T2D, self-reported diabetes, or FG above or equal to 7.0mmol/l, or if there was medical diagnosis of T2D and reported use of medication to lower glucose (i.e. insulin and tablets), but controlled levels of FG (< 7.0mmol/l). Controls were individuals with no medical diagnosis, no use of medication, and levels of FG below 7.0mmol/l. In total, there were 137/4606 women identified as T2D cases, 3219/4606 controls and 1250/4606 missing data. In the dataset of partners there were 124/2001 T2D cases, 1464/2001 controls, and 413/2001 missing data.

Self-reported diabetes in mothers was extracted from two questionnaires completed at 9 years and 11 years after birth of the index child. Medical diagnosis and treatment for T2D were variables extracted from different questionnaires in the mothers available from the gestational period until 19 years of postnatal follow-up. In the partners, medical diagnosis and medication to treat T2D were variables extracted from a self-completed questionnaire 20 years after birth of the index child, but no variable was available for self-reported diabetes.

To avoid biasing results by including potential cases of Type 1 diabetes, participants with T1D (4 mothers and 3 partners) were excluded. A T1D case was defined as a participant whose age at onset of diabetes was below 20 years, who reported the use of insulin to treat diabetes, and who was initially classified as a diabetes case.

Further characterization of T2D cases

Of the 261 T2D cases in total identified in the subsample of middle-age adults in ALSPAC, 21% (n=54) of them had FG below 7.0mmol/l or well-controlled glucose, and in 50% of the cases this was attributed to the use of medication, while for the remaining 50% lower glucose was achieved through diet.

Case control T2D versus other risk categories

According to the American Diabetes Association (ADA), there are three categories to specify increased risk of diabetes based on FG levels: normoglycemic, prediabetic and diabetic[2]. In ALSPAC, 75.0% of the samples were classified as normoglycemic (FG < 5.6mmol/l), another 21.5% of them were prediabetic (FG ≥ 5.6mmol/l and FG < 7.0mmol/l), and the remaining 3.2% samples were classified as diabetic (FG ≥ 7.0mmol/l). Overlapping these categories with the case control definition (Figure 2-1), it was identified that 74.5% (n=4,420) controls were within the normoglycemic

category, while the remaining 20.6% (n=1,222) of them were in the prediabetic category. Regarding T2D cases, 77.7% (n=203) of them were within the diabetic category, but the remaining 22.3% (n=58) cases were grouped within the normoglycemic and prediabetic categories, corresponding these to participants with medical diagnosis of T2D, but well-controlled glucose through medication or diet. Thus, a similar distribution of samples was observed when using three categories of T2D diabetes risk, versus the conventional case control definition, and the former classification allowed us to distinguish controls at higher risk of T2D. Even though analyses throughout this thesis were conducted using the case control definition, when necessary, main results were interpreted based on categories of diabetes risk.

Individuals in ALSPAC were also classified based on different categories of glucose tolerance specified by the ADA in reference to FG levels and 2-h PG. Based on these categories, an individual has impaired glucose tolerance (IGT) if 2-h PG ≥ 7.8mmol/l and 2-h PG < 11.1mmol/l, it has impaired fasting glucose (IFG) if FG ≥ 5.6mmol/l and FG < 7.0mmol/l, diabetes if 2-h PG ≥ 11.1mmol/l or FG ≥ 7.0mmol/, and they are normoglycaemic if 2-h PG < 7.8mmol/l or FG < 5.6mmol/l. For the subsample of ALSPAC, only 0.1% (n=6) participants had IGT, another 0.7% (n=46) had IFG, 1.13% (n=74) participants were diabetics, and the remaining 98.1% (n=6,481) individuals were normoglycaemic (Figure 2-1). When comparing this glucose tolerance classification with the conventional case control definition, there was high discrepancy between the two approaches, especially because 72% of the T2D cases were deemed as normoglycaemic under the new criteria, only 27% of them were accurately allocated to the diabetic group, and none of the T2D cases were seen within the IGT or IFG categories (Figure 2-1). In contrast, controls were overrepresented under the glucose tolerance categories. Thus, using categories of glucose tolerance based on FG and 2-h PG levels did not provide further information of the samples from the conventional case control definition, and this classification was not considered for the interpretation of main results.

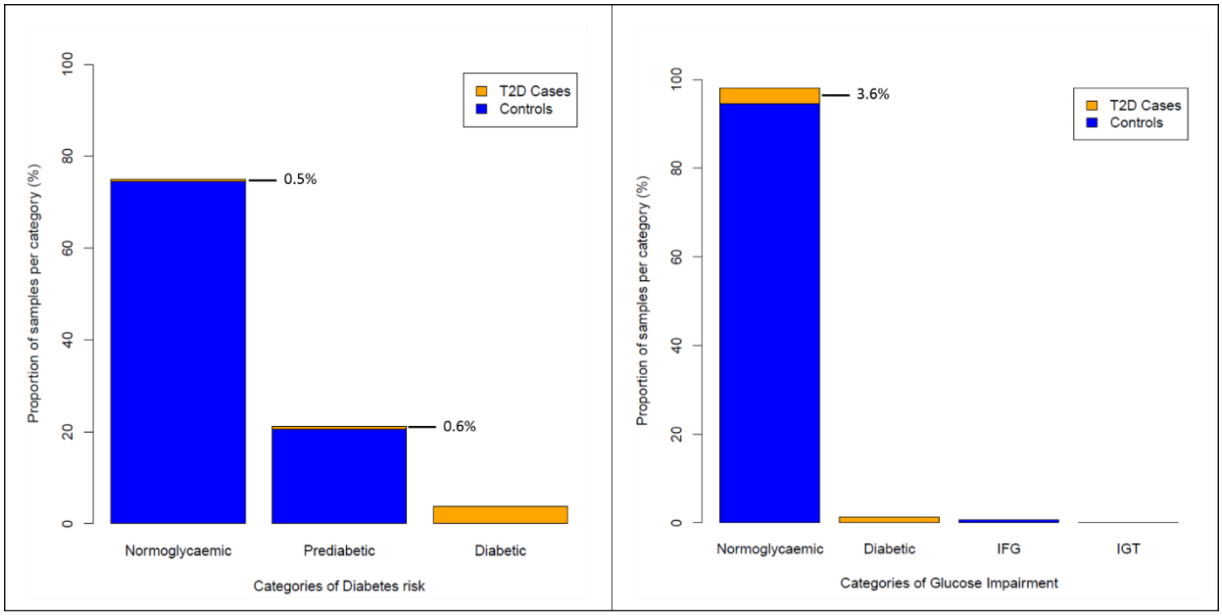


Figure 2-1 Distribution of the subsample of adults in ALSPAC across categories of T2D risk (left-plot: normoglycaemic, prediabetic and diabetic) and glucose tolerance (right-plot: normoglycaemic, diabetic, IFG and IGT) according to ADA (American Diabetes Association) criteria. Overlapping to these categories, is the conventional case control definition of T2D.

2.1.6 The Accessible Resource for Integrative Epigenetic Studies (ARIES)

ARIES is a population-based resource for the study of DNA methylation and the causes and consequences of changes in this marker in health and development[96]. Participants in ARIES are a subset of those included in the ALSPAC study, for whom genome-wide DNA methylation data is available. Samples included in ARIES were approximately 1,018 mother-offspring pairs and 588 fathers (ARIES v4; date released 2017), who were selected from the core ALSPAC sample based on the availability of purified DNA at two time-points in the mothers (antenatal and middle-age), three time-points in the offspring (at birth, 7.5 years and 15.5 years), and a single time-point in the fathers (middle-age). Further detail on the timing of these measurements is illustrated in Figure 2-2. All analyses performed throughout this thesis were focused on peripheral blood DNA methylation from mothers and fathers in the ARIES study, particularly at the middle-age time-point, considering the importance of methylation marks at this stage in the study of prevalent T2D (see section 2.1.7). Further detail on the availability of DNA methylation in other tissues as part of the ARIES project, can be found in the ARIES data resource profile[96].

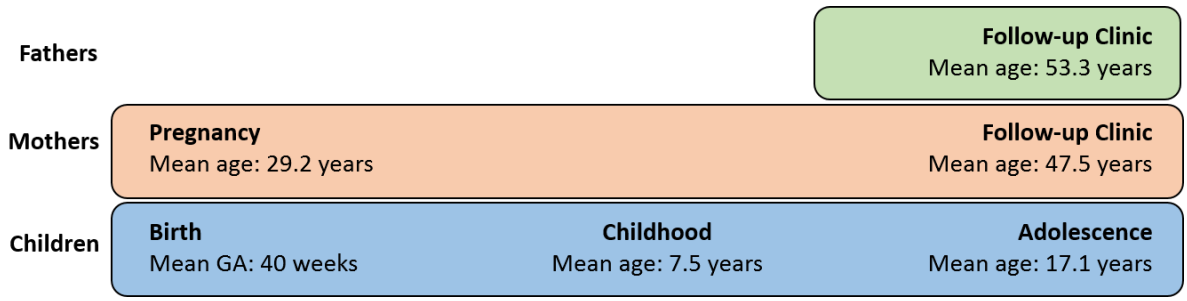


Figure 2-2 Time-points where measures of DNA methylation were available for a subset of participants in ALSPAC included in the ARIES study. GA=gestational age.

2.1.6.1 Assessment of methylation

Laboratory methods

Measurement of DNA methylation in adults from the ARIES study was conducted according to standard procedures reported else-where[96]. Briefly, DNA was extracted from peripheral blood (whole blood or buffy coat), bisulphite converted using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA), followed by the genome-wide measurement of the methylation status of around 485,000 CpG sites using the Illumina Infinium HumanMethylation450K BeadChip (HM450) assay according to the standard protocol[96]. Initial inspection of the quality of the signals obtained was made using GenomeStudio (version 2011.1). For each sample, the HM450 assay reports the level of methylation of a CpG site using a β-value, which is a proportion of the methylated probe intensity divided by the overall intensity in a scale ranging between 0 (unmethylated cytosine) and 1 (completely methylated cytosine). All analyses with DNA methylation were performed using β-values.

Coverage of the HM450 array

Probes covered by the HM450 array target 99% of genes, 96% of CpG islands and non-coding RNA regions of the human genome[116]. Within these regions, probes can be found in the promoter of the gene (~41%), 3'UTR (~3%), within the body of the gene (~31%), or within intergenic regions (~25%) specified by GWAS findings[117]. With respect to CpG islands, probes can be found in the island itself (~31%), or in shore (~23%) (within 2kb from island), shelfs (~31%) (4kb from island), or the open sea region (~36%) (>4kb from island)[117].

Functional normalization using meffil

Functional normalization of methylation data was previously performed as part of the ARIES project using the R package *meffil* (version 1.0.0). Compared to other methods used to process methylation data, functional normalization has the advantage of removing technical variation (i.e. batch effects) between samples, while preserving important biological variation in DNA methylation. The protocol used for functional normalization of ARIES data was described by Min *et al.*[118]. Briefly, IDAT files

provided by Illumina are read into QC objects, which are generated for each sample and contain raw control probe summaries and quantile distribution of raw probe intensities[118]. These QC objects are then used to generate variables summarizing batch effects, to detect poor quality probes based on dye bias and detection P-values, to estimate genotype bias and sex mismatch for the samples, and to generate predicted cell-counts[118]. Cleaned QC objects obtained after excluding samples and CpG sites regarded as outliers, are then quantile normalized across samples, and used to normalize raw probe intensities for each sample. Post-normalization analyses can be performed to further detect outlier samples and exclude them from the methylation dataset before the EWAS.

Pre-processing methylation data

Pre-processing of methylation data was previously performed as part of the ARIES project, and was described before by Relton *et al*.[96], and in unpublished work by Josine Min (Sample QC ARIES data). Briefly, samples that failed QC were excluded and sent back for reanalysis if detection score P-value ≥ 0.01, indicating low confidence to detect difference in the signal between a methylated CpG site and a negative control probe. In addition, samples were excluded if there was genotype mismatch between the SNP-probe (65 SNP-probes on the HM450 array) and genotype array data for the same individual, suggesting sample swapping, or if there was gender mismatch based on the X chromosome methylation. Furthermore, samples were excluded if they had more than 10% CpG sites with low bead-count (n-bead < 3), if there was high IBD or relatedness issues between mother and child pairs, and if samples were technical duplicates. For the CpG sites, they were removed if detection P ≥ 0.01 for more than 5% of the samples, or if the number of beads per probe was lower than 3 for at least 10% of the samples (Sample QC ARIES data, unpublished).

Prediction of white-cell counts from peripheral blood using DNA methylation

Considering that difference in methylation between samples can arise from cellular heterogeneity in peripheral white-blood cells, and because cell counts from whole blood were not available for most of the samples in the ARIES study before the extraction of DNA[96], a *post hoc* prediction of cellular proportions was necessary to account for potential confounding in downstream methylation analyses[96]. Cell proportions for CD8T cells, CD4T cells, Natural Killer cells, B cells, monocytes and granulocytes, were previously estimated as part of the ARIES project using the Houseman algorithm[119] with the reference blood panel *gs35069,* and the function *meffil.cell.count.estimates* from the R package *meffil* [120]. Further detail on the method used for the estimation of cell-counts, can be found else-where[96, 120].

2.1.7 Subsample of ARIES included in the epigenetic study of T2D and glycaemic traits

After pre-processing of the methylation data and excluding samples without complete phenotype data, 482,518 CpG sites and 1,050 participants from ARIES remained in the dataset to conduct the EWAS of T2D (Table 2-3, see Chapter 4). From this subsample, a subset of 1,002 normoglycaemic individuals were included in the EWAS of FG (Table 2-3, see Chapter 5), defining as normoglycaemic participants with FG < 7.0mmol/l, no medical diagnosis of T2D, and no self-reported use of medication to treat diabetes. To conduct the EWAS of 2-h PG, fasting insulin, fasting proinsulin and the HOMA scores, a subset of 622 normoglycemic females (mean age 47.9 years) were included in the analyses. No males were considered for the EWAS in other glycaemic traits as necessary variables were not available in this dataset. Table 2-3 describes the subsamples of ARIES included in the different EWAS analyses, while baseline characteristics of these participants can be found in Chapter 4 and Chapter 5.

Table 2-3 Description of three subsamples from ARIES included in the EWAS of T2D and glycaemic traits. For the EWAS, the final sample included were participants with complete DNA methylation and phenotype data.

Dataset	Analysis	Time-point	N	Mean age (years)
1	EWAS of T2D	FOM1	770	47.5
		TF1-3†	177	42.9
		FOF1	588	53.3
		Total samples in ARIES ‡	1535	47.9
		Samples in ARIES with complete data for T2D and covariates*	**1050**	50.06
2	EWAS of FG (controls)	FOM1	618	47.94
		TF1-3	4	44.28
		FOF1	380	53.31
		Total samples in ARIES	1002	49.97
		Samples in ARIES with complete data for FG and covariates	**1002**	49.97
3	EWAS of 2-h PG, insulin, proinsulin and HOMA scores (controls)	FOM1	618	47.94
		TF1-3	4	44.28
		Total samples in ARIES	622	47.94
		Samples in ARIES with complete data for glycaemic traits and covariates	**622**	47.94

†Opportunistic maternal samples collected in one of three focus clinics for the children when they were 12 to >15 years of age. ‡ Samples in ARIES with or without complete phenotype data. * Total samples included in the EWAS of T2D (cases= 48, controls= 1002).

Statistical analysis

Baseline characteristics in the subsample of the EWAS were described between cases and controls using a t-statistic for continuous variables, while for categorical variables the chi-squared statistic or the p-trend (ordinal variables) were used.

Dealing with missingness

To increase power in the EWAS, missing values for the covariates age, BMI and smoking, were imputed before the analysis. Missing values for age (n=4) and BMI (n=10) were imputed by the mean of the covariate according to the sex and T2D status of the samples with missing data. Missing values for smoking (n=247) were imputed using a methylation score composed of 187 CpG sites that was previously reported by Zeilinger *et al*.[121]. Detail of the method applied to impute missing data in smoking was described before by Elliott *et al*.[122]. The general form of the score is shown in Equation 2-1, and it can be defined as the sum of mean values of methylation across the 187 CpG sites, weighted by the effect size. Because this score had low sensitivity to distinguish between former and never smokers, imputed values of smoking were classified into two categories: smokers or non-smokers. Categories of self-reported smoking were recoded to match with those of imputed smoking. Misclassification error of the score was reported as the percentage of false positives and false negatives identified in the imputed data, in addition to the sensitivity [true positives/ (true positives + false negatives)*100] and specificity [true negatives/ (true negatives + false positives) *100] of the score.

Equation 2-1 Methylation score for smoking

$$Z = \sum_{i}^{n}(meanCpGi - CpGi) \times WeightCpGi + (meanCpGn - CpGn) \times WeightCpGn$$

Where $meanCpG_i$ refers to mean methylation of probe i in ALSPAC samples, CpG_i is the reported mean methylation of probe *i* by Zeilinger *et al*.[121]. Based on absolute values of the effect size, $weightCpG_i$ is the weight of probe *i* calculated using the ratio between the effect size of probe i, divided by the average effect size of all the probes included in the score.

Categories of Glucose tolerance

For the EWAS of glycaemic traits, categories of glucose tolerance defined by the World Health Organization (WHO, 1999)[1] were used to classify control participants at risk of prediabetes: impaired fasting glucose (IFG) if FG ≥ 6.1mmol/l and < 7.0mmol/l, impaired glucose tolerance (IGT) if 2-h PG ≥ 7.8 mmol/l and < 11.1mmol/l, and normal glucose tolerance (NGT) if FG < 6.1mmol/l and 2-h PG < 7.8mmol/l.

2.2 External replication cohorts

To increase the power of findings obtained in the EWAS of T2D and glycaemic traits in ALSPAC, these EWAS were replicated in additional studies, and results were combined across studies via meta-analysis. Studies were selected if participants were of similar ethnic background as individuals in ALSPAC (i.e. Europeans), had similar age structure (i.e. middle-age adults), relatively equivalent proportion of females and males, and if there was no report of additional chronic conditions in these participants that could introduce bias to downstream analyses. In total, three studies and one sub-study were included in the replication of the EWAS in T2D: the cooperative health research in the region of Augsburg (KORA), the Lothian birth cohort of 1936 (LBC1936), and the Rotterdam studies RSIII-1 and RS-Bios. For the replication of the EWAS in glycaemic traits, the only replication study included was the Southall and Brent Revisited (SABRE) study. Important to be highlighted at this point, is that for most of the cohorts only summary data was available to conduct the meta-analysis, but not for SABRE, where I had access to individual level data to perform the EWAS. Further detail of the different cohorts included in the replication of the EWAS of T2D and glycaemic traits, is provided below.

2.2.1 Establishing collaboration across studies

After identifying the cohorts with availability of phenotypic variables of interest and genome-wide measures of DNA methylation, main researchers from each cohort were directly contacted to enquire their willingness to participate in the study of the meta-analysis of EWAS in T2D. For the cohorts who agreed to be involved in the study, an analysis plan was sent explaining the aims of the study, and specific conditions for the analysis, including data pre-processing, code to perform the EWAS, and instructions on how to store and share results to facilitate their use in the meta-analysis.

2.2.2 Description of cohorts

2.2.2.1 Cooperative Health Research in the Region of Augsburg (KORA)

KORA is a research platform for population-based surveys based in the region of Augsburg in southern Germany[123]. KORA was started in 1996 by the German National Research Centre for Environment and Health, to expand from previous surveys conducted in Augsburg as part of the WHO MONICA project, and to promote future studies in epidemiology, health economics and health care research[123]. At baseline, four surveys (S1-S4) were conducted from 1984/85 to 1999/2001, corresponding each survey to completely independent samples[123]. The survey of interest for this study was the S4 survey composed of 4,261 samples, and a subset of 3,080 of them were reinvestigated in the follow-up examination KORA F4 between 2006-2008.

Variables collected from KORA participants included sociodemographic, anthropometric, risk factors, medical history of chronic diseases, use of medication, among others[123]. Further information of the cohort can be found elsewhere[123]. Written informed consent was obtained from all participants, and all the KORA studies have been approved by the ethics committee of the Bavarian Medical Association and the Bavarian commissioner for data protection and privacy. A subsample of 1,719 participants from the KORA F4 examination (aged 32-81years), was included in the replication of the EWAS in T2D. This subsample was composed of 155 T2D cases and 1,564 controls.

2.2.2.2 Lothian Birth Cohorts of 1921 and 1936 (LBC1921 and LBC1936)

LBC1921 and LBC1936 are longitudinal prospective cohort studies based in the Lothian region of Scotland (Edinburgh and surroundings), which were established to understand the general causes of aging and other outcomes of interest [124]. Initially, a cognitive test was undertaken by children attending different schools across Scotland in 1932 and 1947 (mean age 10.9 years). Approximately seventy years later, some of the participants still leaving in the Lothian region were successfully recruited to be part of the LBC1921 and 1936 follow-up studies. Some of the variables recorded were cognitive ability, anthropometric measures, sociodemographic factors, history of disease, in addition to biological samples. Following informed written consent, DNA was extracted from peripheral whole blood for a subset of 514 participants from LBC1921, and 1,004 participants from LBC1936. Ethics approval for this study was granted by the Multi-Centre Research Ethics Committee for Scotland (Wave 1: MREC/01/0/56), and the Lothian Research Ethics Committee (Wave 1: LREC/2003/2/29).

Despite the availability of genotype and DNA methylation data for LBC1921 and LBC1936, there was larger difference in mean age of participants between LBC1921 and ALSPAC (LBC1921 27 years older than in ALSPAC), compared to difference in mean age between LBC1936 and ALSPAC (LBC1936 18 years older than in ALSPAC). Thus, based on having a higher concordance in mean age with ALSPAC, LBC1936 was selected as the replication study for the EWAS in T2D. The subsample of LBC1936 included in the replication of the EWAS was composed of 915 participants, 110 were T2D cases and 805 were controls.

2.2.2.3 Rotterdam Study (RSIII-1 and RS-Bios)

The Rotterdam study is a large prospective population-based cohort study from Rotterdam, The Netherlands, aimed at understanding the determinants, incidence and progression of chronic diseases in the elderly[89]. In 1989, 7,983 females and males aged 55 or over, who lived in the

Ommoord district in the city of Rotterdam, were first contacted to conform the RSI study. Two more cohorts were subsequently established as new participants in the region turned 45 years or older. The RSII (2000) study included 3,011 new samples, and the RSIII (2006) study added another 3,932 samples. To date, six follow-ups are available for RSI, four for RSII, two for RSIII, and the latest follow-up includes 4,000 samples from the RSIV study, which started in 2006. At baseline, home interviews were conducted, and extensive examinations were carried out in the research centre, where metabolic and cardiovascular health factors were measured. DNA was extracted from whole blood using standard procedures[125]. The Rotterdam study was approved by the institutional review board (Medical Ethics Committee) of the Erasmus Medical Centre, and by the Review Board of The Netherlands Ministry of Health, Welfare and Sports[126].

Two sub-studies from the Rotterdam study were included in the replication of the EWAS in T2D, they were the first follow-up of the RSIII cohort (RSIII-1), and RS-Bios, a sub-study that includes participants from the fifth follow-up of the RSI, RSII and RSIII cohorts. For the EWAS in T2D, a subsample of 728 participants from RSIII-1 was included, 74 individuals were T2D cases and 654 were controls, while another 735 participants were included from RS-Bios, 108 individuals were T2D cases and 627 were controls.

2.2.2.4 Southall and Brent REvisited (SABRE) study

SABRE is a UK population-based study initiated in 2008 as a 20-year follow-up based on two previous cross-sectional studies: the Southall and Brent studies. These studies recruited at baseline middle-age participants between 40 and 69 years of age, most of them males of South Asian (n=1,711) or European (n=1,762) origin, living in West London, UK, between 1988 and 1991[122, 127]. The motivation for SABRE was the increasing risk of cardiometabolic diseases in middle-age adults reaching pensionable age, and the underlying ethnic differences observed in the risk of disease[127]. Multiple sociodemographic, disease risk, and cardiometabolic factors, have been measured at baseline and at the follow-up. DNA was extracted from peripheral-blood samples collected at baseline using standard protocols. All participants in SABRE gave written informed consent, and ethical approval for the baseline wave was obtained from Ealing, Hounslow and Spelthorne, Parkside, and the University College London research ethics committees. For the replication of the EWAS of glycaemic traits, a subsample of 382 European males (age range 46 years to 58 years) was used based on the availability of necessary measures at baseline, and good quality DNA.

2.2.3 Comparison of DNA methylation assays used across studies

DNA methylation was quantified using the HM450 array based on bisulphite-converted genomic DNA extracted from whole-blood samples in all the studies included in the replication of the EWAS. In addition, each study applied their own primary sample quality assessment, and pre-processing of methylation data using different normalization methods, and specific criteria to exclude samples and probes from further analyses[89, 121, 122, 128]. In common across studies, was the use of methylation intensities per probe as a continuous measure, with β-values ranging between 0 (completely unmethylated CpG) and 1 (completely methylated CpG). Table 2-4 provides more detail of the different methods used across studies for the assessment and pre-processing of DNA methylation data.

Other differences between studies were observed in the generation of variables to adjust for batch effects in the methylation data, and in the prediction of white blood cell subsets. For instance, correction for batch effects was performed using surrogate variables in ALSPAC (see section 2.2.7 Methods in EWAS), and the first 10 principal components (PCs) in KORA. Specifically, PCs in KORA were obtained from signal intensities of 235 positive control probes in the HM450 array[89]. One disadvantage of using PCs instead of surrogate variables in KORA, was the inability to control for and exclude PCs associated with T2D before the EWAS to avoid over-adjustment of the analysis. Thus, the use of PCs, instead of surrogate variables, was considered a potential source of heterogeneity in results of the EWAS of T2D in KORA. Further inspection was required when conducting the meta-analysis to evaluate the impact of adjusting methylation data for PCs versus surrogate variables.

White-cell counts were predicted using the Houseman algorithm for six cell-types in ALSPAC and in most of the other cohorts, except for SABRE that included an additional cell type by separating granulocytes into eosinophils and neutrophils. Also, differently from ALSPAC, the RS-Bios study provided direct white-cell counts for lymphocytes, monocytes and granulocytes.

Table 2-4 Detail of the methods used for the assessment and pre-processing of DNA methylation across different cohorts included in the replication of EWAS.

	ALSPAC	**KORA**	**LBC1936**	**RSIII-1**	**RS-Bios**	**SABRE**
Methylation assay	HM450	HM450	HM450	HM450	HM450	HM450
Inspection of raw probe intensities	Unknown	GenomeStudio: dye staining, hybridization, nucleotide extension, bisulfite conversion, negative controls and non-polymorphic controls	Manual: dye staining, hybridization, nucleotide extension, bisulfite conversion	Unknown	Unknown	illumina iScan
Normalization	Functional Normalization in *meffil*	Quantile Normalization	Unknown	SWAN	DASEN	Functional Normalization in *meffil*
Background correction	none	Smooth quantile normalization (R package *lumi*)	*minfi*	separate colours	separate colours	unknown
Detection P value cut-off	0.01	0.01	0.01	0.01	0.01	0.01
Sample call rate threshold	95%	none	95%	95%	95%	Unknown
Marker call rate threshold	90%	95%	95%	Unknown	Unknown	Unknown
Sex-mismatch exclusion†	yes	yes	yes	Unknown	Unknown	yes
Genotype-mismatch exclusion‡	yes	Unknown	yes	Unknown	Unknown	Unknown
Nbeads filter	3	3	3	3	3	3
Probes in autosomes after normalization	471,226	473,864	465,861	463,456	419,937	473,172
All covariates used in model (technical and biological)	Age, Sex, Smoking, BMI, Houseman 6 WBC subsets, Surrogate variables	Age, Sex, Smoking, BMI, Houseman WBC subsets, 10 control probe PCs	Age, Sex, Smoking, BMI, Houseman WBC subsets, Surrogate variables	Age, Sex, Smoking, BMI, Houseman WBC subsets, Surrogate variables	Age, Sex, Smoking, BMI, direct count of WBC, Surrogate variables	Age, Smoking, BMI, Houseman WBC subset*, Surrogate variables

†Exclusion of samples with sex mismatch between observed and expected sex, based on methylation levels for probes in the X-Y chromosomes. ‡Exclusion of samples if a mismatch is found when cross-validating genotype data between 65 SNP control probes in the HM450 array, and genotype data in genotyping-chip for the same sample. * Prediction of white cell-counts using the Houseman algorithm for seven cell-types: CD4T, CD8T, NK, B cells, Monocytes, Eosinophils and Neutrophils.

2.2.4 Definition of T2D across studies

There was also variation in the way T2D was defined across cohorts, but in general, the most commonly used criteria for diagnosis were levels of fasting glucose ≥ 7.0mmol/l, and medication to lower glucose in blood. To a less extent, diagnosis of T2D was made based on medical diagnosis, HbA1c ≥ 6.5%, and post-load glucose ≥ 11.1mmol/l (Table 2-5). In addition, in the Rotterdam studies, approximately 1.7% (n=25) participants had non-fasting measures of glucose. Despite this limitation, these samples were included in the analysis as most of them were controls. Diagnosis of T2D based exclusively on HbA1c, was characteristic of the LBC1936. It is important to recognize that differences in case ascertainment across studies, might have introduced some heterogeneity in results of the meta-analysis.

Table 2-5 Outline of main criteria used by each cohort for the diagnosis of T2D.

	Medical diagnosis	FG ≥7.0mmol/l	HbA1c ≥ 6.5%	2-h PG≥ 11.1mmol/l	Medication
ALSPAC	**X**	**X**			**X**
KORA	**X**	**X**		**X**	**X**
LBC1936			**X**		
RSIII-1†		**X**			**X**
RS-Bios†		**X**			**X**

† Some samples had non-fasting glucose measures.

2.2.5 Definition of normoglycemia and assessment of glycaemic traits in SABRE

At baseline, 1,761 European samples were available in SABRE, but 1,643 of them were retained for further analyses based on the availability of necessary phenotypic data, and on the absence of T2D (Figure 2-3). Control samples in SABRE were characterized by having FG < 7.0mmol/l, HbA1c < 6.5%, 2-h PG < 11.1mmol/l, no self-reported use of medication to treat diabetes, and no medical diagnosis of diabetes. From these 1,643 controls, only 382 of them were taken forward for the replication of the EWAS of glycaemic traits considering their availability of methylation data (Figure 2-3).

Assessment of glycaemic traits

Glycaemic measures undertaken in participants in SABRE were assessed using whole blood samples collected at the first visit between 1988-1991. For some of these measures, an overnight fast was required, as it was for fasting glucose and fasting insulin, but not for HbA1c (%), 2-hours glucose (mmol/l), and 2-hours insulin (μIU/L). HOMA scores were calculated using fasting glucose and fasting insulin as described before in section 2.1.2. Glycaemic traits found in common between SABRE and ALSPAC were FG, 2-h PG, fasting insulin, and the HOMA scores. In contrast, HbA1c and 2-h insulin

were traits exclusively measured in SABRE, while no records of fasting proinsulin were available on this dataset.

The haemoglobin A1c (HbA1c), or glycated haemoglobin, was assessed in SABRE samples at the University of Glasgow 20 years after blood sample collection and storage at -80°C, using the Tina-quant Haemoglobin A1c III assay (Roche/Hitachi MODULAR P analyser, Roche Diagnostics, Indianapolis, IN). This protocol has been certified by the NGSP (www.ngsp.org) and standardised to the Diabetes Control and Complications Trial (DCCT) assay. Further detail of the protocol used for the assessment of HbA1c can be found elsewhere[129]. Other glycaemic measures were undertaken at a local health centre after an overnight fast. Fasting glucose and 2-h post load glucose were determined using the hexokinase method (Roche, Basel, Switzerland) in participants with unknown T2D status[130], while fasting insulin and 2-h insulin were measured using the ELISA technique (Boehringer Mannheim, Mannheim, Germany)[130].

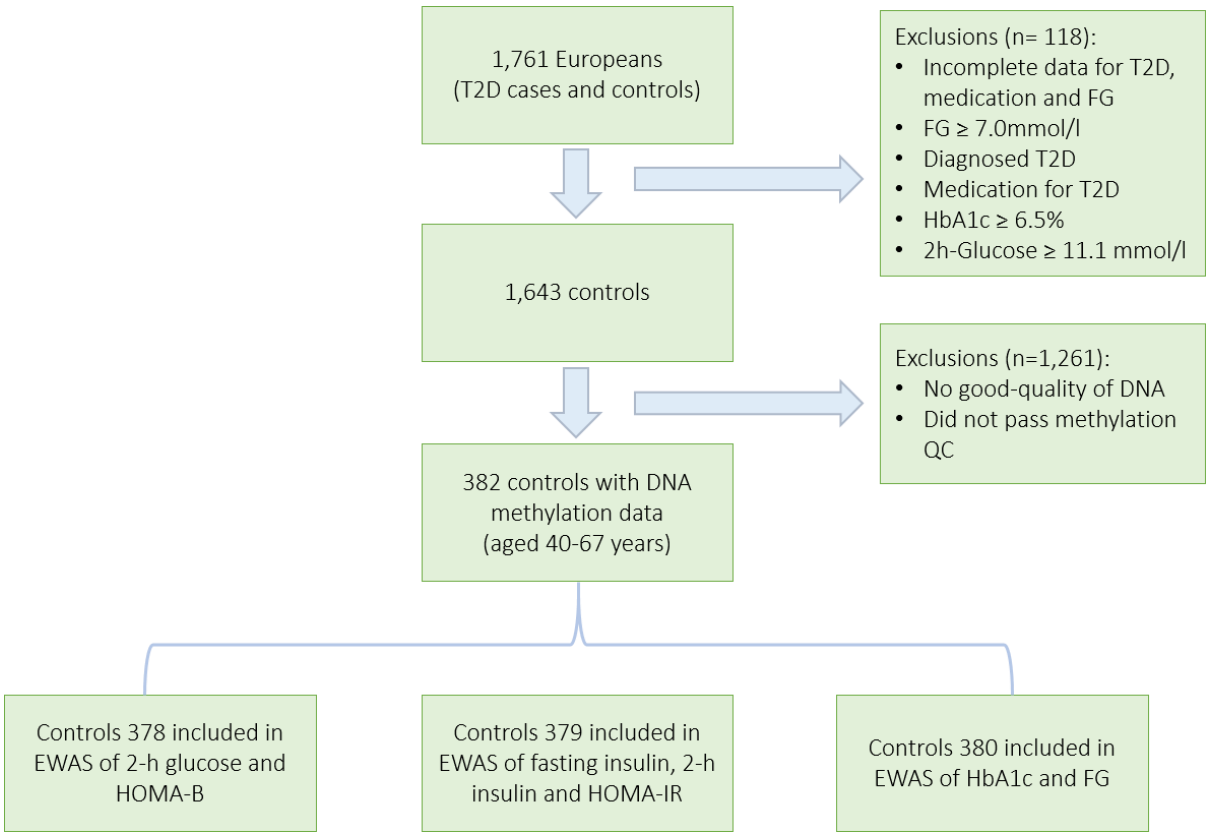


Figure 2-3 Flow-diagram illustrating the process of sample selection of participants in SABRE included in the replication of the EWAS of glycaemic traits.

2.2.6 Comparison of main covariates across studies

As in ALSPAC, additional covariates were required from the replication cohorts to adjust the analyses, or to describe the study population at baseline. In some instances, the definition of these covariates differed from the way they were defined in ALSPAC. For instance, smoking was classified as never, former and current smokers in the Rotterdam studies, LBC1936 and SABRE, but only as non-smokers and smokers in ALSPAC. In KORA, a fourth category for smoking was included for casual smokers, who were defined as participants without daily smoking.

For the variable of socioeconomic status, no records were available in KORA and the LBC1936 studies, while in SABRE, socioeconomic status was defined based on having a manual or non-manual job, which differed from the definition used in ALSPAC (see section 2.1.2). Also different from ALSPAC, physical activity was measured in the Rotterdam studies using a continuous variable for metabolic equivalents of exercise or MET hours per week, while in SABRE physical activity was measured using an exercise score (Kcal/week). General characteristics of participants in the different cohorts included in the replication of the EWAS, can be found in Table 2-6.

2.2.7 Data transformation

To achieve an approximately normal distribution, continuous values of fasting insulin (mmol/l) and the HOMA scores were log-transformed prior to the EWAS in the ALSPAC and SABRE datasets. Log-transformation was also applied for 2-h insulin in SABRE. Measures of fasting proinsulin in ALSPAC were transformed using the reciprocal (i.e. 1/x) in a complete dataset, while the Log2 transformation was used in a second dataset pruned for outliers (see section 2.3.3.1). No transformation was required for continuous measures of FG (mmol/l), 2-h PG (mmol/l) and HbA1c. In addition, one sample in SABRE detected with an FG of 0.29mmol/l, was considered an outlier and this value was set to missing.

2.3 Methods in epigenome-wide association studies

2.3.1 Assessment of data prior to EWAS

The following sections describe post-normalization quality control measures applied to the methylation data to explore potential sources of underlying variation, selection of covariates for the EWAS models, removal of remnant sources of noise in the data, and other steps to prepare for the EWAS.

Table 2-6 Baseline characteristics of participants in ALSPAC and in other five studies included in the replication of the EWAS in T2D and the EWAS of glycaemic traits.

	ALSPAC	KORA (F4)	LBC1936	RSIII-1*	RS-Bios*	SABRE
Sample size	1,050	1,719	915	728	735	382
Country	UK	Germany	Scotland	The Netherlands	The Netherlands	UK
Ethnicity	Europeans	Europeans	Europeans	Europeans	Europeans	Europeans
Study Design	Prospective Birth Cohort	Population based	Population based	Population based	Population based	Population based
Age (yrs)	49.9	61.0	69.6	59.7	67.6	52.3
Sex (% male)	39.5	48.9	50.5	45.4	42.0	100.0
Fasting glucose (mmol/L)	5.4	5.6	---	5.5	5.7	5.4
HbA1c (%)	---	5.6	5.9	---	---	5.5
Fasting insulin (µIU/L)	4.5†	---	---	---	---	7.4
% Fasting	100.0	100.0	100.0	97.5	99.1	100.0
T2D (% cases)	4.6	9.0	12.0	10.2	14.7	---
Body mass index (kg/m^2)	26.8	28.1	27.8	27.5	27.7	26.0
Waist circumference (cm)	89.4	----	----	93.5	94.3	91.5
Smoking ‡						
Never smoked	90.8	43.7	47.0	29.5	34.4	27.0
Former smoker	---	43.86	41.7	43.6	55.0	40.6
Current smoker	9.2	12.5	11.3	26.9	10.6	32.2

*Rotterdam studies (RSIII-1/RS-Bios). **Physical activity defined as a categorical variable in ALSPAC (≤ 4h/w or >4h/w), and as a continuous measure in the Rotterdam studies (MET-h/week) and in SABRE (score in Kcal/week). †Fasting insulin was measured in a subset of 612 normoglycemic females in ALSPAC. ‡ Smoking was categorized as never and current smokers in ALSPAC, while in KORA four categories were available (never/casual/former/current). To match with existent smoking categories in other studies, casual and former smokers in KORA were combined into a single category.

2.3.1.1 Detection of structure in the methylation data

Multi-dimensional scaling (MDS) was used to inform of the presence of structure in the methylation data at first glance, and to determine if known biological variables were driving any underlying structure in the data. Detail of an MDS analysis can be found elsewhere[131]. Briefly, an MDS plot is a representation of the pairwise distance between samples, generally using two dimensions or scales to facilitate the interpretation of results. Distance between samples based on the methylation data, was calculated using the function *dist.matrix* from the R package wordspace (version 0.2-0)[132], while coordinates of the samples in the MDS plot were derived from the matrix of distances using the *cmdscale* function in R version 3.4.1. Finally, MDS plots were generated using the R package ggplot2[133].

To assess the effectiveness of the MDS analysis in capturing the total dimensionality of the data more formally, the *stress*, the *goodness of fit* (GOF), and the R^2 were calculated. Optimal values for these metrics were a *stress* < 0.2, a GOF > 0.6, and an $R^2 > 0.5$ at $p<0.05$. The *stress* value was calculated using the function *isoMDS* in the R package MASS[134], GOF was obtained using the function *cmdscale* in R version 3.4.1, and R^2 was calculated as the squared of the correlation between coordinates of the samples in the MDS space, and secondary distances derived from these coordinates.

2.3.1.2 Selection of covariates

The different EWAS conducted throughout this thesis were adjusted for potential confounders that could be associated with T2D, any of the glycaemic traits, and DNA methylation, to avoid spurious findings in downstream analyses. Potential confounders were selected from established covariates described in the literature in relation to epigenome-wide studies (i.e. age, sex, smoking, cell-type heterogeneity, BMI), some of these factors also identified in association with T2D or DNA methylation in the ALSPAC dataset. To test for associations between confounders and T2D, univariate linear and logistic regressions were conducted with T2D as the independent variable. As an additional analysis, the correlation between lipid measures, SES, and BMI was calculated to determine if adjusting for BMI alone could account for the effect of these other factors, thereby reducing the number of covariates included in the model. Univariate regression analyses were conducted to assess associations between established confounders and average global DNA methylation across all autosomal probes that surpassed QC. Also, an exploratory analysis was conducted to test if T2D was associated with global methylation in an unadjusted regression, and in a regression including an interaction term between T2D and sex.

2.3.1.3 Exclusion of SNP-probes

For the EWAS of glycaemic traits, probes located in the same position as common variants (SNPs), also known as control probes or SNP-probes (n=65 CpG sites), were excluded before the analysis to avoid confounding by the genotype in results of the EWAS. Differently, control probes were not excluded when conducting the EWAS in T2D.

2.3.1.4 Removal of outliers

Extreme values of methylation can arise due to technical factors or rare genetic variants, and if unaccounted for in the analysis, they can lead to spurious results[135]. Trimming of outliers was applied using the Tukey method[136], where outliers were defined as values three times above or below the interquartile range for the upper and lower quartiles of methylation at a CpG site, respectively. This method of trimming was appropriate as it did not depend on distributional assumptions of the data[135, 136]. Identified outliers were set as missing values in the beta-matrix for subsequent analyses.

Removal of outliers was conducted before or after the surrogate variable analysis (SVA). When applied before SVA, missing values were replaced by the median of the probe to allow the SV analysis to be performed, as it is sensitive to missing values.

2.3.1.5 Batch effects: SV analysis

Batch effects are systematic biases in the data that are unrelated to the variable of interest, and that appear as a results of sample handling[91]. Surrogate variable analysis (SVA) is an established method to adjust for batch effects in microarray data, and it was implemented using the R package *sva*[137]. In principle, surrogate variables (SVs) are estimates that capture unexplained sources of variation or noise in highly-dimensional data, such as that generated by DNA methylation arrays[137]. Once generated, SVs can be used as covariates in the analysis to adjust for residual variation. The advantage of using SVs over conventional batch effect variables, is that the former captures technical variability in addition to underlying biological sources of noise in the data, which are independent of the variable of interest in the EWAS model[138].

Two important factors were considered when generating SVs: 1) the number of known variables to be used in the null and full model-matrices for the SVs, which determines the nature of the variation captured by the SVs; and 2) the independence of the SVs from the variable of interest. In the first case, models used to generate SVs included all the covariates utilised in the most adjusted EWAS, meaning that the SVs captured latent variation in the array unrelated to the covariates and the

variable of interest in the EWAS model. To fulfil with the condition of independence, SVs were only included in the EWAS models if there was little evidence that they were associated with T2D or any of the glycaemic traits ($p>0.05$ in multiple linear regressions). For each EWAS model, 10 SVs were calculated in all the cohorts except KORA, which used a similar analysis (PCs) as previously mentioned (2.2.3).

2.3.2 EWAS in T2D

The EWAS of T2D was conducted using multivariable linear regressions following the EWAS pipeline available in *meffil* (version 1.0.0) and using the University of Bristol high performance computer BlueCrystal, phase 3 (bluecrystalp3.bris.ac.uk). From the different models provided in *meffil* to fit results of the EWAS, the one specified was the model considering user-supplied covariates, without generation of SVs or ISVs by the program. In the linear model, untransformed β values (range 0 to 1) were regarded as the outcome and case control T2D was considered the exposure. Covariates were included additively in the model. In addition, knowing that differences in methylation between females and males are present for probes in sex chromosomes, these markers were excluded from the main analysis. On average, 376,820 probes in autosomes were included in the EWAS across studies.

Three linear models were specified for the EWAS: a basic model adjusting for age, sex and independent SVs, a second model additionally adjusting for cell-counts, and a third model additionally adjusting for BMI and smoking. The genomic inflation factor (λ), which is used to identify population stratification or cryptic structure in the data[89], was calculated using the median method, with measures above 1.0 indicating high genomic inflation. Results of the EWAS were not corrected for high genomic inflation due to the homogeneity of the sample (i.e. all European Caucasians).

Correction for multiple testing was applied using the conservative Bonferroni method to account for the number of autosomal markers tested, with significant associations regarded at $p < 1.07 \times 10^{-7}$ (adjusted-$p < 0.05$). Description of top results was made using a borderline significant threshold at $p < 1.0 \times 10^{-5}$. Results of the EWAS were furtherly annotated by specifying the nearest gene for a CpG site, its genomic context and CpG island context using Illumina-provided annotation data, according to hg19/GRCh37 coordinates. Results were interpreted as a difference in methylation (0-1 scale) in T2D cases vs controls. For the top CpG sites identified in the EWAS in ALSPAC, the adjusted-R^2 was calculated to estimate the proportion of variation in methylation explained by T2D and covariates.

To guarantee that parameters of the EWAS used in ALSPAC were similarly applied in the replication studies, a prespecified analysis plan with complete detail of the protocol used in *meffil,* was provided to the responsible analyst in the collaborating cohorts. Results were then submitted in a comma delimited format (CSV) to facilitate further manipulation of the data for annotation purposes, and to conduct the meta-analysis (see section 2.4).

2.3.2.1 Sensitivity analysis

To assess the effect of adjusting for BMI in the association between T2D and DNA methylation, a secondary model was run without adjustment for BMI. Results between the model with and without adjustment for BMI were compared in terms of similarity of top signals, average change in effect-size and P-value. A second sensitivity analysis involved stratifying the population by quartiles of methylation at the strongest locus identified in the EWAS, and comparing difference in mean methylation and in the risk of T2D between Q1 and Q4. Quartiles of methylation were generated in R using the function *quantile*. Difference in methylation across the quartiles was calculated using a two-way ANOVA test, including an interaction term between T2D and the quartile. To determine the risk of T2D, a univariate logistic regression was applied, while presence of a linear trend in the risk of T2D across the quartiles was determined using the *prop.trend.test* function in R version 3.3.3. Results were considered associated at $p<0.05$.

2.3.2.2 Risk-factor analysis and further adjustment of top signals of the EWAS in ALSPAC

At the top CpG sites identified in the EWAS of T2D in ALSPAC, the effect of further adjustment for known T2D risk factors was explored. To do this, different univariate linear regressions were conducted first to test for the association between methylation at the top CpG site and the risk factor. Risk factors identified in strong association with methylation at $p<0.05$, were then included in a multivariable regression model, with methylation as the response variable and T2D and additional risk factors as the predictor variables. Risk factors interrogated were different lipid measures, SBP and DBP, FG, fasting insulin, HOMA scores, c-reactive protein, waist circumference and waist-hip ratio.

2.3.2.3 Results of the EWAS at candidate loci for T2D

Results of the EWAS were inspected for CpG sites located within the region of ten candidate loci for T2D: *TCF7L2, CDKAL1, IGF2BP2, SLC30A8, FTO, PPARG, JAZF1, HNF1B, KCNQ1* and *THADA*. These loci were a subset of 56 loci previously included in a polygenic risk score for T2D validated in ALSPAC samples (see Chapter 3). Candidate loci were selected based on the strength of the evidence from genetic studies (i.e. stronger effect-size and smallest P-value), and for some of them, based on

evidence of differential methylation reported in epigenetic studies of T2D (i.e. all loci except *CDKAL1, IGF2BP2* and *PPARG*)[60, 107]. Plots were generated for each candidate loci to represent the distribution of effect-estimates against the -Log10(P-value) for CpG sites mapping within the region of the candidate loci. P-value and effect estimates were extracted from the fully adjusted EWAS model.

2.3.3 EWAS of glycaemic traits

In the EWASs of FG, 2-h PG, fasting insulin, fasting proinsulin, and the HOMA scores, methylation was included as the exposure variable in multivariable linear regression models. As an additional analysis in ALSPAC, a multivariable logistic regression was used to test the association between methylation as the exposure against T2D, to compare EWAS findings when methylation is treated as the exposure and outcome. Multivariable linear and logistic regressions were performed using the functions *lm* and *glm,* respectively, in R version 3.0.2.

Three adjustment models were implemented in these EWAS: model 1 adjusted for age, sex, independent SVs and cell-counts, model 2 additionally adjusted for smoking, and model 3 additionally adjusted for BMI. For EWAS including only females (EWAS of 2-h PG, fasting insulin and HOMA scores in ALSPAC), or males (the EWAS of glycaemic traits in SABRE), no adjustment was applied for sex. Adjustment for multiple testing was applied using the Benjamini-Hochberg (FDR) method, with associations regarded significant at FDR <0.05. Results were reported in odds ratios (OR) for the EWAS of T2D, and as beta-coefficients for the remaining traits. For phenotypes that were normal-transformed before the analysis, results were back transformed, and beta-coefficients were reported as $[e^{(beta/100)}]$ when using the log-transformation, as $[1 \,/\, (beta * 100)]$ when using the reciprocal, and as $[2^{(beta/100)}]$ when using the Log2 transformation. Results were interpreted as a unit change in the phenotype per 10% increase in methylation, or as the risk of T2D per 10% increase in methylation.

2.3.3.1 Sensitivity analysis

The distribution of glycaemic traits was inspected in the ALSPAC and SABRE datasets to rule out the presence of potential outliers that could introduce bias in downstream analyses. Outliers were considered as samples above the 99th percentile of the total distribution of the trait in the dataset, which were non-biologically plausible measures, or were values extremely outside the normal range for the phenotype. Detected outliers were regarded as missing values, and a sensitivity analysis was performed excluding these samples from the EWAS. Results were compared across analyses based on the similarity in the top signals, change in the magnitude of the effect-estimate and p-value.

2.3.4 Removal of problematic DNA methylation probes

Probes in the HM450 array that were considered problematic based on a list published by Naeem *et al*.[139], were removed from summary results of the EWAS and before the interpretation of the main findings. According to Naeem and colleagues, problematic probes are likely to generate noisy signals and increase the risk of false positives[139], thus highlighting the importance of excluding them from posterior analyses. Probes in the HM450 array were excluded if they had multiple reactivity across the genome, showed low correlation with methylation determined using the whole genome bisulphite sequencing (WGBS), and if probes hybridized to repetitive DNA sequences. In total, 383,104 autosomal probes remained for subsequent analyses in the ALSPAC dataset after excluding Naeem-listed probes.

2.3.5 Visual display of results

Inspection plots

Results were inspected using Manhattan, Q-Q plots, and boxplots only for the top-ten probes identified with the lowest P-value. These graphs were reported in an HTML format using *meffil* and the function "meffil.ewas.report", allowing evaluation of the success of every step of the analysis, particularly for results of the EWAS of T2D, as a different protocol was used for the EWAS of glycaemic traits. More refined Manhattan plots indicating top signals surpassing borderline significance or epigenome-wide significance, and Q-Q plots with confidence intervals and Lambda-value, were created using the R package qqman[140]. Volcano plots, which show the distribution of effect sizes against the -Log10(P-value) for each CpG site in the array, were generated using the function *plot* in R version 3.3.3, and the R package calibrate to annotate probes in the volcano based on their effect-size and level of significance.

Genome browser tracks

A close-up representation of the genomic context of the top signal of the EWAS of T2D in ALSPAC was made using the UCSC Genome Browser (http://genome.ucsc.edu/), depicting the surrounding region of the CpG site of interest in terms of histone marks, CpG islands, genes and DNaseI clusters. In addition, a custom track was included using a bedGraph file to illustrate results of the EWAS in the fully adjusted model, with solid bars representing the score given to each association. The score was calculated as: -Log10(*P-value*) * - (*t-statistic*).

Correlograms and Heatmaps

For the EWAS of glycaemic traits in ALSPAC, and the meta-EWAS of glycaemic traits between ALSPAC and SABRE, correlograms were generated to determine the level of correlation across phenotypes using effect estimates obtained for a subset of CpG sites. CpG sites compared across traits were identified as the strongest signals with the lowest p-value in the meta-EWAS of one of the glycaemic traits analysed. A matrix of correlations between traits across CpG sites was estimated using the Spearman method to account for the non-parametric distribution of effect estimates within each trait. In the correlogram, strength of the correlation and significance of the correlation were represented by colour intensities (more intense if $r > 0.5$, and coloured if significant at $p<0.01$). The correlogram was generated using the R package corrplot [141]. In addition, a heatmap was generated to represent the clustered association between traits with respect to standardised values of effect estimates obtained at the selected CpG sites using Z-scores (mean=0 and SD=1). The heatmap was created using the R package gplots[142] and the function *heatmap.2* available in the same package.

2.4 Meta-analysis of EWAS

Two meta-analyses were conducted: the meta-EWAS of T2D, which included EWAS results from five cohorts (ALSPAC, KORA, LBC1936, RSIII-1 and RS-Bios), and the meta-EWAS of glycaemic traits, which included EWAS results from two cohorts (ALSPAC and SABRE). Most of the methods were the same across the two meta-analyses, but there were some variations with regards to multiple testing correction and further sensitivity analyses in the meta-EWAS of T2D. Different adjustment models were implemented in both meta-EWAS, but from now on, when taking about the main model, this refers to the model adjusted for age, sex, SVs, predicted cell-counts and smoking.

2.4.1 Quality control

Before the meta-analysis, quality control was applied to assess the validity of estimates of the EWAS provided by each cohort, and to ensure their comparability across studies, using the functions provided by the R package QCEWAS (version 1.1-0)[143]. The output of the QC inspection indicated the number of probes that passed QC and those that did not due to missing values for summary statistics, inconsistent units of measurement, probes located in sex chromosomes, and outliers based on pre-established units of measurement of the outcome.

The QCEWAS package also reports inspection plots to summarize the distribution of standard errors, effect estimates, p-values (Q-Q plot), and effect estimates against p-values (i.e. volcano plot) for the individual EWAS. A lambda value was reported for each analysis to identify genomic inflation (high if

λ ≥ 1.0). To compare results across studies, inspection plots were used to show the distribution of effect estimates according to sample-size, and a precision plot using the squared-root of the sample-size against the inverse of the median standard error. Pruned datasets containing markers that surpassed QC, were saved in a new txt.gz file, and were subsequently used to run the meta-analyses.

Expected results from the inspectional plots were:

- Right skewed distribution of the SE, with most of the values close to zero.
- Effect estimates centred around zero, with lower and upper limits based on the expected units of measurement of the outcome (i.e. change in untransformed β-values of methylation, and unit change in the glycaemic traits).
- Correlation between observed and expected p-values equal or close to 1.0 to rule-out data mix-up or formatting error when generating results.
- λ equal or close to one to rule-out population stratification or non-random allocation of p-values. High λ is expected in small studies as p-values are not perfectly distributed at random.
- Similar distribution of effect estimates across studies, with larger studies showing narrower distribution of effects (i.e. better control over outliers).
- Increase in the precision at reporting effect estimates in larger studies compared to smaller studies. If some studies show extreme variance in the precision plot, they can be deemed as outliers.

2.4.2 Meta-analysis

A fixed-effect (FE) inverse variance weighted meta-analysis was conducted in METAL (version 2011-03-25)[144] using the University of Bristol high performance computer BlueCrystal, phase 3 (bluecrystalp3.bris.ac.uk), and EWAS results that had undergone QC. On average, 376,820 autosomal probes were included in the meta-EWAS of T2D, and 376,415 autosomal probes in the meta-EWAS of glycaemic traits. Analyses were corrected for multiple testing using specific methods as previously mentioned (see section 2.3.2 and 2.3.3), and top signals with the lowest p-value identified in the main meta-EWAS model, were taken forward for further analyses.

Heterogeneity across studies was measured using the I^2 statistic, with evidence of high and significant heterogeneity defined as $I^2 > 40\%$ and heterogeneity p-value <0.05, respectively. Top results (at $p<1.0x10^{-5}$) of the meta-analysis were reported using summary tables containing annotation data for the CpG sites (i.e. nearest gene, Chr, position), results of the individual EWAS,

and estimates of the meta-analysis (i.e. effect-estimate, SE, direction of effect, adjusted p-value, heterogeneity statistic (I^2), and heterogeneity p-value). In some cases, summary results were presented by comparing estimates of the meta-analysis across the different adjustment models, based on top-signals identified in the main model. Summary plots of the meta-analysis were generated using Manhattan plots, QQ plots, and volcano plots.

2.4.3 Sensitivity analysis

For the meta-EWAS of T2D, top signals identified across adjustment models were re-analysed using the random-effects (RE) model to allow for potential differences in effect estimates across studies. Compared to the FE model, the RE is more appropriate for obtaining more generalizable results across populations for predictive purposes[145], but it is not recommended for discovery purposes due to the limited power of this model when compared to the FE model[145]. The RE model was conducted in METAL using the DerSimonian-Laird estimator, and the command ANALYZE RANDOM, which was a new function added to the main program by Gibran Hemani (https://github.com/explodecomputer/random-metal/raw/master/executables/metal). Differences between the FE and RE models were described based on the effect estimate, SE and P-value. Further inspection of heterogeneity among studies included the observation of forest plots, heterogeneity statistics, and performing a leave-one-out analysis using the R package metafor[146]. Forest plots and leave-one-out analysis were performed only for top-signals detected in the main model of the meta-EWAS of T2D.

Knowing the potential bias that the KORA study could have introduced in results of the meta-EWAS by adjusting for PCs instead of SVs, a sensitivity meta-EWAS of T2D was performed excluding KORA. Results were compared between the meta-EWAS with and without KORA in terms of similarity in the top signals identified in the main model, average heterogeneity, and statistical significance of main findings.

2.5 Detection of differentially methylated regions

An analysis of differentially methylated regions was applied to results of the fully adjusted model of the EWAS of T2D in ALSPAC (see Chapter 4), and to results of the main model in the meta-EWAS of T2D. The aim in doing a DMR analysis was to reinforce findings from the single-site analysis by identifying regions where groups of differentially-methylated CpG sites are more likely to influence gene expression[44, 91], and to validate results of the single-site analysis by identifying similar patterns of methylation in neighbouring CpG sites[44]. From the different methods currently available to

identify DMRs[44], *comb-p* was selected as the discovery method for its ability to account for the variable spacing of probes within the array, and for the simplicity of its implementation. In addition, *DMRcate* was selected as a validating method for DMRs detected in *comb-p* based on the methodological compatibility between these two approaches[147]. Validation of DMRs through DMRcate was only applied for results of the EWAS of T2D in ALSPAC.

2.5.1 Comb-p

All analyses were conducted in the MRC IEU computer server Epi-Franklin (epi-franklin.epi.bris.ac.uk, MRC IEU, University of Bristol). Two scripts were used, one to convert EWAS results from a comma-delimited format (CSV) to a BED file, and a second script to perform the DMR analysis, specifying the following analytical conditions:

- Autocorrelation factor among p-values above 0.04.
- Identify regions among consecutive CpG sites within 500bp, using a sliding window of 50bp.
- Start a region if P-value lower than 0.05.
- Sidak-significant regions at p<0.05

Based on the protocol described by Pedersen *et al*.[148], the first step of the DMR analysis in *comb-p* involved the calculation of an autocorrelation factor (ACF) between p-values from adjacent CpG sites within 500bp. A minimum ACF of 0.04 was considered across sliding windows of 50bp. The Stouffer-Liptak-Kechris correction (*slk*) was applied to account for the autocorrelation of p-values within a region, generating adjusted p-values for each CpG site. These p-values were then corrected for multiple-testing using the FDR method, and regions enriched in low p-values (at p<0.05) were identified by an algorithm using the FDR, *slk* adjusted p-values or the p-value of the EWAS. After identifying candidate regions, a combined p-value was calculated for the region using the Stouffer-Liptak method, with further adjustment for multiple testing using the Sidak correction. Regions of biological interest in T2D were considered at Sidak < 0.05, where the CpG count per region was equal to or above two CpG sites, and where there was consistency in the direction of effect for sites within the region. Obtained DMRs were described according to genomic coordinates, nearest gene, size (bp), CpG count, percent of the average absolute difference in methylation, index CpG site with the lowest p-value, and Sidak-corrected p-value of the region.

Plots displayed

A Manhattan plot was used to display all the regions identified by *comb-p* before applying a Sidak correction. Regions were highlighted in red and plotted against their chromosome position and

original p-values from the EWAS analysis. In some cases, Sidak significant DMRs were represented by single plots showing the effect-estimates and -Log10(P-values) for CpG sites within the region, based on parameters obtained in the EWAS.

2.5.2 DMRcate

All analyses were conducted in BlueCrystal phase 3 using the R package DMRcate[147]. As in *comb-p, DMRcate* uses a single CpG site analysis first, and retrieves a non-directionality measure associated to each CpG site (i.e. t-statistic) to infer regions of interest[44, 147]. Parameters considered were: a kernel size (σ) of 500bp (i.e. optimal kernel size for a modest cut-off), a bandwidth (λ) of 1kb, a *C* parameter of two, and an FDR cut-off of 0.05. The kernel size is a measure of stringency in the detection of DMRs to control for the probability of identifying false positives. The smaller the kernel size, the higher the stringency. Detail of the protocol used by *DMRcate* can be found elsewhere[147]. The output from the analysis are regions of significance identified by the Stouffer method at p-value < 0.05, according to groups of significant CpG sites located within a 1kb distance from each other. DMRs identified in *DMRcate* that were in overlap with DMRs in *comb-p*, were described based on their genomic coordinates, DMR size, nearest gene, Stouffer-corrected p-value, CpG count and average absolute difference in methylation.

2.6 Association between top CpG sites and T2D risk factors

To determine other mechanisms that can be influencing the association between T2D, the glycaemic traits, and DNA methylation, a risk factor analysis was conducted. For this analysis, the association between methylation at the top CpG sites with the smallest p-value from the meta-EWAS, and established T2D risk factors, was assessed. Associations were measured with methylation as the continuous exposure against the risk factor as the outcome, using univariate linear and logistic regressions in R version 3.3.3.

As a sensitivity analysis, methylation at the top CpG sites was stratified into quantiles using the function *cut* in R version 3.3.3. Difference in mean methylation between quantiles was determined using linear regressions, with further validation of results using an ANOVA test (methylation as the dependent variable). In addition, difference in the distribution of the risk factors between quantiles of methylation was estimated using univariable linear regressions for continuous risk factors, and a chi-square test and ordinal chi-square test for categorical, and categorical ordinal variables, respectively. P-values were corrected for the number of tests applied using Bonferroni correction (*p for trend*, α=0.05/quantiles), with results significant at p<0.05.

2.7 Functional exploration

Top signals obtained in the most-adjusted model of the EWAS of T2D in ALSPAC, the DMR analysis, and the main model of the meta-EWAS of T2D and glycaemic traits, were further explored to assess their biological importance in the pathophysiology of T2D. Top signals were defined as associations achieving epigenome-wide significance (at $p<1.07x10^{-7}$) or borderline significance (at $p<1.0x10^{-5}$). For some of the functional analyses, an arbitrary number of associations with the smallest p-value was chosen to improve the sensitivity of these analyses.

2.7.1 Genomic information of top signals

USCS Genome Browser (hg19/ GRCh37, https://genome.ucsc.edu/) was used to observe the genomic context of identified CpGs in terms of position relative to CpG islands, nearest gene, and surrounding epigenetic features. CpGs were annotated using the Illumina manifest data, with further documentation of gene function and potential impact on disease using Gene Cards (https://www.genecards.org/). Because the influence of the CpG site on gene expression could determine a gene of higher biological importance for the CpG site investigated, a search was done for expression quantitative trait methylation sites or eQTM's with respect to top signals identified in the meta-EWAS analyses (see section 2.7.3).

2.7.2 Enrichment analysis for genomic and epigenomic regulatory elements

To further investigate the genomic context of top CpG sites, different enrichment analyses for regulatory elements were conducted based on a multiple CpG site search. The list of top CpG sites included in this search varied depending on the analysis where these signals came from. For instance, for the EWAS of T2D in ALSPAC, the top 1,000 CpG sites with the smallest p-value were taken forward to be investigated, and this arbitrary number was chosen since for this EWAS there were few CpG sites surpassing epigenome-wide significance or borderline significance. For the meta-EWAS of T2D, only top signals identified at $p<1.0x10^{-5}$ were included in the enrichment search, while in the DMR analysis all the CpG sites within the region were analysed. Web-based platforms used for the genomic-context exploration are described below.

2.7.2.1 Experimentally derived functional element overlap analysis of regions from EWAS (eFORGE)

eFORGE (eFORGE v1.2, release date: 1999-2014, http://eforge.cs.ucl.ac.uk/)[149] is a web-based tool for the analysis and interpretation of EWAS data[149]. For a given list of differentially methylated points (DMPs) of interest, eFORGE generates a set of 1,000 background probes selected to match with input DMPs in their location within genes and CpG islands. Background probes are then

compared to epigenomic features of relevance (i.e. histone marks or DNase I hypersensitive Sites-DHS) using data provided by the Roadmap, BLUEPRINT and ENCODE projects for specific cell-types[149]. The output from this analysis is an enrichment score calculated based on the overlap between input DMPs and the position of DHS regions and various histone marks. Signals of enrichment are corrected for multiple testing using q-values derived from the Benjamini-Yekutieli method, with significant enrichment regarded at q<0.01, and non-significant at q>0.05. Representation of the analysis conducted by eFORGE is illustrated in Figure 2-4.

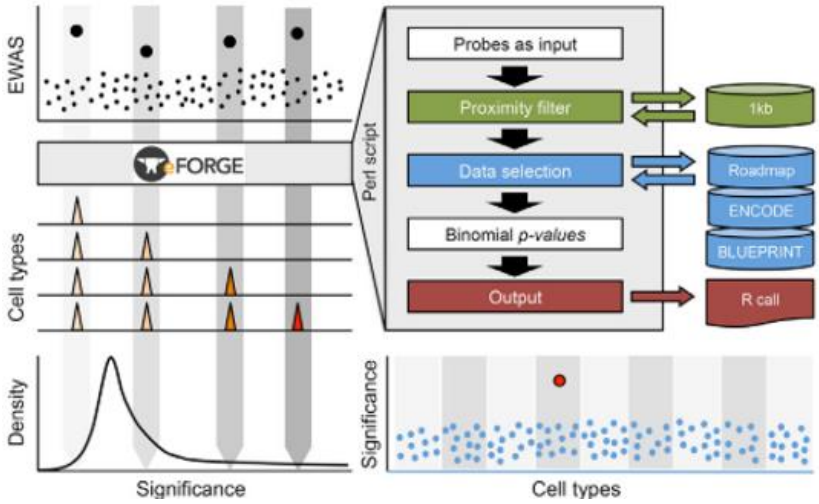
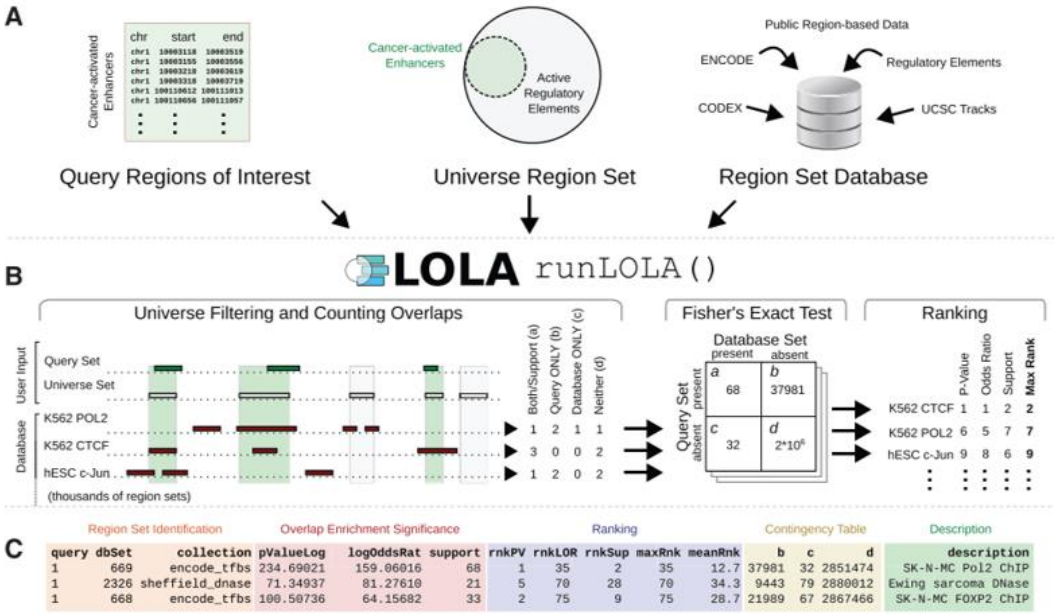


Figure 2-4 Workflow of the analysis conducted by eFORGE to identify enrichment of CpG sites of interest for regulatory elements at specific cell-types. Figure taken with permission from Breeze et al.[149]. eFORGE: a tool for identifying cell type-specific signal in epigenomic data.

2.7.2.2 Locus Overlap Analysis web tool (LOLA)

LOLA (LOLA web version 0c5e2556f, release date: 2015, http://lolaweb.databio.org)[150], another web-based tool, was used to look at the enrichment of target CpG sites and DMRs for functional genomic and epigenomic features reported in different databases (i.e. ENCODE, Codex, UCSC, Cistrome, ROADMAP). Results of the analysis are reported using the -Log10(P-value), Log(OR), or a ranking score derived from the above parameters reflecting the overlap between user-set regions, and regions of different regulatory elements (i.e. transcription factors or TFs, histone marks, transcription start sites or TSS, promoters, DNaseI sites, etc) identified across specific cell-types[150]. The output of the enrichment analysis in LOLA can be presented as a *data.table,* or as bar-plots. A flow-diagram of the analysis conducted in LOLA is illustrated in Figure 2-5, while further detail of the method used was previously reported by Sheffield and Bock[150].



query	dbSet	collection	pValueLog	logOddsRat	support	rnkPV	rnkLOR	rnkSup	maxRnk	meanRnk	b	c	d	description
1	669	encode_tfbs	234.69021	159.06016	68	1	35	2	35	12.7	37981	32	2851474	SK-N-MC Pol2 ChIP
1	2326	sheffield_dnase	71.34937	81.27610	21	5	70	28	70	34.3	9443	79	2880012	Ewing sarcoma DNase
1	668	encode_tfbs	100.50736	64.15682	33	2	75	9	75	28.7	21989	67	2867466	SK-N-MC FOXP2 ChIP

Figure 2-5 Workflow of the enrichment analysis using LOLA. A. input datasets, including regions of interest, background regions, and database of reference (LOLACore). B. determining overlap and level of enrichment in LOLA using the function runLOLA; results are presented as ranking scores. C. Example of ranked results. Diagram taken with permission from Sheffield and Bock. 2016. LOLA: e enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor[150].

2.7.2.3 EpiExplorer (only for DMRs)

The web-based tool EpiExplorer (Epi-explorer, release date: 2012, http://epiexplorer.mpi-inf.mpg.de) was used to improve the genomic annotation of DMRs for histone marks, chromatin states, CpG islands, genomic regions, TF binding sites, DNA methylation level and others, based on results obtained from LOLA. In addition, EpiExplorer provides a tool for gene annotation and Gene Ontology analysis using data from Ensembl Biomart[151]. As in previous tools, different databases (i.e. ENCODE, NIH Roadmap epigenomics, and the UCSC Genome browser) and tissues are available to look at the overlap between user-supplied regions, and regions of multiple regulatory elements. For this particular analysis, the cell-type selected was lymphoblastoid cell-lines, which is a surrogate cell-type for normal peripheral blood cells established after transformation with the Epstein-Barr virus[152]. In contrast to LOLA, EpiExplorer does not compute a formal enrichment analysis, and the output from the exploration are raw values of percent of overlap across regions. Further detail of the methods and tools available in EpiExplorer can be found elsewhere[151], and a workflow of the implementation of EpiExplorer is shown in Figure 2-6.

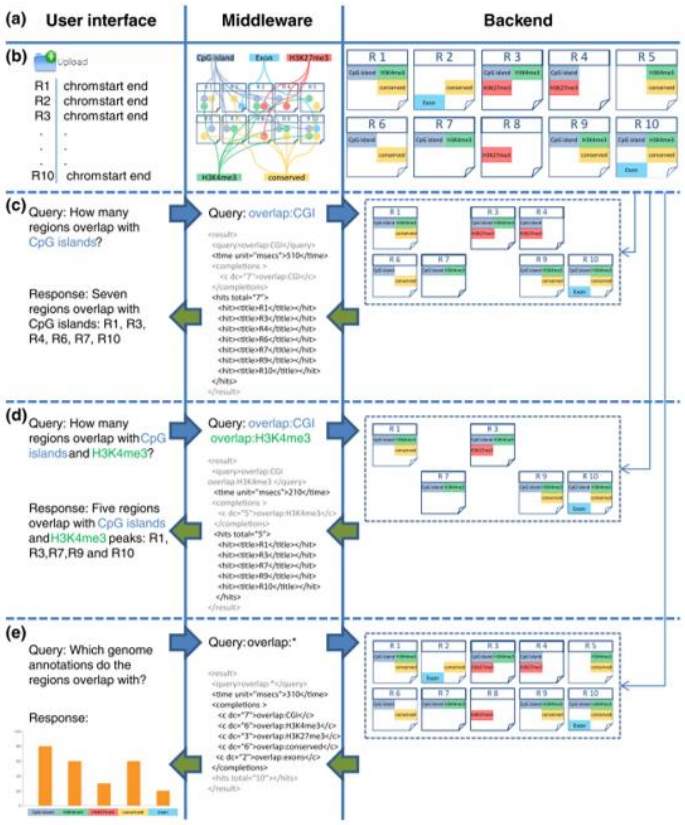


Figure 2-6 Workflow of the enrichment analysis in EpiExplorer. This web-tool uses a three-tier software architecture (user-interface, middleware and backend) to translate user queries into a visual output showing the overlap between regions of interest, and reference regulatory elements retrieved from multiple databases. The search in EpiExplorer is interactive, and new graphs mining the data are generated in response to user-supplied queries. Image taken with permission from Halachev et al. 2012. EpiExplorer: live exploration and global analysis of large epigenomic datasets[151].

2.7.3 Association between methylation and gene expression: eQTM

To determine the gene most likely to be targeted or influenced by methylation at the top CpG sites (at $p<1.0x10^{-7}$) detected in the meta-EWAS, a search for eQTM was performed. eQTM are CpG sites with known influence on gene expression of the nearest gene, and a look up for these sites was done through the Bios QTL browser (https://genenetwork.nl/biosqtlbrowser/), a publicly available repository with data from five Dutch biobanks that included 3,841 whole blood samples[110]. The entire dataset of eQTM was downloaded from the browser, and top CpG sites were compared against the available list of eQTM's. eQTM were regarded associated at FDR < 0.05.

2.7.4 *In silico* comparison of differential gene expression between T2D cases and controls

Difference in gene expression between T2D cases and controls was determined using data extracted from the gene expression omnibus repository GEO (NCBI GEO, last modified: 07-16-2016, https://www.ncbi.nlm.nih.gov/geo/). Genes investigated for their differential expression were those

annotated to index CpG sites detected in the DMR analysis. Data was extracted from GEO using the R packages Biobase (version 2.30.0)[153] and GEOquery (version 2.40.0)[154], with R scripts provided by the GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/). Samples in the GEO datasets were grouped based on their T2D status, or on traits related with T2D e.g. insulin resistance, and unadjusted linear regressions were performed using Log2 transformed values of gene expression as the outcome against T2D. Regression analyses were performed using the R package limma (version 3.26.8), with further correction for multiple testing using the Benjamini-Hochberg method (default method in GEO2R). Results of the association analysis were extracted for the genes of interest, specifying for each gene the beta-coefficient, log2 Fold-change (Log-FC), raw P-value and adjusted P-value. Associations were considered significant at FDR < 0.05.

GEO datasets used were GSE4117 and GSE87005, based on studies conducted by Hayashi *et al.*[155] and Matone *et al*.[156], respectively. In the study by Hayashi *et al*., the effect of laughter on changes in peripheral blood leukocyte gene expression at 18,716 genes, was evaluated between six T2D cases (three affected with diabetic nephropathy) and two controls[155]. Gene expression was measured using the Agilent whole genome oligo microarray[155]. In the second study conducted by Matone *et al.*[156], gene expression in lymphomonocytes was evaluated in 40 healthy participants with different measures of insulin resistance (i.e. HOMA-IR), 20 of them had high HOMA-IR (i.e. regarded as T2D cases), and the other 20 had low HOMA-IR (i.e. regarded as controls)[156]. Gene expression was measured at 321 genes using the Agilent whole human genome microarray[156].

2.7.5 Pathway analysis

Pathway analysis was performed to determine biologically meaningful trends for genes annotated to top CpG sites identified in the association analysis (i.e. single-site and DMR analyses). CpG sites were annotated using the R package lluminaHumanMethylation450kanno.ilmn12.hg19[157], while the enrichment analysis for gene ontology (GO) terms and KEGG pathways was performed using the R package missMethyl[158]. P-values for enrichment were adjusted for multiple testing using the FDR method (significant at FDR<0.05).

2.7.6 Cross-tissue comparison of DNA methylation using publicly available data

To determine the relevance of peripheral blood as a source of methylation markers for T2D and glycaemic traits, an *in-silico* comparison in the levels of methylation was made between blood and other five internal tissues relevant for T2D: liver, skeletal muscle, pancreas, omentum and subcutaneous fat. The dataset interrogated in this analysis was the GEO series GSE4847 (Gene expression omnibus database, https://www.ncbi.nlm.nih.gov/geo/) according to the study

conducted by Slieker *et al*.[159]. Extraction and annotation of meta-data from GEO was done using the R packages Biobase (version 2.30.0)[153] and GEOquery (version 2.36.0)[154], while the Pearson correlation was used to compare levels of methylation between tissues at the top CpG sites identified in the association analysis. Correlation was regarded significant at $p<0.05$.

2.7.7 Comparison between meQTL and GWAS SNPs using publicly available datasets

To determine if the association between methylation, T2D and glycaemic traits, was potentially confounded by common genetics, the overlap between meQTL SNPs and GWAS SNPs for T2D and the glycaemic traits, was investigated. meQTL were considered nominally associated with the trait if the GWAS p-value was <0.05. Overlapping meQTL and GWAS SNPs were compared for similarity in the effect allele and direction of effect.

meQTL for top CpG sites identified in the different association analyses, were looked up in an online catalogue (mQTLdb, analysis date: 14-08-2018, http://www.mqtldb.org/)[109], with data collected from peripheral blood samples of females in ALSPAC at two time-points: antenatal methylation (n=764, mean age 29.2 years) and middle-age methylation (n=742, mean age 47.5 years). In addition, meQTL for top sites were retrieved from the Genetics of DNA methylation consortium (GoDMC, http://www.godmc.org.uk/), the largest consortium available to date for the study of the genetic basis of DNA methylation variation. meQTL from GoDMC were retrieved specifically for top CpG sites detected in the meta-analysis of T2D, our main analysis, after applying for data access. Further description of the GoDMC consortium, can be found in Chapter 3. meQTL were retrieved surpassing the statistical significance threshold of $p<1.0x10^{-7}$ from the mQTLdb database, and $p<1.0x10^{-5}$ from the GoDMC consortium.

Genetic variants associated with T2D (n=711 reported SNPs) were extracted from the GWAS Catalog (release date: 2008, https://www.ebi.ac.uk/gwas), or from one of the largest transethnic GWAS meta-analysis reported in DIAGRAM[30], or they were taken from a list of SNPs identified in strong association with T2D in the DIAGRAM consortium, which were used to generate a polygenic risk score for T2D in ALSPAC samples (see Chapter 3). Genetic variants associated with fasting insulin[160], HOMA-IR and HOMA-B[161], and HbA1c[162], were looked up in the MAGIC consortium, the largest GWAS meta-analysis of glycaemic traits (https://www.magicinvestigators.org/downloads/)[160-162]. Data extracted from MAGIC were European samples without overt diabetes. In addition, genetic variants associated with glycaemic traits were looked up in the GWAS catalog, where 173 SNPs were

reported for fasting glucose, 52 SNPs for fasting insulin, 87 SNPs for HbA1c, 160 SNPs for insulin resistance, 30 SNPs for HOMA-IR, and 22 SNPs for HOMA-B.

2.7.8 Comparison between meQTL and eQTL using publicly available datasets

Shared genetics between methylation and gene expression was investigated by looking at the overlap between meQTL SNPs for top CpG sites identified in the different association analyses, and expression quantitative trait loci (eQTL). meQTL were looked up in the mQTLdb and GoDMC as mentioned above, while eQTL were looked up in the latest GTEx dataset (GTEx_Analysis_v7, https://gtexportal.org)[163] using specific tissues. eQTL were retrieved after multiple testing correction using a statistical significance threshold of Q<0.05. Tissues interrogated for eQTL were peripheral blood, skeletal muscle, liver, omentum, subcutaneous-fat, pancreatic and thyroid tissue. To account for the tissue-specificity of meQTL and eQTL markers, priority was given to observations where the overlap was found between an meQTL and an eQTL from peripheral blood samples. meQTL were considered associated with gene expression if the eQTL Q-value was <0.05.

2.8 Methylation score for glycaemic traits

To predict the proportion of the variance in the glycaemic traits that could be explained by strongest CpG sites (at $p<1.07x10^{-7}$) detected in the meta-EWAS, different methylation scores were calculated using an effect-size weighted linear combination of methylation values (untransformed β-values). The score was validated in control samples in ALSPAC and SABRE. In addition, a methylation score for HbA1c was generated using samples in SABRE for stronger signals identified in this EWAS. For outcomes where the score was calculated in two studies, difference in the mean of the score across studies was calculated using a t-test, or a Kruskal-Wallis test for non-parametric scores.

Calculating the weights

Weights were calculated using absolute values of the effect size reported in the meta-EWAS. For each CpG site in the score, the weight was estimated as the ratio between the effect size for one site, against the average effect size for all the CpG sites included in the score.

Equation 2-2 Estimating weights

$$Wn = \frac{abs(Effectn)}{mean(abs(Effect_1) + abs(Effect_2) + abs(Effectn))}$$

Calculating the score

The score was estimated as the weighted sum of median β-values of methylation for all the CpG sites surpassing epigenome-wide significance in the meta-EWAS.

Equation 2-3 Methylation Score

$$Meth\ Score = CpG_1W_1 + CpG_2W_2 + CpGnWn$$

Where CpG_n is the un-transformed β-value for CpG site *n*, and W_n is the weight allocated to *CpG* site *n* based on the effect size.

Predictive ability of the score

The strength of the score in explaining variation in the glycaemic trait was assessed using different regression models, with the trait as the dependent variable. This analysis was performed to determine if the score alone could explain more variation in the trait than the variation explained by known risk factors. A basic model was constructed including all the risk factors but not the score, and five other models including the score with stepwise adjustment for the risk factors. Performance of models including the score, versus the basic model, was determined using the adjusted-R^2, the root of the mean square error (RMSE), and the statistical significance of the likelihood ratio test (P_{LRT}), all of these were parameters reported in the linear regression. A better fit model was regarded as a model with higher adjusted-R^2, smaller RMSE, and p-value of the likelihood ratio test < 0.05 to indicate significant difference between the basic model, and the model including the score. All analyses were conducted in R version 3.3.3.

Interpretation of results

Results of the association between the score and the glycaemic trait were considered significant at $p<0.05$, and they were interpreted as a unit change in the trait, per unit increase in the score (unstandardized score). Total variance in the trait explained by the score was retrieved from the adjusted-R^2 of the linear regression, and it was described relative to the adjustment model and the sample (i.e. ALSPAC or SABRE) where it was identified.

Chapter 3 Methods in Mendelian randomization

This chapter describes the use of Mendelian randomization to strengthen causal inference when considering the relationship between DNA methylation and T2D. Over recent years, the rapid increase and widespread availability of genotype data to characterize the genetic risk of multiple disease outcomes, has fuelled an increase in the use of Mendelian randomization (MR). This method uses the genotype as an unconfounded an unbiased "causal anchor" to proxy for modifiable factors[164]. Details of the method have been described thoroughly elsewhere[93, 94, 164]. Briefly, an instrumental analysis framework is used to estimate the causal effect of the genotype on an outcome. The method has important assumptions and well documented limitations[94, 161, 164-166]. More recently, the method of MR has been applied to epigenetic data, which as any other molecular biomarker, is susceptible to confounding by common environmental factors[167]. Since genetic variants that correlate with DNA methylation variation have been identified[109, 110](GoDMC consortium), the probability of doing MR to appraise causality in epigenetic pathways has arisen[95, 98, 167, 168].

Methods described in this chapter encompass (1) selection of genetic variants for T2D, (2) generation of a polygenic risk score for T2D, (3) Single sample MR between T2D and DNA methylation using top CpG sites detected in the association analysis, (4) single SNP regressions against T2D, DNA methylation, and confounders, and (5) Bidirectional two sample MR to investigate causality, and direction of causality, in the association between T2D and DNA methylation using top CpG sites detected in the association analysis. The flow-diagram in Figure 3-1 summarizes methods described throughout this chapter.

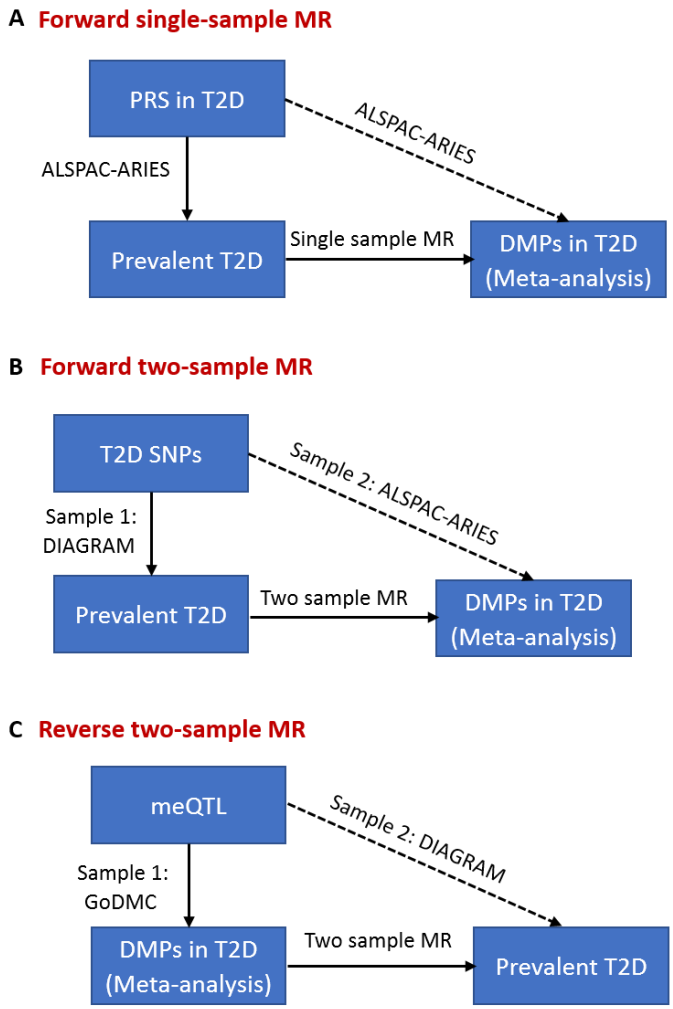


Figure 3-1 Flow-diagram summarizing methods implemented to appraise causality in epigenetics in T2D. The diagram shows the two-steps implemented in the bidirectional MR. The first step evaluates the causal association between T2D as the exposure against DNA methylation as the outcome using a (A) single-sample and a (B) two sample MR. Different from two sample MR, where summary data is required from two independent but comparable samples, single sample MR requires individual level data from a single sample to generate a causal estimate. The second step of the analysis investigates causality in the reverse direction of the association to determine if T2D is a consequence of the variation in methylation. PRS: polygenic risk score for T2D.

3.1 Selecting instruments for T2D

Genetic variants selected to proxy T2D status were extracted from four GWAS meta-analysis of T2D reported in DIAGRAM, the biggest consortium for the study of T2D (Diabetes Genetics Replication and Meta-analysis, http://www.diagram-consortium.org/about.html). The studies selected included Europeans alone [101, 169], or multiple other ethnicities [30, 170]. A proxy for T2D was a common variant that was (1) identified as an index variant in a GWAS meta-analysis, (2) or a variant found within a 99% credible set from an index variant, and with equal or higher posterior probability than the index variant, (3) or a variant identified in the exome mapping close to a well-established locus for T2D, with MAF above 0.05. Two levels of significance were considered in the

selection of genetic proxies: a genome-wide significance threshold with $p< 5.0x10^{-8}$, and a locus-wide significance with $p<1.0x10^{-5}$ (GCTA joint regression model). SNPs extracted from summary data in DIAGRAM were excluded if they had missing data for the risk allele, the reported effect estimate (OR), the p-value of significance, or if the minor allele frequency (MAF) was lower than 0.05.

3.1.1 Data extraction

Summary statistics of SNPs associated with T2D were extracted from the latest GWAS meta-analyses available in DIAGRAM at the time this study was conducted. DIAGRAM studies included were Morris *et al*. 2012, Mahajan *et al*. 2014, Gaulton *et al*. 2015 and Fuchsberger *et al*. 2016. Altogether, a subset of 148 unique SNPs previously identified in strong association with T2D, were retained for further genotyping in a subsample of participants in ALSPAC. Further detail of the GWAS meta-analyses extracted from DIAGRAM can be found in Table 3-1, while association summary statistics for 148 SNPs selected across DIAGRAM studies, are reported in the appendix Table S8-1. Datasets were downloaded from the DIAGRAM data download resource (http://www.diagram-consortium.org/downloads.html).

Table 3-1 Description of four studies reported in the DIAGRAM consortium that were used to extract genetic proxies for T2D.

SNPs	Data source	Population	Discovery sample	Replication sample	Notes
60	Morris *et al.* 2012[101]	Europeans Pakistani (PROMIS)	121,171 T2D cases and 56,862 controls	22,669 T2D cases and 58,119 controls 1,178 T2D cases and 2,472 controls	Genotyping method was the Metabochip array. Meta-analysis of GWAS adjusted for genomic inflation.
34	Mahajan *et al.* 2014[30]	Europeans South Asians East Asians Mexicans and Mexican Americans	12,171 T2D cases, 56,862 controls 6,952 T2D cases, 11,865 controls 5,561 T2D cases, 14,458 controls 1,804 T2D cases, 779 controls	21,491 T2D cases and 55,647 controls	Ancestry-specific GWAS corrected for study-specific covariates and genomic inflation. Trans-ethnic meta-analysis corrected for genomic inflation.
40	Gaulton *et al*. 2015[169]	Europeans	27,206 T2D cases and 57,574 controls	Not reported	Fine mapping of 39 established genetic loci in T2D
14	Fuchsberger *et al.* 2016[170]	Europeans East Asians South Asians African American Hispanics	11,645 T2D cases and 32,769 controls	Not reported	

3.2 Accessing genotype data

Genotyping was completed prior to this project and data were accessed for the purposes of this analysis. Middle-age adults' genetic data for the 148 SNPs of interest in T2D (see section 3.1) was extracted from the ALSPAC GWAS database[112] using the latest genetic imputation datasets available (mothers: release 2015-10-30, fathers: release 2016-11-22). Because imputation was done independently for mothers and fathers in ALSPAC, both datasets were merged retaining the first ten PCs to adjust for genetic structure in downstream analyses. Detail of the method applied for genotyping and imputation of genetic data in mothers in ALSPAC has been described elsewhere[112, 168]. Briefly, genetic data was generated using the Illumina Infinium Human660W-Quad BeadChip array v1.0, and the Illumina GenomeStudio software for genotyping calling (genome build 37). QC measures included removal of SNPs with minor allele frequency (MAF) < 0.01, Hardy-Weinberg equilibrium p-value < 10^{-6}, and missing genotyping rate above 5%. In addition, samples with indeterminate X chromosome heterozygosity, genotyping missingness higher than 5%, and evidence of population stratification were excluded. Data was imputed to the 1,000 Genomes (phase 1, version 3, http://www.internationalgenome.org/about) using Impute2 version 2.2.2, retaining samples with MAF > 0.01 and calling rate > 80%.

Genetic data for the fathers was genotyped using the Illumina HumanCoreExome BeadChip array, and the Illumina GenomeStudio software for variant calling. QC measures included removing SNPs with Hardy-Weinberg equilibrium p-value< 10^{-7}, SNPs failing GenomeStuio QC, and SNPs that were duplicated. Samples with gender mismatch, high or low heterozygosity, genotyping missingness > 5%, contamination, and non-European samples were excluded. Before imputation, genotype data for 3,074 samples (some of them related) was furtherly controlled for variants not included in the 1,000 Genomes, monomorphic SNPs and duplicated sites. Imputation was performed as above. Data from the imputation was retained if MAF> 0.01 and calling rate > 80%.

3.3 Study Population

A total of 1,501 participants in ALSPAC had availability of genotype and DNA methylation data measured at the middle-age time-point (ARIES mothers n=947 mean age=47y, and ARIES fathers n=554 mean age=53y). From this subsample, complete genotype data was obtained for 1,252/1,501 participants: mothers were 867/1,252, and fathers were 385/1,252. In the mothers, extraction of genotype data was done for 139/148 T2D SNPs, and for 135/148 T2D SNPs in the fathers. Datasets were merged in Plink (v2.0) using the dataset of mothers as the reference (i.e. more SNPs were

imputed in this dataset). The final dataset included 1,252 samples with genotype data for 142/148 T2D SNPs. Some SNPs were genotyped in both mothers and fathers (n=132 SNPs), while other SNPs were exclusively genotyped in mothers (n=7 SNPs) or in fathers (n=3 SNPs).

3.4 Further genetic pruning

Additional QC measures after combining genotypic datasets included removal of SNPs with missing genotyping rate > 0.1 (n=16 removed), Hardy-Weinberg equilibrium p-value < 3.97×10^{-4} (i.e. 0.05/126 tests), and MAF < 0.01. Samples were excluded if missing genotyping rate > 0.1. In total, 126 SNPs and 1,252 samples remained in the dataset after QC. Furthermore, allele calling obtained from the imputation was manually inspected to verify that this corresponded to the alleles reported in DIAGRAM for the same SNP. In cases where there was mismatch between datasets, the SNP was looked-up in MR-Base (http://www.mrbase.org/) to establish the correct allele coding based on additional studies including samples from European ancestry. Genetic annotation of T2D SNPs was done using SNP nexus (date accessed: 16-05-2017, http://www.snp-nexus.org/) with data from the UCSC Genome Browser and the human genome build 37 (GRCh37). Genetic annotation was available for 88/126 SNPs, and most of them were in intronic (87%) versus coding regions (29.37%) (Figure 3-2).

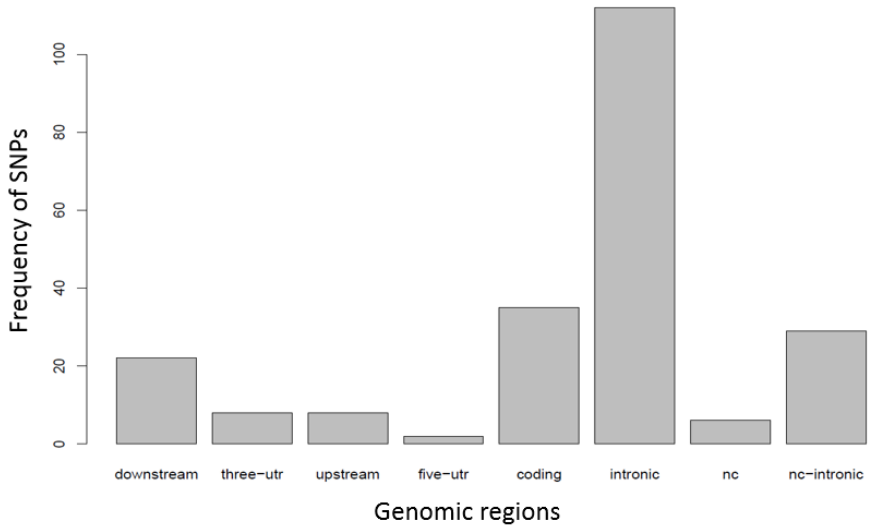


Figure 3-2 Genetic annotation of 126 T2D SNPs extracted from DIAGRAM, and genotyped in 1,252 middle-age participants in ALSPAC. In the x-axis is the genetic context of the SNP, and in the y-axis is the frequency of SNPs in overlap with each genomic region. UTR = untranslated region, nc = non-coding. Plot provided by SNP nexus using annotation data from the UCSC Genome Browser (http://www.snp-nexus.org/).

LD pruning

Further pruning was applied to include only independent T2D SNPs using an LD r^2 of < 0.2 as suggested by Weale[171]. After pruning, 75 SNPs remained in the dataset for further analyses, and this list of independent SNPs did not overlap with another subset of independent SNPs selected to construct the polygenic risk score (PRS) for T2D (see below).

3.5 Genetic proxies versus T2D and potential confounders

Two important assumption of MR studies are that genetic proxies strongly associated with the exposure of interest are used, and proxies selected are unrelated with potential confounders of the main association to avoid horizontal pleiotropy, a form of genetic confounding. These assumptions were tested by performing multivariable linear and logistic regressions to ascertain the strength to which genetic proxies were associated with T2D in ALSPAC, and to verify if they were also associated with potential confounders of the exposure-outcome association. Associations were conducted using an additive genetic model (i.e. AA versus Aa versus aa) with the genotype as the exposure, and T2D and potential confounders (i.e. age, sex, BMI, total cholesterol, HDL, LDL, smoking, physical activity and socioeconomic status) as the outcome. Regressions were adjusted for sex and the first ten genetic PCs to account for population stratification, while correction for multiple testing was done using the Bonferroni method (α=0.05/number of SNPs).

All the regressions were conducted in Plink (v2.0) using the functions *logistic* and *linear*, reporting the effect estimate and the 95% confidence interval. To make sure that the regression coefficient was measuring the effect of the risk allele in T2D as reported in DIAGRAM, a list of reference alleles (A1) was generated to specify which allele needed to be accounted for in the regression. Before the analysis, categorical ordinal variables were transformed to dummy binary variables, while some of the continuous variables were log-transformed (i.e. BMI and HDL). Missing data was observed for T2D status in 412 samples, for PCs in 78 mothers, and for age in 40 fathers. Continuous outcomes were summarized across categories of the genotype using the mean and standard deviation generated by the *qt-means* and *assoc* functions, while categorical variables were summarized using proportions generated by the function *model-trend*.

Power calculation

Power to identify an association between the genotype and T2D was estimated using an online tool for power calculation (http://cnsgenomics.com/shiny/mRnd/)[172]. Input data was a sample-size of 1,252 corresponding to the subsample of middle-age adults in ALSPAC, a proportion of cases of 0.04

(ratio=36 T2D cases/840 T2D cases and controls), a significance threshold of 0.05, the effect estimate obtained for the SNP with the smallest p-value of association with T2D, and a previously reported variation in T2D between 10% and 15% based on the most recent GWAS of T2D[103]. For categorical confounders, power was calculated using the effect estimate obtained for the SNP with the strongest association with the confounder, the proportion of cases was regarded as the number of samples in the testing category, and a hypothetical variation in the outcome explained by the genotype of 0.01.

For continuous confounders, power was calculated using as the observed effect estimate (β_{OLS}) the unadjusted effect for the SNP with the smallest p-value of association with the confounder; the causal effect (β_{yx}) was the adjusted effect estimate for the same SNP; variance in the exposure (σ^2_x) was considered as the frequency of the effect allele; variance in the outcome (σ^2_y) was calculated using the square of the average SD of the confounder across categories of the genotype; the proportion of variance in the confounder explained by the SNP was taken from the R^2 reported in the unadjusted regression. Sample size was regarded as the total number of participants with complete data for the outcome of interest.

3.6 Single sample MR using 2SLS-IV analysis

To strengthen the association between single SNPs and T2D in ALSPAC, SNPs were combined into a polygenic score, which was used to assess the genotype-exposure and genotype-outcome associations in a single sample MR. For this analysis, T2D was considered the exposure and DNA methylation the outcome according to top DMPs identified with the smallest p-value in the association analysis (see Chapter 6). The causal estimate in single sample MR was obtained using a two-stage least square instrumental variable (2SLS-IV) analysis.

3.6.1 Generating a Polygenic Risk Score for T2D

Two polygenic risk scores were generated using the function *score* in Plink to represent the average number of T2D-increasing risk alleles carried by an individual. The dose of the risk allele in the score was weighted by the effect size using effect estimates previously reported in GWAS meta-analyses of T2D in the DIAGRAM consortium. The two polygenic scores differed in the number of SNPs included according to the p-value threshold used. The first score was generated using a p-threshold < $5.0x10^{-8}$ (PRS1 n=56 SNPs), and the second score was generated using a p-threshold < $9.0x10^{-6}$ (PRS2 n=75 SNPs), corresponding to the largest p-value observed among the list of clumped T2D SNPs. The scores were standardized using Z-values, and regressions were conducted to test the association

between the score against T2D, confounders, and methylation sites previously identified in an observational analysis. Results were interpreted as the effect of one SD increase in the score on the risk of T2D, on a unit change in the outcome, or on a 10% increase in untransformed β-values of methylation.

In principle, variants selected to generate a polygenic score should be identified in an independent sample (i.e. training sample) from the sample where the score is intended to be used (i.e. replication sample), this to avoid overfitting the data[173]. Genetic instruments selected to generate the score were the 126 SNPs extracted from different GWAS meta-analyses in T2D that surpassed genotyping QC in ALSPAC. To avoid overestimating the effect of the score in the exposure, SNPs were pruned for high LD using the function *clumping* in Plink (v2.0). This function groups SNPs in high LD into similar LD blocks ($r^2 > 0.2$, range: 250kb), and selects from each LD block the SNP with the smallest p-value of association with the trait, known as the index SNP. Index SNPs obtained after applying clumping were 75 independent SNPs that were used to generate the two polygenic scores for T2D. An example of an LD block and an index SNP is shown in Table 3-2.

Table 3-2 LD block in chromosome nine (chr9:22,133,284-22,134,172) showing the index SNP as the association with the smallest p-value in the region, and other SNPs in high correlation ($r^2 > 0.2$) with the index SNP, with a larger p-value, and located within 250kb from the position of the index SNP.

	SNP	kb	r^2	Alleles	P
Index SNP	**rs10811660**	**0**	**1.000**	**A**	**1.10E-61**
Correlated SNPs	rs10965250	-0.784	1.000	AA/GG	1.80E-25
	rs10811661	0.026	1.000	AC/GT	3.70E-27
	rs10757283	0.104	0.305	AT/GC	3.60E-26

3.6.2 Polygenic score versus T2D and confounders

Logistic and linear regressions were used to investigate the association between the polygenic score, T2D and potential confounders, in an unadjusted model, and in a model adjusted for sex and the first ten genetic PCs. Proportion of variation explained by the polygenic score in T2D and in categorical binary confounders, was calculated using the Nagelkerke's R^2, while for continuous confounders variation was obtained from the R^2 reported in the linear regression. A t-test was used to determine if there was significant difference (at $p<0.05$) in the amount of variation explained between the two regression models. The polygenic score taken forward to investigate the IV-outcome association and to conduct single sample MR, was the score that explained the highest variation in T2D, which was independent of confounders to prevent the inclusion of pleiotropic effects in MR estimates.

3.6.3 Observational IV-Outcome association: PRS against DNA methylation

Before predicting the causal effect of T2D on DNA methylation using a 2SLS-IV analysis, it is necessary to measure the association between the genetic proxy (i.e. polygenic score) and methylation at the top-ranked DMPs identified in the observational analysis (i.e. meta-EWAS of T2D). Observed PRS-DNAm association was estimated by conducting an EWAS with the polygenic risk score (i.e. EWAS of PRS) in the subsample of middle-age adults in ALSPAC, extracting regression estimates for top DMPs identified in the observational analysis. The EWAS of PRS was adjusted for age, sex, the first ten genetic PCs, SVs and predicted cell-counts using the Houseman[119] method (Table 3-3). SVs included in this analysis were unrelated to covariates, and independent of variation in the polygenic score, the exposure of interest in this analysis.

Table 3-3 Description of adjustment models used in the EWAS of the polygenic risk score for T2D.

Model	Description
Basic	Age, sex, 10 genetic PCs, SVs[†]
Adjusted	Age, sex, 10 genetic PCs, SVs, predicted cell-counts[‡]

[†]SVs were generated using all the covariates included across the two adjustment models after verifying their independence from the score. [‡]Houseman cells[119]

Methylation data used in the EWAS of PRS was processed for probe and sample quality as described in Chapter 2. In total, 1,535 samples and 383,104 probes passed quality control. This sample was then subset to 1,078 participants with available measures for the polygenic score to conduct the EWAS of PRS. Since the polygenic score was now replacing case-control T2D as the exposure in the regression model, the subsample of 1,078 participants was not restricted to samples with complete records of T2D status. All regressions were conducted using untransformed methylation beta-values, and to account for potential outliers or influential observations remaining in the methylation data, models were run using robust linear regression (RLR) and the R package MASS[134]. Associations were corrected for multiple testing using the Bonferroni method and a p-threshold of significance lower than 1.31×10^{-7} (α=0.05/383,104 tests).

In addition, findings of the EWAS of PRS were used to (1) identify global variation in methylation attributed to the polygenic score, (2) to conduct a DMR analysis, and (3) to evaluate the correlation between effect estimates of the IV-outcome (i.e. PRS-DNAm) and exposure-outcome (i.e. T2D-DNAm) associations at the top-ranked DMPs. A flow diagram in Figure 3-3 describes analyses derived from the EWAS of PRS.

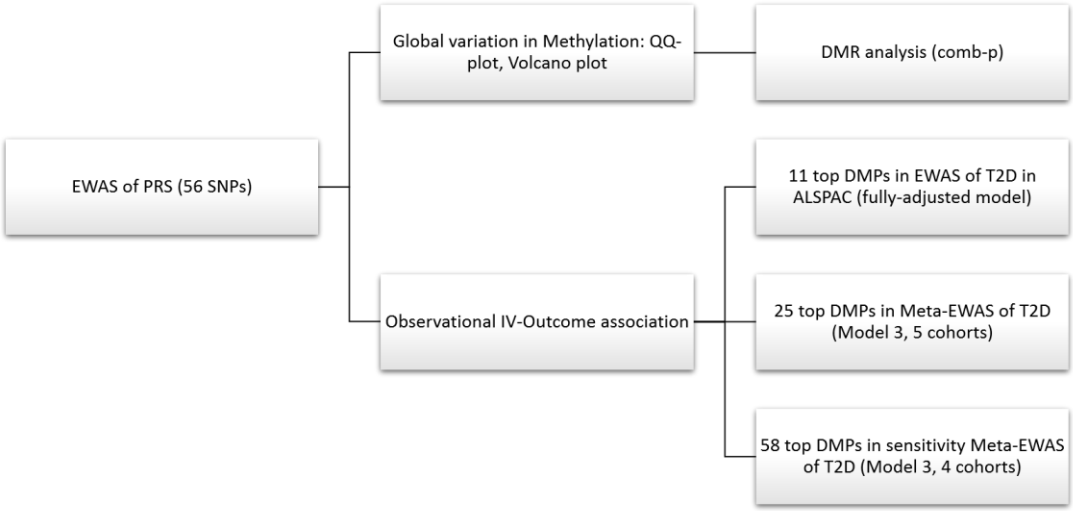


Figure 3-3 Analyses derived from the PRS EWAS.

3.6.3.1 Sensitivity analysis of the observational exposure-outcome association using the PRS

As an additional analysis, a robust linear regression was conducted using case control T2D as the exposure, against methylation at the top ranked DMPs identified in the observational analysis as the outcome. Covariates included in this model were age, sex, smoking, BMI, six predicted cell-counts, and surrogate variables. A second regression model was specified including the polygenic risk score as an additional covariate to determine if the model with the score could capture further T2D-associated difference in methylation, compared to the model using only case control T2D. Results of the regression were corrected for multiple testing using the Bonferroni method (0.05/number of top-ranked DMPs). Difference between models was determined according to change in the effect estimate and p-value of the associations and measuring the correlation between these estimates across models.

3.6.3.2 Investigating PRS-associated DMRs

A DMR analysis was conducted in *comb-p* [148] using association summary statistics obtained in the EWAS of PRS. Parameters of the analysis in *comb-p* have been previously described in Chapter 2. Regions differentially methylated were selected after correction for multiple testing using Sidak p-value < 0.05, where the combined p-value of the region is corrected for n_a/n_r tests: n_a is the number of DMPs initially included in the EWAS, and n_r is the number of DMPs that remained in the region.

3.6.4 Single sample MR using 2SLS-IV analysis

A 2SLS-IV analysis was applied to predict the association between T2D and methylation at the top-ranked DMPs identified in the observational analysis. To perform a 2SLS-IV regression in the context of a single sample MR, individual level data for the instrument, exposure and outcome of interest, needs to be available in participants of the same sample. Since genotype and methylation data was only available in ALSPAC, causal estimates were reported based on findings in this sample.

The first step in the 2SLS-IV analysis was to generate a predicted measure of the exposure using fitted values of the log-odds regression between T2D and the polygenic score. The second step of the analysis was to use fitted values of T2D to recalculate the exposure-outcome association at the top-ranked DMPs. Predicted exposure-outcome estimates were first generated using a standard linear regression, but the disadvantage of this method was the lack of corrected standard errors for the predicted estimates. To surpass this limitation, predicted estimates were calculated using the *ivreg* function available in the R package *AER* [174]. This function provides corrected standard errors for the second step of the IV-regression, and it also provides diagnostic tests to inform of the strength of the instrument (weak-instrument test) and the difference between observed and predicted estimates (endogeneity test or Wu-Hausman test). A p-value < 0.05 for the weak-instrument test and the endogeneity test, indicates presence of a weak instrument in the IV-exposure association, and significant difference between observed and predicted estimates in the exposure-outcome association, respectively.

When interpreting the endogeneity test, it is important not to rely entirely on this test to determine causality of the observational association if the test is non-significant (p-value > 0.05), or in the opposite case, to establish a non-causal association if the test is statistically significant (p-value < 0.05). In practice, it is recommended to make judgements about causality based on results of the IV-regression itself, and the associated confidence intervals of the predicted estimates, and to interpret the endogeneity test in terms of evidence or not of confounding when comparing the observational and the predicted associations.

No covariates were included either in the first stage (IV-exposure), nor in the second stage (fitted exposure-outcome) of the 2SLS-IV analysis, this assuming the presence of unconfounded estimates for the IV-exposure association derived from the use of a genetic proxy for T2D. The total sample included in the 2SLS-IV analysis were 862 middle-age adults in ALSPAC with complete records of T2D,

complete measures of the polygenic risk score, and available DNA methylation data. The formula applied in the IV-regression was the following:

$$y \sim x1 \mid z1 + z2 + z3$$

where *y* is the outcome to be predicted, x_1 is the exposure and/or covariates of the model, also known as endogenous variables, and z_n are the instruments used to predict the outcome. For this analysis, only one instrument was implemented corresponding to standardized values of the polygenic risk score for T2D (see section 3.6.1). When interpreting results, parameters used to measure the strength of the instrument were the Wald test and the significance of the weak-instrument test. The Wald-test has similar interpretation to the F-statistic as a goodness-of-fit parameter. In this sense, a Wald-test higher than 10 indicates a good instrument for the MR analysis as described before by Lawlor and others[94]. Interpretation of the weak-instrument test was mentioned above. Total variation in methylation explained by fitted values of the exposure was the multiple R^2 reported by the 2SLS IV-regression. Results of this MR analysis were interpreted as the effect of a unit increase in the log-odds of the predicted exposure (i.e. fitted values of T2D), on a 10% change in DNA methylation. When comparing between observed and predicted effect estimates, difference in the strength of effect across analyses was measured using absolute values to account for potential conflicts in the effect direction.

3.6.5 Power of the 2SLS-IV analysis

The online power calculator for Mendelian Randomization (http://cnsgenomics.com/shiny/mRnd/) was utilized to estimate power in results of the 2SLS-IV analysis. Parameters required for this calculation are the sample-size, expected level of significance (α=0.05), observed effect estimate (β_{OLS}), predicted effect estimate (β_{yz}), proportion of variance in the exposure explained by the score (R^2_{xz}), variance in the exposure (σ^2_x) and variance in the outcome (σ^2_y).

Power was calculated across all the associations investigated in the MR analysis using average absolute effect estimates. The observed effect was retrieved from results of the observational analysis (meta-EWAS of T2D), predicted effect estimates were obtained from single sample MR, proportion of variance in T2D explained by the score was extracted from the IV-exposure regression, variance in the exposure was regarded as the proportion of cases versus controls in the ALSPAC subsample, and variance in the outcome was calculated from untransformed β-values of

methylation at the DMPs of interest. As an alternative analysis, the sample size required to detect the observed and predicted effect estimates with 80% power and 0.05 significance was calculated.

3.7 Two sample MR in Type 2 diabetes

Detection of an unbiased causal effect between an exposure and an outcome depends on the strength of the genetic effect on the exposure, and on the availability of large samples, normally in the order of tens of thousands, to obtain well-powered results[89]. This later condition is still a limitation in epigenetic studies. To overcome sample-size limitations in MR, a two sample MR allows inference about causality to be made using association summary statistics from two independent but comparable samples, one selected to detect with adequate power the genotype-exposure association, and a second sample selected to detect with similar power the genotype-outcome association. Thus, to strengthen evidence of causality from single sample MR, and to lower the chances of incurring in type II error (false negatives), a two sample MR (2SMR) was implemented to assess the causal effect of T2D on DNAm at the top ranked DMPs identified in the observational analysis. The following sections describe data sources and analyses conducted to obtain summary statistics for the SNP-exposure and SNP-outcome associations, followed by a description of the methods used for the 2SMR in MR-Base (see section 3.7.3). Table 3-4 gives an overview of the cohorts and consortia that provided data to conduct the different analyses.

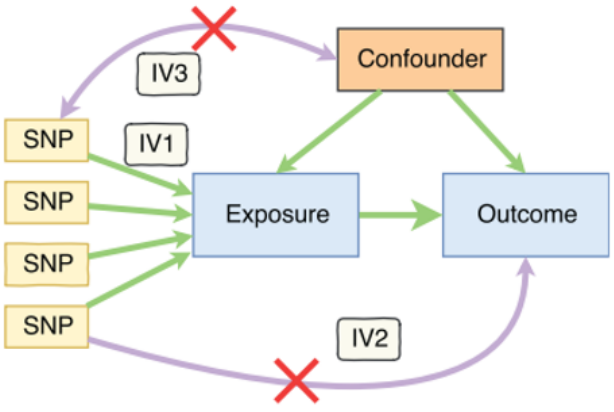


Figure 3-4 Assumptions of MR studies and illustration of the MR framework. The first assumption (IV1) implies that IVs used as proxies should be strongly associated with the exposure, and this is fulfilled by using summary data from well-powered meta-analyses of GWAS, the second assumption (IV2) corresponds to the non-direct association of the IVs with the outcome unless this is through the exposure of interest, and the third assumption (IV3) is the independence of the IVs from potential confounders of the exposure-outcome association. The second and third assumptions are more difficult to be proved and require of the implementation of sensitivity analyses to enhance reliability of MR results. Diagram taken with permission from Hemani et al. 2018 [165].

Table 3-4 Outline of analyses used to investigate the causal effect of T2D on variation in DNA methylation.

Analysis	Method	Cohort/Consortium Included	General requirements	T2D → Methylation analysis
Observational analysis	EWAS of T2D	ALSPAC	All exposure, outcome, and confounding factors	Exposure: Case-control prevalent T2D Outcome: untransformed methylation beta-values
	Meta-EWAS of T2D	ALSPAC + KORA + LBC1936 + Rotterdam-III-1 + Rotterdam-Bios		Confounders: age, sex, predicted cell-counts, SVs and smoking
	Sensitivity meta-EWAS of T2D	ALSPAC + LBC1936 + Rotterdam-III-1 + Rotterdam-Bios		
MR analysis	Two sample MR (forward)	DIAGRAM, ALSPAC, GoDMC	Genotype versus exposure and/or genotype versus outcome association. Confounding factors relevant to each analysis.	Genotype (Z): 62 T2D SNPs Exposure (X): T2D case-control Outcome (Y): Inverse-normal transformed DNA methylation X~Z: DIAGRAM (Covariates: age, sex, PCs, genomic inflation) Y~Z: ALSPAC, GoDMC (Covariates: age, sex, smoking, cell-counts, batch-effects and non-genetic methylation PCs)

3.7.1 Forward 2SMR: T2D as a cause of variation in DNA methylation

3.7.1.1 Genotype-Exposure association: DIAGRAM

Summary statistics for the genotype-exposure association were extracted from four different studies in the DIAGRAM consortium (see section 3.1) in relation to 75 independent SNPs selected to generate the polygenic risk score for T2D in ALSPAC (*clumped* SNPs). From this list of SNPs, 65 remained to be used in the 2SMR analysis based on their availability of data for the genotype-outcome association (see section 3.7.1.3). Effect allele frequencies for these SNPs were extracted from DIAGRAM when available, and when not, they were retrieved from the HapMap project regarding Caucasian samples, making sure that the frequency of the allele reported was that of the effect allele for T2D.

3.7.1.2 Genotype-Confounders association: ALSPAC

Inspection of the association between T2D SNPs and potential confounders was assessed during QC steps for the single sample MR using the subsample of participants in ALSPAC. Therefore, no further inspection was required for the 2SMR.

3.7.1.3 Genotype-Outcome association: ALSPAC and GoDMC

Summary statistics for the association between 75 T2D SNPs and methylation at the top ranked DMPs detected in the observational analysis, were initially requested from the Genetics of DNA Methylation (GoDMC) consortium. However, none of these SNP-CpG associations were included among the candidate list of 120,212,413 SNP-CpG pairs with p-value < 1.0×10^{-5} selected by GoDMC to be meta-analysed across 36 cohorts and reported of summary statistics. Therefore, there were two alternative ways to obtain summary statistics for the genotype-outcome association: (1) to run linear regressions of T2D SNPs versus DMPs of interest in the subsample of ALSPAC, adjusting these regressions for relevant covariates, or (2) to identify proxies (LD > 0.6) for T2D SNPs among SNPs identified in strong association with DMPs of interest in GoDMC data.

Genotype-outcome associations were interrogated first in ALSPAC, keeping in mind the power limitations that arise from this MR by including a less powered sample to assess the genotype-outcome association (ALSPAC, sample-size=1,243), compared to the sample used to determine the genotype-exposure association (DIAGRAM, mean sample-size=152,743). To mitigate this power imbalance and to avoid incurring in type II error, data from GoDMC (mean sample-size=23,430) was used to investigate proxies for T2D SNPs, as explained above.

3.7.1.3.1 SNP-CpG analysis: EWAS of T2D SNPs in ALSPAC

SNP-CpG regressions were conducted using the genotype as the exposure (75 T2D-SNPs) against DNA methylation as the outcome, and the effect of the genotype on methylation was considered per increase in the minor allele count (additive genetic model). Because this analysis was initially set up to measure the effect of the genotype against the whole epigenome, this was regarded as an EWAS of T2D SNPs. Sample included for this analysis were 1,252 (867 females and 385 males) middle-age adults in ALSPAC (mean age 49.1y, range 31y to 75y). Participants in this sample were not excluded based on their T2D status, as it is recommended to keep the largest sample size possible to increase the chances of detecting strong SNP-CpG associations.

The EWAS of T2D-SNPs was conducted following the protocol designed by the GoDMC consortium for meQTL detection. Further documentation of this protocol can be found online (https://github.com/MRCIEU/godmc/wiki). Briefly, all the genetic and epigenetic data was required in the first stage of the process to apply QC measures common to these datasets. Following this, the genotype data was subset for 75 T2D SNPs of interest, and residuals of adjusted methylation values were regressed against the genotype, reporting summary statistics for the top-ranked DMPs detected in the observational analysis and describing strongest SNP-CpG pairs identified in association with middle-age methylation. Further detail of the method applied to conduct the EWAS of T2D SNPs is provided below.

3.7.1.3.1.1 Inspection of genetic and epigenetic data

Genotype data was available for 6,102,837 SNPs in the subsample of middle-age adults in ALSPAC after merging the genetic datasets of mothers and fathers using a consensus method in Plink (Table 3-5). Genotyping rate for 6,102,837 variants was 0.98, and further inspection of the data included plotting imputation quality scores against the MAF to verify that variants remaining in the dataset had MAF > 0.01, and high imputation rate (info> 0.86) (appendix Figure S8-1). Imputation quality scores were available for 6,096,229/6,102,837 variants based on previously reported genotype data for mothers in ALSPAC.

Table 3-5 Example of method used in Plink for merging two genetic datasets.

Dataset 1	Dataset 2	Consensus merging
0/0	0/0	0/0
A/A	0/0	A/A
0/0	A/A	A/A
A/A	A/T	0/0

Genome-wide DNA methylation was available for 482,015 probes, including probes in sex chromosomes and probes reported in the Naeem list [139]. Methylation data was initially pruned for outliers, which were identified as samples with methylation levels ten SD higher than the mean of the probe; outliers were detected after three iterations and replaced by the mean of the probe. Covariates included were age, sex, batch effects (i.e. bisulphite conversion plate), predicted counts for seven Houseman cells, and predicted levels of smoking. Predicted levels smoking were estimated using data from two methylation scores developed by Zeilinger *et al.*[121] and by Elliott *et al.*[122]. Because only complete data was required in the dataset of covariates, missing data for age in a subset of males was replaced using *predictive mean matching* imputation method in the R package MICE [175].

3.7.1.3.1.2 Processing genetic and methylation data

In the genetic dataset, variants were excluded based on MAF < 0.01, calling rate < 0.8, missing genotyping rate > 0.05, and Hardy-Weinberg equilibrium p-value < 10^{-6}. Samples were excluded if cryptic relatedness between pairwise comparisons was higher than 0.125, and if missing genotyping rate was > 0.05. Twenty PCs were generated for the genotype data of unrelated samples using a list of 26,873 independent SNPs identified in the HapMap3 project with LD < 0.1 and MAF higher than 0.2. None of the samples were identified as an outlier based on genetic variation measured by the PCs (Figure 3-5 A), and these PCs were used further down as covariates for the methylation data. In total, 1,243 samples and approximately 5.3 million variants remained in the dataset after QC, total genotyping rate was 0.99, average missing genotyping rate was 0.01, and correlation between observed and expected allele frequencies in the 1,000 Genomes was 0.99. Description of the QC steps applied to the genetic data are summarized in the appendix Table S8-2.

Normalization of the methylation data was applied to avoid false positives in the SNP-CpG analysis, and to minimise the amount of non-genetic residual variation remaining in the data. To achieve this, methylation was first inverse-normal transformed, and then regressed against covariates and twenty genetic PCs previously generated (see above). Covariates and genetic PCs were regarded as fixed effects in the regression model. Residuals of this analysis were then regressed against methylation PCs estimated to account for residual confounding. Methylation PCs were generated using the 20,000 most variable autosomal probes, and they were tested for association with the genotype using a linear regression model in MatrixeQTL. Methylation PCs associated with the genotype at p-value< 10^{-7} were excluded. In total, the first thirteen non-genetic methylation PCs were retained for further analyses. Residuals of the regression between adjusted methylation values and methylation PCs, were used in the SNP-CpG analysis.

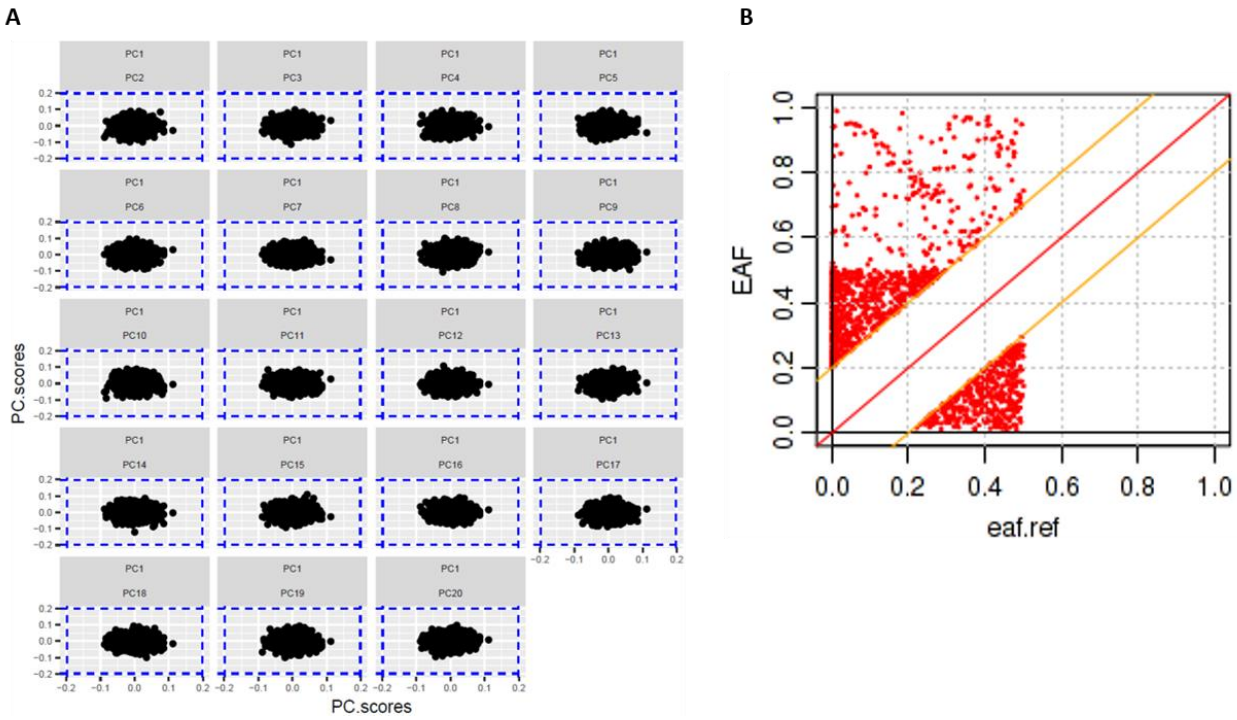


Figure 3-5 Determining genetic outliers for PCs (A) and expected allele frequencies (B) using genetic data from middle-age participants in ALSPAC. A) Twenty PCs were generated for unrelated samples and based on 26,873 independent SNPs retrieved from the HapMap3 project. Outliers were considered as samples outside the blue-dashed line demarking the threshold for genetic variation (SD=7.0). B) Plot of expected (x-axis) versus observed (y-axis) allele frequencies for variants surpassing initial QC in the genetic dataset. Red dots were SNPs with inconsistent allele frequencies with the 1,000 Genomes (Europeans, Phase3), and therefore they were excluded from further analyses (n=1,437 SNPs excluded).

To speed-up the analytical processing of the genetic data and normalization of the methylation data, these analyses were parallelized into 100 chunks, each containing around 52,707 SNPs and 4,821 probes, respectively. Processed genetic and methylation datasets were merged across the different chunks, and merged files were transformed into *traw* files to perform the genotype-methylation regressions using the R package MatrixeQTL[176]. This package enhances the throughput of genotype-methylation analyses by transforming big matrices into sliced objects[176]. The linear regression model used for this analysis was the following:

$$CpG_n = \alpha + \sum_k \beta_k \cdot covariate_k + \gamma \cdot genotype\_additive$$

Where β_k and γ are the effect of covariates and the genotype (i.e. per allele effect) on variation in DNA methylation, respectively.

3.7.1.3.1.3 Positive control analysis

A positive control analysis was run before conducting the SNP-CpG regressions for selected T2D-SNPs. This positive control analysis tested if genetic and methylation data were successfully normalized in previous steps. The meQTL selected as a positive control was the *cis* SNP-CpG pair rs12485195:cg7959070, which has been identified with high significance ($p< 10^{-8}$) across three different cohorts included in the GoDMC consortium, surviving correction for problematic probes based on the Naeem list and some other lists, and for SNPs located in the same position as the probe. In the analysis in ALSPAC, the signal at the *cis* SNP-CpG rs12485195:cg7959070 was successfully replicated with $p=1.02x10^{-133}$ and $\lambda=1.05$, indicating no strong evidence of population stratification.

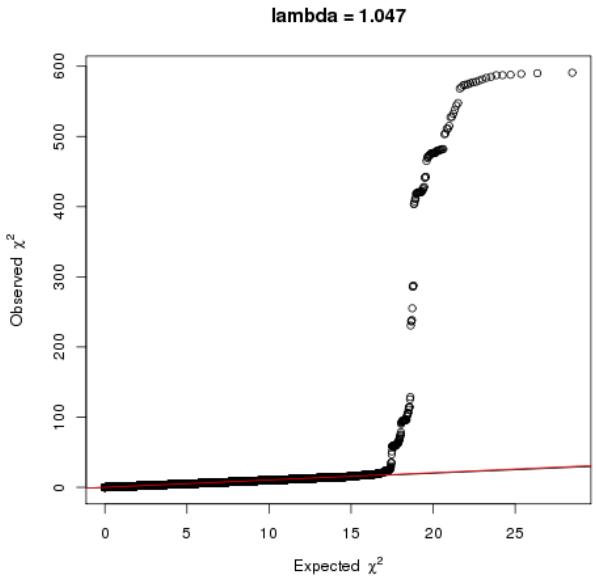


Figure 3-6 QQ-plot of the positive control signal detected at the cis SNP-CpG pair rs12485195:cg7959070 using ALSPAC samples.

3.7.1.3.1.4 Restricting the SNP-CpG analysis to T2D-SNPs

The number of variants used for the SNP-CpG analysis in MatrixeQTL was restricted to 75 independent T2D SNPs included in the polygenic risk score for T2D (see section 3.6.1), which surpassed previous QC steps. The methylation dataset, on the contrary, was assessed genome-widely and not restricted to DMPs of interest. Thus, this analysis was referred to as an EWAS of T2D-SNPs. From the 75 T2D SNPs initially considered, only 65 were present in the pruned genetic dataset for the SNP-CpG analysis.

3.7.1.3.1.5 EWAS of T2D-SNPs

Linear regressions between the genotype for 65 SNPs as the exposure, and residuals of adjusted methylation values for 482,015 probes as the outcome, were conducted in MatrixeQTL using a sample of 1,243 adults in ALSPAC. Associations were equally considered for SNPs located in *cis* (< 1Mb) and *trans* (> 1Mb) from the position of the CpG. No threshold of significance was initially selected to allow obtaining summary statistics for the DMPs of interest. The strongest associations were those surpassing Bonferroni correction (at $p<1.6x10^{-9}$) and pruning for genotype frequencies lower than 0.25 and higher than 0.75, this to avoid including associations with imbalanced genotype frequencies. Variation in methylation explained by the SNP was calculated using total degrees of freedom and the t-statistic, both parameters were reported by MatrixeQTL for the individual associations. Variation was calculated as follows:

$$r^2 = \left(\frac{tstat}{\sqrt{df+(tstat)^2}}\right)^2$$

Due to the large number of associations tested that involved the top-ranked DMPs detected in the observational analysis (n=1,625 associations), summary statistics were only presented for the top-ten strongest SNP-CpG associations with the smallest p-value. For SNP-CpG associations surpassing Bonferroni significance, which did not include DMPs of interest, results are presented in the appendices. Summary statistics for the top ranked DMPs were used in the two sample MR analysis, with methods described next in this chapter.

3.7.2 Conducting 2SMR in MR-Base

Analyses were conducted in the local version of MR-Base using the R package TwoSampleMR (https://github.com/MRCIEU/TwoSampleMR)[177]. MR-Base is a web software that allows the automation of MR methods. The web platform or the local version of MR-Base, allow for multiple statistical methods in MR to be applied using complete summary data from multiple GWAS including a wide range of exposures and outcomes[165]. Alternatively, analyses can be run using existing data-frames if they contain basic parameters, and if they are arranged into the format requested by MR-Base. For analyses conducted in the forward and the reverse 2SMR, existing data-frames were used based on SNP-CpG associations identified in ALSPAC and reported by the GoDMC consortium, and for genotype-T2D associations manually extracted from independent DIAGRAM studies.

Data handling and harmonization

Existing data-frames are uploaded into MR-Base using the functions *format_data* and *read_outcome_data* for the genotype-exposure and genotype-outcome datasets, respectively. Information required from the exposure and outcome datasets are the SNP and outcome IDs, the effect estimate (log-odds scale) of each SNP, SE, effect allele, other allele, effect allele frequency, significance of the IV-exposure and IV-outcome associations, and optionally, units of measurement of the exposure and the outcome and the sample-size.

Once datasets have been uploaded, the next step is to harmonize the data to ensure that the effect allele regarded across datasets is the same. Harmonization of the data included flipping the order of the alleles in the outcome dataset if they were in the same strand but in the wrong order, changing also the direction of effect. When there were strand conflicts, alleles in the outcome dataset were changed to mirror those in the exposure dataset, adjusting allele frequencies accordingly. Harmonization was also applied for palindromic SNPs, which are SNPs that have the same allele-coding in the forward and reverse DNA strand, normally in the form of A/T and G/C alleles[165]. In this case, the reference strand is inferred using allele frequencies, and the direction of effect is changed to reflect that of the effect allele.

Harmonization of palindromic SNPs is more difficult when allele frequencies are close to 0.5 and therefore less reliable. Thus, palindromic SNPs with allele frequencies close to 0.5 were dropped from further analyses. When dealing with incompatible alleles (A/G and A/T), in which no inference of the strand is possible as they correspond to errors in the data or differences in the genome build used to extract the two datasets, these SNPs were dropped from the analysis. In total, three T2D SNPs were dropped from further MR analyses based on strand incompatibility (SNP rs10510110), and on palindromic SNPs with allele frequencies close to 0.5. (SNPs rs944801 and rs7161785).

The level of harmonization applied to the data can be very stringent, and 1) SNPs with strand issues are corrected, but all palindromic SNPs are dropped from the analysis, or it can be less stringent and 2) strand issues are resolved and reference strand is inferred in palindromic SNPs if allele frequencies are not close to 0.5, and finally the less stringent harmonization method 3) assumes that in both, the genotype-exposure and genotype-outcome datasets, the strand used is the forward strand. For MR analyses conducted here, harmonization was applied using the second method.

MR analysis

Default parameters of the *mr* function were used to conduct the 2SMR, allowing for overdispersion of the causal estimates when calculating the weighted linear regression, meaning that the meta-analysis of Wald-ratios was generated using a random-effect and not a fixed effect model (i.e. under dispersion of the data and SE of the causal estimate fixed to one). P-threshold for the causal estimate and the heterogeneity test was $\alpha=0.05$, and results were generated using bootstrapping over 1,000 times. To account for the number of outcomes evaluated (i.e. different top-ranked DMPs), results of the 2SMR were regarded significant after Bonferroni correction for multiple-testing ($\alpha=0.05$/number of DMPs), and borderline significant at $p< 0.05$ (unadjusted P).

Common statistical MR methods were applied including the inverse variance weighted regression (IVW) and the MR-Egger regression. In principle, results of the IVW regression are regarded as main evidence of causality in the absence of horizontal pleiotropy (i.e. when the effect of the SNP on the outcome is not mediated by the exposure), but omission to account for pleiotropy can result in biased causal estimates. The MR-Egger regression, by contrast, accounts for horizontal pleiotropy and estimates directionality of the pleiotropic effect[165, 178]. Thus, MR-Egger can be used as a sensitivity analysis for the IVW regression in the presence of horizontal pleiotropy. Tests reported when using these methods are the heterogeneity test, and the pleiotropy test that is exclusive to the MR-Egger regression. Other statistical methods available are median-based and mode-based methods. Further detail of these MR methods is presented below.

3.7.3 Statistical methods in 2SMR

3.7.3.1 IVW regression

This method is equivalent to combining via meta-analysis results of the Wald-ratio for the causal effect of every SNP on the outcome. Results are combined using a fixed or a random effect meta-analysis, and they are weighted using the inverse of the variance of the outcome effects (i.e. precision of the IV-outcome estimate)[165, 179]. The choice of which IVW model to use depends on the assumptions made. If it is assumed that each SNP has a similar effect on the outcome, and the SE of the estimate is fixed to one, then the fixed effect IVW regression should be implemented. In contrast, if each SNP can have some variation in the mean effect on the outcome, then the random effect IVW regression should be applied. For the random effect model, the estimate reported will be unbiased if there is balanced horizontal pleiotropy, indicating that variation in the mean causal estimate is independent of other effects. By default, MR-Base uses the random effect IVW model, and results were reported according to this model.

As the assumption of no pleiotropy holds for the IVW regression, the intercept of the linear regression between observed effects of the SNPs on the exposure and on the outcome, is forced to pass through zero.

3.7.3.2 MR-Egger regression

The MR-Egger regression represents a variation from the IVW method to account for the effect of horizontal pleiotropy in the SNP-outcome association, giving an unbiased estimate by assuming no correlation between the SNP-exposure effect and the pleiotropic effect (i.e. direct effect of the SNP on the outcome independent of the exposure). This latter assumption is known as the *Instrument effect independent of Direct Effect* (InSIDE) assumption[165, 179]. Since this method allows for unbalanced pleiotropy or directional pleiotropy to exist in the SNP-outcome association, the intercept of the regression between the SNP-exposure and SNP-outcome effects is not constrained to pass through zero, and this intercept can now vary according to the magnitude and direction of the pleiotropic effect[165, 178]. One of the limitations of the MR-Egger regression is the low statistical power and its susceptibility to bias from weak instruments[178]. Causal estimates of the MR-Egger regression were used as a sensitivity analysis for the IVW regression to account for the effect of horizontal pleiotropy.

3.7.3.3 Median-based methods

This additional method provides unbiased causal estimates by allowing at least 50% of the instruments to be valid (i.e. fulfilling the three MR assumptions for IVs), thus taking the median of the effect for all valid SNPs as the final causal effect estimate[165]. For this regression, unbalanced horizontal pleiotropy is permitted, without holding the InSIDE assumption[178]. Furthermore, results are weighted to reflect the effect of the stronger SNPs on the outcome, and the weight is based on the inverse variance of the effect of the SNP on the outcome[165]. Causal estimates were also investigated using median-based methods.

3.7.3.4 Mode-based methods

The mode-based method further relaxes the assumptions of the median-based method and allows the assessment of causality within clusters of SNPs which are grouped by similarity in the estimated causal effect[165]. The cluster with more SNPs is selected to report the final causal effect[165]. The mode estimate can be weighted by the inverse variance of the effect of the SNP on the outcome, and the estimate reported is unbiased if the instruments included in the main cluster are valid instruments. Thus, this method does not account for horizontal pleiotropy.

3.7.4 Diagnostic tests

The following diagnostic tests were conducted, implemented using MR-base:

3.7.4.1 Pleiotropy test

A pleiotropy test was run to estimate the average pleiotropic effect across SNPs using the intercept from the Egger regression, providing the level of significance of the observed pleiotropic effect.

3.7.4.2 Heterogeneity test

A heterogeneity test was run to measure the level of dissimilarity in the causal effect estimated across SNPs, and to provide an idea of horizontal pleiotropy[165]. This test is reported in results of the IVW and MR-Egger regressions using the Cochran's Q estimate and the Rucker's Q estimate, respectively[178].

3.7.4.3 Directionality test

The Steiger test was run to estimate true direction of causality in the predicted exposure-outcome association, with results considered significant at p-value<0.05. This test makes inference about causal direction by comparing the amount of variation explained by the instruments in the exposure and in the outcome[180]. Direction of causality is considered to go from the hypothesized exposure to the outcome if the instruments explain a higher proportion of variation in the exposure than in the outcome [180]. Variation in the exposure (i.e. T2D) and the outcome of interest (i.e. DNAm) was calculated using the functions *get_r_from_lor* and *get_r_from_pn* available in the TwoSampleMR package.

3.7.5 Leave-one-out analysis

This sensitivity analysis was used to detect outlier SNPs, and to determine how robust the analysis was to the effect of these SNPs. The rationale behind conducting a leave-one-out analysis is to establish how much of the total causal effect is influenced by the effect of a single SNP. For this analysis, the combined causal estimate is recalculated after sequentially dropping one SNP at a time from the analysis. A leave-one-out analysis was applied only for the outcomes (i.e. top-ranked DMPs) where the strongest evidence of causality was detected.

3.7.6 Graphical representation of results

3.7.6.1 Scatter plot

A scattered representation of the association between SNP-exposure and SNP-outcome was used, showing the standard errors of these associations, and displaying fitted lines for the different

statistical MR methods used. The slope of the fitted regression line represents the overall causal estimate across SNPs. By using a scatter plot, it is possible to identify SNPs with pleiotropic effects or high heterogeneity.

3.7.6.2 Forest plot

This plot displays as an independent result, the causal estimate in the outcome identified by each SNP used to proxy the exposure, providing for each estimate the 95% CI. The combined causal effect in the outcome using all the SNPs together as a single instrument, is also provided based on the MR method implemented[165]. A variation of this plot is the one-to-many forest plot, where the causal estimate can be displayed for multiple exposures or outcomes according to the MR methods used.

3.7.6.3 Funnel plot

This inspection plot was used to detect horizontal pleiotropy. For each SNP, the causal estimate is plotted against its precision (inverse of the SE of the causal estimate), and the combined causal effect across all SNPs is shown as a vertical line according to the MR method used. Presence of pleiotropic effects is identified based on an asymmetric distribution of the SNPs in the funnel plot, with some SNPs showing extreme protective effects or increased risk effects in the outcome.

3.7.6.4 Volcano plot

This plot is generally used to summarize causal estimates for multiple exposures on one outcome, and it was implemented to visualize results of the reverse 2SMR (i.e. different top-ranked DMPs as the exposure against T2D as the outcome). In a volcano plot, the causal estimate obtained from each exposure is plotted against the significance of the estimate (-Log10 p-value), and these estimates are generally identified using the IVW regression.

3.7.7 Strength of the instruments

To avoid weak instrument bias in results of 2SMR, validity of the no measurement error (NOME) assumption was investigated, which relates to the precision in which estimates of the SNP-exposure association are measured[178]. For the IVW regression, weak instruments were detected using the mean of the F-statistic[178]. In addition, for the MR-Egger regression, weak instruments were identified using the I^2 statistic (I_{GX}^2), with values closer to one indicating less dilution of the causal estimate due to weak instrument bias[178]. The F-statistic was obtained using the samples size, degrees of freedom and p-value reported in the genotype-exposure dataset, implementing the function *qf* available in the R package TwoSampleMR. The I_{GX}^2 statistic was calculated using the *mr_egger* function from the R package MendelianRandomization[181].

3.7.8 Summarizing results of the forward 2SMR

Results of the 2SMR were presented for the IVW, MR-Egger, weighted median, simple mode and weighted mode methods to consider the broad spectrum of assumptions that arise when conducting MR analyses. Associations were taken forward for comparison with observational estimates and further display of inspectional plots, if they were detected with Bonferroni significance or borderline significance in at least one of the MR methods implemented.

3.7.9 Addressing SNP outliers in 2SMR

Even though the implementation of multiple instruments in MR can leverage results of the single instrument analysis by increasing the amount of variation explained in the exposure, and therefore the power to detect a causal association[165], this benefit is balanced out by the liability of detecting horizontal pleiotropic effects and heterogeneity due to invalid instruments. SNPs were regarded as outliers if their effect on methylation was opposite to the effect observed for other variants. To discard the possibility that SNP outliers were due to effect allele coding errors, these SNPs were inspected for palindromic SNPs, which are known to suffer from coding errors in the MR analysis, and their allele annotation was compared to similar risk variants reported in the GWAS catalog, and in summary data from the original studies.

When the presence of coding errors and palindromic SNPs among SNP outliers was ruled out, the next step was to do a lookup of these SNPs in the PheWAS SNP search tool (MR-Base PheWAS (http://phewas.mrbase.org/). This lookup allowed the identification of potential sources of horizontal pleiotropy by determining if alternative outcomes independent of T2D were influencing the SNP-methylation association among SNP outliers. Lookup in PheWAS was restricted to the most common SNP outliers identified for top associations detected with the smallest p-value in MR analyses. Traits surpassing Bonferroni correction in the PheWAS lookup were taken forward to examine their association with methylation at the specific DMP. When identifying putative associations with methylation at p-value< 0.05, these estimates were used to calculate the indirect effect of the SNP on the DMP through the confounder, and the confounder was regarded as a potential mediator if the indirect effect (SNP→confounder→methylation) was in the same direction as the direct effect (SNP→methylation).

3.7.10 Strengthening results of the forward 2SMR by using GoDMC data

To overcome power limitations in the 2SMR by using genotype-outcome estimates from ALSPAC as a second and less powered sample compared to DIAGRAM, recalculating the forward 2SMR using summary data from the GoDMC consortium was considered. GoDMC is the biggest consortium for the study of the genetics of DNA methylation variation, comprising to date more than 20,000 samples (sample range 5,619 to 27,750) gathered across 36 cohorts (GoDMC, date accessed: 14-10-2018, http://www.godmc.org.uk/). Samples included were mainly of Caucasian origin, and mean age of participants was 55.56 years. Due to the larger sample size sustained by GoDMC, estimates of the 2SMR using summary data from this consortium were expected to be more robust, than estimates obtained from ALSPAC.

The identification of meQTL in GoDMC was performed in two analytical phases. In the first phase, each one of the 36 participating cohorts identified all possible meQTL using complete genotype and methylation data, with QC pruning of each dataset as specified in the protocol designed by the consortium for meQTL analysis. Genotype data was imputed to the 1,000 Genomes as the reference panel. meQTL taken forward for further analyses were those identified with p-value<10^{-5} in this first stage of the analysis. meQTL were then meta-analysed across cohorts obtaining a candidate list of 813,603,941 meQTL in *cis* and *trans*. From these meQTL, all those identified in *cis* in at least one dataset (102,965,711 meQTL) and those identified in *trans* in at least two datasets (17,246,702 meQTL) were reanalysed within each cohort in a second phase of the analytical process. Estimates obtained were meta-analysed and results reported based on three different meta-analysis models: fixed-effects, additive random-effects, and multiplicative random-effects models.

From data requested to the GoDMC consortium, strong SNP-CpG pairs were reported for 53/84 top DMPs detected in the observational analysis, and for 75 unique SNPs, none of these corresponding to T2D SNPs included in the polygenic score for T2D (see section 3.6.1). Therefore, the next step was to interrogate the level of correlation between GoDMC SNPs and T2D SNPs to determine if proxies for T2D SNPs could be found among SNP-CpG pairs reported in GoDMC. A proxy was defined as a GoDMC SNP in high correlation (LD > 0.6) with a T2D-SNP. To ascertain correlation among SNPs, pairwise LD metrics were calculated for SNPs located in similar chromosomes using *LDmatrix*, a module available within the web-based application for LD calculation, LDlink (version 3, date accessed: 15-10-2018, https://ldlink.nci.nih.gov/)[182]. In LDlink, LD correlation is calculated using European samples from the 1000 Genomes project Phase 3 as the reference genetic data, with

variants annotated to the human Genome assembly hg19 (GRCh37). SNPs included in the pairwise comparisons were located across the genome.

3.7.11 Power in 2SMR

The web tool mRnd (http://cnsgenomics.com/shiny/mRnd/) was used to assess power in results of the 2SMR, and to calculate the sample-size required to detect a causal effect of T2D on methylation with 95% power and $p<0.05$. The sample-size specified in the power calculator was the mean sample reported across DIAGRAM studies (first sample), and the sample in ALSPAC (second sample); the observed effect estimate was the mean of the absolute effect identified in the observational analysis according to top associations with the smallest p-value detected in the MR analysis; the causal estimate was averaged across different MR methods for the top causal associations detected with the smallest p-value; variance in the exposure (T2D) was 0.02 based on a previous calculation of the variance in T2D explained by a polygenic risk score validated in ALSPAC samples, and variance in methylation was estimated using all the DMPs included in the MR analysis.

3.8 Bidirectional two sample MR: T2D as a consequence of variation in DNA methylation

To tease apart the true direction of causality in the association between T2D and methylation, a bidirectional 2SMR was implemented to investigate the second direction of the association, where methylation is regarded as the exposure and T2D as the outcome. For this analysis, strong meQTL instruments for top-ranked DMPs detected in the observational analysis, were requested from the GoDMC consortium to obtain summary estimates for the genotype-exposure association, while estimates for the genotype-outcome association were extracted from the most relevant (i.e. biggest sample size, largest number of SNPs enquired, and mixed population) GWAS meta-analyses in T2D available in MR-Base. Finally, the true direction of the causal effect was determined for each association by comparing the strength of the causal estimate between the forward and the reverse MR. Table 3-6 describes samples used and analyses performed in the reverse 2SMR. Further detail on the selection of instruments, extraction of genotype-outcome data, and reverse 2SMR analysis in MR-Base, is presented in the following sections.

3.8.1 Selecting Instruments for methylation

Valid instruments were determined based on the strength of their association with DNA methylation. As mentioned before, 53/84 DMPs detected in association with T2D in the observational analysis were successfully instrumented by 75 meQTL reported in the GoDMC

consortium. In GoDMC, the threshold for Bonferroni significance was a p-value<10^{-8} and <10^{-14} for *cis* and *trans* associations, respectively, but because in our analysis only a subset of the epigenome was interrogated (i.e. top-ranked DMPs), the threshold for meQTL significance was relaxed to include all associations with p-value<10^{-5}.

Before the MR analysis, meQTL were inspected to be independent of T2D-SNPs using an LD r^2< 0.01, which is a more stringent LD threshold than the one previously used to select independent T2D-SNPs. The use of independent instruments across the two directions of the bidirectional MR guarantees that the causal estimate obtained is free from bias[183], and it also prevents meQTL for being directly correlated with T2D through pathways independent of methylation itself. The correlation between T2D-SNPs and meQTL was calculated using LDlink (https://ldlink.nci.nih.gov/), and selecting as the reference panel genetic data from European samples in the 1000 Genomes project Phase 3. Four meQTL (SNPs rs35885100, rs150804707, rs796327877 and rs34345524) were identified as indel SNPs, and LD metrics could not be obtained for them in LDlink as this platform does not support the use of indel SNPs. Indel SNPs were not excluded from the dataset of available instruments. As a second QC measure, meQTL located in the same position as the DMP that they were instrumenting were excluded from further analyses. For DMPs with more than one instrument, no manual selection of the meQTL with the smallest p-value was done because this step, also known as clumping, was performed in MR-Base (see below).

Variables reported in the meQTL analysis in GoDMC that were used in the MR were: fixed-effect estimates of the meta-analysis (interpreted as a unit change in inverse-normal transformed residuals of methylation per increase in the effect allele), standard error (SE), SNP (Chr: base-pair position), CpG, effect allele (minor allele), other allele (major allele), effect allele frequency, and p-value of the meta-analysis. Because meQTL SNPs reported in GoDMC were identified using chromosome and genomic position, for the MR analysis it was necessary to extract the rsID number of these SNPs. SNP identifiers were retrieved from SNP-nexus (http://www.snp-nexus.org/)[184] using a batch query and including for each SNP the chromosome, base-pair coordinates, allele coding and DNA strand. If not reported in SNP-nexus, the rsID of the SNPs was manually searched in the UCSC Genome browser using the GRCh37/hg19 human genome assembly. In addition, for some of the meQTL surpassing QC in GoDMC, the association p-value reported was too small and close to cero (n=28 meQTL), indicating highly significant associations. To avoid dropping these meQTL from the MR analysis, p-values equal to cero were replaced with the smallest p-value reported in the meQTL dataset (p-value= $3.5x10^{-202}$).

Table 3-6 Outline of analyses used to investigate the causal effect of variation in DNA methylation on T2D.

Analysis	Method	Cohort/Consortium Included	General requirements	T2D → Methylation analysis
Observational analysis	EWAS of T2D	ALSPAC	All exposure, outcome, and confounding factors	Exposure: Case-control prevalent T2D Outcome: untransformed methylation beta-values
	Meta-EWAS of T2D	ALSPAC + KORA + LBC1936 + Rotterdam-III-1 + Rotterdam-Bios		Confounders: age, sex, predicted cell-counts, SVs and smoking
	Sensitivity meta-EWAS of T2D	ALSPAC + LBC1936 + Rotterdam-III-1 + Rotterdam-Bios		
MR analysis	Two sample MR (reverse direction)	GoDMC, DIAGRAM	Genotype versus exposure and/or genotype versus outcome association. Confounding factors relevant to each analysis.	Genotype (Z): meQTL Exposure (X): DMPs detected in different observational analyses. Outcome (Y): T2D X~Z: GoDMC (Covariates: age, sex, smoking, cell-counts, batch-effects and non-genetic methylation PCs). Y~Z: DIAGRAM (Covariates: age, sex, PCs, genomic inflation). Latest most relevant GWAS meta-analyses in T2D.

3.8.2 Genotype-exposure association

The dataset of genotype-methylation associations was arranged to match with the format requested by MR-Base when using existing data-frames. As a general QC step, input data was inspected for duplicated SNPs, SNPs with unconventional allele coding (i.e. different from A, C, G, T), indel SNPs, and correlated SNPs. To address this latest aspect, clumping was applied on MR-Base by: 1) estimating the LD across SNPs based on correlation structures in European samples from the 1000 Genomes project Phase 3, 2) selecting only independent SNPs based on an LD < 0.01, and 3) taking forward for the analysis variants that represent the strongest association (smallest p-value) with the exposure.

3.8.3 Genotype-outcome association

Summary data for the genotype-T2D association was extracted from the repository of complete GWAS data on MR-Base according to SNPs included in the genotype-methylation dataset. The repository of curated datasets on MR-Base was accessed through the R packages MRInstruments and TwoSampleMR, where data is available for approximately 1,628 traits and 6.1 million variants per study[165]. Variants included in this repository did not necessarily surpass the stringent threshold for genome-wide significance at $p< 5.0x10^{-8}$.

Selecting GWAS studies of relevance

A search was made on MR-Base for all the available studies including the word "diabetes" as a trait. In total, 18 studies were reported under this query in relation to insulin medication to treat diabetes, gestational diabetes, self-reported diabetes without cancer as a comorbidity, early insulin treatment for diagnosed diabetes, and Type 2 Diabetes. From these, only six studies directly related with T2D were taken forward for further inspection. From these, 5/6 studies were reported in the DIAGRAM consortium, and the remaining study was from the UK-Biobank; 4/6 of these studies were from mixed race, and 2/6 included only European samples. The final criteria used for study selection was the number of SNPs with availability of association estimates, sample-size, proportion of cases versus controls, and ethnicity (preferably population from mixed origin). In addition, studies were prioritized if there was no report of comorbidities or secondary phenotypes that could confound the main association. To avoid including studies with overlapping samples, only the most relevant study from DIAGRAM was selected, corresponding to the study by Mahajan *et al*.[30].

Searching for SNP proxies

By default, MR-Base implements a proxy search to increase the number of associations in the exposure dataset that could be identified in the outcome dataset[165]. Proxies are defined as SNPs in the outcome dataset that are in high correlation (LD> 0.8) with a target SNP in the exposure dataset. Reference genetic panel for LD calculation were European samples from the 1000 Genomes Project Phase 3. Once proxies are identified, summary data is retrieved from these associations regarding the effect estimate on the outcome, effect allele of the proxy (allele in phase with the target SNP), and the effect allele of the target SNP[165].

Data Harmonization

Before conducting the MR analysis, genotype-exposure and genotype-outcome datasets were harmonised to adjust for potential inconsistencies in the strand used to report the effect allele across datasets, controlling for wrong effect alleles, palindromic SNPs, and incompatible alleles. Data harmonization guarantees that the effect allele regarded in the exposure dataset, corresponds to the same effect allele in the outcome dataset. Further detail on data harmonization can be found in section 3.7.2 of this Chapter. To avoid errors when merging and harmonizing datasets, it was verified that there were no indel SNPs remaining in the outcome dataset, as this is a common source of error.

3.8.4 Performing the reverse 2SMR

Statistical methods for the reverse 2SMR coincided with those implemented in the forward 2SMR and were conditioned to the number of instruments available for each exposure. Considering that some of the predicted exposure-outcome associations had duplicated MR results because of using two different datasets for extracting genotype-outcome data for a similar instrument, the final causal estimate reported was the one with the strongest evidence of causality based on the smallest p-value.

Diagnostic tests applied when including multiple instruments per exposure were the pleiotropy test and the heterogeneity test; the Steiger test (see section 3.7.4) was suitable for exposures with a single instrument or multiple instruments. The leave-one-out analysis was only relevant when there were multiple instruments per exposure, which was not common in the reverse 2SMR. Inspection plots remained the same as those applied in the forward 2SMR, except for a volcano plot which was included to represent the effect of multiple exposures (i.e. top-ranked DMPs) on one outcome (i.e. T2D).

3.8.5 Power in the reverse 2SMR

As in the forward MR, power to detect a causal association between DNA methylation and T2D was calculated using the web tool mRnd (http://cnsgenomics.com/shiny/mRnd/). In addition, the minimum sample size required to detect a similar causal effect with 95% power and $p<0.05$ was estimated.

3.8.6 Determining true direction of effect in the bidirectional MR

Regarding DMPs with available results in both directions of the MR analysis, for which significant or borderline significant causal associations were detected, the true hypothesized exposure was determined based on the direction of the association where the strongest causal effect was detected (smallest p-value). In addition, considering that for each direction of the MR the method that provided the strongest result was different, and that each method underlines specific assumptions in terms of validity of instruments and consideration for heterogeneity and pleiotropy, the two baseline methods used to compare estimates across MR analyses were the IVW regression and the Wald estimate. Visually, results were compared between the causal analysis (i.e. single sample MR, forward and reverse 2SMR) and the observational analysis using a forest plot generated with the R package forestplot.

Chapter 4 Epigenetic analysis of prevalent type 2 diabetes

This chapter presents evidence of the association between T2D and genome-wide DNA methylation in middle-age adults from the ALSPAC cohort using a single CpG site analysis (EWAS) and a differentially methylated regions (DMR) analysis. DMR results were validated using an alternative method, and results of the EWAS and the DMR analysis were supported via functional exploration of main findings. To improve the power to detect associations in the EWAS, replication was undertaken in four European studies: LBC1936, KORA, and two cohorts from the Rotterdam Study, Rotterdam-III-1 and Rotterdam-Bios.

Aims of the chapter

1. Provide evidence of DNA methylation markers associated with prevalent T2D in ALSPAC using a single-CpG site analysis.
2. Identify differentially methylated regions associated with T2D and validate the evidence using an alternative DMR method.
3. Interpret evidence of the EWAS and DMR analysis using functional exploration of main results to provide insight into new molecular pathways related to the pathogenesis of T2D.
4. Strengthen evidence of the single-CpG site analysis by replication of the EWAS in comparable European cohorts.

Motivation to undertake an epigenome-wide association study

Previous EWAS in T2D have been conducted using European samples[46, 62, 63, 65, 66], but there are few examples where this analysis has been done using prevalent T2D as the exposure of interest to investigate the effect of reverse causation in this association[46, 65, 66]. Furthermore, few of the EWAS in Europeans include findings from population-based studies in the UK[61, 62], where prevalence of T2D is among the lowest according to global records in 2015[22]. Thus, the motivation to undertake this analysis was to investigate the association between genome-wide DNAm and prevalent T2D in a subsample of middle-age participants in ALSPAC, a UK-based cohort study. T2D and DNA methylation were variables collected cross-sectionally from participants in ALSPAC, and because DNA methylation was measured close to the time of disease onset, I hypothesize that observed differences in methylation could be affected by the ongoing occurrence of the disease (reverse causation), and not the opposite. Thus, in this first Chapter, I present results of the association between prevalent T2D as the exposure against DNA methylation as the outcome, aiming at identifying novel associations with T2D in samples from this UK-based study. In a second Chapter

(Chapter 6), I investigate results of the opposite direction of the association, where DNA methylation is regarded as the exposure and T2D as the outcome.

4.1 Baseline characteristics of the subsample of ALSPAC/ARIES

In total, 1,305 middle-age participants in ALSPAC/ARIES were eligible for this study. Mean age in this subsample was 49.96 years (SD=5.4), 94.10% participants were of European origin, 39.46% of the total sample were males, and T2D was reported in 3.68% samples. Mean measures of fasting glucose, BMI and waist-circumference were 5.42 mmol/l, 26.84 kg/m^2 and 89.36 cm, respectively. Further characteristics of participants in ALSPAC/ARIES are described below in Table 4-1.

Table 4-1 Baseline characteristics of participants in the subsample of ALSPAC/ARIES (n=1,305). Participants included had DNA methylation and phenotypic data available.

N= 1,305	**Mean (SD)**	**% (number)**	**Missing (%)**
Age	49.96 (5.41)	---	0.31
Ethnicity [% white]	---	94.10 (1,228)	5.52
Fasting Glucose (mmol/l)	5.42 (1.08)	---	3.14
Body mass index (kg/m^2)	26.84 (4.75)	---	0.77
Waist circumference (cm)	89.36 (13.24)	---	0.46
Waist-hip ratio (cm)	0.86 (0.09)	---	0.46
Systolic BP (mmHg)	123.61 (14.37)	---	1.00
Diastolic BP (mmHg)	74.39 (10.38)	---	1.00
Family History of CHD [Yes %] [a]	---	11.95 (156)	6.97
Serum Total Cholesterol (mmol/L)	4.82 (0.93)	---	3.98
Triglycerides (mmol/l)	1.19 (0.66)	---	3.14
HDL cholesterol (mmol/l)	1.40 (0.35)	---	3.14
LDL cholesterol (mmol/l)	3.04 (0.82)	---	3.14
Fasting C-reactive protein (mg/l)	1.95 (2.70)	---	
Sex [male %]	---	39.46 (515)	0.00
Estimated white cell subset [b]			
CD4T	0.17 (0.06)	---	0.00
CD8T	0.02 (0.03)	---	0.00
Natural Killer Cells	0.20 (0.05)	---	0.00
B-cells	0.09 (0.03)	---	0.00
Monocytes	0.07 (0.03)	---	0.00
Gran	0.50 (0.09)	---	0.00
Medication for T2D [Yes %]	---	1.07 (14)	7.51
T2D [% cases]	---	3.68 (48)	19.54
Family History of Diabetes [Yes %]	---	7.13 (93)	15.79
Smoking [%]			18.93
Never smoker	---	42.91 (560)	
Former smoker	---	31.49 (411)	
Current smoker	---	6.67 (87)	
Physical activity [less than 4h/week %]	---	53.95 (704)	35.25
Socioeconomic status [%]			16.63
High income	---	45.98 (600)	
Middle income	---	26.51 (346)	
Low income	---	10.88 (142)	

[a] CHD: coronary heart disease. [b] Predicted cell-count for white-blood cells using the Houseman method[119].

4.1.1 Addressing missingness in the ALSPAC/ARIES subsample

Missing values in the subsample of 1,305 participants in ALSPAC/ARIES were observed for T2D and some of the covariates including age (n=4), BMI (n=10) and smoking (n=247). Imputation of missing data for age and BMI was done as explained before in Chapter 2, but missing data for T2D was not imputed as this variable was derived from multiple other records (see Chapter 2 "Definition of T2D"), which complicates the process of imputation. Missing values for smoking were imputed using a methylation score previously reported by Zeilinger *et al*.[121], with methods described by Elliott *et al.*[122]. The score was calculated in the complete subsample of ALSPAC/ARIES, but imputation was only applied for samples with missing data in self-reported smoking.

Values of the methylation score ranged between -9.46 and 26.69, and the cut-off used to distinguish smokers from non-smokers was estimated at 3.52 and derived from categories of self-reported smoking (Figure 4-1). This cut-off was calculated as the average point between 500 different nodes generated by a random forest algorithm to accurately split the score based on proximal data points[122, 185]. According to this threshold, never smokers were regarded as participants with a methylation score lower than 3.52, and smokers were participants with a score higher or equal to 3.52. The score had low sensitivity to distinguish between never smokers and former smokers, reason why these two categories were combined when imputing smoking data. For participants with unknown self-reported smoking status (n=247), the score was able to reallocate 211 of them in the non-smoker's group, and the remaining 36 in the smoker's group. When recalculating the cut-off in the score based on imputed categories of smoking (i.e. non-smokers and smokers), the new threshold was identified at 4.91 (Figure 4-1). A higher margin in the score to distinguish samples based on predicted categories of smoking, might reflect the fact that former smokers have been now included in the same category as never smokers, thus increasing the mean of the score in the non-smoker's subgroup.

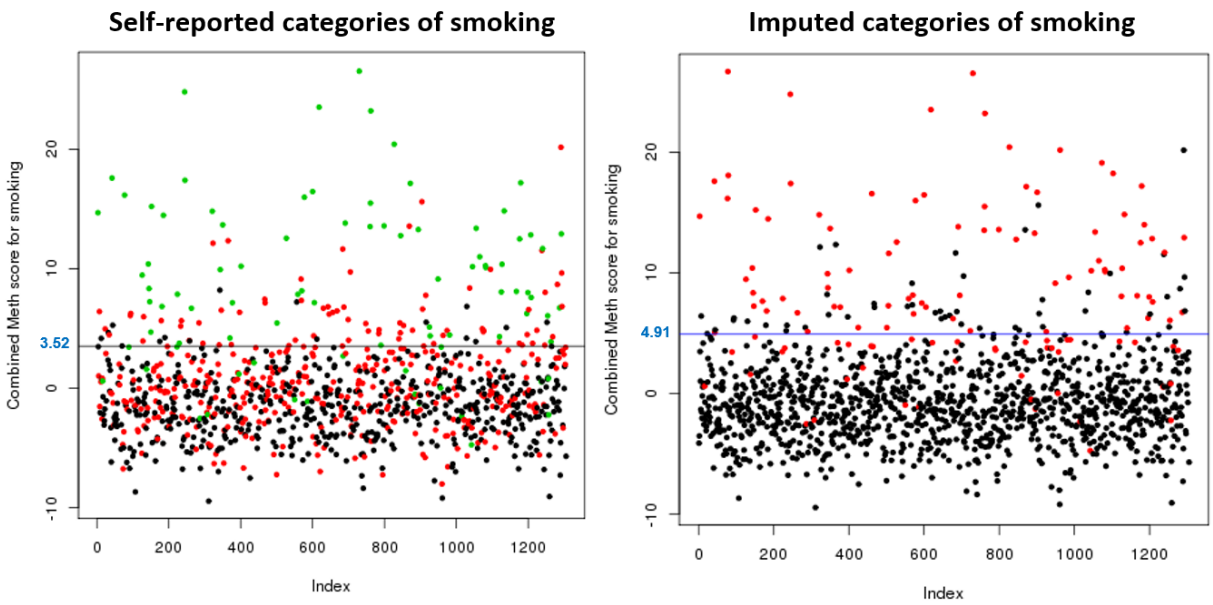


Figure 4-1 Methylation score for smoking calculated for participants in ALSPAC/ARIES (n=1,305). The score is coloured based on categories of self-reported smoking (i.e. never-smokers in black, former smokers in red and current smokers in green) or imputed smoking (i.e. non-smokers in black and smokers in red). The horizontal line across the plot corresponds to the cut-off used to distinguish between smokers and never-smokers, which was calculated using a random forest algorithm. Two different cut-offs were observed, one when using categories of self-reported smoking (y=3.52), and another when using categories of imputed smoking (y=4.91).

With respect to the distribution of the score across categories of self-reported smoking, it was found that mean of the score was significantly lower in never smokers (mean=-1.58, SD=2.67), compared to former (mean=0.49, SD=3.78) and current smokers (mean=8.61, SD=6.78) according to a $p<0.001$ obtained across comparisons (Figure 4-2). Misclassification error was regarded as the percentage of false-positive and false-negatives captured by the score when predicting smoking. To facilitate the calculation of misclassification error, self-reported smoking was recoded into two categories by combining never smokers and former smokers into non-smokers. On average, the percentage of misclassification error obtained with the score was 16.71, and 9.27% of the samples with imputed data for smoking were false positives, and another 24.14% were false negatives. The score had a sensitivity of 75.86% to detect smokers, and a specificity of 90.43% to detect non-smokers. A higher specificity of the score for detecting non-smokers might have been related with the intensity of smoking among participants with self-reported smoking, in which case if the intensity of smoking was low (i.e. casual smokers) among former smokers, then the score was more prone to classify them as never smokers rather than as smokers.

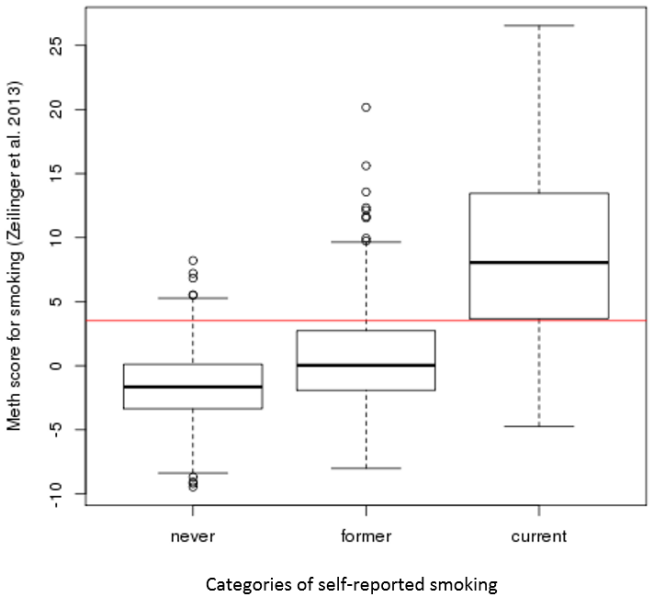


Figure 4-2 Distribution of the methylation score for smoking across categories of self-reported smoking (i.e. never, former and current smoker) in the subsample of adults in ALSPAC/ARIES (n=1,305). As shown in the plot, the score was able to capture differences in methylation between smoking groups, where mean of the score was significantly lower in never smokers compared to former smokers (mean difference=-2.07, p<0.001), and in never smokers compared to current smokers (mean difference=-10.19, p<0.001). Horizontal red line is the cut-off in the score estimated at 3.52 to differentiate smokers from non-smokers. Methylation score was generated based on data reported by Zeilinger et al. [121].

4.1.2 Subset of participants in ALSPAC/ARIES included in the EWAS in T2D

After replacing missing data, a subset of 1,050 participants from the initial subsample of ALSPAC/ARIES remained to conduct the EWAS in T2D. From these, 1,002/1,050 were controls and 48/1,050 were T2D cases (Table 4-2). Comparing between disease groups, there were no differences in age, sex, estimated cell-counts, family history of diabetes or smoking status (Table 4-2). Although most of the samples included in the EWAS were Europeans, there were three T2D cases and 54 controls from a non-European background, a factor to be considered when interpreting genomic inflation (i.e. lambda) in results of the EWAS. Strong differences among cases and controls were detected in levels of fasting glucose (3.18 mmol/l higher in cases, p<0.01), BMI (3.65 kg/m^2 higher in cases, p<0.01), waist-circumference (11.32 cm larger in cases, p<0.01), systolic (6.64 mmHg higher in cases, p<0.01) and diastolic blood pressure (2.87 mmHg higher in cases, p<0.01) (Table 4-2). Also, differences among groups were detected in all the lipid measures, and in fasting levels of C-reactive protein (0.56 mg/l higher in cases, p<0.01). In terms of life-style factors, there were no differences between cases and controls in smoking status or in physical activity, but there were differences in socioeconomic position (p=0.02), where a larger proportion of cases were found in the middle-income group, while controls were predominant in the high-income group.

Table 4-2 Baseline characteristics of the subset of participants in ALSPAC/ARIES included in the EWAS of T2D (n=1,050). Continuous variables were described using the mean and SD, while categorical variables were described using the percent per category and sample size.

	Controls (n=1,002)	**Cases (n=48)**	**P-value**
Age	49.97 (5.28)	51.94 (7.40)	0.17
Ethnicity [% white]	94.61 (948)	93.75 (45)	0.25
Fasting Glucose (mmol/l)	5.27 (0.46)	8.45 (3.00)	<0.01
Body mass index (kg/m^2)	26.44 (4.51)	30.09 (5.52)	<0.01
Waist circumference (cms)	88.31 (12.96)	99.63 (15.65)	<0.01
Waist-hip ratio (cm)	0.86 (0.09)	0.92 (0.10)	<0.01
Systolic BP (mmHg)	122.81 (13.99)	129.45 (15.93)	<0.01
Diastolic BP (mmHg)	73.95 (10.41)	76.82 (12.48)	0.04
Family History of CHD [Yes %]	13.03 (122)	20.45 (9)	0.16
Serum Total Cholesterol (mmol/L)	4.84 (0.92)	4.28 (0.83)	<0.01
Triglycerides (mmol/l)	1.15 (0.63)	1.62 (0.70)	<0.01
HDL cholesterol (mmol/l)	1.42 (0.35)	1.17 (0.29)	<0.01
LDL cholesterol (mmol/l)	3.06 (0.81)	2.57 (0.85)	<0.01
Fasting C-reactive protein (mg/L)	1.96 (2.79)	2.52 (3.05)	<0.01
Sex [male %]	37.92 (380)	18.50 (25)	0.05
Estimated white cell subset			
CD4T	0.17 (0.05)	0.16 (0.05)	0.61
CD8T	0.02 (0.03)	0.03 (0.04)	0.08
Natural Killer Cells	0.19 (0.05)	0.19 (0.05)	0.88
Beta cells	0.09 (0.03)	0.09 (0.03)	0.7
Monocytes	0.07 (0.03)	0.08 (0.03)	0.05
Gran	0.51 (0.08)	0.50 (0.09)	0.33
Medication for T2D [Yes %]	0.00 (0)	28.26 (13)	<0.01
Family History of Diabetes [Yes %]	8.56 (75)	12.50 (5)	0.39
Smoking [%]			
Never smoker	91.02 (912)	85.42 (41)	0.19
Smoker	8.98 (90)	14.58 (7)	
Physical activity [less than 4h/week %]	83.24 (586)	76.92 (20)	0.40
Socioeconomic status [%]			
High income	57.64 (479)	35.14 (13)	0.02
Middle income	30.45 (253)	48.65 (18)	
Low income	11.91 (99)	16.22 (6)	

4.2 Selecting adjustment covariates for the EWAS

Since factors that influence DNA methylation and T2D can bias this association, it was necessary to adjust for potential confounders. Adjustment variables included in the analysis were well-established EWAS covariates including age, sex, smoking, predicted cellular composition[62, 63, 65, 88] and BMI[89, 90], most of them are also well-known risk factors for T2D (see Chapter 1). The association between these confounders and T2D was investigated in the subsample of ALSPAC/ARIES using multiple univariate linear and logistic regressions with T2D as the predictor variable. Similarly, the association between average DNA methylation as the exposure and T2D and potential confounders as the outcome was tested. In an additional analysis, sex was included as an interaction term in the association between average DNA methylation and T2D because risk of T2D can be influenced by sex (see Chapter 1 "Major risk factors"), and because there is substantial evidence supporting difference

in methylation in response to sex[44]. Selecting the appropriate number of independent covariates for the EWAS was fundamental to avoid over-adjustment of the model and removal of important variation between T2D cases and controls.

4.2.1 Association between T2D and potential confounders

T2D was strongly associated with age, BMI, and the predicted cell-count for CD8T cells at p<0.05 (Table 4-3). Borderline association was detected between T2D and sex and predicted cell-count for monocytes (p=0.05), but no association was detected between T2D and smoking (p=0.19) or any of the remaining white cell subsets (p>0.05) (Table 4-3). Based on their association with T2D, potential adjustment covariates for the EWAS were age, BMI and some of the predicted white cells, while there was less evidence that sex and smoking could be used as relevant covariates. However, previous studies have consistently reported sex and smoking as important covariates in EWAS[62, 66, 88], as well as age[46, 63, 66] and BMI[46, 65, 66]. Therefore, age, sex, BMI, smoking and predicted-cell counts (Houseman method) were all included as established covariates in the EWAS of T2D.

Table 4-3 Association between T2D and potential confounders in the subsample of ALSPAC/ARIES (n=1,050). Estimate is the regression coefficient showing the effect of T2D on the covariate. For categorical variables, this estimate is in the log(odds) scale. N is the number of samples included in each regression.

	Estimate	SE	95% CI	P-value	N
Age	1.971	0.797	(0.408, 3.534)	0.014	1,050
Sex (females vs males)	0.576	0.296	(-0.004, 1.157)	0.052	1,050
BMI (kg/m^2)	3.656	0.673	(2.336, 4.975)	<0.001	1,050
Total Cholesterol (mmol/L)	-0.553	0.137	(-0.820, -0.285)	<0.001	1,029
LDL (mmol/L)	-0.493	0.120	(-0.728, -0.259)	<0.001	1,050
[a]HDL (mmol/L)	-0.191	0.035	(-0.260, -0.121)	<0.001	1,050
[b]Physical activity (h/week)	-0.399	0.476	(-1.332, 0.535)	0.403	730
[c]Smoking	0.548	0.424	(-0.282, 1.378)	0.196	1,050
[d]SES	0.921	0.351	(0.233, 1.609)	0.009	868
CD4T	-0.007	0.008	(-0.023, 0.008)	0.356	1,050
CD8T	0.011	0.004	(0.003, 0.019)	0.009	1,050
Natural Killer	-0.001	0.008	(-0.016, 0.014)	0.879	1,050
Beta cells	-0.002	0.004	(-0.009, 0.007)	0.703	1,050
Monocytes	0.008	0.004	(1.1x10^{-5}, 0.017)	0.050	1,050
Granulocytes	-0.010	0.012	(-0.034, 0.015)	0.427	1,050

[a] HDL measures were normalized before the analysis using the natural logarithm. [b] Physical activity was defined as number of hours per week dedicated to do exercise, and it was categorized as less than or equal to 4h/week versus more than 4h/week. [c] Smoking was defined as non-smokers and smokers. [d] Socioeconomic status was categorized as low-, middle- and high-income range based on the level of education.

T2D was also strongly associated with other covariates including total cholesterol, HDL, LDL, and socioeconomic status, but no association was identified between T2D and physical activity (p=0.40) in the subsample of ALSPAC/ARIES. To determine if lipid measures and socioeconomic status could have also been included as covariates in the model, their independence from BMI was tested by estimating the correlation between these factors and BMI. Table 4-4 shows that various lipid measures and socioeconomic status were correlated with BMI, with the highest absolute correlation detected between BMI and LDL (r=-0.42, p<0.001), and the lowest absolute correlation between BMI and total cholesterol (r=0.09, p<0.001). Based on the above, and to avoid over-adjustment of the regression model, lipid measures and socioeconomic status were not included as covariates in the EWAS as their effect was partially proxied by BMI.

Table 4-4 Correlation between BMI and various lipid measures and socioeconomic status in the subsample of ALSPAC/ARIES (n=1,050). Even though there was a weak correlation between BMI and lipid traits and socioeconomic status (r< 0.50), these correlations were significant, suggesting non-independence between these factors and BMI. r: correlation coefficient, N: sample-size in the correlation analysis. SES: socio-economic status.

Outcome	r	P-value	95% CI	N
Total cholesterol (mmol/L)	0.097	0.002	(0.036, 0.157)	1,029
LDL (mmol/L)	-0.415	<0.001	(-0.464, -0.364)	1,050
HDL (mmol/L)	0.131	<0.001	(0.071, 0.189)	1,050
[a]SES	0.114	0.001	(0.048, 0.179)	868

[a] Correlation between socioeconomic status and BMI was approximated using a Pearson correlation between fitted values of BMI, extracted from the regression between BMI and socioeconomic status, and observed values of BMI.

4.2.2 Association between average DNA methylation and T2D and covariates

Average DNA methylation in T2D cases and controls was 0.468 (SD=0.01) and 0.469 (SD=0.01), respectively, and no significant difference was observed between these values (mean difference= -0.001, 95%CI=-0.001, 0.004, p=0.33). Furthermore, there was no difference in mean levels of methylation by sex (mean difference=-3.82x10^{-5}, 95%CI=-0.001, 0.001, p=0.95). When stratifying the association between T2D and average DNA methylation by sex, it was found that among T2D cases, methylation was on average 0.003 higher in males compared to females, but there was no strong evidence of an interaction between T2D and sex (SE=0.003, p=0.35). In addition, average DNA methylation was tested against smoking (estimate=-1.0x10^{-4}, SE=0.001, p=0.89), BMI (estimate=1.22x10^{-4}, SE=6.99x10^{-5}, p=0.08) and fasting glucose (estimate=-0.005, SE=0.005, P=0.35), but none of these association surpassed the threshold of significance at p<0.05.

In conclusion, average DNA methylation is not a good measure to identify potential differences associated with T2D, or with any of the risk factors here addressed. This means that a more specific method needs to be applied to capture variation in methylation in response to these factors, and an EWAS offers this advantage by looking at the association between individual CpG sites in the array, and the trait of interest.

4.3 Identifying structure in methylation data via multi-dimensional scaling

Multi-dimensional scaling (MDS) is a method to visualize the spatial arrangement of samples based on similarities in methylation pattern between them. This method was used as an inspection mechanism to (1) inform of the presence of structure in the methylation data at a first glance, and to (2) determine if known biological variables can be driving any underlying pattern in the data. No metric or estimate was derived from this visual inspection. Analyses were conducted using probes in autosomes that surpassed QC, and further detail of the method used can be found in Chapter 2.

In general, the MDS plot showed that samples in ALSPAC/ARIES were grouped into a single cluster based on their similarities in average levels of methylation, and there was no indication of outliers or samples with extreme values of methylation (Figure 4-3). Thus, results of the MDS plot suggested no structure in the methylation data. Some metrics were used to evaluate performance of the MDS analysis including the *stress*, the goodness-of-fit (GOF) and the R^2. The value obtained for the *stress* was 24.4, the GOF was 0.22 (per dimension of MDS plot), and the estimated R^2 was 0.75 ($F=3069.41$, $p< 0.05$). These results suggest that the MDS analysis was not able to capture the whole dimensionality of the methylation data (*stress* > 0.2), but it was able to represent a good proportion of the variation in the data according to the R^2, but not to the GOF (GOF <0.6). Looking at any underlying pattern in the data according to levels of different factors, there was no suggestion that samples were grouped in the main cluster of the MDS plot based on sex, smoking, ethnicity, T2D status or batch effects (i.e. bisulphite conversion plate) (see Figure 4-4).

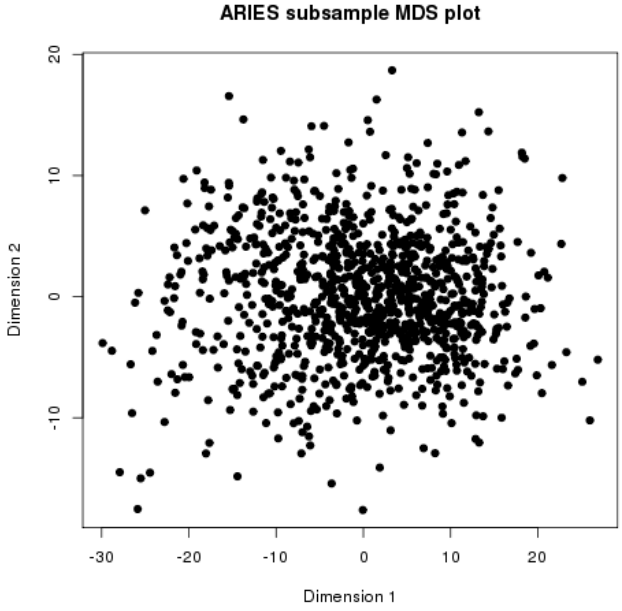


Figure 4-3 MDS plot showing the spatial distance between samples in ALSPAC/ARIES (n=1,050) based on their average values of methylation. No independent clustering was identified, suggesting no structure in the methylation data.

4.4 Batch Effects

A well-established method for the adjustment of batch effects and unmeasured sources of variation in the methylation data was implemented by using a surrogate variable analysis. Further detail of the method can be found in Chapter 2. For the EWAS, ten SVs were initially generated for probes in autosomes using previously defined covariates of the model (i.e. age, sex, BMI, smoking and predicted cell-counts), and testing that these SVs were independent of T2D, the variable of interest in this analysis. In total, seven independent SVs were used as covariates in the EWAS model, and three SVs identified in strong association with T2D (SV2: estimate=0.011 and p=0.016, SV5: estimate=0.010 and p=0.027, SV8: estimate=-0.011 and p=0.013) were excluded from further analyses.

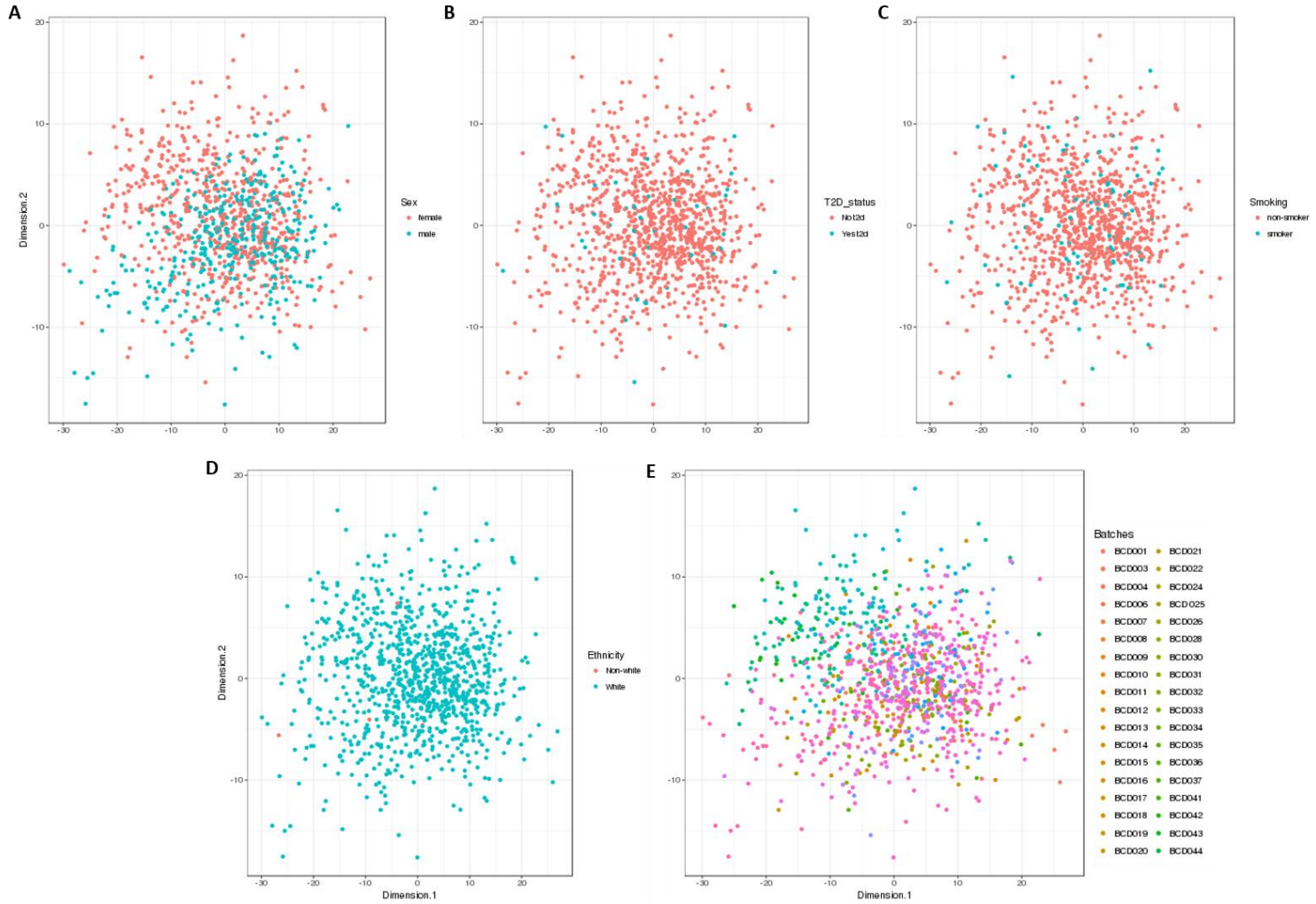


Figure 4-4 Multi-dimensional scaling representing the spatial distance between samples in ALSPAC/ARIES (n=1,050) based on their average levels of methylation. Coordinates of the samples in the MDS were overlapped with categories of different factors to reveal any underlying pattern in the data. Grouping factors considered were: A) sex (female, male), B) T2D status (No t2d, Yes t2d), C) smoking status (non-smoker, smoker), D) ethnicity (non-white, white), and E) batch effects (bisulphite conversion plate, n=107 batches).

4.5 EWAS of T2D in the subsample of ALSPAC/ARIES

An epigenome-wide association analysis was conducted using probes in autosomes, and three adjustment models (Table 4-5). Problematic probes reported by Naeem *et al.*[139] (see Chapter 2) were excluded from the EWAS. Associations were considered significant after multiple testing correction using the Bonferroni method (p-value< 1.07×10^{-7}). Comparison across models of the top-ten associations with the smallest p-value obtained in each EWAS model, is presented in the appendix Table S8-3.

Table 4-5 Description of models implemented in the EWAS in T2D.

EWAS model	Factors included
Minimally adjusted	Age, sex, [a] SVs
Cell adjusted	Age, sex, SVs, [b] predicted cell-counts
Fully adjusted	Age, sex, SVs, predicted cell-counts, BMI, [c] smoking

[a] Surrogate variables included in the EWAS were 7 independent SVs. [b] Predicted count for six white cells: CD4T, CD8T, monocytes, granulocytes, B-cells and natural killer cells, using the method reported by Houseman *et al*.[119]. [c] Smoking was coded as non-smokers and smokers.

4.5.1 Minimally adjusted EWAS model

Strong evidence of association between T2D and methylation was detected at DMP cg10870892 mapping to the region of the *CTTN* gene (estimate=-0.05, SE=0.01, $p=6.24 \times 10^{-8}$). Additional associations with p-value in the order of 10^{-7} were detected at the DMP cg24605023 in *CADPS* (estimate=-0.03, SE=0.01, $p=4.84 \times 10^{-7}$) and the DMP cg15986668 in *NFYC* (estimate=-0.06, SE=0.01, $p=7.0 \times 10^{-7}$) (Table 4-6).

Table 4-6 Top-ten DMPs detected in the EWAS of prevalent T2D using a minimally adjusted model (covariates: age, sex and SVs). CpG context: position of the CpG site relative to the nearest CpG island (island, shore, shelf, open sea); Beta: regression coefficient; SE: standard error; P-value: unadjusted p-value of significance; Bonferroni: adjusted p-value. Associations were considered significant at p-value < 1.07×10^{-7} or at Bonferroni p< 0.05.

CpG	Chr	Gene	CpG context	Beta	SE	P-value	Bonferroni
cg10870892	11	*CTTN*	Open sea	-0.050	0.009	6.24×10^{-8}	0.024
cg24605023	3	*CADPS*	Open sea	-0.031	0.006	4.84×10^{-7}	0.185
cg15986668	1	*NFYC*	N_Shore	-0.064	0.013	7.00×10^{-7}	0.268
cg17749033	17	*Unannotated*	S_Shore	-0.019	0.004	1.29×10^{-6}	0.495
cg25341923	17	*KRTAP4-7*	Open sea	-0.016	0.003	1.74×10^{-6}	0.667
cg19823491	2	*OTX1*	Island	-0.006	0.001	1.79×10^{-6}	0.686
cg26353859	12	*SLC16A7*	Open sea	0.031	0.007	2.99×10^{-6}	1.000
cg04016326	12	*GRIN2B*	N_Shore	-0.054	0.012	3.44×10^{-6}	1.000
cg05575921	5	*AHRR*	N_Shore	-0.036	0.008	3.61×10^{-6}	1.000
cg03206717	3	*SLC25A38*	Island	-0.003	0.001	4.38×10^{-6}	1.000

The effect direction of the associations detected varied relative to the DMP under analysis, but when looking at signals with borderline evidence of association based on a p-value<10^{-5}, there was some suggestion that T2D cases were predominantly hypomethylated compared to controls (Table 4-6 and Figure 4-5). Also, among these suggestive signals of the EWAS, the absolute value of the effect size was smaller than 0.06 (Figure 4-5).

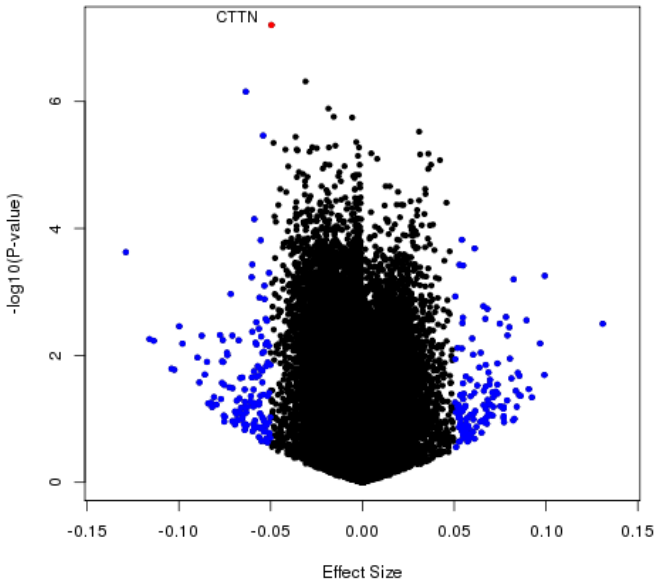


Figure 4-5 Volcano plot showing the distribution of the effect-size against p-values for associations detected in the minimally-adjusted EWAS of T2D in the subsample of ALSPAC/ARIES (n=1,050). Each dot represents a CpG site included in the EWAS after QC, blue dots are CpG sites with an absolute effect-size greater than 0.05 and those in red are the CpG sites that surpassed Bonferroni correction (p<1.07x10^{-7}). X-axis: effect-size (regression coefficient), and y-axis: -Log10 (p-value). One CpG was identified with EWAS significance at the DMP cg10870892 in the CTTN gene.

4.5.2 Cell-adjusted EWAS model

None of the associations identified in this model reached EWAS significance (Table 4-7). P-value of significance for the signal previously identified at the DMP in *CTTN* was attenuated after adjustment for cell counts (minimally-adjusted p=6.24x10^{-8} vs cell-adjusted p=6.40x10^{-7}), without a change in the magnitude of the effect estimate (minimally- and cell-adjusted estimate=-0.05). The association with the smallest p-value in this model was detected at the DMP cg14045803 in *STARD10* (estimate=-0.01, SE=0.002, p=2.70x10^{-7}). Other association identified with p-value in the order of 10^{-7} was at the DMP cg15986668 in *NFYC* (estimate=-0.07, SE=0.01, p=4.61x10^{-7}).

Table 4-7 Top-ten DMPs detected in the EWAS of T2D additionally adjusted for six white cells (covariates: age, sex, SVs and predicted cell-counts). Associations were considered significant at p-value < 1.07x10^{-7} or Bonferroni p< 0.05.

CpG	Chr	Gene	CpG context	Beta	SE	P	Bonferroni
cg14045803	11	*STARD10*	Island	-0.011	0.002	2.70x10^{-7}	0.103
*cg15986668	1	*NFYC*	N_Shore	-0.065	0.013	4.61x10^{-7}	0.177
*cg10870892	11	*CTTN*	Open sea	-0.045	0.009	6.40x10^{-7}	0.245
*cg19823491	2	*OTX1*	Island	-0.006	0.001	1.62x10^{-6}	0.621
*cg04016326	12	*GRIN2B*	N_Shore	-0.055	0.012	2.71x10^{-6}	1.000
cg26652413	19	*CPAMD8*	N_Shore	-0.022	0.005	2.96x10^{-6}	1.000
cg00204249	17	*DNAH17*	S_Shelf	-0.014	0.003	3.29x10^{-6}	1.000
*cg05575921	5	*AHRR*	N_Shore	-0.036	0.008	3.72x10^{-6}	1.000
cg14290451	6	*RPL10A*	Island	-0.004	0.001	4.38x10^{-6}	1.000
*cg03206717	3	*SLC25A38*	Island	-0.003	0.001	4.76x10^{-6}	1.000

*CpG sites also showed borderline evidence of association with T2D at p<10^{-5} in the minimally adjusted model.

Overall, there was consistency in the magnitude and direction of effect for signals detected in common between the minimally- and the cell-adjusted models (see Table 4-6 and Table 4-7). The volcano plot in Figure 4-6 shows that for associations with the smallest p-value in the EWAS, T2D cases were predominantly hypomethylated compared to controls.

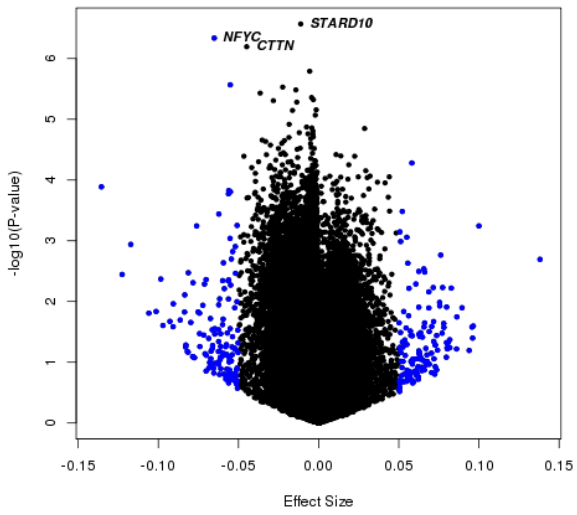


Figure 4-6 Volcano plot showing the distribution of effect-sizes against p-values for associations detected in the EWAS additionally adjusted for predicted cell-counts in the subsample of ALSPAC/ARIES (n=1,050). Each dot represents a CpG site included in the EWAS after QC. Blue dots are CpG sites with an absolute effect-size greater than 0.05. X-axis: effect-size (regression coefficient), and y-axis: -Log10 (p-value). None of the CpG sites reached EWAS significance at p<1.07x10^{-7}, and the association with the smallest p-value was detected at DMP cg14045803 in the STARD10 gene.

4.5.3 Fully adjusted EWAS model

Strong evidence of association was detected between T2D and methylation at the DMP cg15986668, mapping to the region of the *NFYC* gene (p=5.48x10^{-8}) (Table 4-8, see appendix Figure S8-3). This association was independent of the effect of BMI and smoking, and of SNPs located in the probe-

binding region. At the DMP in *NFYC*, T2D cases were on average 0.07 (SE=0.01) hypomethylated compared to controls (mean methylation in cases= 0.46, SD=0.10, and mean in controls= 0.53, SD=0.09, see Figure 4-7). The goodness-of-fit of the model revealed that T2D and covariates explained approximately 11.29% of the variation in methylation at the DMP in *NFYC*. The second strongest association in this EWAS was detected at DMP cg14045803 in *STARD10* (estimate=-0.01, SE=0.002, $p=1.39x10^{-7}$) (see appendix Figure S8-3). The goodness-of-fit of the model showed that T2D and covariates explained around 40.79% of the variation in methylation in the DMP in *STARD10*.

Remaining top-ten signals of the EWAS were identified with p-value in the order of 10^{-6}, and most of them were consistently detected across the different adjustment models. Unique signals of the fully-adjusted model were detected at the DMP cg02307288 in *TRPC7* ($p=5.54x10^{-6}$) and the DMP cg04656330 in *PNKD* ($p=7.96x10^{-6}$) (Table 4-8). Of interest in this model was the removal of a borderline association previously detected in the *AHRR* locus (DMP cg05575921), which has been widely reported in association with smoking (Table 4-8).

Table 4-8 Top-ten DMPs detected in the EWAS of T2D using a model additionally adjusted for BMI and smoking (covariates: age, sex, SVs, predicted cell counts, BMI and smoking). Associations were considered significant at p-value < $1.07x10^{-7}$ or at Bonferroni p< 0.05.

CpG	Chr	Gene	CpG context	Beta	SE	P	Bonferroni
*cg15986668	1	*NFYC*	N_Shore	-0.071	0.013	$5.48x10^{-8}$	0.021
*cg14045803	11	*STARD10*	Island	-0.012	0.002	$1.39x10^{-7}$	0.053
*cg10870892	11	*CTTN*	Open sea	-0.045	0.009	$1.13x10^{-6}$	0.431
*cg26652413	19	*CPAMD8*	N_Shore	-0.023	0.005	$2.51x10^{-6}$	0.963
*cg00204249	17	*DNAH17*	S_Shelf	-0.015	0.003	$2.76x10^{-6}$	1.000
*cg03206717	3	*SLC25A38*	Island	-0.003	0.001	$2.95x10^{-6}$	1.000
*cg19823491	2	*OTX1*	Island	-0.006	0.001	$2.99x10^{-6}$	1.000
cg02307288	5	*TRPC7*	Open sea	-0.038	0.008	$5.54x10^{-6}$	1.000
*cg04016326	12	*GRIN2B*	N_Shore	-0.054	0.012	$5.71x10^{-6}$	1.000
cg04656330	2	*PNKD*	Island	-0.002	$3.52x10^{-4}$	$7.96x10^{-6}$	1.000

* CpG sites also showed borderline evidence of association with T2D at $p<10^{-5}$ in the cell-adjusted model.

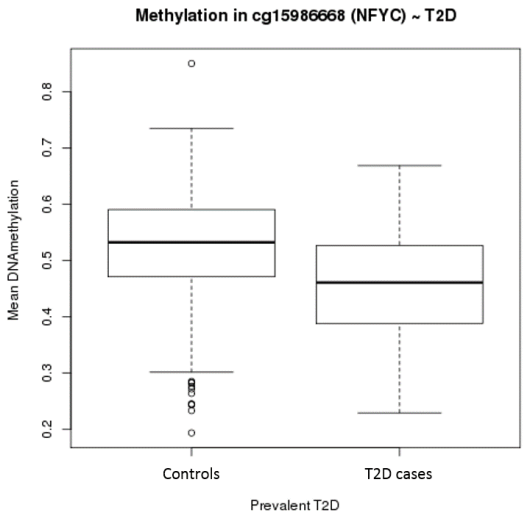


Figure 4-7 Difference in DNA methylation between T2D cases and controls at the DMP in NFYC (cg15986668). Mean DNA methylation corresponds to average beta values, which was 7.1% (SE= 0.01, p=5.48x10^{-8}) lower in T2D cases compared to controls.

As in previous models, the distribution of effect sizes was predominantly negative for DMPs with the smallest p-value, indicating general hypomethylation in T2D cases compared to controls. For top-ranking associations, absolute effect size ranged between 0.002 and 0.071 (Figure 4-8).

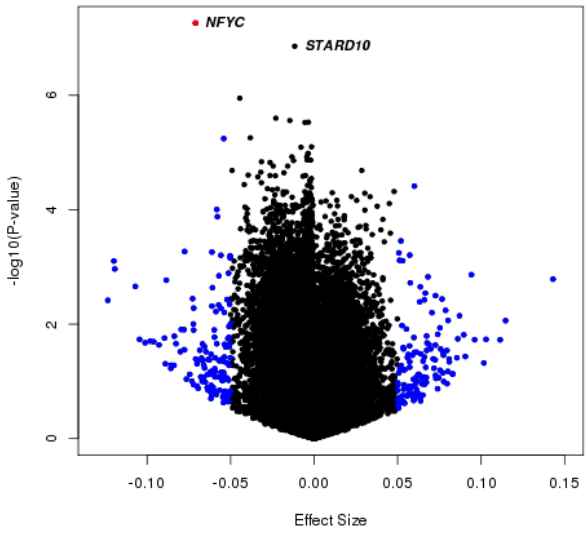


Figure 4-8 Volcano plot showing the distribution of effect sizes against p-values for associations detected in the most-adjusted EWAS additionally adjusted for BMI and smoking, using the subsample of ALSPAC/ARIES (n=1,050). Each dot represents a CpG site included in the EWAS after QC. Blue dots are CpG sites with an absolute effect-size greater than 0.05. X-axis: effect-size (regression coefficient), and y-axis: -Log10 (P-value). The strongest association was identified at the DMP cg15986668 in NFYC (p=5.48x10^{-8}), and the second strongest association was identified at the DMP cg14045803 in STARD10 (p=1.39x10^{-7}).

4.5.4 Sensitivity Analyses

Following the EWAS analyses conducted above, I performed a number of sensitivity analyses.

Analysis	Aim
Sensitivity analysis for BMI	Investigate the effect of adjusting for BMI in the EWAS by comparing estimates obtained with and without BMI as a covariate.
Association between T2D and quartiles of methylation at *NFYC*	Determine if the association between T2D and *NFYC* remains significant after stratifying the sample by quartiles of methylation at *NFYC*. In addition, evaluate the risk of T2D across quartiles.
Risk-factor analysis and further adjustment of the association at *NFYC*	Investigate potential mechanisms explaining the association between T2D and DNAm at the DMP in *NFYC*, with further adjustment of this association for risk-factors in correlation with *NFYC* methylation.

4.5.4.1 Sensitivity analysis for BMI

To determine the effect of adjusting for BMI in results of the EWAS, a secondary analysis was conducted by excluding BMI from the regression model, and calculating the correlation between effect estimates obtained with and without adjustment for BMI, in addition to reporting the average change in the effect estimate and p-value for top associations (smallest p-value) identified in common between models.

Overall, in the model without adjustment for BMI, none of the associations identified surpassed Bonferroni correction at $p<1.07x10^{-7}$. The strongest association was detected at the DMP in *STARD10* (estimate=-0.011, SE=0.002, $p=3.07x10^{-7}$), and other associations with p-value in the order of 10^{-7} were identified at the DMP cg15986668 in *NFYC* (estimate=-0.064, SE=0.013, $p=5.78x10^{-7}$) and the DMP cg10870892 in *CTTN* (estimate=-0.045, SE= 0.008, $p=7.58x10^{-7}$). Comparing effect estimates between this model and the fully adjusted model, there was weak evidence of correlation between effect estimates across models (rho=0.09, p<0.001).

Of the top 14 DMPs identified with $p<10^{-5}$ in the unadjusted model for BMI, nine of them overlapped with top DMPs detected in the fully adjusted model (Table 4-9). At these top nine DMPs, adjustment for BMI had little effect: median absolute change in the effect estimate before and after adjustment for BMI was 1.87% (relative percentage change ranged from 9.86 decrease to 1.87 increase), and only 3/9 sites changed by 3% or more. At almost half of the top DMPs, adjustment for BMI increased the magnitude of the effect estimate, and in all the top associations adjustment for BMI increased the standard error (i.e. decreased precision). P-value in the order of 10^{-7} was identified in 3/9 top

DMPs in the unadjusted model, and in 2/9 top DMPs in the fully adjusted model. For the strongest associations (smallest p-value) detected in the fully-adjusted model (DMPs in *NFYC* and *STARD10*), adjustment for BMI had the largest effect at the DMP in *NFYC*, where the absolute effect estimate increased by 9.2% and the p-value decreased from 10^{-7} to 10^{-8} (i.e. higher significance) after adjustment for BMI (Table 4-9).

In conclusion, this sensitivity analysis demonstrated that BMI was a confounder of the association between T2D and DNAm based on results of top-ranking DMPs detected with the smallest p-value in the EWAS. Evidence showed that on average adjustment for BMI increased the absolute effect estimate, decreased precision of the effect estimates, and had a small impact in the significance of the associations. For the DMP in *NFYC*, which was the strongest signal detected in the fully adjusted model, adjustment for BMI had the largest effect by increasing the magnitude of the absolute effect estimate, while increasing the strength of this association.

Table 4-9 Comparison of EWAS results for top DMPs with the smallest p-value detected in common between the model with and without adjustment for BMI.

CpG	Chr	Gene	Unadjusted model for BMI			Fully adjusted model			% change†
			Beta	SE	P	Beta	SE	P	
cg15986668	1	*NFYC*	-0.065	1.28E-02	5.78E-07	-0.071	1.30E-02	5.48E-08	9.86
cg04656330	2	*PNKD*	-0.002	3.47E-04	7.58E-06	-0.002	3.52E-04	7.96E-06	1.12
cg19823491	2	*OTX1*	-0.006	1.16E-03	1.70E-06	-0.006	1.17E-03	2.99E-06	1.08
cg03206717	3	*SLC25A38*	-0.003	7.18E-04	4.84E-06	-0.003	7.26E-04	2.95E-06	3.46
cg10870892	11	*CTTN*	-0.045	8.97E-03	7.58E-07	-0.045	9.09E-03	1.13E-06	0.23
cg14045803	11	*STARD10*	-0.011	2.17E-03	3.07E-07	-0.012	2.20E-03	1.39E-07	4.38
cg04016326	12	*GRIN2B*	-0.055	1.17E-02	2.78E-06	-0.054	1.19E-02	5.71E-06	1.87
cg00204249	17	*DNAH17*	-0.014	3.04E-03	2.47E-06	-0.015	3.08E-03	2.76E-06	0.88
cg26652413	19	*CPAMD8*	-0.022	4.78E-03	3.49E-06	-0.023	4.85E-03	2.51E-06	2.83

†Percentage change of the absolute effect estimate (beta) between the unadjusted model, and the fully adjusted model (including BMI).

4.5.4.2 T2D versus quartiles of methylation at NFYC

In an additional analysis, differences in methylation between T2D cases and controls were further determined by stratifying methylation at the *NFYC* locus into quartiles. In addition to this, it was investigated if the risk of T2D varied across quartiles of *NFYC*. Quartiles were arranged so that they represented the increasing risk of T2D. Thus, mean methylation at the bottom quartile was the highest, while mean methylation at the top quartile was the lowest.

Considering differences in methylation between T2D cases and controls within the quartiles, significant difference at $p<0.05$ was only detected at the top quartile, where cases were on average

2.9% (SE=0.01) hypomethylated compared to controls (95%CI= -0.050, -0.007, p=0.002) (Table 4-10). An analysis of variance further revealed that differences in mean methylation between cases and controls were also significant across the quartiles based on a p-value of 0.04 for the interaction between T2D and the quartiles (Table 4-10). In addition, it was identified that participants in the upper quartile of *NFYC* had approximately five times higher risk of T2D compared to participants in the lower quartile (OR=4.49, 95%CI=11.16, 1.81, p=0.001), and the p-value for trend further suggested that the risk of T2D increased linearly from the bottom to the top quartile (p-trend= $2.38x10^{-5}$). In conclusion, this quartile analysis further demonstrated that T2D is associated with hypomethylation of *NFYC*, and that the risk of the disease increases with a decrease in methylation at this locus.

Table 4-10 Association between quartiles of methylation at the DMP cg15986668 in NFYC and T2D. Within each quartile it is reported the mean and SD of methylation in T2D cases and controls, difference in mean methylation, and the p-value from a regression analysis. A two-way ANOVA was used to compare mean difference in methylation between cases and controls across the quartiles.

	Q1 (n=263) Mean (SD)	**Q2 (n=262) Mean (SD)**	**Q3 (n=262) Mean (SD)**	**Q4 (n=263) Mean (SD)**
T2D cases	0.624 (0.026)	0.562 (0.025)	0.502 (0.016)	0.379 (0.052)
Controls	0.634 (0.036)	0.558 (0.017)	0.499 (0.017)	0.408 (0.051)
Mean difference	-0.010	0.003	0.003	-0.029
SE	0.015	0.008	0.005	0.011
P-value	0.997	0.999	0.999	0.002
Two-way ANOVA†	MS=0.003	SS=0.010	F=2.799	p=0.039

†Parameters of the ANOVA: mean square (MS), sum-of-squares (SS), F-value (F) and p-value for the interaction between T2D and the quartile.

4.5.4.3 Correlation between risk factors for T2D and methylation at NFYC

To determine potential mechanisms influencing the association between T2D and the DMP in *NFYC*, it was investigated the correlation between different risk factors for T2D and *NFYC*. Results demonstrated that methylation in *NFYC* was positively correlated with c-reactive protein (r=0.02, p=0.02), but inversely correlated with fasting glucose (r=-0.13, $p=2.98x10^{-5}$) and HOMA-IR (r=-0.10, p=0.01) (see appendix Table S8-4). No correlation was identified between methylation in *NFYC* and fasting insulin, HOMA-B, various lipid measures, waist circumference, waist-hip ratio, and systolic and diastolic blood pressure (see appendix Table S8-4). These findings suggest that one possible mechanism by which hypermethylation of *NFYC* could lower the risk of T2D, is by acting directly upon the levels of glucose in blood. In contrast, the mechanism linking c-reactive protein with methylation in *NFYC* was independent of T2D (adjusted-p=0.02), suggesting that different biological processes could be determining the association between C-reactive protein, methylation in *NFYC*, and T2D (see appendix Figure S8-2). Even though c-reactive protein was not considered a covariate

in the EWAS, it was identified as a true confounder due to its association with both, T2D and methylation at the *NFYC* locus.

To determine the effect of risk factors correlated with *NFYC* in the association between methylation at this locus and T2D, further adjustment was applied for these risk factors. In a first model, the association at *NFYC* was adjusted for age, sex, fasting glucose and c-reactive protein as covariates, regarding associations at $p<0.05$. After further adjustment, the association at *NFYC* remained significant ($p=9.2x10^{-4}$), and T2D cases were on average 5.7% (SE=0.02) hypomethylated compared to controls. The effect detected in this association was attenuated compared to the effect detected in the main EWAS model (T2D cases 7.1% hypomethylated vs controls, see Table 4-8). In a second model, the association at *NFYC* was additionally adjusted for HOMA-IR, and only borderline association was identified in this model ($p=0.047$). As with the first model, further adjustment for HOMA-IR reduced the effect estimate at *NFYC* by almost 30% compared to the effect identified in the main EWAS model (see Table 4-8). Adjustment for HOMA-IR was only possible for a subset of mothers in ALSPAC/ARIES (n=645) with available measures of HOMA-IR and other covariates.

In conclusion, the risk factor analysis showed that methylation at *NFYC* was correlated with some established risk factors for T2D, and the association between T2D and *NFYC* remained significant after adjustment for these risk factors, except for HOMA-IR. However, there is no sufficient evidence to demonstrate that T2D is causally associated with hypomethylation of the CpG in *NFYC* because residual variation can still be present regarding unmeasured confounders. Therefore, to reinforce evidence from the observational analysis, it is necessary to implement Mendelian randomization, which methods and results are described in Chapter 3 and Chapter 7.

4.6 Functional exploration of top signals identified in the EWAS

Various functional analyses were conducted for top CpG sites with the smallest p-value detected in the fully-adjusted EWAS in T2D aiming at: identifying eQTM, describing the genomic location of these top sites, reporting the presence of regulatory epigenomic features and genetic variants associated with methylation at these sites (meQTL), and their overlap with genetic determinants of T2D and related outcomes. Finally, a pathway analysis was conducted to determine biological mechanisms linking the association between methylation and T2D.

4.6.1 Identifying eQTM's for top T2D-associated DMPs in the EWAS

Using the Bios QTL browser, a lookup was performed to interrogate if top CpG sites with the smallest p-value ($p<1.0x10^{-5}$) detected in the fully adjusted EWAS model, have been previously associated with gene expression of the nearest gene. However, none of the top CpG sites of interest was previously identified as a *cis*-eQTM in the Bios QTL dataset. Thus, functional description of the two strongest CpG sites detected in the EWAS at the DMP cg15986668 and cg14045803, was based on the position of the nearest gene.

4.6.2 Genomic context of DMP cg15986668 in NFYC

Using the UCSC Genome Browser (version: GRCh37/hg19, release date: 02-2009, analysis date: 24-04-2018, http://www.epigenomebrowser.org), it was investigated the genomic context of the strongest signal of the EWAS identified at DMP cg15986668, which mapped near the region of the *NFYC* gene, 1500bp upstream the transcription start site (TSS) of *NFYC*, and within the region of the antisense non-coding RNA *NFYC-AS1* (see Figure 4-9). *NFYC* is the nuclear transcription factor gamma subunit C gene, a conserved trimeric protein that binds with specificity to 5'-CCAAT-3' DNA sequences in the promoter region of many genes[186], acting as a repressor or activator depending on interacting cofactors[186]. One pathway related to this gene was the regulation of cholesterol biosynthesis by *SREBPF (sterol regulatory element binding transcription factor 1)*[186], but there was no further evidence relating the function of *NFYC* to the pathophysiology of T2D. According to GTEx data, higher median expression of this gene was detected in the thyroid (42.39 TPM), spleen (41.61 TPM), whole blood (32.29 TPM), among other tissues[163]. TPM was the unit used for the measurement of gene expression, and it stands for transcripts per million. Evidence of the ubiquitous expression of *NFYC* goes in line with the function of this gene as a conserved transcription factor.

Regarding distance from the nearest CpG island, the CpG cg15986668 was in a shore region, approximately 2kb upstream the nearest CpG island spanning the transcription start site and promoter region of *NFYC* (Figure 4-9). Based on reported data on K562 cell lines (leukaemia cells), levels of the histone mark H3K27Ac, which is a signature of transcription activation[187], were low upstream the position of the CpG in *NFYC*, but they increased downstream this site, towards the promoter of *NFYC* [188]. In addition, there were two DNaseI hypersensitive clusters next to this CpG, which are regions of accessible DNA[189]. Furthermore, this CpG was partially methylated in K562 cells [188, 190]. Multiple transcription factors overlapped within the region of the CpG in *NFYC*, one of them

was *TCF7L2* according to data in several carcinoma cell lines *(HeLa, PANC-1, HEK-293, MCF-7, HCT*)[188, 190].

Altogether, the genomic context of the top signal of the EWAS suggested that the CpG in *NFYC* was in a region of open DNA based on the presence of DNaseI hypersensitive clusters, which was transcriptionally active based on the incremental enrichment of the histone mark H3K27Ac downstream the position of the CpG, and based on the partial level of methylation of this marker in K562 cells. The function of *NFYC*, a highly conserved transcription factor, was only linked to T2D via a pathway related to the biosynthesis of cholesterol.

4.6.3 Shared genetics between DNA methylation in *NFYC* and T2D

Common genetic variants in association with methylation at the *NFYC* locus were looked up in the meQTL database (analysis date: 24-04-2018, http://www.mqtldb.org/search.htm) to identify common genetics between these, and genetic variants previously reported in association with T2D and other related outcome. The time-point used to retrieve meQTL was middle-age to coincide with the time-point where the signal at *NFYC* was detected in association with T2D. *Cis* meQTL were compared to reported risk variants for T2D (n= 711 reported SNPs) and diabetes-related traits based on data available in the GWAS Catalog (release date: 2008, analysis date: 24-04-2018, https://www.ebi.ac.uk/gwas). Related traits with T2D searched were fasting glucose fasting insulin, HbA1c, insulin resistance, HOMA-IR, HOMA-B, and three T2D complications: diabetic retinopathy (n=34 reported SNPs), diabetic nephropathy (n=6 reported SNPs) and diabetic foot (n=5 reported SNPs). In total, 131 *cis* meQTL were interrogated for CpG cg15986668, but none of them overlapped with SNPs reported in the GWAS Catalog for T2D, T2D related traits, or T2D complications.

Alternatively, it was searched for traits directly associated with *cis* meQTL for *NFYC*, using for this the GWAS Catalog and the Genetic Association of Complex Diseases and Disorders database (GAD), a repository accessed through the SNP nexus platform (release date: 2008, analysis date: 24-04-2018, http://snp-nexus.org). According to the GWAS Catalog, the meQTL rs4660456 was borderline associated ($p=4.0 \times 10^{-6}$) with platelet count in 115 HIV controls from African ancestry[191], but no other trait was reported in association with meQTL for *NFYC*. Based on data from the GAD repository, 88/131 meQTL for *NFYC* were associated with renal cell carcinoma based on a candidate loci study[192]. However, no information was provided on this study for the effect size and strength of these associations.

In conclusion, there was no indication of shared genetics between DNA methylation in *NFYC* and T2D and other related outcomes. Because there was no previous evidence supporting an association between the CpG in *NFYC* and T2D, and because there was no obvious connection between this locus and T2D from this functional inspection, no further conclusions were drawn of the mechanisms linking *NFYC* and T2D.

4.6.4 Functional exploration of DMP cg14045803 at STARD10 locus

The second strongest marker detected in the EWAS (fully adjusted model) was at DMP cg14045803 (Bonferroni adjusted p=0.05), mapping to the region of *STARD10*. Evidence showed that hypomethylation at this DMP was suggestively associated with prevalent T2D (estimate=-0.012, SE=0.002). The associated gene to this DMP is the steroidogenic acute regulatory protein-related lipid transfer domain-containing 10, a cytosolic protein with a lipid binding domain that facilitates the intracellular transport of lipids between organelles[193]. *STARD10* has binding affinity for phosphatidylcholine and phosphatidylethanolamine phospholipids[193]. Pathways related to this gene were the positive regulation of peroxisome proliferator activated receptor (PPAR) signalling pathway[194], and a pathway related to the biosynthesis of glycerophospholipids[186]. Looking at tissue-specific gene expression, GTEx data reported higher transcripts for *STARD10* in the liver (281.7 TPM), testis (240.7 TPM), stomach (230.9 TPM), and visceral adipose tissue (140.3 TPM), respectively[163], and this evidence was in line with the metabolic function of *STARD10* as an intracellular transporter of lipids. Common genetic variants in *STARD10* have been reported in association with risk of T2D by Imamura *et al*.[195], in association with acute insulin response by Wood *et al*.[196], and in association with differential expression of *STARD10* in pancreatic islets of T2D cases versus controls by Carrat *et al*.[197]. In the study by Carrat and colleagues, they reported that genetic variants in *STARD10* increased the risk of T2D by downregulating the expression of this gene in β-cells[197].

Regarding the genomic context of the CpG in *STARD10*, it was located within a CpG island in the 5'UTR of the same gene, and there was no SNP overlapping the DNA binding-site of this CpG. In addition, this site was unmethylated in most cell lines reported by the ENCODE project, except for HeLa cells (cervical cancer cell lines). There were some DNaseI hypersensitive clusters overlapping the region of the CpG in *STARD10*, but there was low representation of the histone mark H3K27Ac, which is a signature of transcription activation[187].

Figure 4-9 Genomic context of the strongest signal detected in the EWAS of T2D using the UCSC Genome Browser. Different tracks were displayed as available in the genome browser and described to the right-hand side of the figure. The CpG site of interest is highlighted in solid background within the track named CpG sites. A custom track was included and named "EWAS of T2D" to represent results of the EWAS in the fully-adjusted model conducted in participants in ALSPAC/ARIES. For this track, the length of the bar represents the score given to the association at each CpG site based on the level of significance reported by the P-value, and the direction of the association reported by the t-statistic. The score was generated as the product of the log10(P-value) and the sign of the t-statistic (i.e. directional P-value). For the genomic region displayed, the strongest association was detected at DMP cg15986668 close to the NFYC gene (beta=-0.07, SE=0.01, P=5.48x10^{-8}). Image provided by the UCSC Genome Browse, 2017. Available at: http://genome.ucsc.edu/ (accessed: 27 April 2018).

Epigenetic features for the CpG in *STARD10* suggested that this site was in a region of open DNA, where it was found primarily unmethylated. In addition, because this CpG was within a CpG island located downstream the TSS of *STARD10*, it was possible that it could play a role in regulating the expression of *STARD10.* According to the literature, unmethylated CpG islands outside the promoter of a gene and within the gene body, have regularly unknown function and are considered orphan CpG islands[198]. Orphan CpG islands can function as alternative promoters and regions of transcriptional initiation[198]. This latter concept was supported by the presence of a binding site for RNA Polymerase-II subunit A (POLR2A) within the CpG island containing the CpG of interest in *STARD10*. To recall, POLR2A is the main RNA protein required for gene transcription in eukaryotic cells[199].

Even though there was evidence supporting an association between genetic variants in *STARD10* and T2D, by the time this study was conducted there was no literature supporting an association between DNA methylation in *STARD10* and T2D. As before, meQTL for the target DMP were retrieved from the meQTL database, aiming to identify an overlap between the meQTL and genetic risk factors for T2D previously reported in *STARD10*. Since meQTL for this DMP were all *trans* meQTL (meQTL > 1Mbp apart from the CpG site), no overlap could be identified between *trans* meQTL and T2D SNPs in the *STARD10* gene. Furthermore, *trans* meQTL for the DMP in *STARD10* were identified in association with methylation at birth and childhood, which were not the time-points of interest in this study. Thus, *trans* meQTL for *STARD10* were not valid genetic proxies for middle-age methylation.

Summary of meQTL inspection at NFYC and STARD10, and further considerations for the causal analysis

MeQTL identified for the CpG in *NFYC* did not overlap with risk variants for T2D and T2D related traits reported in the GWAS Catalog. Regarding meQTL for the CpG in *STARD10*, there was no overlap between these and risk variants for T2D reported in the same gene. Even though meQTL for *NFYC* were not associated with T2D and related outcomes, they were taken forward to investigate the causal direction of the association between T2D and methylation at *NFYC* using a bidirectional Mendelian randomization. Methods and results of the causal analysis are described in Chapter 3 and Chapter 7, respectively. Considering that meQTL for the CpG in *STARD10* were not valid genetic instruments, it was necessary to request data from the Genetics of DNA methylation consortium (GoDMC, http://www.godmc.org.uk/) to improve power to detect *cis* meQTL for this locus. Further detail of the GoDMC consortium can be found in Chapter 3.

4.6.5 Enrichment for regulatory elements among top CpG sites identified in the EWAS

This section conducts functional analysis using the top 1,000 CpG sites with the smallest p-values obtained from the fully adjusted EWAS model. This arbitrary number of CpG sites was selected to improve the output of the functional analysis considering the small number of associations obtained with $p<1.0x10^{-5}$ (n=11 DMPs) in the EWAS. The p-value of association for the CpG sites included in this analysis ranged between $5.48x10^{-8}$ and $1.04x10^{-3}$. Functional analysis comprised the description of various regulatory elements including DNAse I hypersensitive sites and histone marks at specific cell-types and tissues, and description of the genomic context of top CpG sites of interest.

Overlap between top CpG sites and DNaseI hypersensitive sites and the histone mark H3K27me3, was assessed using the platform eFORGE, with data reported by the ROADMAP epigenomics project (eFORGE v1.2, release date: 1999-2014, analysis date: 01-05-2018, http://eforge.cs.ucl.ac.uk/). Results suggested that the majority of the strongest CpG sites of the EWAS resided in DNaseI hypersensitive regions, and this finding was consistent across different human tissues like blood, pancreas, small intestine, skin and gastric tissue (see Figure 4-10). DNaseI hypersensitive sites are regions commonly associated with accessible DNA, able to concentrate active regulatory elements and therefore, potentially active for transcription[149]. In contrast, top CpG sites were depleted of regions enriched in the histone mark H3K27me3, and this was a common characteristic across different tissues (Figure 4-10). H3K27me3 is a signal associated with inactive promoters and gene inactivation[200, 201].

Based on the evidence above, it is possible that the top 1,000 most associated sites from the EWAS reside in regions of open DNA, potentially active for transcription. Furthermore, because representation of these signals was similar across tissues, it is possible that regulatory effects of these CpG sites in blood, which was the discovery tissue, are similar in other tissues more relevant for T2D, like muscle and pancreas.

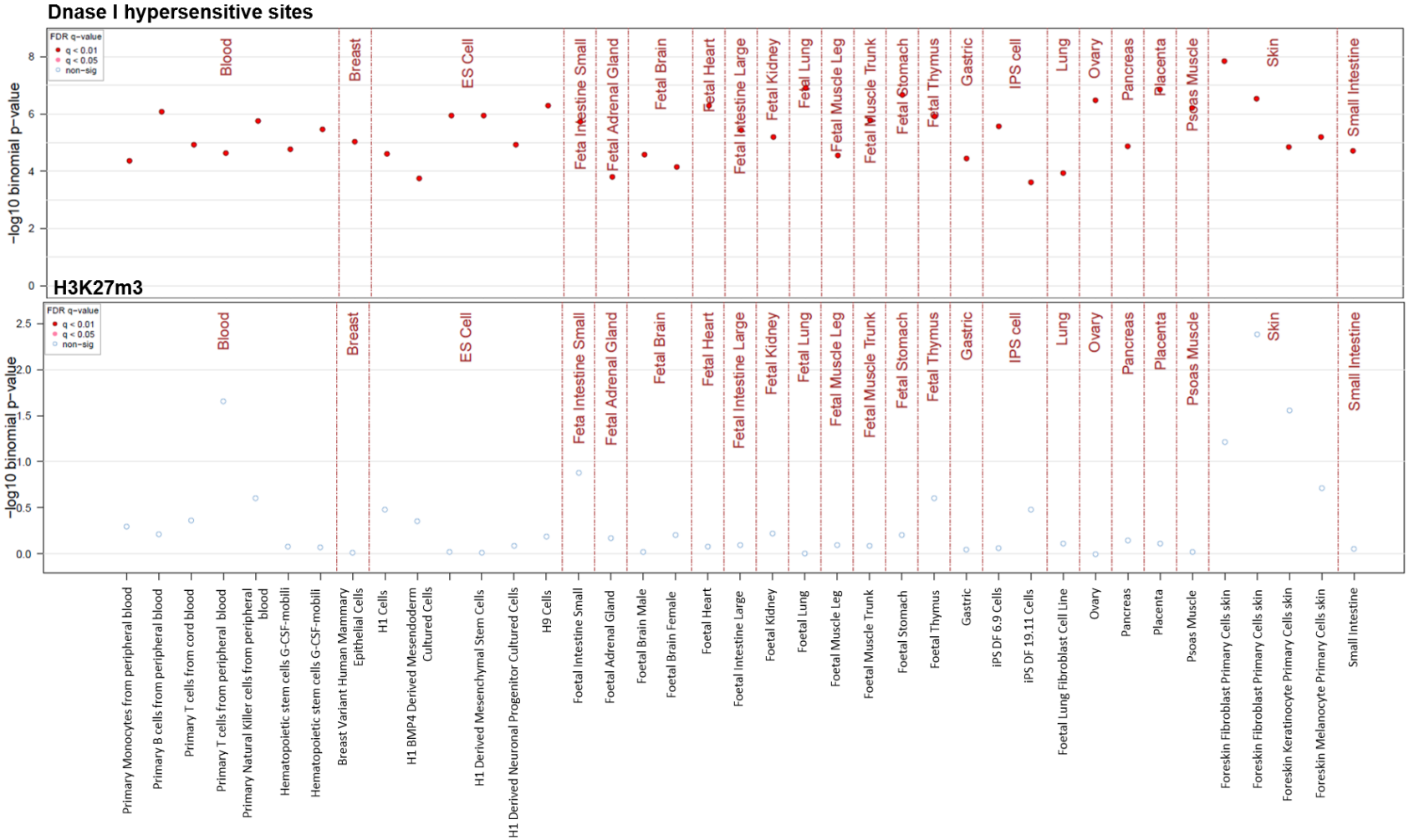


Figure 4-10 Tissue-specific overlap between DNaseI hypersensitive sites (top panel) and H3K27me3 (lower panel), and top 1000 CpG sites identified in the EWAS of T2D using the eFORGE platform. Data in eFORGE was extracted from the ROADMAP project. Red dots represent sites of overlap between CpG sites and the regulatory element under inspection (q < 0.01), while blue dots represent sites of non-overlap (q > 0.05) according to the specific tissue. Image provided by eFORGE [149].

Further annotation of top 1,000 CpG sites for functional elements was conducted using the Locus Overlap Analysis web tool (LOLA web version 0c5e2556f, release date: 2015, analysis date: 03-05-2018, http://lolaweb.databio.org). Based on this analysis, it was possible to determine that top signals of the EWAS were enriched in DNA binding sites for the transcription factor CBFB (core-binding factor subunit beta) in leukaemia cells, Pol2 (RNA polymerase II subunit A) in lymphoblastoid and B cell-derived cell lines, ELF1 (E74 like ETS transcription factor) in lung cancer and liver carcinoma cell-lines, TAF1 (TATA-box binding protein associated factor 1) in ovarian cancer cells, and SP1 (specificity protein 1) in lymphoblastoid cells (see Figure 4-11). In terms of histone marks, enrichment was found for H3K4me3 in prostate cancer and breast cancer cells, H3K9ac in breast cancer cell-lines, and for H3K9me3 in prostate cancer cells (Figure 4-11). Looking at the genomic distribution of top associations in the EWAS, these sites were most likely found within introns and exons, and to a less extent within intergenic regions and core promoters. Relative to their distance from the nearest transcription start site (TSS), most of the top CpG sites were found within 1kb to 1Mb upstream or downstream the nearest TSS.

Overall, the enrichment analysis for regulatory elements in LOLA suggested that strongest CpG sites with differential methylation in T2D, were more likely to overlap with regions targeted by different transcription factors, and these regions were possibly located far from the transcription start site of genes, within introns or exons rather than at the intergenic and core promoter regions. The observed enrichment for histone marks associated with transcriptional activation (H3K9ac) and elongation (H3K4me3) in CpG sites of interest in T2D, reinforced the idea that they could have been in transcriptionally active regions. However, most of the evidence provided by LOLA referred to carcinoma cells, and these regulatory signals could differ from normal peripheral blood cells.

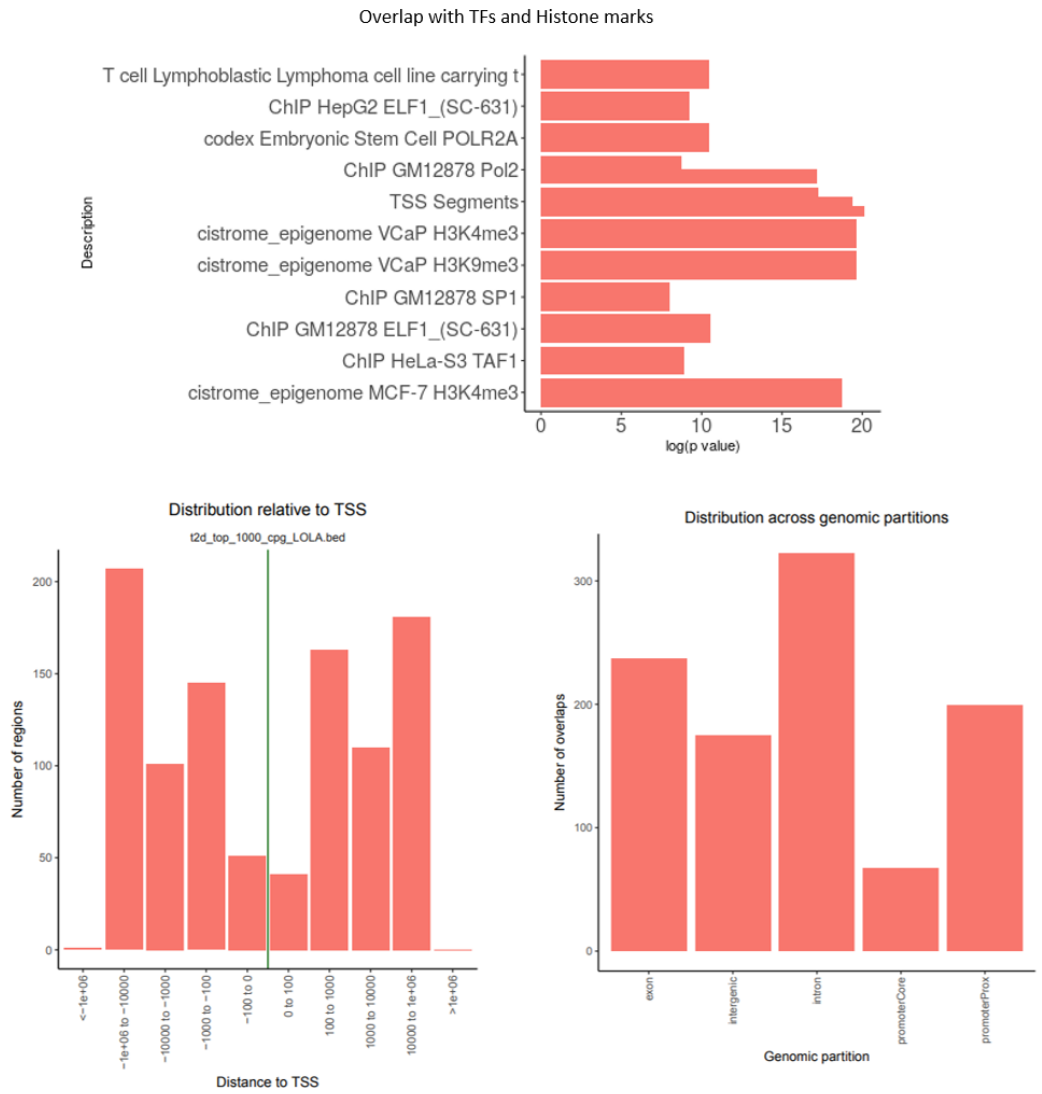


Figure 4-11 Regulatory elements and genomic regions overlapping with the position of top CpG sites of the EWAS according to a functional inspection conducted in LOLA. Top panel shows the p-value score given to each transcription factor and histone mark matching with the position of target CpG sites. The label to the left of the bar-plot describes the dataset of origin, cell-type, and type regulatory element found in enrichment. Bottom plots show the distribution of target CpG sites with respect to transcription start sites (bottom left), and various genomic annotations (bottom right); from left to right: exon, intergenic, intron, promoter core and proximal promoter. Image provided by LOLAweb (http://lolaweb.databio.org/)[150].

4.6.6 Gene-set enrichment analysis

As in section 4.6.5, the top 1000 CpG sites ranked by p-value from the fully-adjusted EWAS were extracted and a gene-set enrichment analysis was conducted using the DAVID Bioinformatics Resource (DAVID version 6.8, released date: 2003-2018, accessed date: 30-04-2018, https://david.ncifcrf.gov/) to identify enrichment for any particular biological processes. One of the benefits of using a long list of genes in an enrichment analysis, is to overcome power limitations to detect strongly-enriched terms[202]. In the original publication of this bioinformatic tool, the authors

mentioned that by using a longer list of sensibly selected genes, one can improve the statistical power of findings and enhance the sensitivity for specific terms[202]. In total, 760 unique genes mapping to the strongest CpG sites of the EWAS were used in the functional analysis in DAVID. Background genes were selected from the total number of genes present in the human genome. Significantly enriched terms were selected using a Bonferroni corrected p-value < 0.05.

Of the 321 terms reported by DAVID, only 16 surpassed Bonferroni correction. Significance of Bonferroni-corrected terms ranged between 7.03×10^{-10} and 0.045, and the fold enrichment for the strongest pathways ranged between 1.19 to 2.86; fold enrichment is a measure of the proportion of input genes found within a pathway, versus the proportion of the total number of genes in the human genome allocated to the same pathway. Most of the enriched terms identified were related to protein class, cellular location, protein-protein interaction and protein function (see Table 4-11). In detail, the strongest term was the "phosphoprotein" protein-class, followed by a term for "cytoplasm" location, "alternative splicing", "protein binding" activity, "disease mutation", "active-site: proton acceptor", "protein kinase" function, among others (see Table 4-11).

Even though DAVID provided an easy way to investigate biological processes enriched in genes of interest, there was a considerable amount of redundancy in the terms reported. In addition, among the top terms identified, there was predominance for general processes related with protein structure and function, rather than more specific terms related to pathways in T2D. Overall, gene enrichment analysis using top signals of the EWAS was not informative of processes related with T2D.

Table 4-11 Functional annotation of top terms enriched for genes related to the strongest CpG sites identified in the fully-adjusted EWAS of T2D in ALSPAC. Total number of unique input genes was 760. Count is the number of input genes included in a term. Ratio is the percent of input genes versus total number of background genes included in a pathway. Fold enrichment measures the magnitude of enrichment of input genes for certain pathway. Adjusted p-value is the Bonferroni corrected p-value for overrepresentation of input genes in a biological process

Term	Count	Ratio	Fold Enrichment[b]	Bonferroni
Phosphoprotein	367	0.34	1.33	$7.03x10^{-10}$
Cytoplasm [a]	235	0.22	1.45	$5.07x10^{-8}$
Alternative splicing	435	0.40	1.22	$1.85x10^{-7}$
GO: Cytoplasm [a]	251	0.23	1.34	$2.63x10^{-5}$
GO: Protein binding	380	0.35	1.19	$2.63x10^{-4}$
Disease mutation	130	0.12	1.52	$3.61x10^{-4}$
Protein kinase, ATP binding site [a]	33	0.03	2.42	$1.00x10^{-2}$
Pleckstrin homology domain	26	0.02	2.53	$1.00x10^{-2}$
Protein kinase-like domain [a]	41	0.04	2.15	$1.00x10^{-2}$
Active site: proton acceptor	47	0.04	2.05	$1.00x10^{-2}$
Protein kinase, catalytic domain [a]	38	0.03	2.18	$2.00x10^{-2}$
Pleckstrin homology domain [a]	26	0.02	2.68	$2.00x10^{-2}$
Domain: Pleckstrin homology domain [a]	24	0.02	2.86	$3.00x10^{-2}$
Domain: Protein kinase [a]	36	0.03	2.20	$4.00x10^{-2}$
Repressor	39	0.04	1.96	$4.00x10^{-2}$
Splice variant	318	0.29	1.20	$5.00x10^{-2}$

[a] Redundant term reported in DAVID. [b] Fold enrichment above 1.5 is considered interesting. However, if the term contains a small number of input genes, it is usual to have high fold enrichment values which are less reliable than values obtained from terms with high number of input genes.

4.6.7 Summary of EWAS results and functional analysis using top associated CpG sites

In the EWAS of T2D, strong evidence of association was identified at a DMP in *NFYC,* while borderline association was detected at a DMP in *STARD10*. A sensitivity analysis for BMI revealed that this was a confounder of the association between T2D and *NFYC*. DMPs in *NFYC* and *STARD10* were novel methylation markers in association with prevalent T2D. From the genetic side, GWA studies have reported an association between genetic variants in *STARD10* and T2D, but there are no studies reporting an association between genetic variants in *NFYC* and T2D. Results of the EWAS at *NFYC* and *STARD10* suggested that cases of T2D were on average hypomethylated compared to controls. Further inspection of the association at *NFYC* showed that this surpassed additional adjustment for fasting glucose, fasting insulin, HOMA-IR and C-reactive protein. An additional sensitivity analysis stratifying the sample by quartiles of methylation at *NFYC,* revealed that difference in mean methylation between cases and controls was only significant at the top quartile, where methylation at *NFYC* was the lowest. In addition, when comparing across quartiles, the absolute difference in methylation between the groups was significantly higher in the upper quartile compared to the lower quartiles, suggesting a positive and strong interaction between T2D and the quartile of methylation. In line with this evidence, it was shown that the risk of T2D increased in a

linear manner from the lower quartile to the top quartile of methylation, indicating that the risk of T2D was strongly associated with hypomethylation of the *NFYC* locus.

A functional analysis revealed that the CpG in *NFYC* was in a region of open DNA based on the presence of DNase I hypersensitive sites. This region was potentially active for transcription considering the increasing levels of the histone mark H3K27Ac downstream the position of the CpG and towards the promoter of *NFYC*. GTEx data did not support an enrichment of *NFYC* transcripts in tissues relevant to T2D, and *cis* meQTL for this DMP were not related to risk variants previously reported for T2D, or T2D related traits. Considering top signals of the EWAS, a functional analysis revealed that they were more likely to overlap with DNase I hypersensitive sites, binding-sites for different transcription factors, and with histone marks related with transcriptionally active regions (H3K4me3 and H3K9Ac). Top CpG sites of the EWAS were depleted from regions enriched in the histone mark H3K27me3, a signal associated with transcriptional repression. Enrichment for DNase I sites and H3K27me3 in top signals of the EWAS was similar across different tissues, suggesting that blood could be a good proxy to study the distribution of these regulatory elements in more relevant tissues for T2D.

4.7 EWAS at candidate loci for type 2 diabetes

Results of the EWAS in T2D were inspected for CpG sites located in the region of ten candidate loci for T2D that were extracted from an initial list of 56 risk loci used to generate a polygenic risk score for T2D in ALSPAC samples (see Chapter 3). According to results of the EWAS for the strongest CpG site in each candidate loci (Table 4-12), there was no evidence that CpG sites located within the region of GWAS loci for T2D, were associated with differential methylation in T2D (Bonferroni adjusted $p > 1.07 \times 10^{-7}$). This finding suggests that there is not necessarily an overlap in the signals identified between genetic and epigenetic studies of T2D, implicating different mechanisms of action between these two processes in their association with T2D. Even though some of the candidate loci inspected have been associated with differential methylation in T2D in other studies (i.e. *TCF7L2, SLC30A8, FTO, HNF1B, JAZF1, KCNQ1* and *THADA*), null associations detected in this study may be due to power constraints because of the small sample size used, or due to differences in the way that differential methylation was assessed between studies, or to a lack of replication of signals previously detected in these candidate loci.

Table 4-12 Results of the EWAS in T2D for the strongest CpG sites located within ten candidate loci for T2D. Estimates for each CpG site correspond to the fully-adjusted EWAS conducted in ALSPAC (cases=48, controls=1,002). Probes: total number of CpG sites included in the region of the candidate loci; Top DMP: CpG site with the smallest p-value of association with T2D; SE: standard error; Bonferroni: adjusted p-value from the EWAS. N: number of samples included in the analysis for each CpG site.

Locus	Chr	Probes	Top DMP	Beta	SE	P-value	Bonferroni	N
TCF7L2	10	81	cg07591090	0.011	0.004	0.011	1.00	1035
CDKAL1[a]	6	56	cg22626973	0.012	0.005	0.008	1.00	1050
IGF2BP2[a]	3	36	cg09746170	-0.001	0.001	0.010	1.00	1026
SLC30A8	8	7	cg26687497	-0.017	0.005	0.002	1.00	1039
FTO	16	29	cg01485549	-0.011	0.004	0.011	1.00	1048
PPARG[a]	3	23	cg06573644	-0.003	0.001	0.002	1.00	1037
JAZF1	7	56	cg14491535	-0.029	0.012	0.015	1.00	1050
HNF1B	17	26	cg05110178	-0.008	0.003	0.008	1.00	1043
KCNQ1	11	288	cg06960356	-0.005	0.001	0.002	1.00	994
THADA	2	48	cg20341942	-0.001	0.000	0.005	1.00	1049

[a] Genetic loci without evidence of differential methylation in T2D according to Toperoff *et al*.[60] and Dayeh *et al*.[79]

Representation of the genomic distribution of CpG sites against the effect size and -log10(p-value) obtained in the EWAS is shown in Figure 4-12 for *TCF7L2* (Chr10, n=81 probes) and *FTO* (Chr16, n=29). These two loci represent the strongest genetic evidence in association with T2D according to GWA studies[30], and some of the CpG sites within the region of *TCF7L2* and *FTO* have been reported with differential methylation in T2D according to Toperoff *et al*.[60] and Dayeh *et al.*[79]. Similar plots for the remaining candidate loci listed in Table 4-12 can be found in the appendix Figure S8-4.

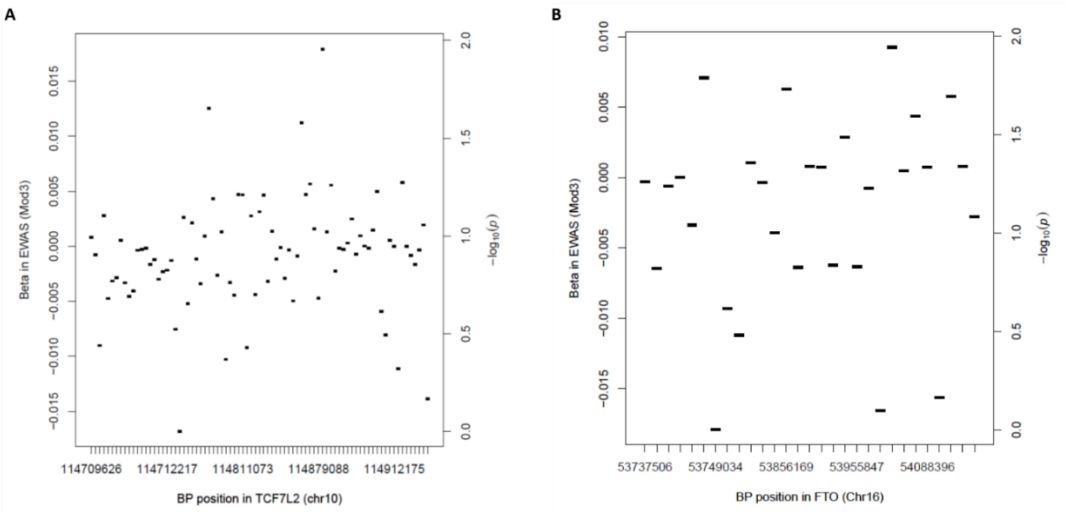


Figure 4-12 Estimates of the EWAS for CpG sites found in the region of TCF7L2 and FTO. Effects size and -log10(p-value), were taken from results of the most adjusted EWAS conducted in ALSPAC (n=1,050). In the x-axis, genomic coordinates of CpG sites within candidate genes, and in the y-axis, beta coefficient (left-hand side) and -log10 (p-value) (right-hand side). None of the DMPs within TCF7L2 or FTO reached EWAS significance at a p< $1.07x10^{-7}$ or -log10(p-value) ≥ 7.0.

4.8 Analysis of differentially methylated regions in type 2 diabetes

As mentioned in Chapter 2, an analysis of differentially methylated regions (DMRs) has the advantage of providing a broader perspective of the methylation context associated with a phenotype of interest when compared to the single CpG analysis[42]. In addition, a DMR analysis can help to reinforce evidence from the single-site analysis, since it is expected that patterns of methylation observed in a single CpG site, can also be detected in nearby CpG sites[44]. It is also known that regional changes in methylation are more likely to influence chromosome conformation and gene expression in comparison to changes in methylation at single CpG sites[42].

4.8.1 DMRs associated with T2D using comb-p

A DMR analysis was conducted in *comb-p* with methods described elsewhere[148]. This analysis identified 13 DMRs strongly associated with T2D at Sidak<0.05, mapping to 12 different genes and including 53 DMPs (Figure 4-13, appendix Table S8-5). From these DMRs, 12 were informative regions based on a CpG count equal or above two CpG sites. None of these informative DMRs overlapped with the genomic position of top associated DMPs detected in the fully adjusted EWAS. At the top-ranking DMRs (smallest Sidak), genomic size ranged between 23bp and 267bp, DMP count ranged between three and nine DMPs, and average absolute difference in methylation between T2D cases and controls ranged between 0.16% and 6.0%. Most of the DMRs were hypomethylated in T2D cases versus controls, except for the DMR mapping to the *NCRNA00028* gene, which was identified as the strongest DMR with Sidak p-value=1.25×10^{-6} (Table 4-13). Further detail of DMPs mapping within top DMRs, estimates reported in the fully adjusted EWAS for these sites, and estimates from the DMR analysis, are presented in the appendix Table S8-5.

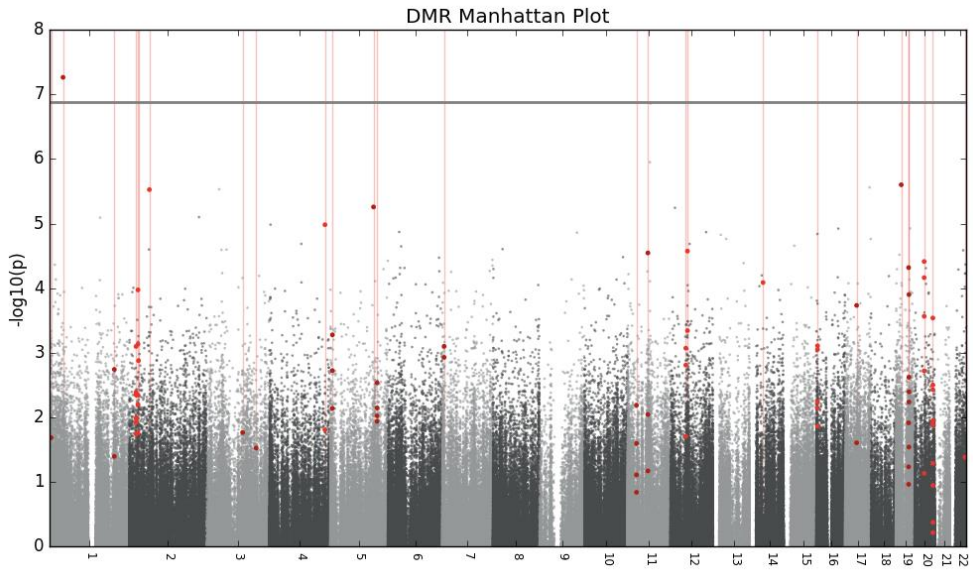


Figure 4-13 Manhattan plot showing the genomic location of 27 DMRs (red vertical lines) initially identified by comb-p, but only 12 of them surpassed Sidak significance at p<0.05. Red dots represent DMPs within the DMRs. X-axis shows the chromosome position of the DMR, and y-axis is the -log10(p-value) for the CpG site as reported in the fully adjusted EWAS.

Of interest in this analysis was the detection of a DMR mapping 200bp upstream the transcription start site of *APOB*, and partially overlapping with a CpG island located in the promoter of this gene (UCSC Genome Browser track view, analysis date: 10-05-2018, http://genome.ucsc.edu/). According to the literature, apolipoprotein-B is the main regulatory protein that constitute the surface of chylomicrons, LDL and VLDL particles[203, 204]. *APOB* expression is confined to the small intestine and the liver (GTEx Portal V7, https://www.gtexportal.org/home/gene/APOB), and has a molecular function in the packaging of triglycerides and cholesterol into VLDL particles, and in the cellular signalling and catabolism of LDL particles[204]. Elevated concentration of *APOB* in plasma has been associated with higher risk of coronary arterial disease[204]. In GWA studies, genetic variants in *APOB* have been reported in association with levels of total cholesterol, LDL cholesterol, and age-related disease endophenotypes (GWAS catalog, https://www.ebi.ac.uk/gwas/). The relationship between *APOB* and T2D relies on the fact that plasma levels of apolipoproteins, in general, tend to be irregular in patients with metabolic disorders like T2D[204]. Abnormal levels of apolipoproteins can affect the metabolism of lipids, and a common condition observed in patients with T2D is atherogenic dyslipidaemia[205], a disorder where levels of HDL in blood are significantly lower relative to levels of LDL and triglycerides [205]. In addition to this, some studies have reported that serum levels of *APOB* were able to predict future risk of hypertension and T2D in women from a prospective observational study [206].

4.8.2 Functional Exploration of DMRs associated with T2D

Various functional analyses were conducted to determine how difference in methylation within regions might be related with T2D. As with the single CpG site analysis, it was first determined the enrichment of DMRs for epigenetic regulatory elements, and their genomic annotation. Secondly, it was investigated tissue-specificity in the expression of genes mapping to DMRs, and difference in the expression of these genes between T2D cases and controls using publicly available datasets. Furthermore, an meQTL search was conducted for the strongest CpG sites within each top DMR, and it was investigated if meQTL have also been identified as eQTL. Lastly, a cross-tissue comparison in the levels of methylation was conducted using CpG sites within top DMRs to establish the relevance of blood as a source of methylation markers in T2D. Potential pathways associated with CpG sites within DMRs were also interrogated using a gene-set enrichment analysis.

4.8.2.1 Enrichment for regulatory elements among top DMRs associated with T2D

Using eFORGE and a cut-off Q-value < 0.05 for significant enrichment, there was no suggestion that DMPs identified within DMRs overlapped with regions enriched in DNase I hypersensitive sites, or peaks of the histone mark H3K27me3 at the specific tissues analysed. Considering other histone marks, some overlap was identified between DMPs within DMRs and peaks of H3K27me3 in embryonic stem cells (Q-value= 0.53), and peaks of H3K4me1 in psoas muscle (Q-value= 0.53). The histone marks H3K27me3 and H3K4me1 are associated with inactive promoters[207] and with enhancers[200], respectively. Instead of using DMPs to identify enrichment for regulatory elements, a second approach based on genomic regions was implemented using LOLAweb. For this analysis, regions were separated between those with average hypomethylation in T2D (n=11 DMRs), from one region with average hypermethylation in T2D. Background regions were defined based on 27 DMRs initially identified by *comb-p* with and without Sidak significance.

Table 4-13 Summary of 12 DMRs identified in strong association with T2D using comb-p, and based on results of the fully-adjusted EWAS conducted in the subsample of ALSPAC/ARIES.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Index DMP	P_{region}	Sidak
2	27,485,967-27,486,134	*SLC30A3*	167	5	1.96	↓	cg23151303	$1.01x10^{-8}$	$2.32x10^{-5}$
2	21,266,947-21,267,114	*APOB*	167	6	3.26	↓	cg03350299	$1.19x10^{-7}$	$2.72x10^{-4}$
5	146,832,182-146,832,357	*DPYSL3*	175	2	1.65	↓	cg18635723	$3.29x10^{-6}$	$7.17x10^{-3}$
5	6,447,235-6,447,258	*UBE2QL1*	23	3	1.61	↓	cg12035880	$3.01x10^{-6}$	$4.89x10^{-2}$
7	5,609,731-5,609,898	*Unannotated*	167	2	0.61	↓	cg05281338	$8.71x10^{-6}$	$1.98x10^{-2}$
11	63,974,772-63,974,956	*FERMT3*	184	3	0.16	↓	cg01447914	$7.65x10^{-6}$	$1.58x10^{-2}$
12	53,591,756-53,591,767	*ITGB7*	11	2	3.37	↓	cg04972065	$3.70x10^{-7}$	$1.28x10^{-2}$
12	48,298,924-48,298,993	*VDR*	69	3	0.32	↓	cg13865595	$7.26x10^{-6}$	$3.95x10^{-2}$
16	3,507,460-3,507,583	*NAT15*	123	5	2.51	↓	cg00484396	$6.95x10^{-8}$	$2.17x10^{-4}$
19	41,256,647-41,256,914	*C19orf54*	267	5	0.39	↓	cg26015947	$3.51x10^{-8}$	$5.04x10^{-5}$
20	30,073,399-30,073,577	*NCRNA00028*	178	5	6.00	↑	cg02991085	$5.82x10^{-10}$	$1.25x10^{-6}$
20	57,427,274-57,427,504	*GNAS*	230	9	1.39	↓	cg06065549	$1.23x10^{-5}$	$2.02x10^{-2}$

CpG count: number of DMPs detected within the DMR. %Meth: Percent of the average absolute difference in methylation between T2D cases and controls for CpG sites in the region. Direction: regional effect-direction in the association between T2D and DNA methylation. Index DMP: CpG site in a region with the strongest evidence of association with T2D as reported in the EWAS. Pregion: p-value of the region calculated using the Stouffer-Liptak correction. Sidak: level of significance to establish regions of interest. DMRs were selected as significant based on Sidak < 0.05.

With respect to DMRs hypomethylated, most of them overlapped with introns and exons (66.6%), and to a less extent, with intergenic regions and core promoters (16.6%) (see Figure 4-14). Regarding position from the nearest transcription start site (TSS), a great proportion of the DMRs were located downstream the nearest TSS (66.6%) versus those located upstream the TSS (33.3%), and in both cases it was more likely to find top DMRs distant from the position of the TSS (10kb-1Mb, 20.8% DMRs) rather than near the TSS (100bp-1kb, 8.3% DMRs) (Figure 4-14). The only DMR hypermethylated in T2D (Chr20, *NCRNA00028* gene) was in an intergenic region, between 10kb to 1Mb upstream the nearest TSS. Three transcription factors had binding sites within this DMR, and they were the zinc finger *ZNF143*, *P300* and *TAF1*, which were signals reported by ENCODE in relation to leukaemia cell lines (for *ZNF143*) and neuroblastoma cell lines (for *P300* and *TAF1*). There was also evidence that the DMR in *NCRNA00028* was in a region of repressed chromatin based on the enrichment found for the histone mark H327Kme3 in leukaemia cell lines (LOLA-Cistrome database, https://doi.org/10.1186/gb-2011-12-8-r83); H327Kme3 is a signal related with transcriptional silencing.

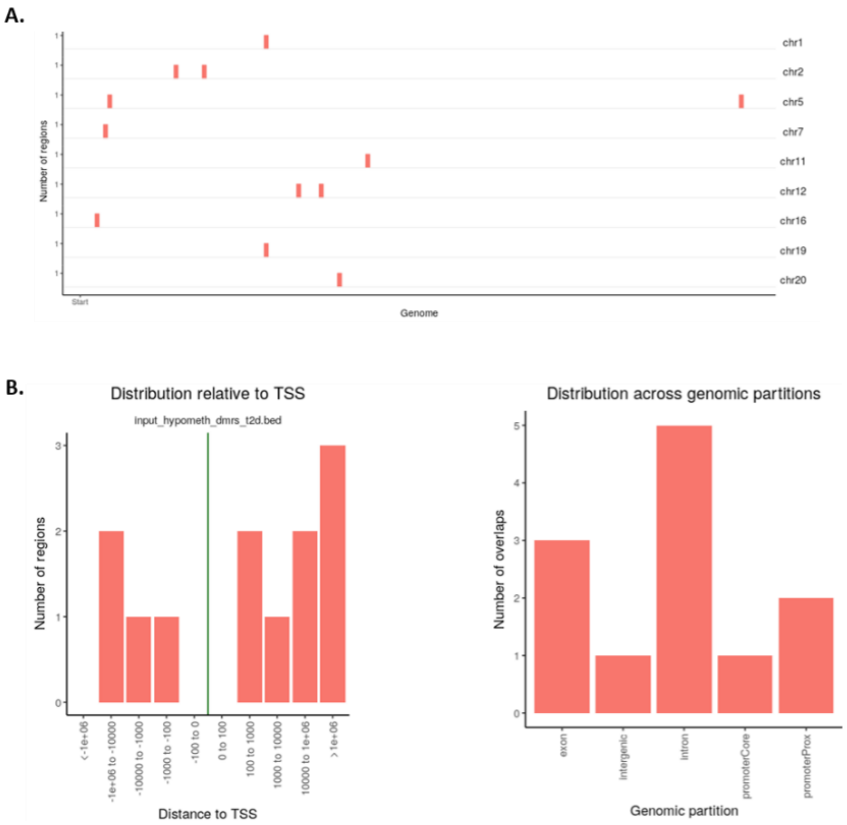


Figure 4-14 Enrichment analysis for regulatory elements and genomic context of top DMRs associated with T2D as reported by LOLA. DMRs analysed were those associated with hypomethylation in T2D. A) Distribution of DMRs by chromosome. B) Distribution of DMRs relative to TSS (left), and genomic context (right). From left to right, genomic context refers to exons, intergenic regions, introns, core promoters (30bp upstream the TSS) and proximal promoters (100-200bp upstream the TSS). Image provided by LOLAweb (http://lolaweb.databio.org/)[150].

Using genomic coordinates for top DMRs associated with T2D to identify enrichment for histone marks and transcription factor binding sites: analysis in Epi-explorer

Using lymphoblastoid cell-lines as surrogate cells for normal peripheral blood cells, the analysis in Epi-explorer suggested that hypomethylated DMRs in T2D were enriched for histone marks associated with enhancers (H3K4me1, 83% DMRs), compared to the proportion of them associated with repressed chromatin (H3K27me3, 58.3% DMRs). In line with this evidence, results of the chromatin state revealed that hypomethylated DMRs were enriched in regions of active promoters (41.7% DMRs), but depleted from regions of heterochromatin (8.3% DMRs) and Polycomb repressed signals (16.7% DMRs) (Table 4-14).

Table 4-14 Overlap between DMRs hypomethylated in T2D and histone marks and chromatin states based on the analysis in Epi-explorer. Annotation for histone marks and chromatin state segmentation was retrieved from ENCODE based on signals detected in lymphoblastoid cell-lines. The level of stringency selected was an overlap of at least 50% between the annotated region and the query region. A DMR could have overlapped simultaneously with different histone marks and chromatin states. This table shows the percentage of overlap with each mark, and the corresponding number of DMRs in brackets.

Histone Mark	Role in Transcription [a]	% DMRs	Chromatin states	% DMRs
H3K4me1	+	83.3 (10)	Active promoters	41.7 (5)
H3K4me2	+	75.0 (9)		
H4K20me1	-	75.0 (9)	Poised promoters	25 (3)
H2AZ	+	66.7 (8)		
H3K4me3	+	66.7 (8)		
H3K9ac	+	58.3 (7)		
H3K27me3	-	58.3 (7)	Polycomb repressed [b]	16.7 (2)
H3K9me3	-	41.7 (5)	Heterochromatin	8.3 (1)
H3K27ac	+	41.7 (5)	Transcriptional transition	8.3 (1)
H3K79me2	+	41.7 (5)		
H3K36me3	+	16.7 (2)		

[a] Histone marks with positive effects in transcription (+), and those with a repressing role in transcription (-). [b] Transcription factor associated with repressed promoters.

Based on ENCODE data, it was identified that hypomethylated regions in T2D were enriched in binding sites for the transcription factors PAX5-C20 (33.3%), Pol2 (25%), BCLAF1 (25%), CTCF (16.7%), SP1 (16.7%), TAF1 (16.7%), ZBTB33 (16.7%), and some others detected in lymphoblastoid cell-lines (Table 4-15). A plot summarizing the distribution of DMRs hypomethylated in T2D across different regulatory regions is presented in Figure 4-15. For the single DMR hypermethylated in T2D mapping to the *NCRNA00028* gene, Epi-explorer did not provide further annotation for regulatory elements relative to data reported by LOLA.

Table 4-15 Percentage of overlap between hypomethylated DMRs in T2D and DNA binding sites for transcription factors reported in lymphoblastoid cell lines using data from ENCODE. Genome coverage is the proportion of the genome covered by the transcription factor; overlap with DMRs is the percentage of DMRs overlapping with DNA binding sites for suggested TFs.

Transcription Factor binding sites	Genome coverage	Overlap with DMRs
PAX5-C20	0.6%	33.3%
BCLAF1	0.7%	25.0%
Pol2	0.9%	25.0%
CTCF	5.6%	16.7%
Sp1	0.5%	16.7%
TAF1	0.1%	16.7%
ZBTB33	0.2%	16.7%
ETS1	0.1%	8.3%
BCL3	0.4%	8.3%
TCF12	0.4%	8.3%
POU2F2	0.2%	8.3%

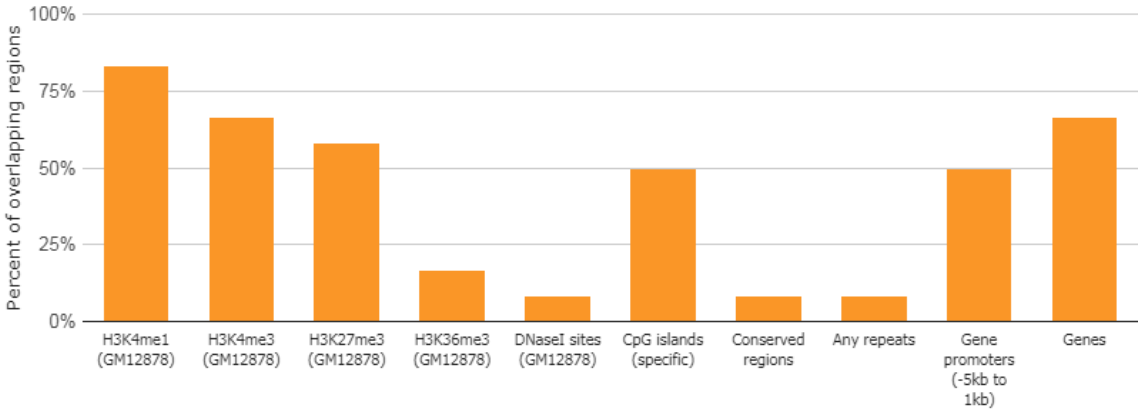


Figure 4-15 Distribution of DMRs hypomethylated in T2D across different epigenetic regulatory elements and genetic regions. Plot provided by Epi-explorer using data from lymphoblastoid cell lines (GM12878) [151].

4.8.2.2 Identifying eQTM for top DMPs within T2D-associated DMRs

A search for eQTM associated with the top DMP within each DMR was conducted to determine if methylation at these sites influenced gene expression of the gene in *cis*. This lookup was conducted using the Bios QTL browser as mentioned before. None of the 12 index CpG sites within T2D-associated DMRs was identified as an eQTM according to data from Bios QTL. Thus, further description of these sites within top DMRs is based on the nearest gene.

4.8.2.3 Differential expression of the nearest gene annotated to T2D-associated DMRs

A multi-gene query was submitted to the GTEx Portal to compare levels of gene expression across different tissues for genes annotated to DMRs in T2D. This search revealed that *GNAS* (GNAS complex locus) was the gene with the highest level of expression across tissues (transcripts per million > 160), and it is known that a DMR in the 5' exon of transcripts of this imprinted gene,

correlates with the transcript expression (GeneCards v4.7.1 Build 10, accessed date: 17-05-2018, https://www.genecards.org). Another gene with detectable levels of expression (TPM > 15) across several tissues was *DPYSL3*, except for the liver, whole blood and pancreas (see Figure 4-16). Considering peripheral blood alone, high expression was observed for the gene *FERMT3* (TPM=279.7), which is a protein-coding gene important for the migration, adhesion and differentiation of hematopoietic cells (GeneCards v4.7.1 Build 10, accessed date: 17-05-2018, https://www.genecards.org). In terms of other tissues more relevant to T2D, high expression was detected for *APOB* in the liver (TPM=348.4). The remaining genes annotated to DMRs had relatively low levels of expression in relevant tissues (TPM < 15).

In an additional analysis, it was investigated difference in gene expression between T2D cases and controls in reference to genes mapping to DMRs associated with T2D. Gene expression data was obtained from two gene expression datasets available in the omnibus repository (last modified: 07-16-2016, date accessed: 16-05-2018, https://www.ncbi.nlm.nih.gov/geo/). Further detail of the method implemented can be found in Chapter 2. This analysis revealed that none of the genes annotated to DMRs in T2D had significantly altered expression after adjustment for multiple testing (q > 0.05). In the first dataset used, the gene with the highest fold-change in expression was identified at *DPYSL3* (Beta=-4.49, Log_2Fold-change=-0.15, adjusted-p=0.17), while the gene with the lowest fold-change in expression was *GNAS* (Beta=-5.23, Log_2Fold-change=-0.01, adjusted-p=0.99). In the second dataset analysed, the gene with the highest fold-change in expression was *GNAS* (Beta=-3.40, Log_2Fold-change=-0.29, adjusted-p=0.81), and the gene with the lowest fold-change in expression was *UBE2QL1* (Beta=-5.33, Log_2Fold-change=0.001, adjusted-p=0.99).

In summary, evidence of gene expression across tissues revealed that some of the genes mapping to DMRs had higher expression in other tissues rather than in peripheral blood. For instance, the expression of *APOB* was higher in the liver and in the small intestine compared to peripheral blood. Higher expression in blood was only detected for *FERMT3*, which function relates to the mobility of hematopoietic cells. Regarding differential expression of genes annotated to DMRs in peripheral blood of T2D donors, the association analysis revealed that none of these genes had levels of expression significantly altered between diseased and non-diseased participants. Therefore, differential methylation did not overlap with changes in expression for these genes.

One possible reason for the non-overlap between methylation and gene expression in genes annotated to the DMRs, is related to the position of the DMR with respect to the gene. According to

the literature, a higher impact of methylation on gene expression is expected when the DMR or the DMP is located in the promoter region of the gene[208], or near the TSS[209], but genetic annotation for the top associated DMRs revealed that only one out of 13 DMRs was within the region of a core promoter, while the majority of them were within introns (46.6%). Another possible explanation for the lack of correlation between methylation and gene expression, is the potential influence of the DMR on gene expression of a more distant gene. Distant influence of methylation on gene expression is likely to happen when the DMR is within the region of an insulator, a repressor or an enhancer[208], which constitute distant cis-regulatory elements for a gene. One way to investigate if a DMR could influence gene expression of a more distant gene, is to search for meQTL for the strongest CpG within the DMR (see Table 4-17), and to determine the overlap between meQTL and eQTL, and the gene targeted by the eQTL.

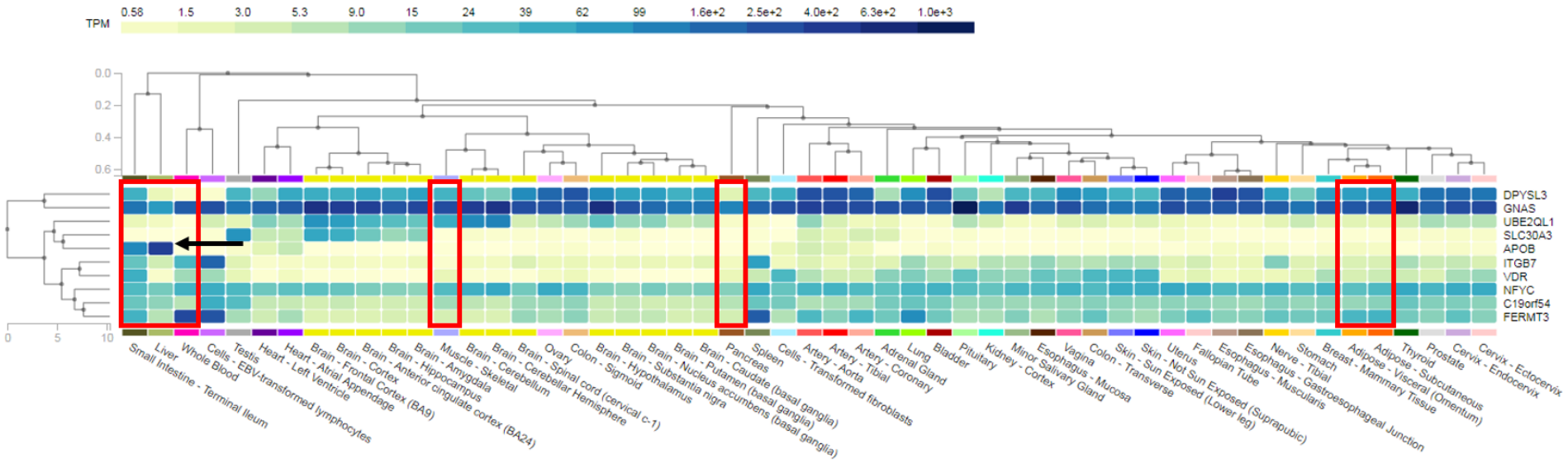


Figure 4-16 Clustered heatmap showing difference in gene expression across tissues for genes annotated to DMRs associated with T2D. On the right-hand side, the list of genes annotated to DMRs, and at the bottom, a list of tissues for which gene expression was available in the Genotype-Tissue Expression (GTEx) Project. Legend at the top shows the magnitude of gene expression detected for each gene across tissues, reported in transcripts per million (TPM); darker colours represent higher levels of expression compared to lighter colours. Highlighted in red are tissues of relevance for T2D. From left to right they are small intestine, liver, whole blood, skeletal muscle, pancreas, omentum and sub-cutaneous fat. Black arrow demarks two relevant tissues where there was high expression of APOB. The cluster analysis for tissues at the top, and for genes on the left-hand side, represents the level of similarity in the profile of gene expression among tissues and genes, respectively. Genes and tissues connected by a same node are more similar to each other, than genes or tissues connected via a distant node. This heatmap was obtained from the GTEx Portal on 17-05-2018, https://www.gtexportal.org/home/multiGeneQueryPage).

4.8.2.4 Identifying overlap between meQTL and eQTL signals associated with index CpG site within T2D-associated DMRs

meQTL for the strongest CpG site within each DMR were retrieved from the meQTL database, considering associations significant at a $p<10^{-7}$. meQTL were available for most of the CpG sites of interest, except for sites within DMRs in *SLC30A3*, *VDR*, *C19orf54* and *GNAS*. Most of the meQTL detected were *cis*-meQTL, and some *trans*-meQTL were identified in association with CpG sites in *ITGB7* and *FERMT3*. Overlap between meQTL SNPs and eQTL was interrogated using the latest dataset from GTEx (GTEx_Analysis_v7) at specific tissues (see Chapter 2).

In total, 19 eQTL were detected in overlap with meQTL in seven tissues: one eQTL was detected in liver, two in blood, other two in omentum, three in pancreas, four in muscle, five in subcutaneous fat, and other five eQTL in thyroid tissue (Table 4-16). The CpG with the highest number of meQTL overlapping eQTL was cg00484396, identified within the DMR in the *NAT15* gene. The overlapping eQTL for cg00484396 were identified across tissues, and they were associated with gene expression of the DMR-gene *NAT15*, and other nearby genes: *AC006111.1* (RNA gene), *ZNF597,* and *DNASE1* (Table 4-16). Another CpG with several meQTL overlapping with eQTL across tissues was the CpG cg02991085, identified within the DMR in *NCRNA00028*. Overlapping eQTL for cg02991085 were associated with gene expression of the DMR-gene *NCRNA00028* (non-coding RNA), and the nearby gene *DEFB124* (Table 4-16). Further evidence of CpG sites within DMRs potentially influencing gene expression of more distant genes at specific tissues, is presented in Table 4-16.

Table 4-16 Summary of eQTL overlapping with meQTL identified for some of the top CpG sites found within DMRs. meQTL data was retrieved from peripheral blood samples, and eQTL were looked up at specific tissues. meQTL were regarded significant at $p<1.0x10^{-7}$, while eQTL were significant at Q-value < 0.05.

Chr	DMR	Gene	CpG	Chr:Pos	eQTL	Chr:Pos	Gene	meQTL-P	eQTL-Q	Tissue
1	41,156,730-41,156,731	*NFYC*	cg15986668	1:41,156,730	rs2744808	1:41,177,943	*RP4-739H11.4*	$3.19x10^{-9}$	$2.38x10^{-3}$	Muscle
					rs1327887	1:41,243,428	*NFYC*	$2.94x10^{-8}$	$7.37x10^{-3}$	Muscle
					rs2744798	1:41,169,120	*NFYC*	$2.00x10^{-9}$	$2.66x10^{-4}$	Subcut. Fat
2	21,266,947-21,267,114	*APOB*	cg03350299	2:21,266,960	rs4665178	2:21,321,721	*APOB*	$5.93x10^{-22}$	$1.06x10^{-13}$	Subcut.Fat
16	3,507,460-3,507,583	*NAT15*	cg00484396	16:3,507,460	rs757270	16:3,534,451	*NAT15*	$1.08x10^{-54}$	$3.65x10^{-4}$	Blood
					rs4010630	16:3,538,990	*NAT15*	$1.02x10^{-28}$	$6.34x10^{-7}$	Pancreas
					rs2379830	16:3,538,918	*NAT15*	$4.86x10^{-54}$	$1.47x10^{-24}$	Muscle
					rs2379828	16:3,538,665	*NAT15*	$4.86x10^{-54}$	$1.09x10^{-3}$	Liver
					rs9926609	16:3,540,962	*NAT15*	$1.44x10^{-51}$	$1.44x10^{-51}$	Omentum
									$2.83x10^{-11}$	Subcut.Fat
					rs11077345	16:3,537,240	*NAT15*	$3.63x10^{-53}$	$1.44x10^{-19}$	Thyroid
					rs1639150	16:3,747,204	*AC006111.1*	$4.86x10^{-54}$	$2.02x10^{-2}$	Pancreas
					rs250528	16:3,478,834	*AC006111.1*	$2.91x10^{-48}$	$7.69x10^{-3}$	Subcut.Fat
					rs12925683	16:3,723,046	*AC006111.1*	$2.45x10^{-9}$	$1.86x10^{-2}$	Thyroid
					rs28603	16: 3,449,479	*ZNF597*	$1.11x10^{-11}$	$1.76x10^{-2}$	Pancreas
					rs37827	16:3,488,357	*ZNF597*	$1.47x10^{-48}$	$3.64x10^{-8}$	Thyroid
					rs11861770	16:3,664,009	*DNASE1*	$3.42x10^{-8}$	$2.04x10^{-2}$	Thyroid
20	30,073,399-30,073,577	*NCRNA00028*	cg02991085	20:30,073,537	rs6058172	20:30,187,310	*NCRNA00028*	$6.91x10^{-8}$	$2.85x10^{-2}$	Blood
					rs717064	20:30,066,356	*DEFB124*	$2.70x10^{-11}$	$6.33x10^{-21}$	Muscle
									$8.46x10^{-3}$	Omentum
					rs1543438	20:30,061,053	*DEFB124*	$4.72x10^{-8}$	$1.68x10^{-9}$	Subcut.Fat
									$3.71x10^{-2}$	Thyroid

CpG: strongest CpG detected within each of the significant DMRs identified by *comb-p*. Chr:Pos : genomic coordinates corresponding to the CpG (index CpG in the DMR) and the eQTL SNP. Gene: gene with altered expression. meQTL-P: p-value associated with the meQTL, as reported in the meQTL database. eQTL-Q: Q-value of significance reported by GTEx for the eQTL. Tissue: tissue where difference in gene expression was reported by GTEx.

4.8.2.5 *Cross-tissue comparison in the levels of methylation using index CpG sites within T2D-associated DMRs*

To determine the relevance of blood as a source of methylation markers in T2D, the level of correlation between mean methylation in peripheral blood and mean methylation in five different tissues relevant to T2D was investigated using for this the most significant DMP detected within each DMR. As DNA methylation was not available for internal tissues in samples in ALSPAC at the time this study was conducted, comparison of methylation levels across tissues was performed using a publicly available dataset downloaded from the Gene Expression Omnibus repository[210] (NCBI GEO database, last modified: 07-16-2016, date accessed: 16-05-2018, https://www.ncbi.nlm.nih.gov/geo/). Further detail on the method used for cross-tissue comparison is described in Chapter 2. Internal tissues compared against blood were liver, muscle, omentum (visceral peritoneum), pancreas and sub-cutaneous fat. DMPs selected for the cross-tissue comparison are listed in Table 4-17.

Table 4-17 List of top DMP within each DMR included in the cross-tissue comparison.

Chr	DMR	Gene	DMP
2	27,485,967-27,486,134	*SLC30A3*	cg23151303
2	21,266,947-21,267,114	*APOB*	cg03350299
5	146,832,182-146,832,357	*DPYSL3*	cg18635723
5	6,447,235-6,447,258	*UBE2QL1*	cg12035880
7	5,609,731-5,609,898	*Unannotated*	cg05281338
11	63,974,772-63,974,956	*FERMT3*	cg01447914
12	53,591,756-53,591,767	*ITGB7*	cg04972065
12	48,298,924-48,298,993	*VDR*	cg13865595
16	3,507,460-3,507,583	*NAT15*	cg00484396
19	41,256,647-41,256,914	*C19orf54*	cg26015947
20	30,073,399-30,073,577	*NCRNA00028*	cg02991085
20	57,427,274-57,427,504	*GNAS*	cg06065549

Overall, correlation was high and strong between methylation in blood and methylation in omentum ($r=0.91$, $P=1.8\times10^{-5}$), sub-cutaneous fat ($r=0.83$, $p=3.9\times10^{-4}$), liver and muscle ($r=0.77$, $p=2.1\times10^{-3}$) (Figure 4-17). Strong but lower correlation was identified between blood and pancreas ($r=0.59$, $p=0.04$). Looking at similarity in methylation across tissues, it was evident that sub-cutaneous fat was the tissue with the highest similarity in methylation with other tissues (level of correlation ranged between 0.83 and 0.98), whilst blood was the tissue with less similarity in methylation with other tissues (level of correlation ranged between 0.59 and 0.91).

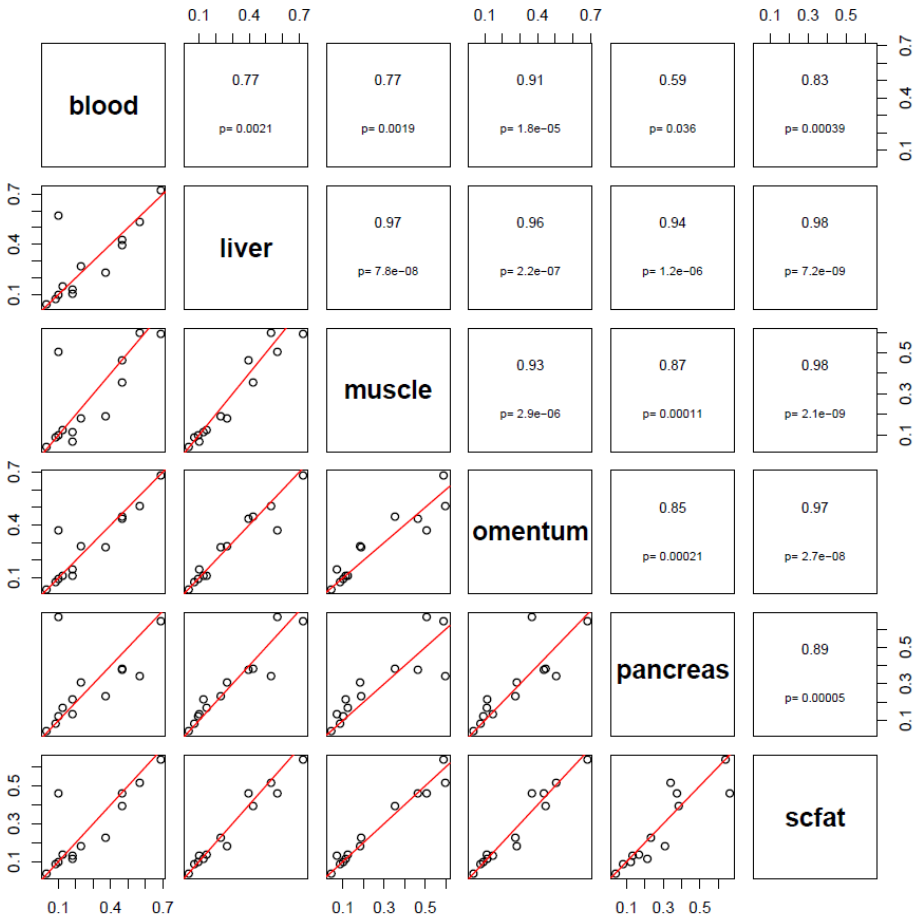


Figure 4-17 Correlation in the levels of methylation at 12 DMPs across six different tissues relevant for T2D. DMPs included in the cross-tissue comparison were the strongest DMPs identified within 12 DMRs detected in T2D. DNA methylation data from peripheral and internal tissues was extracted from the GEO dataset GSE48472 based on a study conducted by Slietker et al. 2013 [159].Scfat: subcutaneous fat.

4.8.2.6 Identifying biological pathways enriched for genes mapping to T2D-associated DMRs

The database of GO terms and KEGG pathways was interrogated to determine if genes annotated to T2D-associated DMRs were enriched for biological pathways related with T2D. In contrast to the gene-enrichment analysis used for top associated DMPs in the EWAS (i.e. DAVID), in the DMR analysis gene-enrichment was implemented using the R package missMethyl to correct for the differential representation of CpG sites within genes included in the methylation array. Further detail of the method is described in Chapter 2. In general, none of the terms reported by GO survived FDR correction for pathway enrichment. Top three terms identified by GO were related with the "regulation of alcohol biosynthetic process", "regulation of steroid biosynthetic process", and "platelet aggregation", all of these terms identified with FDR =1.00.

Compared to GO, five pathways reported by KEGG were identified with FDR significance (FDR < 0.05), and they were related with "Endocrine and other factor-regulated calcium reabsorptions",

"Dilated cardiomyopathy (DCM)", "Parathyroid hormone synthesis, secretion and action", "platelet activation" and "tuberculosis". Summary of top pathways reported in the gene-enrichment analysis in KEGG are described in Table 4-18. Some other pathways detected with borderline enrichment in KEGG were related with the pathophysiology of T2D: "fat digestion and absorption" (FDR=0.26), "cholesterol metabolism" (FDR=0.26), and "regulation of lipolysis in adipocytes" (FDR=0.26).

Calcium reabsorption and the synthesis and release of the parathyroid hormone (PTH) are two processes that belong to the same metabolic pathway, and they were reported by KEGG as two of the strongest terms in relation to T2D-associated DMR genes. Calcium reabsorption occurs primarily in the kidney, and it is a process partially regulated by PTH[211]. Likewise, synthesis of PTH is controlled by levels of calcium in serum, where hypocalcemic states stimulate the rapid synthesis and secretion of PTH from the parathyroid gland[211]. The association between these two processes and T2D is based on the presence of insulin resistance and reduced insulin-stimulated glucose transport, in response to increased levels of intracellular free calcium in patients with primary hyperparathyroidism[212], a condition that arises when the parathyroid gland undergoes hyperplasia and overactivity due to low levels of calcium in serum, among other factors[211]. Patients with long-term diabetes mellitus show lower levels of PTH in serum and higher levels of calcium in plasma compared to controls[212]. On the contrary, patients in whom primary hyperparathyroidism precedes T2D, or they occur together, the mechanism goes from increased levels of PTH and higher levels of free intracellular calcium, to impaired glucose intake, insulin resistance, and T2D[212].

Table 4-18 Top 20 pathways identified in the KEGG database for genes mapping near index DMPs identified within T2D-associated DMRs. Background Genes: total number of genes in the KEGG pathway, DM: number of genes with differential methylation, P: corrected p-value for enrichment using the FDR method (p<0.05).

Pathway	Background genes	DM genes	FDR
Endocrine and other factor-regulated calcium reabsorption	46	2	0.03
Dilated cardiomyopathy (DCM)	85	2	0.04
Parathyroid hormone synthesis, secretion and action	103	2	0.04
Platelet activation	122	2	0.04
Tuberculosis	154	2	0.04
Human papillomavirus infection	298	2	0.24
Vitamin digestion and absorption	23	1	0.26
Intestinal immune network for IgA production	37	1	0.26
Fat digestion and absorption	39	1	0.26
Mineral absorption	46	1	0.26
Antigen processing and presentation	52	1	0.26
Vasopressin-regulated water reabsorption	43	1	0.26
Cholesterol metabolism	49	1	0.26
Ovarian steroidogenesis	49	1	0.26
Vibrio cholerae infection	48	1	0.26
Cocaine addiction	46	1	0.26
Regulation of lipolysis in adipocytes	53	1	0.26
Bile secretion	69	1	0.26
Renin secretion	64	1	0.26
Long-term depression	58	1	0.26

4.8.3 Summary of functional exploration on T2D-associated DMRs

The annotation of DMRs for regulatory regions and genetic context revealed that most of the DMRs hypomethylated in T2D were identified within introns and exons, and to a less extent within promoters and intergenic regions. With respect to CpG islands, 50% of the DMRs hypomethylated overlapped with CpG islands. Enrichment for histone marks in these DMRs revealed that there was an overrepresentation of histone marks associated with enhancers and active chromatin, and a low representation of marks associated with inactive chromatin and transcriptional repression. Also, within hypomethylated DMRs, binding sites were detected for the transcription factors *PAX-C20*, *BCLAF1*, *Pol2*, among others. A single DMR was found hypermethylated in Chr20 and mapping to the region of the non-coding RNA gene *NCRNA00028*. According to signals detected in leukaemia cells, this DMR overlapped with histone marks associated with repressive chromatin states. Transcription factors with binding sites within this DMR were *ZNF143*, *P300* and *TAF1*.

An eQTM search revealed that there was no evidence that index CpG sites within T2D-associated DMRs were associated with gene expression of the gene in *cis* (not an eQTM). Thus, further functional description of these CpG sites was based on the nearest gene. Tissue-specificity in the level of expression was observed for genes annotated to DMRs, and only *FERMT3* was highly

expressed in peripheral blood. An in-silico analysis showed that there was no significant difference in the levels of expression of DMR genes between T2D cases and controls when using peripheral blood samples. An overlap between eQTL and meQTL for index CpG sites within DMRs, revealed that methylation at these sites could influence gene expression either in the DMR gene, or in genes located nearby, and this evidence was obtained at specific tissues. Cross-tissue comparison in the levels of methylation for the strongest CpG within each DMR, suggested a high correlation between methylation in blood and methylation in other internal tissues relevant to T2D. The highest correlation was identified between blood and omentum, and the lowest correlation was identified between blood and pancreas. A gene-enrichment analysis suggested that genes annotated to T2D-associated DMRs were significantly enriched for pathways related to the reabsorption of calcium, synthesis of the parathyroid hormone, platelet activation, dilated cardiomyopathy and tuberculosis, all of them with potential influence on T2D pathophysiology. Even though these observations did not surpass threshold for statistical significance, some pathways related with the metabolism of lipids and immunological mechanisms in response to infection were observed among top pathways in the enrichment analysis.

4.9 Validation of DMRs in comb-p

Even though DMRs obtained with *comb-p* were regarded as the main evidence, *DMRcate* was implemented as an alternative method to validate the discovery regions. Including a validation step for the DMR analysis was possible due to the variety of methods currently available for DMR identification[44]. However, the downside of this is that different methods use different definitions of regions without a clear consensus on which method performs best, which makes it more difficult to attempt replication of results, even when a similar training sample is used. As in *comb-p, DMRcate* uses a single CpG site analysis first, and retrieves a non-directionality measure associated to each CpG site (i.e. *t-statistic*) to infer regions of interest based on genomic locations, and on the implementation of a method to combine CpG site measures across the region.

4.9.1 Identifying replication between DMRs in comb-p and DMRs in DMRcate

Implementation of DMRcate was described in Chapter 2, and further detail of the method can be found elsewhere[147]. Regions in DMRcate were adjusted for multiple testing using the FDR correction, and DMRs were considered associated at Stouffer p-value<0.05. DMRcate did not identify regions strongly associated with T2D after Stouffer correction, suggesting that this DMR approach uses a more stringent method to select regions of interest (i.e. Stouffer) in comparison to comb-p (i.e. Sidak). Despite being non-significant, seven of the top DMRs reported by DMRcate were identified in

common with DMRs in comb-p (see Table 4-19); Stouffer p-value for these seven DMRs ranged between 0.34 and 0.82. DMRs in common were annotated to the genes *NCRNA00028, DPYSL3, NAT15, UBE2QL1, ITGB7, APOB,* and an unannotated DMR located in Chr7 (see Table 4-19). For these DMRs, coordinates reported by DMRcate were wider than coordinates reported by comb-p, indicating that DMRs in DMRcate included more DMPs within each region (see Table 4-19). Overall, the strongest DMR identified by DMRcate in borderline association with T2D was annotated to the non-coding RNA gene *NCRNA00028* (Stouffer= 0.34, size=177bp). Mean absolute difference in methylation between T2D cases and controls in this DMR was 3.4%, and the DMP count in this region was five DMPs. A plot for the DMR in *NCRNA00028*, and for the remaining six DMRs identified in common between DMRcate and *comb-p,* is presented in the appendix Figure S8-6.

4.10 Replicating the EWAS in type 2 diabetes using three European studies

Even though the DMR analysis provided more top-ranking associations with T2D relative to the single CpG site analysis, it was not possible to validate DMR results using a second method due to power limitations in the EWAS, or due to technical differences between the methods used for DMR analysis. In addition, studies reporting DMRs are less common than those reporting results from an EWAS. Therefore, more attention was put into following-up results of the EWAS rather than the DMR analysis. Because no final conclusion could be drawn from the EWAS in ALSPAC (i.e. power limitations), replication of the EWAS in additional European cohorts was necessary. Results of the replication were further used in a meta-analysis to summarize evidence across cohorts, and to increase power to detect associations with T2D at the CpG site level (see Chapter 6). Thus, the final section of this chapter is dedicated to describing main results of the replication of the EWAS in T2D using three European studies: KORA, LBC1936 and the Rotterdam studies RSIII-1 and RS-Bios. As in ALSPAC, the EWAS in additional cohorts was performed using four adjustment models: a minimally adjusted model, a model adjusted for cell heterogeneity, another additionally adjusted for smoking, and a fully adjusted model including BMI. Results of the EWAS are presented after exclusion of probes in the Naeem list.

Table 4-19 DMRs in T2D identified in common between comb-p and DMRcate. Even though none of the DMRs in the analysis in DMRcate surpassed Stouffer correction at p<0.05, some of the top-ranking DMRs in DMRcate overlapped with DMRs identified with Sidak significance in comb-p. Coordinates of the DMR correspond to those reported in the analysis in DMRcate.

				Comb-p (discovery method)				DMRcate (validation method)			
Chr	DMR	Index CpG	Gene	Size	CpG count	% Meth	Sidak	Size	CpG count	% Meth	Stouffer
2	21265912-21267334	cg03350299	*APOB*	167	6	3.26	$2.72x10^{-4}$	1422	14	2.24	0.64
5	146832182-146832933	cg18635723	*DPYSL3*	175	2	1.65	$7.17x10^{-3}$	751	7	1.00	0.74
5	6446895-6447257	cg12035880	*UBE2QL1*	23	3	1.61	$4.89x10^{-2}$	362	4	1.05	0.63
7	5609731-5610264	cg05281338	*Unannotated*	167	2	0.61	$1.98x10^{-2}$	533	3	0.22	0.71
12	53591398-53591766	cg04972065	*ITGB7*	11	2	3.37	$1.28x10^{-2}$	368	4	1.95	0.79
16	3507460-3508546	cg00484396	*NAT15*	123	5	2.51	$2.17x10^{-4}$	1086	10	0.77	0.82
20	30073399-30073576	cg02991085	*NCRNA00028*	178	5	6.00	$1.25x10^{-6}$	177	5	3.40	0.34

Index CpG: strongest CpG identified within a region based on the p-value reported in the fully-adjusted EWAS; CpG count: number of CpG sites detected within a DMR; %Meth: percent of the average absolute difference in methylation detected between T2D cases and controls within a region; Sidak: method used by comb-p to correct for multiple testing (Sidak < 0.05); Stouffer: method used by DMRcate to correct for multiple testing (Stouffer < 0.05).

4.10.1 Epigenetics of T2D in KORA

The subsample of KORA included in the EWAS comprised 1,719 participants from the fourth follow-up of this study (KORA F4). Further detail of this subsample was provided earlier in Chapter 2. Briefly, participants in KORA were on average 61 years of age, the proportion of females and males was approximately the same, and the total number of T2D cases observed was 155, the largest number reported across the three independent studies.

4.10.1.1 Summary of EWAS results in KORA

Even though KORA was the largest study used for the replication of the EWAS, accounting with a large number of T2D cases, detection of an association between methylation and T2D was not possible in this dataset for any of the adjustment models implemented (Table 4-20, see appendix Figure S8-7). Considering results of the most adjusted model, the strongest signal was identified at the DMP cg11696475 in *GNG4* (Table 4-20).

Table 4-20 Top-ten DMPs detected in the most adjusted EWAS of T2D conducted in participants in KORA (n=1,719). Model adjusted for age, sex, 10 PCs, 6 predicted cell-counts, BMI and smoking (non-smoker, smoker).

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg11696475	*GNG4*	1	1stExon	-0.001	1.66×10^{-4}	6.01×10^{-7}
cg10950524	*MAD1L1*	7	Body	-0.033	0.007	2.21×10^{-6}
cg02976539	*SLC9A3R1*	17	Body	0.008	0.002	2.73×10^{-6}
cg24377329	*SLC37A4*	11	TSS200	-0.007	0.001	3.34×10^{-6}
cg10780164	*CALY*	10	5'UTR	0.013	0.003	5.09×10^{-6}
cg14768946	*STAT1*	2	1stExon	0.002	0.000	5.60×10^{-6}
cg11949207	*C1orf66*	1	TSS200	-0.005	0.001	7.71×10^{-6}
cg06596743	*MON1B*	16	Body	0.004	0.001	1.21×10^{-5}
cg16898425	*ITIH1*	3	TSS200	-0.006	0.001	1.61×10^{-5}
cg01622006	*Unannotated*	6	Undefined	0.004	0.001	1.72×10^{-5}

The proportion of probes hypomethylated and hypermethylated in association with T2D was similar in the EWAS in KORA. Stronger probes with $p<10^{-5}$ had the smallest effect-size, which ranged between -0.001 and 0.01. Comparing results of the EWAS between ALSPAC and KORA, there was no similarity in the top-ten signals detected with the smallest p-value between studies across models (see appendix Table S8-6).

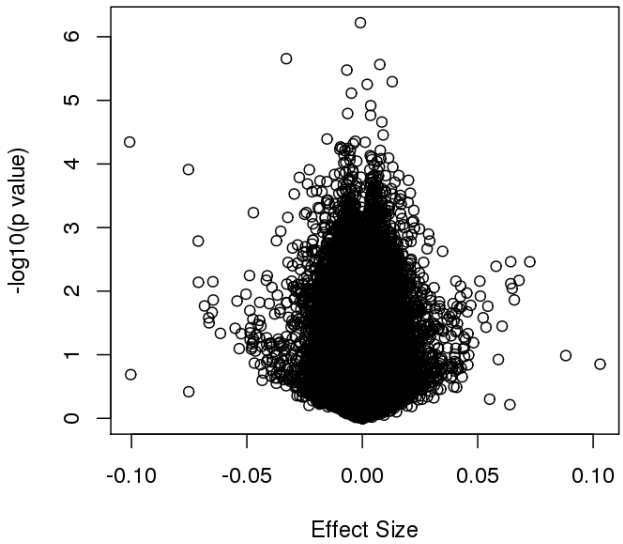


Figure 4-18 Volcano plot showing the distribution of effect-sizes (x-axis) against the -log10(p-value) (y-axis) for results of the most-adjusted EWAS conducted in KORA (n=1,719). None of the associations in this EWAS surpassed epigenome-wide significance at Bonferroni corrected p<1.07x10^{-7}. Top signal with the smallest p-value was detected at the DMP cg11696475 in GNG4.

4.10.2 Epigenetics of T2D in LBC1936

In total, the subsample of LBC1936 included in the EWAS was composed of 110 T2D cases and 804 controls. Further characteristics of this subsample and the method used for the assessment and QC of the methylation data, were described earlier in Chapter 2. Briefly, participants in LBC1936 were on average older than participants in ALSPAC, there was a much more equal proportion of females and males in this dataset, and there were more T2D cases than in ALSPAC, potentially related with the older age of LBC1936 participants.

4.10.2.1 Minimally adjusted EWAS model

Adjusting for age, sex and eight SVs, a strong association was detected at the DMP cg06500161, mapping to the region of the *ABCG1* gene (estimate=0.03, SE=4.0x10^{-3}, p=1.94x10^{-8}) (Table 4-21). There is compelling evidence of the association between higher methylation at *ABCG1* and future risk of T2D[62]. This marker has been also identified in previous EWAS using traits related with T2D, demonstrating that an increase in methylation at the DMP in *ABCG1* was positively associated with increasing levels of fasting glucose and fasting insulin, and with higher HOMA-IR score [88]. Importantly, previous studies have shown that the association between *ABCG1* and T2D or T2D-related traits, is not independent of BMI[62, 88]. An additional association at *ABCG1* (cg27243685) was identified with p-value in the order of 10^{-6} in this EWAS model. No further association surpassed Bonferroni correction in the analysis in LBC1936.

Table 4-21 Top-ten DMPs detected in the EWAS of T2D conducted in a subsample of participants from the LBC1936 study (n=915). Results based on a minimally adjusted model (covariates: age, sex and 8 SVs).

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg06500161	*ABCG1*	21	Body	0.025	0.004	$1.94x10^{-8}$
cg27243685	*ABCG1*	21	Body	0.017	0.004	$1.07x10^{-6}$
cg20068400	*Unannotated*	11	undefined	-0.036	0.007	$1.38x10^{-6}$
cg17055821	*C17orf75*	17	TSS1500	0.030	0.006	$1.64x10^{-6}$
cg12194745	*BAHCC1*	17	Body	0.018	0.004	$2.06x10^{-6}$
cg13555278	*EXTL1*	1	1stExon	0.017	0.004	$3.29x10^{-6}$
cg07051796	*ZFHX3*	16	Body	-0.012	0.003	$3.81x10^{-6}$
cg03312117	*Unannotated*	2	undefined	-0.043	0.009	$4.07x10^{-6}$
cg09371351	*HSD3B2*	1	5'UTR	0.038	0.008	$4.59x10^{-6}$
cg08515811	*TBC1D16*	17	Body	-0.013	0.003	$4.67x10^{-6}$

4.10.2.2 Cell-adjusted EWAS model

After adjustment for predicted cell-counts, no attenuation was observed for the strongest association detected in *ABCG1* (Table 4-22), and only small changes were identified in the effect size and the significance of this association. A second signal was identified with p-value in the order of 10^{-6} at the DMP cg19693031 in *TXNIP,* which is another marker widely detected in epigenetics of T2D[62, 65, 66]. Methylation at *TXNIP* has been previously reported in association with incident and prevalent T2D [62, 64, 66], and in inverse association with levels of Hb1Ac, fasting glucose[66], and the HOMA-IR score[64]. Results of the EWAS in LBC1936 showed that T2D cases were on average hypomethylated at *TXNIP* compared to controls (Table 4-22), and this result is in agreement with current evidence.

Table 4-22 Top-ten DMPs detected in the EWAS of T2D using a model adjusted for cells in participants in the LBC1936 cohort. Modell adjusted for age, sex, predicted cell-counts and 8 SVs.

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg06500161	*ABCG1*	21	Body	0.024	0.004	$5.89x10^{-8}$
cg19693031	*TXNIP*	1	3'UTR	-0.026	0.006	$2.68x10^{-6}$
cg09371351	*HSD3B2*	1	5'UTR	0.039	0.008	$2.77x10^{-6}$
cg07051796	*ZFHX3*	16	Body	-0.012	0.003	$3.51x10^{-6}$
cg27243685	*ABCG1*	21	Body	0.016	0.003	$4.89x10^{-6}$
cg17055821	*C17orf75*	17	TSS1500	0.026	0.006	$4.97x10^{-6}$
cg20068400	*Unannotated*	11	undefined	-0.029	0.006	$6.70x10^{-6}$
cg15127702	*EMID2*	7	Body	-0.019	0.004	$7.35x10^{-6}$
cg13565670	*FBRSL1*	12	Body	0.010	0.002	$8.96x10^{-6}$
cg03312117	*Unannotated*	2	undefined	-0.040	0.009	$1.15x10^{-5}$

4.10.2.3 Fully adjusted EWAS model

Following adjustment for BMI and smoking (i.e. never, former and current smoker), none of the associations tested surpassed Bonferroni correction at p<0.05, and the signal previously detected at *ABCG1* was markedly attenuated (29.5% decrease in the effect size, and an increase in the p-value from 10^{-8} to 10^{-4}). Evidence from a sensitivity analysis without adjustment for BMI suggested that BMI, rather than smoking, was confounding the association at *ABCG1* (see appendix Table S8-7). Strongest association with p-value in the order of 10^{-7} was identified at the DMP cg07051796 in *ZFHX3* (Table 4-23). The DMP in *TXNIP* was also identified among top-ranking associations with the smallest p-value ($p<10^{-5}$) in this EWAS (Table 4-23).

Table 4-23 Top ten DMPs detected in the most adjusted EWAS of T2D conducted in participants in the LBC1936 cohort. Model adjusted for age, sex, 8 SVs, 6 predicted cell-counts, BMI and smoking (never, ever, current smoker).

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg07051796	*ZFHX3*	16	Body	-0.013	0.003	$3.47x10^{-7}$
cg09371351	*HSD3B2*	1	5'UTR	0.042	0.009	$1.23x10^{-6}$
cg13565670	*FBRSL1*	12	Body	0.011	0.002	$1.65x10^{-6}$
cg17055821	*C17orf75*	17	TSS1500	0.027	0.006	$4.47x10^{-6}$
cg22077313	*Unannotated*	4	undefined	0.022	0.005	$5.36x10^{-6}$
cg20068400	*Unannotated*	11	undefined	-0.030	0.007	$7.02x10^{-6}$
cg17384323	*Unannotated*	4	undefined	0.012	0.003	$9.98x10^{-6}$
cg21740964	*FAM50B*	6	TSS1500	-0.033	0.007	$1.28x10^{-5}$
cg19693031	*TXNIP*	1	3'UTR	-0.025	0.006	$1.75x10^{-5}$
cg16611005	*FOXB1*	15	TSS1500	-0.003	0.001	$2.10x10^{-5}$

A volcano plot revealed an approximately similar distribution of probes with hypomethylation and hypermethylation in response to the effects of T2D (Figure 4-19). However, most of the probes with p-value < 10^{-5} were hypermethylated in T2D cases compared to controls (Figure 4-19). Absolute effect estimate in this analysis ranged between 0.10 and 0.08.

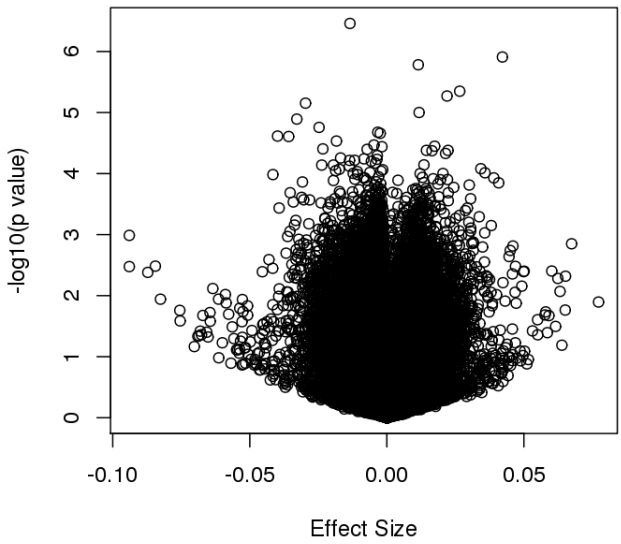


Figure 4-19 Volcano plot showing the distribution of effect-sizes (x-axis) against the -log10(P-value) (y-axis) for results of the most adjusted EWAS conducted in the subsample of LBC1936. None of the CpG sites plotted surpassed epigenome-wide significance at Bonferroni corrected p<0.05. Top signal with smallest p-value was detected at the DMP cg07051796 in ZFH3X.

Summary of epigenetics of T2D in LBC1936 and comparison with results of the EWAS in ALSPAC

In summary, the EWAS of T2D in the subsample of LBC1936 revealed no markers associated with T2D after adjustment for important covariates. The top signal of the EWAS, when ranking by p-value, was detected at DMP cg07051796 in *ZFHX3,* where T2D cases were on average hypomethylated compared to controls. In less adjusted models, a strong association was detected at DMP cg06500161 in *ABCG1*, but this association did not surpass adjustment for BMI, corroborating a possible mediating role of BMI in differences in methylation at *ABCG1* associated with T2D. Some evidence of association was detected at DMP cg19693031 in *TXNIP,* another important marker in epigenetic studies of T2D, after adjustment for cells, BMI, and smoking. Results showed that T2D cases were on average hypomethylated at *TXNIP* compared to controls, and this result is consistent with previous evidence at this locus. Comparing results of the EWAS in LBC1936 versus ALSPAC, there was no similarity among top-ten signals identified with the smallest p-value between studies across models (appendix Table S8-7).

4.10.3 Epigenetics of T2D in the Rotterdam Study RSIII-1

Participants of the third cohort of the Rotterdam study who were examined at baseline (RSIII-1), were included in this analysis. Details of the sub-sample of RSIII-1 were provided earlier in Chapter 2**.** Briefly, a subsample of 728 participants in RSIII-1 were selected to conduct the EWAS in T2D. On average these participants were 60 years old (SD=8.21), the proportion of females to males was relatively similar, and the number of T2D cases was 74, which is larger than the number of cases observed in ALSPAC.

4.10.3.1 Summary of EWAS results in RSIII-1

No association was identified below p=5.0x10^{-7} in any of the adjustment models implemented (Table 4-24, appendix Figure S8-9). An important signal was detected with p-value <10^{-5} at the DMP cg00574958 in *CPT1A* after adjustment for cells and smoking, showing that T2D cases were on average 1.8% (SE=0.004, p=6.16x10^{-6}) hypomethylated compared to controls in *CPT1A*. Previous evidence showed an inverse association between methylation at *CPT1A* and BMI in an African American[213] and an Arab[67] population, and in the study conducted by Al Muftah *et al*.[67], they found that increased methylation at *CPT1A* was protective against the risk of T2D. The strongest signal of the EWAS in RSIII-1 was detected at the DMP cg16330965 in *SNAPC5,* with p-value in the order of 10^{-7}. Table 4-24 shows top-ten signals with the smallest p-value identified in the most-adjusted EWAS; strongest associations detected for the remaining models are summarised in the appendix Table S8-8.

Table 4-24 Top-ten DMPs detected in the most adjusted EWAS of T2D conducted in participants in the Rotterdam Study Cohort III at baseline. Model adjusted for age, sex, 8 SVs, 6 predicted cell-counts, BMI and smoking (non-smoker, smoker).

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg16330965	*SNAPC5*	15	TSS1500	-0.013	0.003	5.00x10^{-7}
cg14278808	*LOC157627*	8	TSS1500	0.020	0.004	1.04x10^{-6}
cg02484673	*JPH3*	16	Body	-0.012	0.003	2.01x10^{-6}
cg05887092	*PGS1*	17	Body	-0.013	0.003	3.20x10^{-6}
cg21477861	*PLCD3*	17	Body	-0.004	0.001	5.79x10^{-6}
cg08121984	*APOC1P1*	19	TSS200	-0.012	0.003	6.82x10^{-6}
cg12986726	*CEBPB*	20	TSS1500	0.009	0.002	7.40x10^{-6}
cg13212575	*MAEL*	1	Body	-0.008	0.002	9.47x10^{-6}
cg07264682	*Unannotated*	10	Undefined	-0.022	0.005	1.02x10^{-5}
cg07416844	*Unannotated*	3	Undefined	-0.013	0.003	1.20x10^{-5}

A volcano plot showed that the proportion of probes hypomethylated in T2D cases versus controls was higher among associations surpassing the threshold of significance at p<10^{-5} (Figure 4-20). The effect-size in this analysis ranged between -0.11 and 0.11. Comparing top-ten strongest associations identified in the EWAS between RSIII-1 and ALSPAC, there was no similarity in top signals detected between studies across models (see appendix Table S8-8).

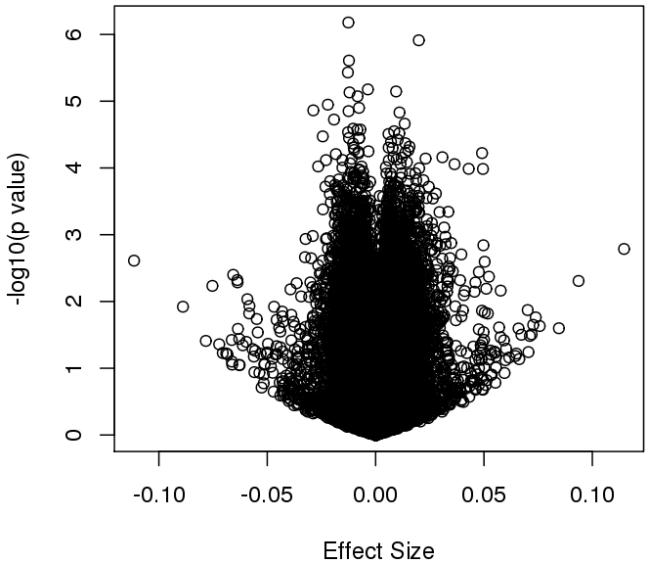


Figure 4-20 Volcano plot showing the distribution of effect-sizes (x-axis) against -log10(P-value) (y-axis) for results of the most-adjusted EWAS conducted in the Rotterdam Study RSIII-1.None of the CpG sites in this analysis surpassed epigenome-wide significance at $p < 1.07x10^{-7}$. Top signal with smallest p-value was detected at DMP cg07051796 in SNAPC5.

4.10.4 Epigenetics of T2D in the Rotterdam-Bios study

The Rotterdam-Bios cohort, which is a sub-study of the Rotterdam study, was included as a fourth independent cohort. Rotterdam-Bios comprises participants that were recruited in the Rotterdam cohort II third follow-up (RSII-3), and in the Rotterdam cohort III second follow-up (RSIII-2). Further detail of this study and baseline characteristics of the subsample included in the replication of the EWAS, was provided earlier in Chapter 2. Briefly, the subsample of RS-Bios included in the EWAS was composed of 723 participants, mean age in this subsample was 68 years, the proportion of females was slightly larger than that of males, and there were more cases of T2D in this study compared to ALSPAC.

4.10.4.1 Summary of EWAS results in RS-Bios

From the four adjustment models applied, none of the association identified in the RS-Bios study surpassed the threshold for statistical significance at $p<1.07x10^{-7}$ (Table 4-25, see appendix Figure S8-10). Based on results of the most adjusted model, the strongest signal with p-value in the order of 10^{-6} was detected at DMP cg16339915 in *TIFAB* (Table 4-25). Further detail of top-ten associations with the smallest p-value identified in the most-adjusted EWAS in RS-Bios, is presented in Table 4-25.

Table 4-25 Top-ten DMPs detected in the most adjusted EWAS of T2D conducted in participants in the Rotterdam Bios sub-study (n=723). Model adjusted for age, sex, 9 SVs, 6 predicted cell-counts, BMI and smoking (non-smoker, smoker).

CpG	Loci	Chr	Genomic context	Beta	SE	P-value
cg16339915	*TIFAB*	5	Body	-0.007	0.002	$3.32x10^{-6}$
cg24795867	*WNT5B*	12	Body	-0.008	0.002	$6.61x10^{-6}$
cg16575444	*CX3CL1*	16	Body	-0.008	0.002	$7.19x10^{-6}$
cg11983038	*Unannotated*	13	undefined	-0.025	0.006	$9.82x10^{-6}$
cg14491707	*CACNA1B*	9	3'UTR	-0.015	0.003	$1.05x10^{-5}$
cg02635644	*AGRN*	1	Body	-0.003	0.001	$1.13x10^{-5}$
cg02859537	*AKT1S1*	19	5'UTR	0.012	0.003	$1.49x10^{-5}$
cg16565002	*RBMS3*	3	Body	-0.011	0.003	$1.50x10^{-5}$
cg24512093	*ROBO1*	3	Body	-0.011	0.003	$1.59x10^{-5}$
cg25250358	*PLOD2*	3	5'UTR	-0.015	0.004	$1.78x10^{-5}$

Distribution of probes hypomethylated and probes hypermethylated in association with T2D was approximately similar (Figure 4-21). However, there was an over-representation of probes hypomethylated for the strongest associations of the EWAS detected at $p<10^{-5}$ (Figure 4-21). Effect-sizes in the most-adjusted model ranged between -0.07 and 0.07. As mentioned earlier for other analyses, there was no overlap in the top signals detected with the smallest p-value between RS-Bios and ALSPAC (see appendix Table S8-9).

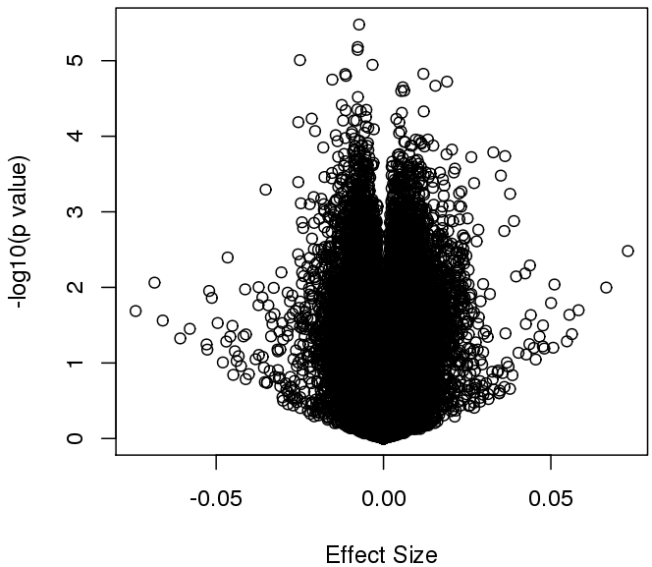


Figure 4-21 Volcano plot showing the distribution of effect-sizes (x-axis) against the -log10(P-value) (y-axis) for results of the most-adjusted EWAS conducted in participants in the Rotterdam-Bios study (RS-Bios). None of the associations identified surpassed epigenome-wide significance at $p<1.07x10^{-7}$. Top signal with the smallest p-value was detected at DMP cg16339915 in TIFAB.

4.10.5 Comparison of the association at the CpG in *NFYC* across studies

The association at cg15986668 in *NFYC* was detected with epigenome-wide significance in the analysis in ALSPAC (see section 4.5.3), but this association did not surpass Bonferroni correction at

p<0.05 in any of the additional studies when looking at results of the most adjusted EWAS model (Table 4-26). Thus, there was weak evidence of replication of the association at *NYFC* across studies. Direction of effect was similar between ALSPAC and LBC1936, but this effect was in opposite direction for the remaining studies. The absolute effect size was on average 6.5% larger in ALSPAC compared to the other studies, and the unadjusted p-value was in the order of 10^{-8} in ALSPAC, and in the order of 10^{-1} in the additional cohorts (Table 4-26).

Table 4-26 Association between T2D and methylation at cg15986668 in NFYC across five European cohorts. Estimates correspond to the fully adjusted EWAS model (covariates: age, sex, SVs, predicted cell counts, BMI and smoking). P is the unadjusted p-value. Associations were considered significant at $p<1.07x10^{-7}$.

Study	Cases	Controls	Beta	SE	P
ALSPAC	48	1002	-0.071	1.30E-02	5.48E-08
KORA	154	1563	0.002	5.96E-03	7.02E-01
LBC1936	110	803	-0.013	9.48E-03	1.57E-01
RSIII-1	73	650	0.006	6.93E-03	4.25E-01
RS-Bios	74	654	0.007	6.75E-03	2.90E-01

4.11 Chapter summary

This Chapter presented evidence of the single CpG site analysis in participants in the subsample of ALSPAC/ARIES, as well as in four replicating cohorts. Looking across studies, there was no overlap in the top-ranking signals detected in the EWAS, even though the analysis was conducted under similar conditions. The reasons for this could include population-specific differences in methylation because of a different environment, differences in age between studies, differences in the way T2D was defined, the presence of comorbidities that were not accounted for in the analysis, or to a combination of these factors.

In the analysis in ALSPAC/ARIES, it was demonstrated that a DMP in *NFYC* was the strongest evidence of association with T2D, showing that participants affected with T2D had on average lower methylation at the DMP in *NFYC* when compared to controls. The goodness-of-fit of the model showed that T2D and covariates explained around 11.29% of the total variation in methylation at the *NFYC* locus, and the strength of this association was not confounded by BMI, fasting glucose or levels of C-reactive protein. However, the functional exploration of this marker did not suggest any biological process related with the pathophysiology of T2D. Since the observational association between T2D and methylation at *NFYC* may still be biased by unmeasured confounders, a causal

analysis was then planned for this association using the principles of Mendelian randomization (see Chapter 7).

It was also shown that the DMR analysis in ALSPAC/ARIES had more power to detect strong associations in comparison to the EWAS, but DMRs were not validated using an alternative method due to conflicts in the way that regions are defined and selected across different methods. From the 12 DMRs identified in association with T2D, one of them was annotated to *APOB*, a protein-coding gene related with the metabolism of lipids. To my knowledge, no signal has been reported to this gene in any GWAS and EWAS of T2D. Functional exploration of DMRs showed that genes annotated to DMRs were enriched in pathways related to the reabsorption of calcium and the synthesis of the parathyroid hormone, this for pathways representative for T2D. In terms of regulatory elements, DMRs were enriched in histone marks associated with transcriptionally active regions (H3K4me1), and there was evidence of binding-sites for transcription factors within the DMRs. Although there was no evidence of differential gene expression between T2D cases and controls for genes annotated to the DMRs, there was some suggestion that DMRs could have some influence on expression of nearby genes due to the presence of eQTL overlapping meQTL associated with some of the top CpG sites within the DMRs.

In the replication of the EWAS, LBC1936 was the only study where a signal was detected in strong association with T2D, corresponding to DMP cg06500161 in *ABCG1* which has been reported to be associated with T2D and related traits in the literature[46, 62-64]. However, the association between T2D and *ABCG1* methylation was confounded by the effect of BMI in this subsample. To improve the applicability of results obtained in the replication of the EWAS, a meta-analysis was proposed to summarize this evidence, with results provided in Chapter 6.

Chapter 5 DNA methylation as a predictor of glycaemic traits

Introduction and aims of the Chapter

In Chapter 4, one CpG site and several DMRs were identified as being differentially methylated in adults with prevalent T2D, compared to disease-free controls. The main aim in the present chapter is to investigate the role of DNA methylation in glycaemic traits used in the diagnosis of T2D. DNA methylation (as the exposure) is investigated in relation to glycaemic traits in adults *without* diagnosed diabetes at the time DNA methylation was measured. As a secondary analysis, the role of DNA methylation (as the exposure) is examined in relation to T2D (as the outcome) in adults *with* the disease at the time DNA methylation was measured. The purpose of this second analysis was not to predict T2D using DNA methylation, but to compare results of the EWAS using DNA methylation as an exposure and an outcome in relation to prevalent T2D.

Previous epigenetic studies on T2D outnumber those in glycaemic traits. However, it is important to characterize patterns of methylation in relation to glycaemic traits in disease-free controls to understand which mechanisms might trigger future risk of T2D. From the epigenetic studies conducted on glycaemic traits[46, 64, 73, 87, 88], the most recent is a systematic review of evidence of DNA methylation in association with HbA1c and fasting glucose[46], where the authors attempted to replicate markers identified with study-specific Bonferroni significance in a subsample of 100 diabetes-free participants from the Dutch population-based Lifelines study. Commonly identified methylation sites in association with HbA1c and fasting glucose were *ABCG1*, *CPT1A*, *SREBF1* and *TXNIP*[87, 88], which were also sites identified in association with T2D based on studies using peripheral blood DNA methylation. Markers taken forward for replication were 10 CpG sites associated with HbA1c in adipose tissue, and 21 CpG sites associated with fasting glucose in peripheral blood samples. Replication was achieved for CpG sites in *ABCG1* and *CCDC57* in relation with fasting glucose[46]. None of the CpG sites identified for HbA1c in adipose tissue was replicated at Bonferroni significance using peripheral blood samples[46].

Apart from blood, other studies in pancreatic islets from non-diabetic donors have reported the correlation between HbA1c levels and DNAm[79], and between elevated glucose concentrations and increased average DNA methylation across genomic regions in pancreatic cells[214], without identifying significant changes at the single CpG site level[214]. In addition, Rönn and colleagues found that average DNA methylation in adipose tissue from non-diabetic donors was negatively correlated with

increased levels of HbA1c[73]. In the same study, associations at the single CpG site level were reported in 711 CpG sites, with 14% of these sites showing positive and the remaining 86% sites showing negative correlation between adipose tissue DNA methylation and HbA1c[73]. Of interest in the study by Rönn *et al*.[73] was that some of the sites correlated with increased levels of HbA1c in non-diabetes donors (30/711 sites), overlapped with CpG sites previously identified with differential methylation in adipose tissue between T2D cases and disease-free controls.[73] For most of the CpG sites in overlap across studies, there was consistency in the direction of effect, indicating that differential methylation in relation to elevated HbA1c levels in the normal range, can determine pathways associated with future risk of disease.

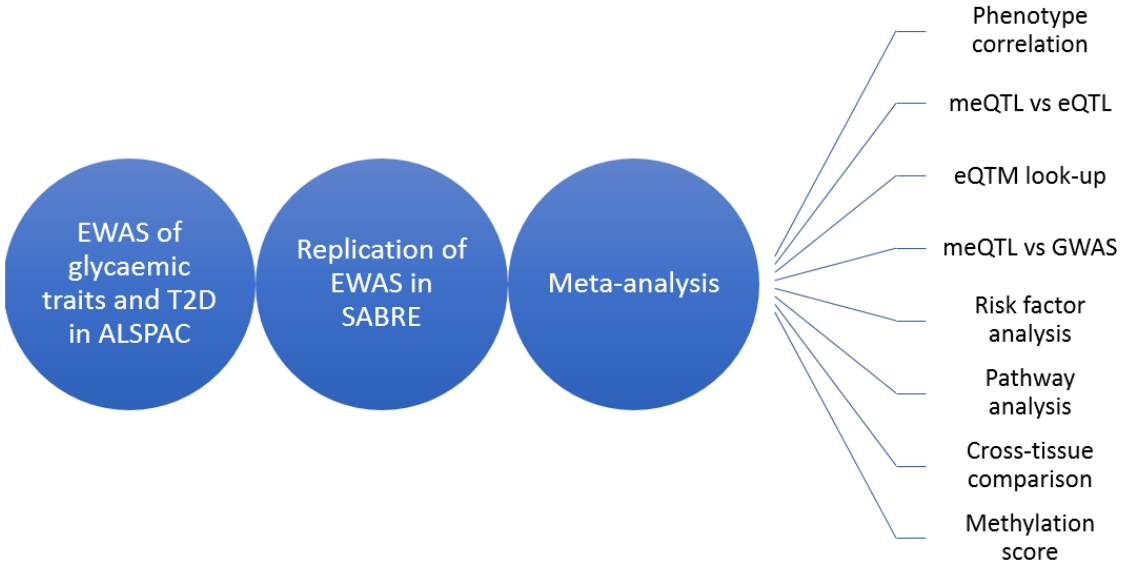
The association between DNA methylation as a predictor of prevalent T2D has been considered in previous studies[79, 215], where important biomarkers have been obtained. For instance, Toperoff *et al.*[60] identified methylation markers mapping to *FTO, TCF7L2, THADA, JAZF1, SLC30A8* and *KCNQ1* in peripheral blood samples of participants with prevalent T2D. Furthermore, Dayeh *et al.*[79] identified difference in methylation in relation to prevalent T2D at *ABCG1* in adipose tissue*,* at *PHOSPHO1* and *TXNIP* in skeletal muscle*,* and at *SREBF1* in pancreatic islets. Since changes in DNA methylation measured at the same time of disease occurrence can be biased by the effects of reverse causation, these markers are not good predictors of future risk of T2D, but indicators of disease state. An ideal approach to obtain predictive markers for T2D will be to measure DNAm at baseline prior to disease occurrence using longitudinal studies, but these studies tend to be more expensive and difficult to conduct compared to cross-sectional studies. More recently, some studies have shown that methylation markers detected prior to disease occurrence, can also be captured in relation to prevalent T2D[46, 64-66], reflecting the stability of methylation levels at these sites after disease occurrence[46].

The research described in this chapter (Figure 5-1) aims to contribute to our understanding of the role of DNA methylation in pathways to T2D by:

1) Investigating the association between DNA methylation and different glycaemic traits in middle-aged participants from two European cohorts (ALSPAC and SABRE), both individually and in a meta-analysis.
2) Investigating the role of methylation as a mediator of the SNP-glycaemic trait and SNP-gene expression associations, by using publicly available data from GWAS meta-analyses of glycaemic traits, meQTL and eQTL data.

3) Assessing the functional relevance of T2D- and glycaemic trait-related methylation sites by implementing gene enrichment analysis and eQTM inspection.
4) Determining the proportion of variation in glycaemic traits explained by patterns of differential methylation using a weighted methylation score.

Figure 5-1 Flow-diagram summarizing different analyses conducted in this Chapter.



5.1 *Methods implemented in this chapter*

This section shows an overview of the data and methods implemented to conduct the analyses on glycaemic traits. Further detail of the assessment of glycaemic traits in the specific cohorts, pre-processing of the methylation data, methods in EWAS, meta-analysis of EWAS, and construction of a methylation score, was provided earlier in Chapter 2.

5.1.1 Samples

Three subsamples from ALSPAC were selected for the analyses, one that included T2D cases and controls which was used in the association between DNA methylation (exposure) and T2D (n=1050, 48 cases and 1002 controls); a second subsample derived from the first one, which included normoglycemic and non-medicated participants for the EWAS of fasting glucose (FG; n=1002 controls); finally, the third subsample were 622 normoglycemic females included in the EWAS of 2-h glucose, fasting insulin and proinsulin, HOMA-IR and HOMA-B. The subsample from SABRE included in the EWAS of glycaemic traits were 382 normoglycemic males with available measures for the traits of interest, in addition to HbA1c and 2-h insulin. Fasting proinsulin was not measured in SABRE samples.

5.1.2 Models and variables

Multivariable linear regressions were applied to conduct the EWAS of glycaemic traits, and a multivariable logistic regression for the EWAS of T2D. In addition, a sensitivity analysis was conducted for the EWAS of fasting proinsulin in ALSPAC to account for outliers in the distribution of the outcome in this subsample. For each of the glycaemic traits, there were three adjustment models (Table 5-1). Sex was considered a covariate in the EWAS of T2D and FG in ALSPAC, but not in the EWAS of other glycaemic traits, where the associations were tested only in females (ALSPAC) and males (SABRE). Fasting insulin, HOMA-IR and HOMA-B were log-transformed before the analysis to achieve an approximately normal distribution. Proinsulin was transformed using the reciprocal. Associations identified in the EWAS of glycaemic traits in ALSPAC and SABRE were combined using a fixed-effect meta-analysis, with correction for multiple testing using the Benjamini-Hochberg method. Associations were regarded statistically significant at FDR< 0.05, and results were interpreted as a unit change in the phenotype per 10% increase in methylation.

Table 5-1 Summary of regression models implemented in the EWAS of T2D (ALSPAC) and glycaemic traits (ALSPAC and SABRE).

Model	Description
M1	Age, sex, surrogate variables [a] and 6 Houseman cells [b].
M2	Age, sex, surrogate variables, 6 Houseman cells and smoking.
M3	Age, sex, surrogate variables, 6 Houseman cells, smoking and BMI.

[a] Surrogate variables were calculated using the R package *sva*, and the effect measured by these variables was independent of the outcome investigated. [b] Predicted cell-counts using the Houseman method[119]. Predicted cells were CD4T, CD8T, Natural killer cells, B-cells, Monocytes and Granulocytes in ALSPAC, while an additional cell-type was included in SABRE by separating granulocytes into eosinophils and neutrophils.

5.1.3 Methylation score

Traits identified in strong association (at FDR<0.05) with differential methylation in the meta-analysis between ALSPAC and SABRE, were taken forward to determine the proportion of variance in the outcome explained by top-ranking CpG sites (smallest p-value) using a methylation score. The strength of the score in predicting the outcome was compared relative to established risk factors. In addition, a score was generated for traits uniquely measured in SABRE using the strongest associations detected at FDR<0.05 in the EWAS.

5.2 Study population in ALSPAC

Baseline characteristics of the subsample of 1050 middle-age adults in ALSPAC included in the EWAS of T2D, with DNAm as the exposure, was previously described in Chapter 4, while baseline characteristics of normoglycemic participants included in the EWAS of FG, can be found in Table 5-2.

Briefly, 62% of the total subsample were females, median age of these participants was 49 years, and median FG, BMI, waist-circumference and total cholesterol were 5.24mmol/l, 25.68kg/m^2, 87.25cm and 4.77mmol/l, respectively (see Table 5-2). Categories of glucose tolerance defined by the World Health Organization (WHO, 1999) were incorporated to identify disease-free participants at risk of pre-diabetes (see Chapter 2). Of the total subsample, 5% participants had impaired fasting glucose (IFG), but no distinction between impaired glucose tolerance (IGT) and normal glucose tolerance was possible because 2-h glucose was not measured in the males' subset in ALSPAC.

Baseline characteristics of 622 normoglycemic females included in the EWAS of other glycaemic traits is presented in Table 5-2. Briefly, these females were between 35 and 60 years of age, median values of 2-h glucose, fasting insulin, fasting proinsulin, HOMA-IR and HOMA-B, were 4.25mmol/l, 4.49µIU/ml, 2.47pmol/l, 1.05 and 55.14, respectively (Table 5-2); based on categories of glucose tolerance, 98% of these samples had normal glucose tolerance, and the remaining 2% had impaired fasting glucose (IFG).

5.3 EWAS of type 2 diabetes and glycaemic traits in ALSPAC

Table 5-3 shows results of the EWAS for the different phenotypes evaluated across three adjustment models. A p-value < $1.0x10^{-5}$ was used to describe top-ranking associations, while borderline significance was regarded at FDR ≤ 0.10. Associations identified were independent of SNPs present in the probe-binding region, and of problematic probes reported by Naeem *et al.*[139]. Taking together results of the EWAS across phenotypes and models, most associations with genome-wide significance and borderline significance were detected in model 2 after adjustment for smoking. In this model, there was strong evidence of an association between differential methylation and HOMA-B (intergenic CpG sites cg23436042 and cg20391220), HOMA-IR (cg06500161 in *ABCG1*) and fasting insulin (cg06500161 in *ABCG1*). In addition, after adjustment for smoking, 12 CpG sites were identified in association with different traits at FDR < 0.10: CpG cg03826430 in *PAOX* was associated with 2-h glucose, CpG cg24272697 in *HOXA3* was associated with fasting insulin and HOMA-IR, the intergenic CpG cg24192505 was associated with fasting proinsulin, and methylation levels of nine CpG sites were associated with HOMA-B (Table 5-3). The only association that remained significant at FDR<0.05 after further adjustment for BMI was between the intergenic CpG cg20391220 and HOMA-B, while three other associations with 2-h glucose, fasting proinsulin and HOMA-B, were identified at FDR<0.10 in the model adjusted for BMI. Manhattan and QQ-plots highlighting main findings of the different EWAS are presented in the appendix Figure S8-11 and Figure S8-12.

Table 5-2 Baseline characteristics of the subsample of normoglycemic participants included in the EWAS of fasting glucose (n=1002 females and males), and in the EWAS of other glycaemic traits (n=622, only females) in ALSPAC. Continuous variables were described using the median and the first and third quartiles. Categorical variables were described using the percentage of samples per category.

	EWAS of FG (n=1002)		**EWAS of T2D-traits (n=622)**	
	Median (25^{th}; 75^{th})	%	Median (25^{th}; 75^{th})	%
Age (years)	49.0 0 (46.00; 53.00)	-	48.00 (45.00; 50.00)	-
Cholesterol (mmol/L)	4.77 (4.18; 5.43)	-	4.59 (4.02; 5.09)	-
Triglycerides (mmol/L)	0.98 (0.73; 1.37)	-	0.88 (0.67; 1.17)	-
HDL cholesterol (mmol/L)	1.37 (1.15;1.61)	-	1.44 (1.21; 1.72)	-
LDL cholesterol (mmol/L)	3.01 (2.50; 3.59)	-	2.89 (2.37; 3.43)	-
Fasting glucose (mmol/L)	5.24 (4.96; 5.55)	-	5.15 (4.87; 5.39)	-
2-hours serum glucose (mmol/l)	-	-	4.25 (4.02; 4.58)	-
Fasting Insulin (µIU/ml)	-	-	4.49 (3.16; 6.67)	-
Proinsulin (pmol/L)	-		5.56 (4.44; 7.88)	
HOMA-IR [a]	-	-	1.05 (0.69; 1.52)	-
HOMA-B [b]	-	-	55.14 (40.14;77.88)	
Glucose Tolerance status [c]				
NGT	-	NA	-	98.00
IFG	-	5.0	-	2.00
IGT	-	NA	-	0.00
Combined IFG and IGT	-	NA	-	0.00
Waist circumference (cm)	87.25 (78.35; 96.25)	-	81.05 (74.96; 89.76)	-
BMI (kg/m^2)	25.68 (23.52; 28.46)	-	25.14 (22.59; 28.16)	-
No Medication [%]	-	100	-	100
Sex [% female]	-	62.08	-	100
Smoking [% smoker]	-	8.98	-	8.84
Physical activity [% < 4h/week]	-	83.24	-	89.06
Socioeconomic status [%]				
High income	-	57.64	-	54.98
Middle income	-	30.44	-	32.48
Low income	-	11.91	-	12.54
Predicted cell-counts				
CD4T	0.17 (0.13;0.20)	-	0.17 (0.14; 0.21)	-
CD8T	<0.01 (<0.01;0.03)	-	<0.01 (<0.01; 0.02)	-
Natural Killer Cells	0.19 (0.17; 0.23)	-	0.19 (0.16; 0.23)	-
B-cells	0.09 (0.07; 0.11)	-	0.09 (0.08; 0.12)	-
Monocytes	0.07 (0.05; 0.08)	-	0.06 (0.05; 0.08)	-
Gran	0.50 (0.45; 0.56)	-	0.52 (0.47; 0.58)	-

[a] HOMA-IR: homeostasis model-assessment for insulin resistance. This value was calculated as: [fasting plasma glucose(mmol/l) x fasting insulin (µIU/ml)] / 22.5 [27]. For the EWAS, this variable was log transformed to approximate a normal distribution. [b] HOMA-B: homeostasis model-assessment for β-cell disfunction. This value was calculated as: 20 x fasting insulin (µIU/ml) / fasting plasma glucose (mmol/l) - 3.5 [114]. [c] Glucose tolerance status was defined using the WHO criteria (1999). IFG: impaired fasting glucose if FG ≥ 6.1 and FG < 7.0 mmol/l. IGT: impaired glucose tolerance if FG < 7.0mmol/l and 2h-glucose ≥ 7.8 mmol/l and 2h-glucose < 11.1mmol/l. NGT: normal glucose tolerance if FG < 6.1mmol/l and 2h-Glucose < 7.8 mmol/l. Measures of 2-hours serum glucose (mmol/l), fasting insulin (µIU/ml), HOMA-IR and HOMA-B, were only available in the subset of control females.

No association surpassing FDR<0.10 was identified between methylation and fasting glucose, prevalent T2D, and the sensitivity analysis for proinsulin (excluding outliers), in any of the three adjustment models (Table 5-3). Of interest in the EWAS of prevalent T2D with DNAm as the exposure, was the replication of an association in the CpG cg1598668 in *NFYC*. Methylation at this site was previously identified as the strongest marker in the EWAS of prevalent T2D with DNAm as the outcome (see Chapter 4). Results in the present analysis suggested that a 10% increase in methylation at the CpG in *NFYC* was associated with an average reduction in 8% (OR: 0.92, 95%CI: 0.88-0.95) in the risk of prevalent T2D at a p-value in the order of 10^{-6}. Direction of this association was consistent with our previous finding showing that T2D cases were on average hypomethylated at the CpG in *NFYC* (estimate=-0.07, SE=0.01, p=5.48×10^{-8}) compared to disease-free controls (see Chapter 4). Overall, evidence at the CpG in *NFYC* suggested that hypermethylation of this site could have a protective effect against the risk of T2D, whilst hypomethylation was associated with prevalence of T2D. Furthermore, in the present EWAS of T2D, an association was identified with p-value in the order of 10^{-6} to 10^{-4} at the CpG cg06500161 (*ABCG1*), where a 10% increase in methylation at *ABCG1* was associated with an average increase in 17% (OR: 1.17, 95%CI: 1.09-1.25, p=4.72×10^{-6}) in the risk of prevalent T2D. Direction of effect of the association at *ABCG1* is in agreement with previous studies reporting that hypermethylation at this locus is associated with higher risk of T2D[46, 62-64]. Top-ranking signals with the smallest p-value identified in the EWAS of T2D and in the EWAS of FG, are described in the appendix Table S8-10 and Table S8-11. Associations identified with p< 1.0×10^{-5} in the sensitivity analysis of proinsulin are presented in Table 5-3.

The following sections describe: (1) a sensitivity analysis in the EWAS of proinsulin to determine the impact of outliers in the effect-size and p-value of the associations identified between a complete analysis, and a sensitivity analysis excluding outliers (see section 5.3.1), (2) the impact of adjusting for BMI in the effect-size and p-value of the associations detected (see section 5.3.2), (3) evidence of overlap between glycaemic traits by measuring the correlation in regression coefficients identified between EWAS for specific CpG sites (see section 5.3.3), and finally, (4) investigating potential reasons why it was detected an over-representation of negative effects in most of the EWAS among probes with low significance and large effect-sizes (see section 5.3.4).

Table 5-3 Results of the EWAS of glycaemic traits conducted in a subset of normoglycemic participants in ALSPAC (n=1002 and 622). The EWAS of T2D included 1050 samples, 48 T2D cases and 1002 controls. Three adjustment models were applied to account for specific covariates. Associations reported were identified with p < $1.0x10^{-5}$ in at least one model.

						Model 1 (age, SVs, predicted cells)			Model 2 (age, SVs, predicted cells, smoking)			Model 3 (age, SVs, predicted cells, smoking, BMI)		
Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Beta[b]	P	FDR	Beta	P	FDR	Beta	P	FDR
Fasting glucose[a]	cg01099300	*Intergenic*	10	Intergenic	Open sea	-0.30	$3.49x10^{-6}$	0.71	-0.30	$3.63x10^{-6}$	0.72	-0.30	$2.15x10^{-6}$	0.35
	cg23274377	*BPNT1*	1	TSS200	Open sea	0.50	$4.05x10^{-6}$	0.71	0.50	$4.27x10^{-6}$	0.72	0.50	$2.20x10^{-6}$	0.35
	cg17540765	*RECQL5*	17	Body	S_Shelf	-0.20	$4.50x10^{-6}$	0.71	-0.20	$4.58x10^{-6}$	0.72	-0.20	$1.14x10^{-6}$	0.35
	cg17219086	*Intergenic*	6	Intergenic	Open sea	0.10	$1.05x10^{-5}$	0.98	0.10	$9.94x10^{-6}$	1.00	0.02	$2.52x10^{-5}$	1.00
	cg26234543	*TMEM17*	2	Body	N_Shore	0.03	$1.25x10^{-5}$	0.98	0.03	$1.29x10^{-5}$	1.00	-0.10	$5.46x10^{-6}$	0.64
	cg03693099	*CEL*	9	TSS1500	Open sea	0.04	$3.81x10^{-5}$	1.00	0.04	$3.58x10^{-5}$	1.00	-0.20	$7.28x10^{-6}$	0.69
T2D	cg15986668	*NFYC*	1	TSS1500	N_Shore	0.92	$3.40x10^{-6}$	0.75	0.92	$3.88x10^{-6}$	0.83	0.92	$1.86x10^{-6}$	0.87
	cg06500161	*ABCG1*	21	Body	S_Shore	1.16	$6.99x10^{-6}$	0.75	1.17	$4.72x10^{-6}$	0.83	1.14	$1.73x10^{-4}$	0.87
2-h glucose	cg03826430	*PAOX*	10	5'UTR	Island	-1.30	$1.52x10^{-7}$	0.07	-1.30	$1.58x10^{-7}$	0.07	-1.30	$1.12x10^{-7}$	0.05
	cg05339942	*HRNBP3*	17	5'UTR	Open sea	0.20	$1.73x10^{-6}$	0.41	0.20	$1.68x10^{-6}$	0.39	0.20	$2.37x10^{-6}$	0.56
	cg08817540	*HHLA2*	3	TSS1500	Open sea	0.10	$1.17x10^{-5}$	1.00	0.10	$1.05x10^{-5}$	1.00	0.10	$6.92x10^{-5}$	1.00
Fasting insulin	cg06500161	*ABCG1*	21	Body	S_Shore	10.30	$6.02x10^{-9}$	$2.83x10^{-3}$	10.30	$7.02x10^{-9}$	$3.30x10^{-3}$	10.50	$7.32x10^{-6}$	0.65
	cg24272697	*HOXA3*	7	5'UTR	N_Shelf	10.70	$6.58x10^{-7}$	0.15	10.70	$2.71x10^{-7}$	0.06	10.50	$7.32x10^{-6}$	0.65
	cg17340655	*DDHD2*	8	TSS200	Island	6.20	$1.33x10^{-6}$	0.16	6.20	$1.29x10^{-6}$	0.17	7.20	$1.23x10^{-4}$	0.73
	cg00606312	*KMO*	1	TSS200	Open sea	10.40	$1.25x10^{-6}$	0.16	10.40	$1.41x10^{-6}$	0.17	10.30	$8.82x10^{-5}$	0.73
	cg25690958	*SPNS2*	17	Body	Island	9.90	$2.37x10^{-6}$	0.19	9.90	$1.92x10^{-6}$	0.18	9.90	$1.48x10^{-4}$	0.73
	cg23436042	*Intergenic*	5	Intergenic	Open sea	10.30	$5.24x10^{-6}$	0.27	10.30	$4.68x10^{-6}$	0.28	10.20	$3.26x10^{-5}$	0.65
	cg11927233	*NPM1*	5	Body	S_Shore	10.30	$2.18x10^{-6}$	0.19	10.30	$4.78x10^{-6}$	0.28	10.10	$1.24x10^{-2}$	0.90
	cg16570129	*RUFY1*	5	Body	Island	10.10	$4.90x10^{-6}$	0.27	10.10	$5.76x10^{-6}$	0.30	10.10	$2.62x10^{-5}$	0.65
	cg04870212	*Intergenic*	1	Intergenic	Open sea	8.50	$1.40x10^{-5}$	0.40	8.50	$6.50x10^{-6}$	0.30	8.80	$9.94x10^{-5}$	0.73
	cg11822932	*LMO2*	11	1stExon	Open sea	10.20	$1.45x10^{-5}$	0.40	10.20	$6.94x10^{-6}$	0.30	10.10	$5.52x10^{-4}$	0.81
Fasting proinsulin	cg24192505	*Intergenic*	6	Intergenic	open sea	-0.40	$1.62x10^{-7}$	0.08	-0.40	$1.66x10^{-7}$	0.08	-0.50	$2.79x10^{-6}$	0.44
	cg23015341	*TFAP2B*	6	3'UTR	Island	-0.20	$2.02x10^{-6}$	0.38	-0.20	$2.23x10^{-6}$	0.42	-0.30	$5.49x10^{-5}$	1.00
	cg00858483	*DENR*	12	Body	S_Shore	0.50	$3.12x10^{-6}$	0.38	0.50	$3.17x10^{-6}$	0.42	0.60	$1.03x10^{-5}$	0.97
	cg14171486	*Intergenic*	1	Intergenic	open sea	-0.20	$4.63x10^{-6}$	0.38	-0.20	$4.83x10^{-6}$	0.42	-0.30	$5.47x10^{-5}$	1.00
	cg00980592	*Intergenic*	7	Intergenic	open sea	0.10	$4.74x10^{-6}$	0.38	0.10	$5.16x10^{-6}$	0.42	0.10	$1.02x10^{-7}$	0.05
	cg08379158	*Intergenic*	2	Intergenic	N_Shelf	-0.20	$4.87x10^{-6}$	0.38	-0.20	$5.35x10^{-6}$	0.42	-0.30	$2.58x10^{-5}$	1.00
	cg26503877	*ETS1*	11	Body	N_Shore	0.10	$7.31x10^{-6}$	0.47	0.10	$7.39x10^{-6}$	0.47	0.20	$4.88x10^{-5}$	1.00
	cg18967021	*Intergenic*	1	Intergenic	open sea	-0.30	$8.01x10^{-6}$	0.47	-0.30	$7.93x10^{-6}$	0.47	-0.30	$6.75x10^{-7}$	0.16
	cg03444340	*ITGA4*	2	Body	open sea	0.40	$1.52x10^{-5}$	0.79	0.40	$1.51x10^{-5}$	0.79	0.40	$7.60x10^{-6}$	0.89

Table 5-3(Continued)

Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Model 1 (age, SVs, predicted Cells)			Model 2 (age, SVs, predicted Cells, smoking)			Model 3 (age, SVs, predicted Cells, smoking, BMI)		
						Beta[b]	P-value	FDR	Beta	P-value	FDR	Beta	P-value	FDR
Fasting proinsulin (sensitivity)[c]	cg24192505	*Intergenic*	6	Intergenic	Open sea	10.20	$5.71x10^{-7}$	0.27	10.20	$5.81x10^{-7}$	0.27	10.10	$9.32x10^{-6}$	1.00
	cg08379158	*Intergenic*	2	Intergenic	N_Shelf	10.30	$2.12x10^{-6}$	0.50	10.30	$2.31x10^{-6}$	0.54	10.30	$1.21x10^{-5}$	1.00
	cg14171486	*Intergenic*	1	Intergenic	Open sea	10.30	$3.37x10^{-6}$	0.53	10.30	$3.48x10^{-6}$	0.55	10.30	$2.38x10^{-5}$	1.00
	cg25962358	*PCDHAC2*	5	1stExon	S_Shore	10.10	$7.85x10^{-6}$	0.80	10.10	$8.09x10^{-6}$	0.81	10.10	$5.51x10^{-5}$	1.00
	cg01213645	*HECW1*	7	Body	Open sea	10.30	$2.01x10^{-5}$	0.86	10.30	$2.10x10^{-5}$	0.87	10.30	$5.58x10^{-6}$	1.00
	cg00980592	*Intergenic*	7	Intergenic	Open sea	9.50	$1.74x10^{-4}$	1.00	9.50	$1.85x10^{-4}$	1.00	9.50	$7.52x10^{-6}$	1.00
HOMA-IR	cg06500161	*ABCG1*	21	Body	S_Shore	10.30	$7.40x10^{-9}$	$3.48x10^{-3}$	10.30	$8.61x10^{-9}$	$4.05x10^{-3}$	10.20	$8.10x10^{-5}$	0.88
	cg24272697	*HOXA3*	7	5'UTR	N_Shelf	10.70	$5.87x10^{-7}$	0.14	10.80	$2.55x10^{-7}$	0.06	10.60	$6.89x10^{-6}$	0.87
	cg00606312	*KMO*	1	TSS200	Open sea	10.50	$9.01x10^{-7}$	0.14	10.50	$1.01x10^{-6}$	0.16	10.30	$6.37x10^{-5}$	0.88
	cg25690958	*SPNS2*	17	Body	Island	9.90	$2.75x10^{-6}$	0.25	9.90	$2.26x10^{-6}$	0.27	9.90	$1.71x10^{-4}$	0.88
	cg17340655	*DDHD2*	8	TSS200	Island	6.10	$3.25x10^{-6}$	0.25	6.10	$3.18x10^{-6}$	0.27	7.10	$2.82x10^{-4}$	0.90
	cg11927233	*NPM1*	5	Body	S_Shore	10.30	$1.66x10^{-6}$	0.20	10.30	$3.49x10^{-6}$	0.27	10.10	$9.67x10^{-3}$	0.97
	cg19699090	*Intergenic*	6	Intergenic	S_Shore	10.20	$7.40x10^{-5}$	0.49	10.20	$8.51x10^{-5}$	0.47	10.20	$3.99x10^{-6}$	0.87
	cg01704999	*SLC24A1*	15	Body	open sea	9.60	$5.06x10^{-4}$	0.61	9.60	$7.58x10^{-4}$	0.56	9.50	$7.77x10^{-6}$	0.87
HOMA-B	cg23436042	*Intergenic*	5	Intergenic	Open sea	10.30	$1.26x10^{-7}$	0.03	10.30	$1.05x10^{-7}$	0.03	10.20	$6.99x10^{-7}$	0.11
	cg20391220	*Intergenic*	15	Intergenic	S_Shelf	10.30	$8.28x10^{-8}$	0.03	10.30	$1.26x10^{-7}$	0.03	10.30	$5.22x10^{-8}$	0.02
	cg05497107	*RUFY1*	5	Body	Island	7.90	$1.21x10^{-6}$	0.09	7.80	$5.19x10^{-7}$	0.06	8.20	$8.53x10^{-6}$	0.28
	cg06281265	*MUC5B*	11	Body	Island	10.60	$1.28x10^{-6}$	0.09	10.60	$1.55x10^{-6}$	0.08	10.50	$1.36x10^{-6}$	0.16
	cg04870212	*Intergenic*	1	Intergenic	Open sea	8.70	$4.33x10^{-6}$	0.10	8.70	$2.16x10^{-6}$	0.08	8.90	$1.58x10^{-5}$	0.28
	cg21437157	*Intergenic*	8	Intergenic	N_Shelf	10.20	$4.20x10^{-6}$	0.10	10.30	$2.26x10^{-6}$	0.08	10.20	$2.35x10^{-7}$	0.06
	cg13340126	*SLC9A3*	5	Body	Island	10.40	$1.23x10^{-6}$	0.09	10.40	$2.40x10^{-6}$	0.08	10.30	$2.72x10^{-5}$	0.32
	cg11859607	*ADCYAP1*	18	Body	Island	10.30	$3.36x10^{-6}$	0.10	10.30	$2.42x10^{-6}$	0.08	10.20	$2.22x10^{-4}$	0.42
	cg17340655	*DDHD2*	8	TSS200	Island	6.70	$2.35x10^{-6}$	0.10	6.70	$2.47x10^{-6}$	0.08	7.40	$7.90x10^{-5}$	0.38
	cg02604018	*Intergenic*	3	Intergenic	N_Shelf	9.60	$1.83x10^{-6}$	0.10	9.60	$2.56x10^{-6}$	0.08	9.60	$1.42x10^{-5}$	0.28
	cg06500161	*ABCG1*	21	Body	S_Shore	10.20	$2.45x10^{-6}$	0.10	10.20	$2.65x10^{-6}$	0.08	10.10	$3.01x10^{-3}$	0.49
	cg20579012	*CACNA1C*	12	3'UTR	S_Shelf	10.40	$3.19x10^{-6}$	0.10	10.40	$3.47x10^{-6}$	0.10	10.30	$8.70x10^{-5}$	0.38
	cg20390515	*SNTG2*	2	TSS1500	Island	11.30	$4.98x10^{-6}$	0.11	11.40	$3.83x10^{-6}$	0.10	11.20	$7.07x10^{-6}$	0.28
	cg24335149	*PLA2G5*	1	5'UTR	Open sea	10.30	$5.85x10^{-6}$	0.11	10.30	$4.37x10^{-6}$	0.10	10.20	$6.70x10^{-4}$	0.45
	cg21276413	*GABRA5*	15	1stExon	Island	11.30	$4.07x10^{-6}$	0.10	11.30	$4.45x10^{-6}$	0.10	11.10	$3.98x10^{-6}$	0.23
	cg02241759	*NR1I2*	3	Body	N_Shore	10.20	$1.16x10^{-5}$	0.14	10.20	$4.71x10^{-6}$	0.10	10.10	$7.93x10^{-5}$	0.38

Table 5-3(Continued)

Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Model 1 (age, SVs, predicted Cells)			Model 2 (age, SVs, predicted Cells, smoking)			Model 3 (age, SVs, predicted Cells, smoking, BMI)		
						Beta[b]	P-value	FDR	Beta	P-value	FDR	Beta	P-value	FDR
HOMA-B	cg11822932	*LMO2*	11	1stExon	Open sea	10.19	$1.17x10^{-5}$	0.14	10.20	$5.22x10^{-6}$	0.11	10.14	$3.04x10^{-4}$	0.42
	cg15043106	*CRK*	17	Body	Open sea	10.52	$1.65x10^{-5}$	0.18	10.55	$5.61x10^{-6}$	0.11	10.43	$6.92x10^{-5}$	0.36
	cg16985255	*Intergenic*	6	Intergenic	Open sea	10.35	$4.28x10^{-6}$	0.10	10.35	$5.82x10^{-6}$	0.11	10.29	$1.56x10^{-5}$	0.28
	cg07815386	*RNF149*	2	TSS1500	S_Shore	9.02	$7.35x10^{-6}$	0.12	9.02	$6.46x10^{-6}$	0.11	9.23	$8.83x10^{-5}$	0.38
	cg26296653	*STIM1*	11	Body	Open sea	10.26	$5.92x10^{-6}$	0.11	10.26	$7.07x10^{-6}$	0.12	10.16	$1.49x10^{-3}$	0.47
	cg18764107	*LBR*	1	1stExon	Island	8.96	$1.01x10^{-5}$	0.14	8.95	$7.86x10^{-6}$	0.12	9.16	$7.77x10^{-5}$	0.38
	cg18232548	*DDC*	7	Body	Island	8.12	$1.07x10^{-5}$	0.14	8.10	$7.95x10^{-6}$	0.12	8.56	$2.33x10^{-4}$	0.42
	cg23228858	*TTBK1*	6	Body	N_Shelf	10.34	$2.46x10^{-5}$	0.19	10.36	$8.36x10^{-6}$	0.12	10.30	$4.64x10^{-5}$	0.35
	cg01692572	*PUF60*	8	Body	N_Shore	9.56	$1.93x10^{-5}$	0.18	9.54	$8.56x10^{-6}$	0.12	9.60	$1.38x10^{-5}$	0.28
	cg24272697	*HOXA3*	7	5'UTR	N_Shelf	10.54	$2.09x10^{-5}$	0.18	10.56	$9.16x10^{-6}$	0.12	10.42	$2.08x10^{-4}$	0.42
	cg01919011	*JDP2*	14	3'UTR	Open sea	10.67	$1.83x10^{-5}$	0.18	10.69	$9.19x10^{-6}$	0.12	10.49	$3.45x10^{-4}$	0.42
	cg20314038	*CSMD2*	1	Body	Island	9.57	$9.49x10^{-6}$	0.14	9.57	$9.96x10^{-6}$	0.13	9.70	$5.48x10^{-4}$	0.44
	cg03775372	*FRMD4A*	10	Body	Open sea	10.30	$2.71x10^{-5}$	0.19	10.30	$1.63x10^{-5}$	0.15	10.30	$2.64x10^{-6}$	0.22
	cg26923779	*HECTD1*	14	TSS1500	Island	7.23	$3.35x10^{-5}$	0.19	7.22	$2.86x10^{-5}$	0.18	7.33	$6.06x10^{-6}$	0.26
	cg26783146	*VASN*	16	5'UTR	S_Shore	10.27	$1.22x10^{-4}$	0.26	10.28	$1.10x10^{-4}$	0.23	10.28	$9.67x10^{-6}$	0.28

[a]EWAS of FG was conducted in a subsample of 1,002 female and male participants. Results are interpreted as the effect of 10% increase in methylation on a unit change in the phenotype, or as the risk of prevalent T2D per 10% increase in methylation. [b]Beta coefficients were back-transformed to the original units of the outcome using [exp(beta/100)] for outcomes normalized using the natural-logarithm; when using the reciprocal transformation, coefficients were back-transformed using [1/ (beta x 100)], and when using the Log2 transformation, coefficients were back-transformed using $[2^{(beta/100)}]$. P-value is the unadjusted p from the regression analysis, and the adjusted p-value is presented after correction for multiple testing using the FDR method (FDR<0.05). [c] Sensitivity analysis conducted in the EWAS of fasting proinsulin by excluding participants with measures of proinsulin above the 99th percentile of the total distribution of this phenotype in the sample (fasting proinsulin ≥ 30.85pmol/L, n=5 subjects).

5.3.1 Impact of removing outliers on associations between methylation and fasting proinsulin

Outliers of proinsulin were regarded as participants with proinsulin levels above or equal to 30.85pmol/L (n=5 participants), corresponding this to the 99th percentile of the distribution of the outcome in the subsample of females in ALSPAC (Figure 5-2). Based on the literature, measures of fasting proinsulin considered as extreme values could represent legitimate biological readings, and they could characterize participants at higher risk of T2D[12]. Elevated production of proinsulin at baseline reflects deterioration of β-cell function, and it is considered a marker of insulin resistance[12]. Pfutzner and Forst[13] reported that reference values of proinsulin at baseline were below 11pmol/L, whilst values above this level were indicative of β-cell disfunction and insulin resistance[13]. In the subsample of middle-age females in ALSPAC, measures of fasting proinsulin ranged between 3.50 and 80.73pmol/L, and 58 participants had proinsulin levels above the normal range. Despite this, fasting proinsulin was not considered a criterion to select disease-free participants.

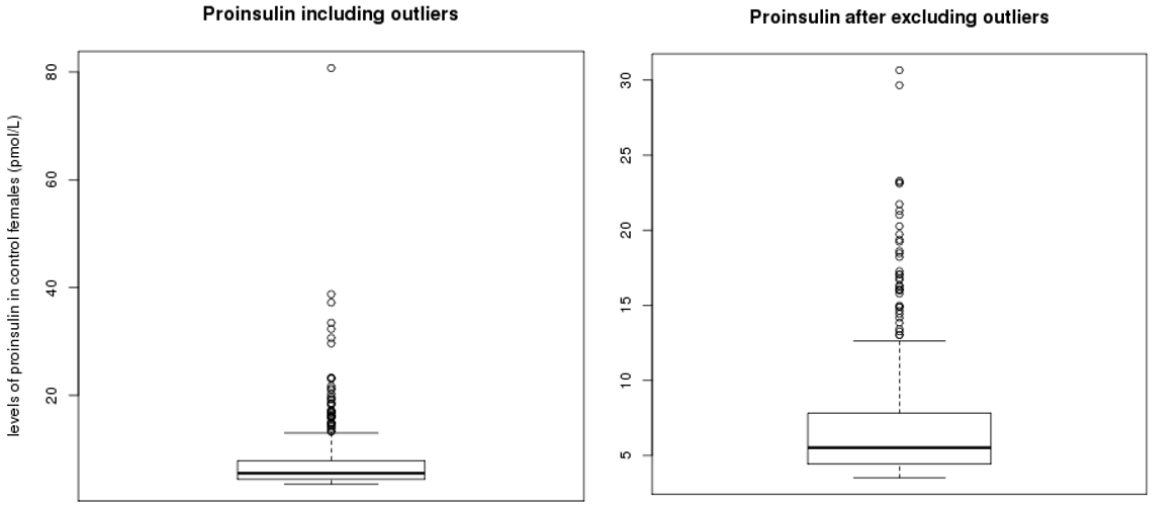


Figure 5-2 Distribution of fasting proinsulin (pmol/L) in a sample of normoglycemic females in ALSPAC including outliers of proinsulin (left), and in the same sample after exclusion of outliers (right).

According to results of the EWAS in proinsulin, none of the signals identified in the main EWAS or in the sensitivity analysis surpassed epigenome-wide significance (see Table 5-3). Differences in top-ranking signals with p-value<10^{-5} were detected across analyses (Figure 5-3). For association detected in common between the sensitivity analysis and the complete analysis at the intergenic CpG cg00980592 and cg01213645 in *HECW1*, the effect of excluding outliers was towards increasing the effect-size but reducing the statistical significance of the associations. Changes in the effects size and p-value of the associations between analyses, can also be due to the method used to transform the data before the EWAS. For the main EWAS of proinsulin, the outcome was transformed using the

reciprocal, while for the sensitivity analysis a Log2 transformation was applied. In conclusion, knowing that extreme values of proinsulin could represent legitimate measures, and that exclusion of outliers required the use of a different scale to measure values of proinsulin in the EWAS, results of the analysis including all samples were regarded as the main evidence.

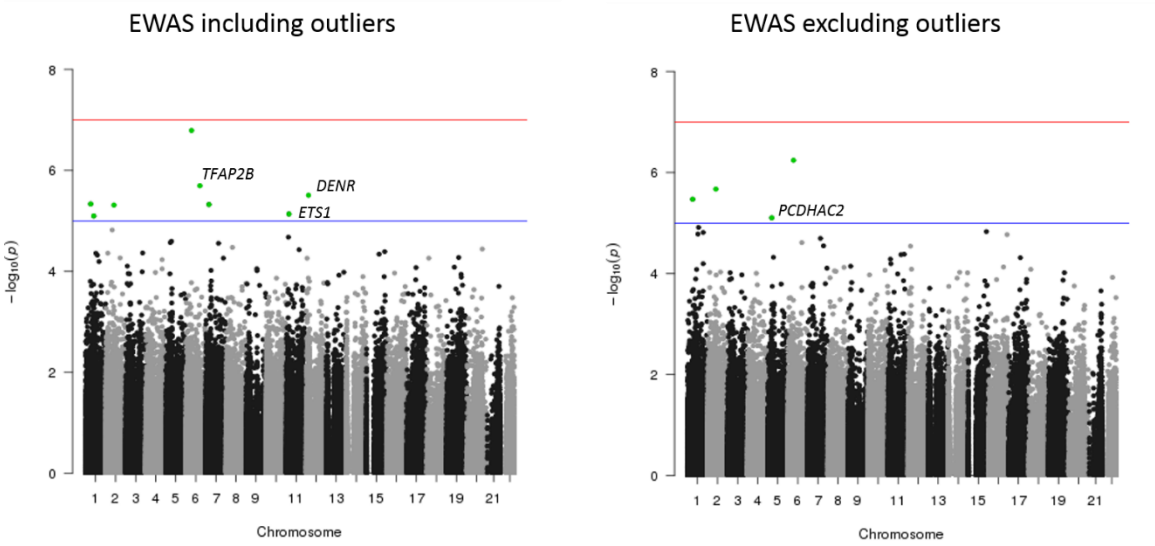


Figure 5-3 Manhattan plot comparing results of the EWAS of fasting proinsulin using a dataset including outliers (left), and a dataset pruned for outliers (right). EWAS conducted in a subset of normoglycemic females in ALSPAC (n=622). In total, 490 samples were included in the complete analysis, whilst 485 were included in the sensitivity analysis.

5.3.2 Impact of adjusting for BMI in associations between methylation and glycaemic traits

The methylation level of the intergenic CpG site cg20391220 remained associated with HOMA-B after adjustment for BMI (FDR=0.02), without detecting a change in the effect size between this model, and the model adjusted for smoking (model 2). After adjustment for BMI, some associations were newly detected with borderline significance at FDR=0.10: the cg03826430 in *PAOX* was associated with 2-h glucose, the intergenic CpG cg00980592 was associated with fasting proinsulin, and the intergenic CpG cg21437157 was associated with HOMA-B. Because for these associations no difference in the effect size was detected between the model adjusted for BMI and the model adjusted for smoking (model 2), BMI was not regarded as a confounder or a mediator in the association between methylation and HOMA-B, 2-h glucose, and fasting proinsulin.

5.3.3 Correlation in effect estimates between phenotypes

To evaluate the level of correlation between phenotypes, regression coefficients obtained in the EWAS of each trait were compared based on a subset of 11 top CpG sites identified with the smallest p-value in the EWAS of HOMA-B (model 2). Methylation sites included in the correlation analysis

were the intergenic CpG sites cg23436042, cg20391220, cg04870212, cg21437157 and cg02604018, cg05497107 in *RUFY1*, cg06281265 in *MUC5B*, cg13340126 in *SLC9A3*, cg11859607 in *ADCYAP1*, cg17340655 in *DDH2*, and cg06500161 in *ABCG1*.

Strong positive correlation was found between most of the glycaemic traits compared (Figure 5-4), except for fasting proinsulin, which was inversely correlated with all the traits and was not correlated with T2D or 2-h glucose ($p>0.01$) (Figure 5-4). In addition, 2-h glucose showed the weakest positive correlation with other traits, and no correlation was identified between T2D and any of the glycaemic traits (Figure 5-4). A cluster analysis using a heatmap (see Chapter 2 for more on heatmaps) showed that effect estimates identified in the EWAS of HOMA-B were closely related with those identified in the EWAS of fasting insulin and HOMA-IR at the specific CpG sites compared. However, less similarity was observed between effect estimates of HOMA-B, and effect estimates of fasting glucose, 2-h glucose and fasting proinsulin (Figure 5-4). T2D was distantly related to other phenotypes.

Results obtained in the correlation and in the cluster analysis, were consistent with the underlying biological characteristics of the phenotypes evaluated and the way they are measured. For instance, effect estimates detected in the HOMA scores were more correlated and closely linked to effect estimates of fasting insulin, than to effect estimates of fasting glucose and 2-h glucose, and this is because the HOMA scores rely partly on measures of fasting insulin for their estimation. HOMA scores are also dependent on levels of fasting glucose, but based on this analysis, their correlation was higher with fasting insulin than with fasting glucose.

The inverse correlation between proinsulin and other glycaemic traits was either a true observation, or the result of using a reciprocal transformation to normalize values of proinsulin before the EWAS. To test this latter concept, it will be necessary to perform the EWAS using untransformed values of proinsulin, and to measure again the correlation among traits. A quick inspection using back-transformed effect estimates of proinsulin showed consistency in the direction of the correlation between proinsulin and the remaining traits. If this inverse correlation was a true observation, the comparison between proinsulin and fasting insulin was different from what was expected knowing that proinsulin is the precursor molecule of insulin, and both molecules are secreted by β-cells in the pancreas[12, 13]. An alternative explanation for this result would be that in non-diabetic participants, levels of circulating proinsulin are lower than those of fasting insulin, with no insulin resistance (lower HOMA-IR) or β-cell disfunction (higher HOMA-B), thus supporting an inverse correlation

among traits as the one observed. According to the literature, levels of intact proinsulin are an indirect measure of insulin resistance, β-cell disfunction and T2D, as elevated levels of proinsulin are secreted in serum when there is an impairment in the secretory capacity of β-cells[12, 13].

Lack of correlation and distant association between T2D and other traits at the specific CpG sites can be explained by different factors. The first one is that for T2D methylation was measured in cases and controls, but only using controls in the EWAS of glycaemic traits. Therefore, further variation in methylation was considered in the EWAS of T2D by including T2D cases in the analysis. The second possible reason for the dissimilarity observed between T2D and other traits, is that the effect measured in this EWAS was risk rather than unit change in the outcome. Therefore, there was heterogeneity in the measurement of the outcome between the EWAS of T2D and the EWAS of glycaemic traits.

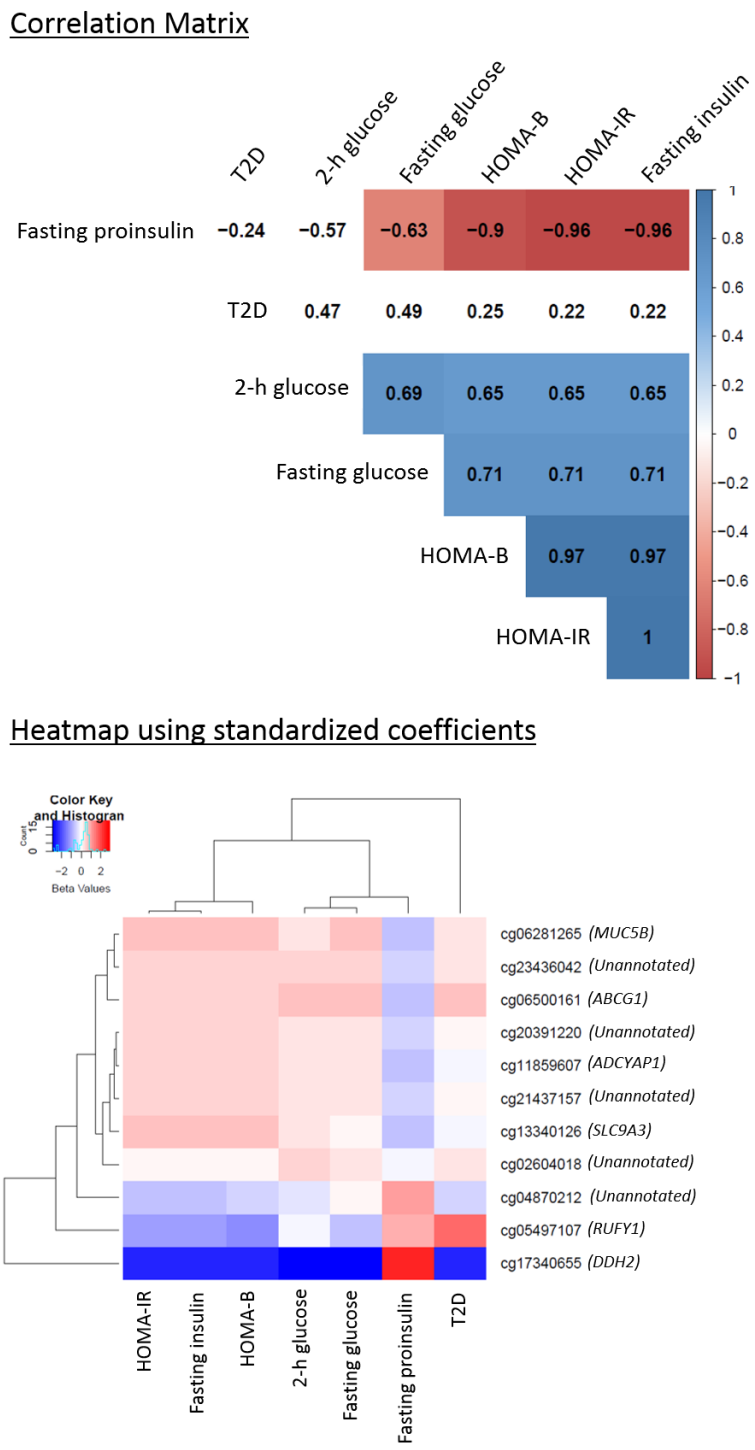


Figure 5-4 Correlogram and heatmap showing the level of correlation and similarity observed between effect estimates obtained in the EWAS of glycaemic traits and T2D in ALSPAC. Effect estimates compared across traits were 11 top-ranking CpG sites initially identified in the EWAS of HOMA-B (model 2). In the correlogram (top plot), significant correlations at $p < 0.01$ are represented by coloured boxes, while non-significant correlations are indicated by blank boxes; the strength of the correlation is indicated by the intensity of the colour (legend on the right).The heatmap (bottom plot) shows the level of similarity between traits for the CpG sites compared. Traits connected by a similar node in the dendrogram at the top, are more similar (i.e. fasting glucose and 2-h glucose) than traits connected by more distant nodes (i.e. fasting glucose and HOMA-B).

5.3.4 Overrepresentation of negative effect estimates in the EWAS of glycaemic traits and T2D in ALSPAC

A volcano plot showing the distribution of effect estimates against the -Log10(p-value) (see Figure 5-5) for EWAS results, revealed that there was an overrepresentation of negative effects among probes with low significance and large effect estimates, and this was common to all the outcomes evaluated. The observed distribution of effect estimates was unusual and different from the expected distribution of effects in a volcano plot, as it has been described elsewhere[143]. Possible explanations for this abnormal distribution include errors in the analysis, small sample size and outliers in the methylation or outcome variables. Effect estimates were centred around zero and standard errors were right-skewed, suggesting no evidence of errors in the analysis. However, sample sizes were small (n ranging 622 to 1050) and there was evidence of outliers in the distribution of methylation among probes with large negative effect estimates (defined as probes with $p>1.0x10^{-5}$ and untransformed effect estimate <-20 for most traits, <-2.5 for proinsulin, and <-200 for T2D). Outliers were detected for most of the glycaemic traits, except for FG, 2-h glucose and T2D (Table 5-4).

Table 5-4 Identifying outliers (extreme methylation values) among probes with large negative effects and low significance in the EWAS of T2D and glycaemic traits in ALSPAC.

	Lower effect estimate	Transformed effect estimate	Min p-value	Probes	Min-beta	Max-beta	Extreme beta
Fast. glucose	-20.0	-0.20	$1.0x10^{-5}$	649	0.01	0.02	0
2hours glucose	-20.0	-0.20	$1.0x10^{-5}$	1321	0.01	0.03	0
Fast. insulin	-20.0	0.82	$1.0x10^{-5}$	3879	0.01	0.98	1
Fast. proinsulin	-2.5	$-4.0x10^{-3}$	$1.0x10^{-5}$	2061	0.01	0.99	22
HOMA-IR	-20.0	-0.20	$1.0x10^{-5}$	4329	0.01	0.98	1
HOMA-B	-20.0	-0.20	$1.0x10^{-5}$	2837	0.01	0.98	1
T2D	-200.0	0.14	$1.0x10^{-5}$	1053	0.01	0.02	0

Lower effect estimate: lower bound of the negative effect estimate; transformed effect estimate: lower bound of the negative effect estimate transformed-back to the natural units of measurement of the outcome; Min p-value: minimum p-value considered to retrieve probes with large negative effects; Probes: number of probes with large p-values and large negatives effects; Min-beta and Max-beta: minimum and maximum median value of methylation among probes with large negative effects; extreme beta: number of probes with median values of methylation three times below the interquartile-range for the distribution of methylation in probes with large negative effects.

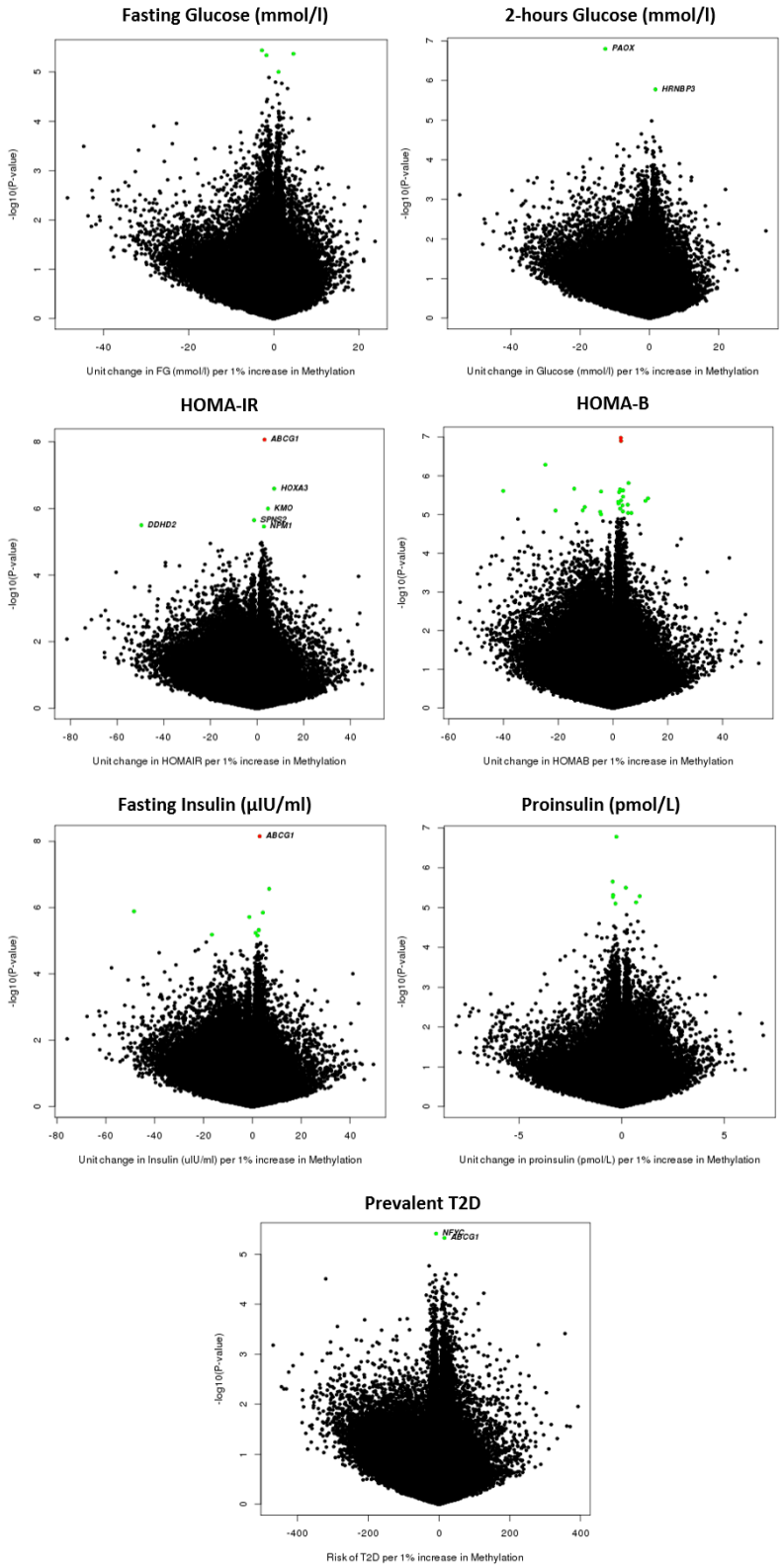


Figure 5-5 Volcano plots showing the distribution of effect estimates against the -log10(p-value) for associations detected in the EWAS glycaemic traits and T2D in ALSPAC. Results correspond to the model adjusted for age, SVs, 6-Houseman cells and smoking. Associations were regarded borderline significant at -Log10(p-value) ≤ 5.0 or $p < 1.0x10^{-5}$ (green-dots), and significant at -Log10(P-value) ≥ 7.0 or $p < 1.07x10^{-7}$ (red-dots). Probes with large negative effects and small significance, were overrepresented across these EWAS.

5.4 Replication of the EWAS of glycaemic traits in SABRE

Results of the EWAS of glycaemic traits in ALSPAC suggested power limitations to detect strong associations due to a reduced sample-size. Thus, the next step was to replicate the different EWAS in a second cohort with similar characteristics as participants in ALSPAC. Replication was conducted in a subsample from the Southall and Brent areas in London-study, SABRE. Results of the EWAS in ALSPAC and SABRE were combined via meta-analysis. Further detail of the cohort, assessment of the methylation data, measurement of glycaemic traits, available covariates, and selection of samples, was presented earlier in Chapter 2.

5.4.1 Baseline characteristics of the subsample

The SABRE subsample comprised 382 normoglycemic males, aged between 40 and 67 years of age, with methylation data profiled at their middle-age, and with availability of measures for different glycaemic traits and relevant covariates. Baseline characteristics of this subsample are presented in Table 5-5. Briefly, median values of total cholesterol, FG, 2-h glucose, HbA1c, fasting insulin, 2-h insulin, HOMA-IR and HOMA-B in SABRE were 6.08 mmol/l, 5.33mmol/l, 4.83mmol/l, 5.52%, 7.35 IU/L, 19.75 IU/L, 0.80 and 72.40, respectively (Table 5-5).

Stratification of the subsample by categories of glucose tolerance (WHO, 1999) revealed that 8.12% of these participants had impaired fasting glucose (IFG), another 1.31% of them had impaired glucose tolerance (IGT), 9.4% participants had either impaired fasting glucose or impaired glucose tolerance, and 89.53% of the total subsample had normal glucose tolerance (NGT) (Table 5-5).

Table 5-5 Baseline characteristics of a subsample of normoglycemic males in SABRE included in the EWAS of glycaemic traits (n=382).Continuous variables were described using the median and the 25th and 75th percentile of the distribution of the outcome, while categorical variables were summarized using the proportion of samples per category of the outcome.

	EWAS of glycaemic traits (n= 382)	
	Median (25th; 75th percentile)	%
Age (years)	51.00 (46.00; 58.00)	-
Sex [% male]	-	100
Cholesterol (mmol/l)	6.08 (5.34; 6.72)	-
Triglycerides (mmol/l)	1.39 (0.99; 2.03)	-
HDL cholesterol (mmol/l)	1.25 (1.05; 1.49)	-
LDL cholesterol (mmol/l) [a]	3.98 (3.38; 4.55)	-
Fasting glucose (mmol/l)	5.33 (5.03; 5.67)	-
2-hours serum glucose (mmol/l)	4.83 (4.10; 5.65)	-
Fasting Insulin (IU/L)	7.35 (4.78; 10.60)	-
2-hours Insulin	19.75 (11.40; 36.30)	-
HbA1c (%)	5.52 (5.32; 5.74)	-
HOMA-IR	0.80 (0.50; 1.20)	-
HOMA-B	72.40 (54.95; 92.90)	-
Glucose Tolerance status [%] [b]		
NFG	-	89.53
IFG	-	8.12
IGT	-	1.31
Combined IFG and IGT	-	9.40
Waist circumference (cm)	90.70 (84.20; 98.75)	-
BMI (kg/m^2)	25.58 (23.75; 28.09)	-
Systolic BP (mmHg)	121.00 (111.00; 132.00)	-
Diastolic BP (mmHg)	77.00 (70.00; 83.00)	-
Diabetes [% Yes]	-	0.00
Medication [% non-treated]	-	100
Smoking [%]		
never smoker	-	26.96
Ex-smoker	-	40.58
Current-smoker	-	32.20
Physical activity [MJ/wk score]	10.77 (6.56; 16.56)	-
Socioeconomic status [%]		
manual	-	45.55
non-manual	-	11.78
others	-	42.67
Predicted Cell-counts		
CD4T	0.14 (0.11; 0.18)	-
CD8T	<0.01 (<0.01; 0.02)	-
Natural Killer Cells	0.15 (0.13; 0.18)	-
B-cells	0.07 (0.06; 0.08)	-
Monocytes	0.10 (0.09; 0.12)	-
Eosinophils	<0.01 (<0.01; <0.01)	-
Neutrophils	0.56 (0.51; 0.62)	-

[a] LDL was calculated using the Friedewald Formula (1972): LDL= Total Cholesterol - HDL - (Triglycerides/2.2)[216]. Lipid measures were taken during fasting and measured in mmol/l. [b] Glucose tolerance status was defined using WHO criteria (1999). IFG: impaired fasting glucose if FG ≥ 6.1 and FG < 7.0 mmol/l. IGT: impaired glucose tolerance if FG < 7.0mmol/l and 2h-glucose ≥ 7.8 mmol/l and 2h-glucose < 11.1mmol/l. NGT: normal glucose tolerance if FG < 6.1 and 2h-Glucose < 7.8 mmol/l.

5.4.2 Main findings of the EWAS in SABRE

Table 5-6 summarizes top associations identified with $p<1.0x10^{-5}$ in the EWAS glycaemic traits in SABRE. Similar to analyses in ALSPAC, associations identified in SABRE were independent of probes located in sex-chromosomes, probes with SNPs in the probe-binding region, and probes reported in the Naeem list as problematic probes (see Chapter 2). Overall, it was evident that in SABRE there was less power to detect strong associations than in ALSPAC due to the smaller sample analysed in this study (n=382). In SABRE, no association was detected at FDR<0.05 or FDR<0.10 for fasting glucose, fasting insulin, 2-h glucose, 2-h insulin, and the HOMA scores, but strong associations at FDR<0.05 were identified between HbA1c and methylation at the CpG sites cg12671247 (*RAD1*), cg26316702 (*TEKT4*), cg13583414 (*LZTS1*), all of them surpassing adjustment for age, batch effects, predicted cell-counts, smoking and BMI. Figure 5-6 summarises results of the EWAS in HbA1c.

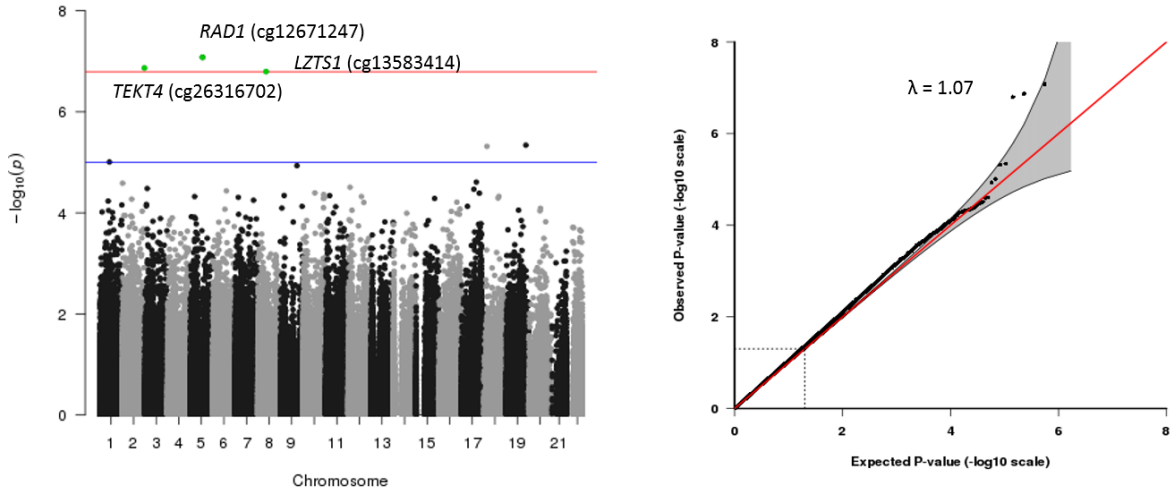


Figure 5-6 Manhattan (left) and QQ-plot (right) showing results of the EWAS of HbA1c conducted in a subsample of males in SABRE (n=382). Results correspond to a model adjusted for age, 9 SVs, 7 Houseman cells and smoking (never, former and current smoker). Manhattan: red line represents the Benjamin-Hochberg threshold of significance at $p < 1.60x10^{-7}$ or FDR < 0.05, while blue-line represents the threshold of borderline significance at $p< 1.0x10^{-5}$. Three CpG sites were identified with FDR significance. The QQ-plot shows the distribution of observed versus expected p-values under the null hypothesis of no associations. The lambda reported for this EWAS suggested weak evidence of genomic inflation (λ ~ 1.0). Diagonal red line represents the line of null associations, and three markers significantly deviated from this line, demonstrating strong evidence of association with HbA1c.

Looking at the distribution of expected versus observed p-values, there was no suggestion of high genomic inflation in results of the EWAS in SABRE (λ ranged between 0.95 to 1.07), except for two traits where the lambda was considerably higher than 1.0: λ of 1.14 in the EWAS of 2-h glucose, and λ of 1.11 in the EWAS of HbA1c. A volcano plot showed very small evidence of skewness in the distribution of effect estimates across EWAS, and this was different from results obtained in ALSPAC

(see section 5.3.4). In SABRE, phenotypes with the highest skewness in the distribution of the effect estimates were 2-h insulin and HbA1c (appendix Figure S8-13). Traits that were meta-analysed between ALSPAC and SABRE were fasting glucose, fasting insulin, 2-h glucose, HOMA-IR and HOMA-B. Contrary, the EWAS of HbA1c and 2-h insulin were not meta-analysed as these traits were only measured in SABRE, and fasting proinsulin was only measured in ALSPAC. Even though not derived from a meta-analysis, probes detected in strong association with HbA1c in SABRE were combined in a weighted methylation score to investigate the proportion of variance in HbA1c explained by methylation at the top-ranking CpG sites of this EWAS (see section 5.14).

Table 5-6 Main findings of the EWAS of glycaemic traits in the subsample of SABRE (n=382 males). Additional traits analysed in SABRE that were not considered in ALSPAC were HbA1c and 2-h insulin. Fasting proinsulin was not available in SABRE. Top associations were identified with p< 1.0x10^{-5} in at least one of the three adjustment models implemented. Highlighted in bold are associations surpassing genome-wide significance at FDR<0.05.

						Model 1 (age, SVs, predicted cells)			Model 2 (age, SVs, predicted cells, smoking)			Model 3 (age, SVs, predicted cells, smoking, BMI)		
Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Beta	P-value	FDR	Beta	P-value	FDR	Beta	P-value	FDR
Fasting glucose	cg03145924	*Intergenic*	1	intergenic	N_Shore	-0.30	7.80x10^{-6}	0.94	-0.30	5.68x10^{-6}	0.69	-0.30	5.59x10^{-6}	0.68
	cg19581424	*EML4*	2	TSS200	Island	0.80	7.91x10^{-6}	0.94	0.70	2.86x10^{-5}	0.94	0.70	1.23x10^{-4}	0.93
	cg22667294	*GFPT2*	5	Body	open sea	0.30	3.43x10^{-5}	1.00	0.30	1.22x10^{-5}	0.72	0.30	8.62x10^{-6}	0.68
	cg12754421	*GRIA4*	11	TSS200	N_Shore	-0.60	2.31x10^{-5}	1.00	-0.60	9.89x10^{-6}	0.72	-0.60	6.22x10^{-6}	0.68
	cg23620184	*PPP1R3E*	14	TSS1500	S_Shore	-0.30	5.81x10^{-6}	0.94	-0.30	5.60x10^{-6}	0.69	-0.30	1.18x10^{-5}	0.76
	cg00179070	*COL1A1*	17	TSS200	Island	0.40	4.61x10^{-6}	0.94	0.40	5.45x10^{-6}	0.69	0.40	6.10x10^{-6}	0.68
	cg06560588	*RUVBL2*	19	Body	S_Shelf	-0.40	1.29x10^{-4}	1.00	-0.50	2.97x10^{-5}	0.94	-0.50	8.20x10^{-6}	0.68
2-h glucose	cg18828883	*EFNA3*	1	Body	Island	-3.30	6.78x10^{-5}	1.00	-3.70	1.68x10^{-6}	0.26	-3.70	1.84x10^{-6}	0.29
	cg00719108	*ABCB6*	2	Body	N_Shelf	-5.70	2.76x10^{-6}	0.65	-4.90	7.24x10^{-5}	0.44	-4.80	1.09x10^{-4}	0.41
	cg09593402	*LNPEP*	5	TSS1500	N_Shore	-2.30	6.58x10^{-4}	1.00	-2.80	1.16x10^{-5}	0.44	-2.80	6.71x10^{-6}	0.33
	cg02391713	*Intergenic*	6	intergenic	Island	0.90	1.31x10^{-4}	1.00	1.00	6.10x10^{-6}	0.41	1.00	7.84x10^{-6}	0.33
	cg10162464	*ZNF282*	7	5'UTR	Island	-2.90	9.88x10^{-4}	1.00	-3.10	1.36x10^{-5}	0.44	-3.20	6.82x10^{-6}	0.33
	cg14034325	*HNRNPF*	10	5'UTR	N_Shore	0.60	1.04x10^{-5}	0.94	0.60	6.92x10^{-6}	0.41	0.60	4.62x10^{-6}	0.33
	cg21591624	*SLC6A5*	11	Body	Island	-1.10	1.73x10^{-5}	0.94	-1.20	5.69x10^{-6}	0.41	-1.20	3.85x10^{-6}	0.33
	cg22501608	*IDH3A*	15	Body	open sea	-0.70	1.18x10^{-5}	0.94	-0.80	1.05x10^{-6}	0.25	-0.80	1.05x10^{-6}	0.25
	cg10413513	*A2BP1*	16	Body	open sea	1.10	1.78x10^{-5}	0.94	1.20	6.49x10^{-6}	0.41	1.20	4.20x10^{-6}	0.33
HbA1c	cg07830269	*FMN2*	1	Body	open sea	-0.10	3.31x10^{-6}	0.22	-0.10	9.85x10^{-6}	0.47	-0.10	1.35x10^{-5}	0.57
	cg26316702	***TEKT4***	2	Body	S_Shore	**-0.30**	**1.94x10^{-7}**	**0.03**	**-0.30**	**1.35x10^{-7}**	**0.03**	**-0.30**	**1.85x10^{-7}**	**0.03**
	cg16382047	*GPR55*	2	TSS200	open sea	-0.20	6.44x10^{-6}	0.28	-0.20	5.35x10^{-5}	0.60	-0.20	6.38x10^{-5}	0.60
	cg02365780	*NPAS2*	2	5'UTR	S_Shore	-0.10	9.79x10^{-6}	0.35	-0.10	2.59x10^{-5}	0.60	-0.10	2.55x10^{-5}	0.60
	cg12671247	***RAD1***	5	Body	open sea	**-0.20**	**9.99x10^{-8}**	**0.03**	**-0.20**	**8.32x10^{-8}**	**0.03**	**-0.20**	**4.86x10^{-8}**	**0.02**
	cg13583414	***LZTS1***	8	Body	S_Shelf	**0.20**	**1.21x10^{-7}**	**0.03**	**0.20**	**1.60x10^{-7}**	**0.03**	**0.20**	**2.13x10^{-7}**	**0.03**
	cg07211259	*PDCD1LG2*	9	TSS200	open sea	0.20	6.04x10^{-6}	0.28	0.20	1.16x10^{-5}	0.50	0.20	1.44x10^{-5}	0.57
	cg13249876	*ARHGAP12*	10	TSS200	Island	-0.90	8.96x10^{-6}	0.35	-0.90	4.03x10^{-5}	0.60	-0.90	4.67x10^{-5}	0.60
	cg12802356	*SEH1L*	18	5'UTR	Island	-2.60	2.86x10^{-6}	0.22	-2.50	4.83x10^{-6}	0.25	-2.50	5.77x10^{-6}	0.30
	cg06698707	*C19orf21*	19	Body	S_Shore	-0.10	1.33x10^{-5}	0.37	-0.10	4.57x10^{-6}	0.25	-0.10	4.96x10^{-6}	0.29
Fasting insulin	cg26223536	*Intergenic*	1	intergenic	open sea	10.70	4.27x10^{-6}	1.00	10.70	3.57x10^{-6}	1.00	10.50	8.67x10^{-5}	1.00
	cg18383603	*CLDN19*	1	3'UTR	N_Shelf	9.30	4.34x10^{-5}	1.00	9.30	2.17x10^{-5}	1.00	9.30	6.60x10^{-6}	1.00
	cg07146535	*Intergenic*	1	intergenic	Island	17.20	3.99x10^{-4}	1.00	17.10	3.63x10^{-4}	1.00	18.50	9.21x10^{-6}	1.00
	cg23100428	*SNAI1*	20	Body	S_Shore	10.50	3.04x10^{-5}	1.00	10.50	1.65x10^{-5}	1.00	10.50	1.47x10^{-6}	0.70

Table 5-6(Continued)

Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Model 1 (age, SVs, predicted cells)			Model 2 (age, SVs, predicted cells, smoking)			Model 3 (age, SVs, predicted cells, smoking, BMI)		
						Beta	P-value	FDR	Beta	P-value	FDR	Beta	P-value	FDR
2-hours Insulin	cg00982271	*ECE1*	1	TSS1500	S_Shore	7.70	$2.18x10^{-5}$	0.32	7.70	$2.23x10^{-5}$	0.34	7.50	$1.53x10^{-6}$	0.13
	cg07246449	*SLC27A3*	1	1stExon	Island	8.00	$4.42x10^{-5}$	0.38	7.90	$1.11x10^{-5}$	0.34	7.90	$4.19x10^{-6}$	0.16
	cg14198172	*HBXIP*	1	TSS200	S_Shore	4.80	$2.89x10^{-6}$	0.25	4.60	$7.65x10^{-7}$	0.23	5.10	$8.32x10^{-6}$	0.16
	cg05042697	*NOL10*	2	TSS1500	S_Shore	12.00	$5.44x10^{-6}$	0.25	12.00	$6.36x10^{-6}$	0.34	11.70	$5.42x10^{-5}$	0.24
	cg03953709	*NR4A2*	2	TSS1500	Island	7.50	$1.45x10^{-6}$	0.25	7.50	$1.54x10^{-6}$	0.23	7.70	$5.27x10^{-6}$	0.16
	cg11458642	*ATR*	3	Body	Island	7.20	$1.09x10^{-4}$	0.45	7.00	$1.66x10^{-5}$	0.34	7.00	$8.20x10^{-6}$	0.16
	cg21176026	*GUCY1A3*	4	5'UTR	S_Shelf	10.30	$6.98x10^{-6}$	0.25	10.40	$1.91x10^{-6}$	0.23	10.30	$1.61x10^{-6}$	0.13
	cg01203651	*Intergenic*	4	Intergenic	open sea	9.20	$8.06x10^{-6}$	0.25	9.30	$9.80x10^{-6}$	0.34	9.30	$6.56x10^{-6}$	0.16
	cg26460247	*KLHL32*	6	5'UTR	Island	7.80	$5.40x10^{-5}$	0.40	7.70	$2.07x10^{-5}$	0.34	7.50	$7.90x10^{-7}$	0.13
	cg01190522	*Intergenic*	6	Intergenic	Island	7.90	$9.80x10^{-5}$	0.45	7.90	$6.96x10^{-5}$	0.44	7.70	$2.68x10^{-6}$	0.16
	cg00221035	*C1GALT1*	7	1stExon	Island	8.10	$1.10x10^{-4}$	0.45	8.00	$1.42x10^{-5}$	0.34	8.00	$7.11x10^{-6}$	0.16
	cg25245261	*GSDMD*	8	1stExon	Island	7.40	$6.42x10^{-5}$	0.40	7.40	$3.40x10^{-5}$	0.37	7.30	$4.04x10^{-6}$	0.16
	cg21511069	*UBAP2*	9	TSS1500	Island	6.70	$6.00x10^{-6}$	0.25	6.60	$1.98x10^{-6}$	0.23	6.90	$1.13x10^{-5}$	0.18
	cg25945303	*KLF6*	10	TSS1500	Island	7.70	$8.31x10^{-5}$	0.44	7.60	$2.73x10^{-5}$	0.35	7.30	$9.09x10^{-7}$	0.13
	cg05479554	*CHCHD8*	11	TSS1500	Island	6.80	$3.25x10^{-6}$	0.25	6.80	$2.73x10^{-6}$	0.26	7.10	$1.20x10^{-5}$	0.18
	cg00430080	*GLB1L2*	11	Body	S_Shore	9.10	$4.41x10^{-6}$	0.25	9.10	$5.55x10^{-6}$	0.34	9.20	$4.64x10^{-5}$	0.23
	cg25634032	*LOC100130987*	11	Body	N_Shelf	8.90	$1.89x10^{-5}$	0.32	8.90	$7.51x10^{-6}$	0.34	8.90	$1.05x10^{-5}$	0.18
	cg23989912	*USP35*	11	TSS1500	Island	7.70	$1.09x10^{-4}$	0.45	7.70	$6.07x10^{-5}$	0.43	7.50	$5.72x10^{-6}$	0.16
	cg09168808	*Intergenic*	12	Intergenic	N_Shore	9.50	$2.42x10^{-5}$	0.32	9.50	$3.19x10^{-5}$	0.37	9.50	$6.34x10^{-6}$	0.16
	cg23005387	*CCND2*	12	TSS200	N_Shore	7.70	$5.85x10^{-5}$	0.40	7.70	$4.48x10^{-5}$	0.39	7.60	$6.63x10^{-6}$	0.16
	cg27022663	*SALL1*	16	3'UTR	S_Shore	9.20	$5.68x10^{-5}$	0.40	9.20	$1.44x10^{-5}$	0.34	9.20	$7.95x10^{-6}$	0.16
	cg05550919	*ACLY*	17	5'UTR	Island	7.70	$2.95x10^{-5}$	0.33	7.90	$5.39x10^{-5}$	0.43	7.70	$7.08x10^{-6}$	0.16
	cg00284005	*PIGN*	18	TSS1500	Island	7.80	$4.63x10^{-4}$	0.54	7.60	$5.69x10^{-5}$	0.43	7.40	$4.55x10^{-6}$	0.16
	cg02878102	*ZNF320*	19	TSS1500	open sea	8.60	$1.53x10^{-5}$	0.32	8.60	$9.43x10^{-6}$	0.34	8.70	$1.54x10^{-5}$	0.18
	cg07806715	*NAPA*	19	5'UTR	Island	8.00	$2.70x10^{-5}$	0.32	7.90	$1.34x10^{-5}$	0.34	7.80	$6.97x10^{-7}$	0.13
	cg18134732	*ZNRF3*	22	TSS1500	N_Shore	16.40	$1.23x10^{-5}$	0.28	16.00	$2.38x10^{-5}$	0.34	16.20	$4.52x10^{-6}$	0.16
	cg16841327	*Intergenic*	22	Intergenic	Island	15.40	$7.86x10^{-5}$	0.43	15.70	$3.45x10^{-5}$	0.37	15.90	$7.38x10^{-6}$	0.16

Table 5-6(Continued)

Phenotype	CpG	Gene	Chr	Genetic Context	CpG island context	Model 1 (age, SVs, predicted Cells)			Model 2 (age, SVs, predicted Cells, smoking)			Model 3 (age, SVs, predicted Cells, smoking, BMI)		
						Beta	P-value	FDR	Beta	P-value	FDR	Beta	P-value	FDR
HOMA-IR	cg22047013	*PLEKHG4B*	5	Body	Island	9.80	$6.49x10^{-5}$	1.00	9.80	$1.13x10^{-5}$	1.00	9.80	$3.91x10^{-6}$	0.92
	cg01044961	*DIO3*	14	TSS1500	Island	11.00	$2.86x10^{-5}$	1.00	11.10	$1.65x10^{-5}$	1.00	11.00	$9.31x10^{-6}$	1.00
	cg23100428	*SNAI1*	20	Body	S_Shore	10.50	$3.18x10^{-5}$	1.00	10.50	$1.72x10^{-5}$	1.00	10.50	$1.47x10^{-6}$	0.70
HOMA-B	cg04994405	*DNAJC6*	1	Body	Island	10.40	$8.34x10^{-7}$	0.32	10.40	$1.79x10^{-6}$	0.84	10.40	$1.94x10^{-6}$	0.92
	cg21566642	*Intergenic*	2	Intergenic	Island	10.10	$1.34x10^{-6}$	0.32	10.10	$4.62x10^{-2}$	1.00	10.10	$9.17x10^{-2}$	1.00
	cg01940273	*Intergenic*	2	Intergenic	Island	10.20	$3.14x10^{-6}$	0.49	10.10	$4.30x10^{-2}$	1.00	10.10	$1.32x10^{-1}$	1.00
	cg05951221	*Intergenic*	2	Intergenic	Island	10.20	$5.30x10^{-6}$	0.63	10.10	$6.55x10^{-2}$	1.00	10.10	$6.95x10^{-2}$	1.00
	cg10156077	*Intergenic*	6	Intergenic	open sea	9.70	$6.63x10^{-6}$	0.63	9.70	$1.77x10^{-5}$	1.00	9.80	$3.56x10^{-5}$	1.00
	cg06708215	*Intergenic*	10	Intergenic	open sea	9.50	$2.20x10^{-5}$	0.93	9.50	$2.14x10^{-5}$	1.00	9.50	$6.59x10^{-6}$	1.00

Results are interpreted as the effect of 10% increase in methylation on a unit change in the outcome. Beta coefficients were transformed to the original units of the outcome using [exp(beta/100)] for outcomes normalized using the natural logarithm. P-value is the unadjusted-p, and adjusted p-values are reported after FDR correction (FDR<0.05).

5.5 Meta-analysis of EWAS of glycaemic traits

A fixed-effect inverse variance weighted meta-analysis was conducted in METAL (version 2011-03-25)[144] to summarize evidence of the individual EWAS performed for fasting glucose, fasting insulin, 2-h glucose, HOMA-IR and HOMA-B, using ALSPAC and SABRE. The main difference between the two studies was in the proportion of males and females, where in ALSPAC only females were included in most of the analyses, except for the EWAS of FG (38% males), while in SABRE all the analyses were conducted in males (see section 5.4.1). In addition, differences between studies were detected in age and in the mean of the outcomes meta-analysed (Table 5-7). No significant difference was observed in BMI across studies. Total sample size included in the different meta-EWAS ranged between 980 and 1384. The overall sample size-weighted mean of age, BMI, FG, fasting insulin, 2-h glucose, HOMA-IR and HOMA-B, was 51 years, 26.33 kg/m^2, 5.30mmol/l, 4.19 µIU/mL, 4.54mmol/l, 1.16 and 71.66, respectively (Table 5-7).

Table 5-7 Comparison of baseline characteristics across ALSPAC and SABRE for variables considered in the meta-EWAS of glycaemic traits. ALSPAC1 refers to the subsample of 1002 normoglycemic females and males included in EWAS of FG, and ALSPAC2 refers to the subsample of 622 normoglycemic females included in the EWAS of other glycaemic traits. Continuous outcomes were summarized using the mean and standard deviation (SD).

	ALSPAC1	ALSPAC2	SABRE	P	Weighted mean	Total
Ethnicity	European	European	European		-	-
Sex [% males]	0.38	0.00	100.00		-	-
Age	49.96 (5.28)	47.94 (4.14)	52.26 (7.14)	<0.01[c]	50.59	-
BMI	26.44 (4.51)	25.97 (4.77)	26.04 (3.65)	0.81[c]	26.33	-
N continuous FG [a]	1,002	622	382		-	1384
FG	5.27 (0.46)	5.16 (0.40)	5.37 (0.51)	<0.01[c]	5.30	-
N continuous 2-h glucose	-	601	379		-	980
2-h glucose	-	4.31 (0.39)	4.91 (1.21)	<0.01[d]	4.54	-
N continuous Fasting insulin	-	619	380		-	999
Fasting Insulin (SD)	-	1.53 (0.58)	8.52 (5.58)	<0.01[d]	4.19	-
N continuous HOMA-IR	-	619	380		-	999
HOMA-IR	-	1.28 (0.61)	0.97 (0.62)	<0.01[d]	1.16	-
N continuous HOMA-B	-	619	379		-	998
HOMA-B	-	67.71 (0.54)	78.11 (33.15)	<0.01[d]	17.66	-
Glucose Tolerance						
Total N NFG	-	588	342	<0.01[d]	-	-
Total N IFG + IGT [b]	52	14	38		-	-

[a] Fasting glucose was the only glycaemic trait in ALSPAC where measures were available for females and male. Therefore, the subsample composed of females and males was the one included in the meta-analysis of FG. [b] IFG and IGT were combined to reach the minimum number of samples per glucose-tolerance category to apply a χ^2 test. [c] P-value corresponding to the comparison of means between ALSPAC1 (males and females) and SABRE. [d] P-value corresponding to the comparison of means between ALSPAC2 (only females) and SABRE.

5.5.1 QC inspection before meta-EWAS

EWAS results from ALSPAC and SABRE were passed through a quality control (QC) pipeline (see Chapter 2 for a detailed description) before inclusion in the meta-analysis to determine validity in the parameters assessed in the individual EWAS. The QC reports (summarised in Table 5-8) showed that EWAS estimates in SABRE were less precise (larger distribution) than those in ALSPAC (see appendix Figure S8-14), as expected given the smaller sample size in this study. In general, there was small evidence of genomic inflation in the EWAS across studies. The largest lambda was seen in the EWAS of fasting insulin in ALSPAC (λ=1.50). After applying QC, 374,901 autosomal probes remained in the methylation dataset in ALSPAC to conduct the meta-analysis of FG, and another 374,313 probes for the meta-analysis of other glycaemic traits. After QC in SABRE, 376,320 autosomal probes were retained for the meta-analysis.

Table 5-8 QC report before the meta-analysis for estimates obtained in the EWAS of glycaemic traits conducted in ALSPAC and SABRE. Results of the EWAS correspond to model 2, adjusted for age, (sex optional), SVs, predicted cell-counts and smoking. Beta refers to the regression coefficient, Min-P is the smallest P-value detected in the EWAS, Lambda is the genomic inflation factor (high if λ>1.0), and Median SE is the median of the distribution of the standard error.

Study	Trait	Probes after QC	Min-beta	Max-beta	Median SE	Min-P	Lambda
ALSPAC	Fasting glucose	374,901	-48.47	23.76	0.51	3.63×10^{-6}	0.95
	Fasting insulin	374,313	-87.19	43.37	0.89	1.54×10^{-7}	1.50
	2-h glucose	374,313	-54.79	33.57	0.59	1.58×10^{-7}	0.95
	HOMA-IR	374,313	-91.64	46.57	0.95	3.57×10^{-7}	1.46
	HOMA-B	374,313	-67.93	35.55	0.83	1.33×10^{-7}	1.08
SABRE	Fasting glucose	376,320	-100.53	111.26	1.26	5.45×10^{-6}	0.97
	Fasting insulin	376,320	-107.31	94.96	1.38	3.57×10^{-6}	0.96
	2-h glucose	376,320	-293.89	269.50	2.88	1.05×10^{-6}	1.11
	HOMA-IR	376,320	-109.87	93.65	1.38	1.11×10^{-5}	0.96
	HOMA-B	376,320	-82.55	79.64	0.95	1.79×10^{-6}	0.95

5.5.2 Main results of the meta-EWAS of glycaemic traits

A total of 376,415 autosomal probes were included in the meta-analysis. Most of the associations with genome-wide significance (Bonferroni correction for 376,415 tests $p < 1.33 \times 10^{-7}$ or FDR<0.05) and borderline significance ($p < 1.0 \times 10^{-5}$) were identified in model 1 (i.e. 9 and 51 sites, respectively), compared to those identified in model 2 (i.e. 8 and 44 sites, respectively) and model 3 (i.e. 3 and 17 sites, respectively) (Table 5-9). Figure 5-7 illustrates this reduction in the number of signals across models by means of a volcano plot generated for results of the meta-EWAS of fasting insulin and HOMA-IR, two of the outcomes where a higher number of significant associations were detected across models, in addition to HOMA-B.

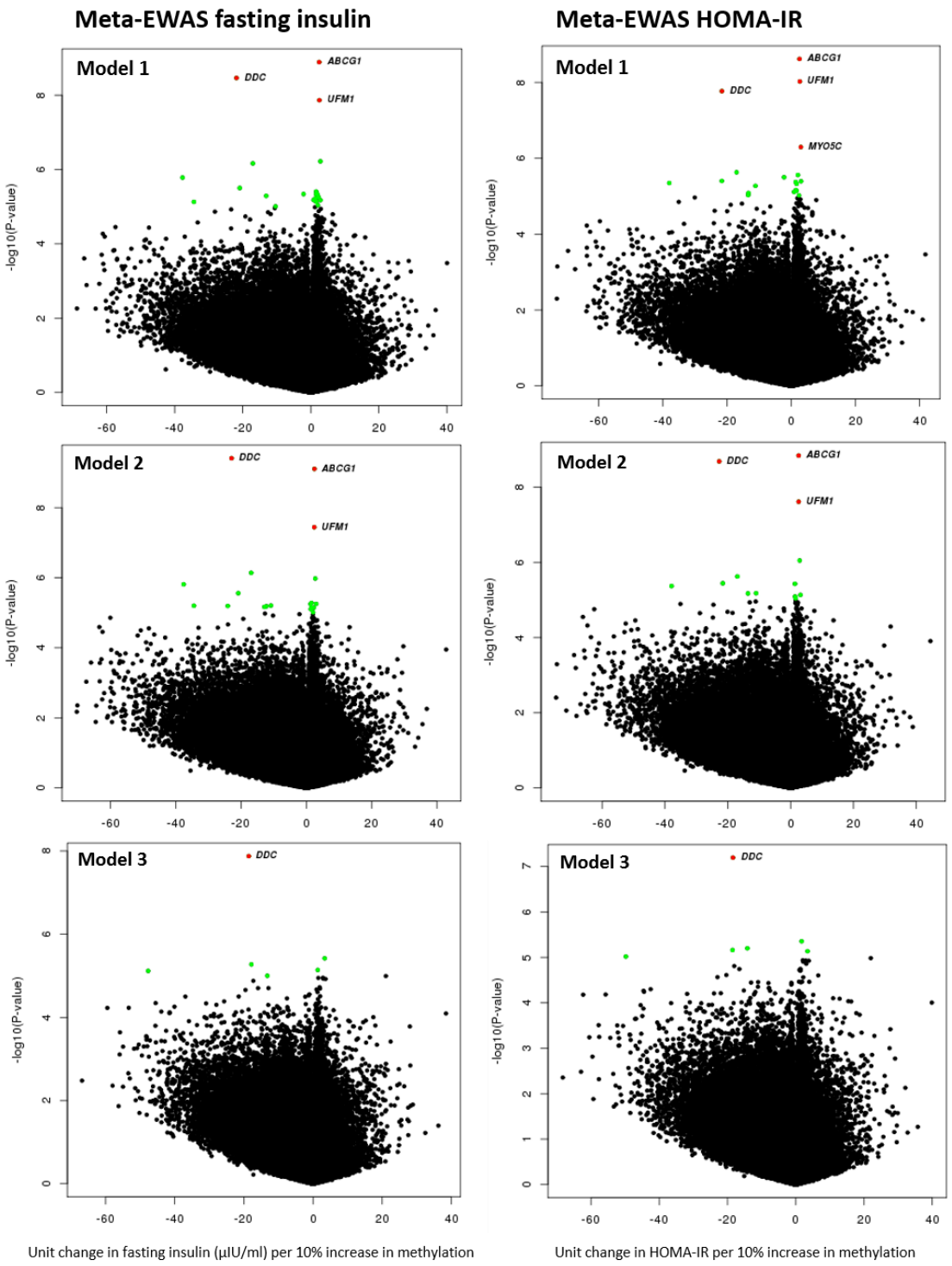


Figure 5-7 Volcano plots showing difference in the strength of the associations detected between adjustment models in the meta-EWAS of fasting insulin and HOMA-IR, two of the phenotypes with the highest number of associations detected with genome-wide significance and borderline significance across models. Associations were regarded significant at $p < 1.33x10^{-7}$ or FDR < 0.05 and are represented by red dots; borderline significant associations were regarded at $p < 1.0x10^{-5}$ and are represented by green dots.

Table 5-9 Summary of results of the meta-analysis of glycaemic traits.

Outcome	N	λ (m1)	FDR hits (m1)	Borderline (m1)	λ (m2)	FDR hits (m2)	Borderline (m2)	λ (m3)	FDR hits (m3)	Borderline (m3)
Fasting glucose	1384	0.95	0	5	0.96	0	7	0.95	0	4
2-h glucose	980	0.95	0	3	0.98	0	3	0.95	0	2
Fasting insulin	999	1.37	3	21	1.40	3	20	1.30	1	6
HOMA-IR	999	1.35	4	19	1.37	3	13	1.28	1	6
HOMA-B	998	1.05	2	12	1.06	2	9	1.06	1	4
Total		1.13	9	51	1.15	8	44	1.10	3	19

m1: basic model adjusted for age, SVs, predicted cells (Houseman method) and smoking. Sex was only included as a covariate in the EWAS of fasting glucose in ALSPAC. m2: model furtherly adjusted for smoking. Smoking categories were non-smoker and smoker in ALSPAC, and never, former and current smoker in SABRE. m3: fully adjusted model, additionally adjusted for BMI (kg/m^2). N: total number of samples included in the meta-EWAS; FDR hits: number of associations identified with EWAS significance (FDR < 0.05 or $p< 1.33x10^{-7}$). Total lambda refers to the average lambda obtained across EWAS within adjustment models, and total FDR and Borderline refers to the total number of signals identified with genome-wide and borderline significance within each model.

The highest number of associations with genome-wide significance and borderline significance were identified in model 2; thus, findings from this model were taken forward for further interpretation and functional exploration of meta-EWAS results (see sections 5.7 to 5.13). The Manhattan plot in Figure 5-8 illustrates main results of the meta-EWAS of glycaemic traits obtained in model 2. With respect to other parameters derived from results of the meta-analysis, average lambda across EWAS was higher in model 2 (mean λ =1.15) compared to model 3 (mean λ =1.10) and model 1 (mean λ =1.13), and fasting Insulin and HOMA-IR were the outcomes where the highest lambda was reported across models (mean λ fasting insulin= 1.35 and mean λ HOMA-IR=1.33). In terms of significance, traits with the strongest evidence of association with methylation were fasting insulin, HOMA-IR and HOMA-B, while no association was detected for fasting glucose and 2-h glucose (Table 5-9, Figure 5-8). Comparing between traits across models, there was some overlap in top-ranking signals (at $p<1.0x10^{-5}$) detected for fasting insulin, HOMA-IR and HOMA-B, but no overlap was identified between these signals, and top signals identified for fasting glucose and 2-h glucose (Table 5-10, more on correlation between traits in section 5.7).

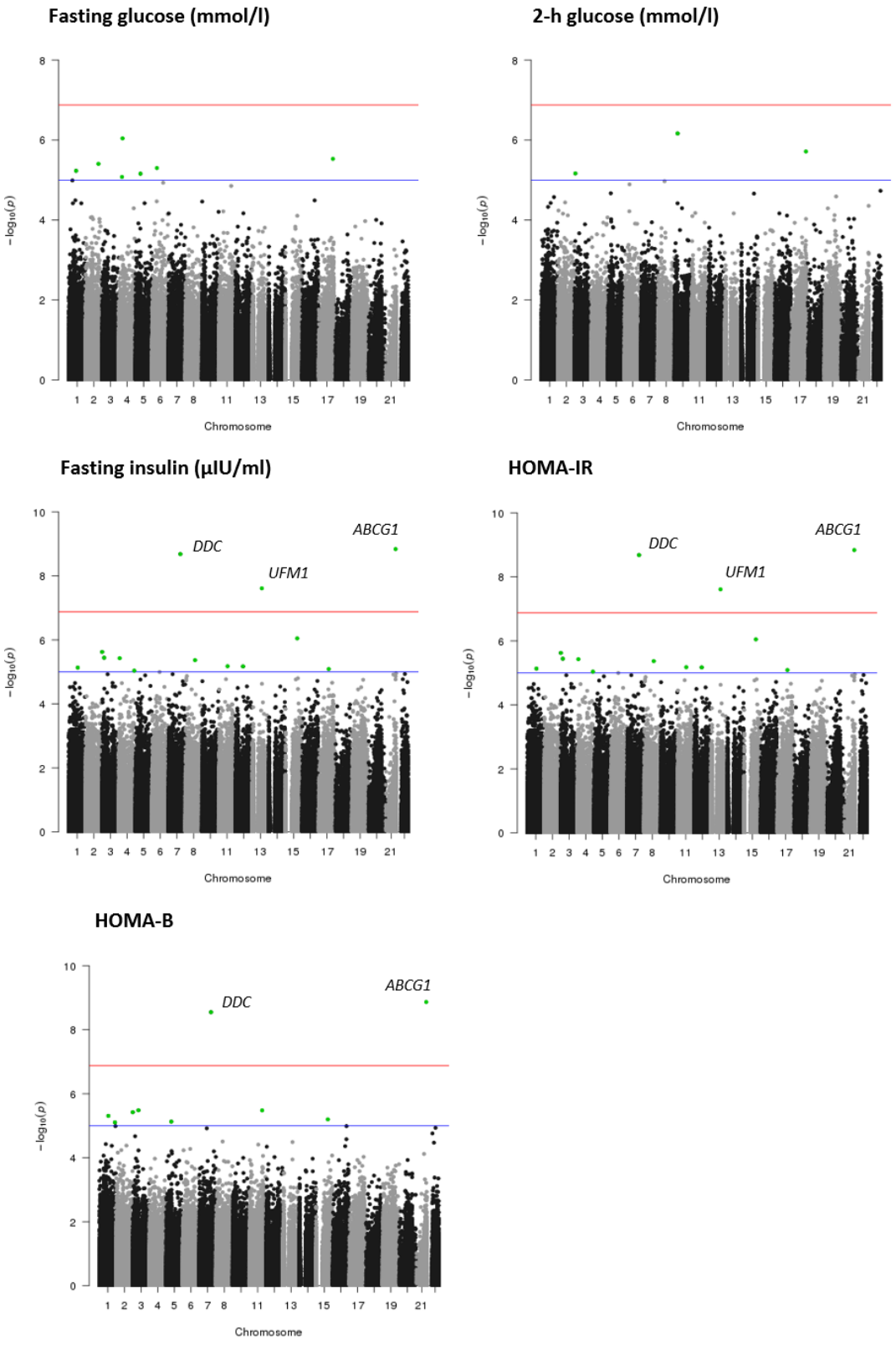


Figure 5-8 Manhattan plot showing meta-EWAS results for glycaemic traits analysed in individual EWAS in ALSPAC (n=1002 and 622) and SABRE (n=382 males) using normoglycemic participants. Results correspond to the model adjusted for age, SVs, predicted Houseman cells, and smoking. Blue-line is the threshold of borderline significance at $p=1.0x10^{-5}$, and red-line is the threshold of meta-EWAS significance at $p=1.33x10^{-7}$ or FDR < 0.05.

5.5.2.1 Results of the main model of the meta-EWAS of glycaemic traits

Table 5-10 summarizes top signals identified in the meta-EWAS of glycaemic traits across adjustment models, using for the cross-model comparison top signals identified with $p<1.0x10^{-5}$ in model 2. In this main model adjusted for smoking, three markers were identified in strong association with fasting insulin and HOMA-IR, corresponding to CpG sites cg18232548 (*DDC*), cg06500161 (*ABCG1*) and cg19750657 (*UFM1*). The CpG sites in *DDC* and *ABCG1* were also found in strong association with HOMA-B. Associations identified at the CpG in *DDC* surpassed further adjustment for BMI. Suggestive associations with FDR<0.1 were identified in model 2: cg10343442 (*TFG*), cg06192883 (*MYO5C*) and cg17340655 (*DDHD2*) in the meta-EWAS of fasting insulin, and the CpG in *MYO5C* was also detected in the meta-EWAS of HOMA-IR (Table 5-10). No association at genome-wide significance or borderline significance was detected in the meta-EWAS of fasting glucose and 2-h glucose (Table 5-10), and this was consistent with results of the individual EWAS for these traits, where only borderline association was identified between a CpG in *PAOX* and 2-h glucose in ALSPAC (see section 5.3).

In general, low heterogeneity was identified between studies across traits for most of the top-ranking associations detected with the smallest p-value in the meta-EWAS. However, high heterogeneity was detected in specific associations: $I^2 > 40\%$ was observed in 1/3 top associations for 2-h glucose, in 5/20 top associations for fasting insulin, in 4/13 top associations for HOMA-IR, and in 3/9 top associations for HOMA-B; for these associations, p-value for heterogeneity was < 0.05 in 2-h glucose, in 1/5 associations in fasting insulin, in 2/4 associations in HOMA-IR, and in 2/3 associations in HOMA-B.

Considering associations detected with borderline significance (at $p<1.0x10^{-5}$) in model 2, there was a relatively small change in the outcome per 10% increase in methylation (see appendix Figure S8-15). For instance, a median absolute change in 0.12mmol/l was detected for fasting glucose per 10% increase in methylation (i.e. 12% absolute change, range: 31% decrease to 10.5% increase). For the remaining traits, median absolute change in the outcome per 10% increase in methylation was 0.2mmol/l for 2-h glucose (range: 11.7% decrease to 16.4% increase), 10.1 μIU/ml for fasting insulin (range: 6.8 μIU/ml to 10.3 μIU/ml increase), 10.13 units in the HOMA-IR score (range: 6.9 to 10.3 increase), and 10.2 units in the HOMA-B score (range: 8.4 to 10.5 increase).

Table 5-10 Results of the meta-EWAS of five glycaemic traits using results of the individual EWAS in ALSPAC and SABRE. Associations compared across models were identified with p< 1.0x10^{-5} in the model adjusted for age, SVs, predicted-cells and smoking.

				Model 1[a] (age, SVs, predicted cells)			Model 2 (age, SVs, predicted cells, smoking)			Model 3 (age, SVs, predicted cells, smoking, BMI)		
Phenotype	CpG	Chr	Gene	Beta	SE	P	Beta	SE	P	Beta	SE	P
Fasting glucose	cg22724847	1	*OR14C36*	0.08	0.02	1.00x10^{-5}	0.08	0.02	5.87x10^{-6}	0.08	0.02	1.10x10^{-5}
	cg26234543	2	*TMEM17*	-0.11	0.02	4.77x10^{-6}	-0.11	0.02	3.94x10^{-6}	-0.11	0.02	2.84x10^{-6}
	cg06690548	4	*SLC7A11*	-0.13	0.03	1.10x10^{-5}	-0.13	0.03	8.37x10^{-6}	-0.11	0.03	2.05x10^{-4}
	cg07147166	4	*Intergenic*	-0.30	0.06	1.38x10^{-6}	-0.31	0.06	9.08x10^{-7}	-0.29	0.06	2.24x10^{-6}
	cg05468458	5	*Intergenic*	-0.11	0.03	8.19x10^{-6}	-0.12	0.03	6.93x10^{-6}	-0.11	0.03	2.60x10^{-5}
	cg17219086	6	*Intergenic*	0.11	0.02	4.49x10^{-6}	0.10	0.02	5.00x10^{-6}	0.10	0.02	1.02x10^{-5}
	cg17540765	17	*RECQL5*	-0.15	0.03	7.03x10^{-6}	-0.16	0.03	2.95x10^{-6}	-0.17	0.03	7.39x10^{-7}
2-h glucose	cg08817540	3	*HHLA2*	0.06	0.01	8.96x10^{-6}	0.06	0.01	6.83x10^{-6}	0.05	0.01	4.55x10^{-5}
	cg03826430	10	*PAOX*	-1.20	0.23	3.07x10^{-7}	-1.17	0.24	6.84x10^{-7}	-1.16	0.23	4.44x10^{-7}
	cg05339942	17	*HRNBP3*	0.16	0.03	2.40x10^{-6}	0.16	0.03	1.93x10^{-6}	0.16	0.03	2.76x10^{-6}
Fasting insulin	cg04311473	1	*Intergenic*	10.30	10.07	1.15x10^{-5}	10.31	10.07	5.57x10^{-6}	10.24	10.06	5.72x10^{-5}
	cg01212284	3	*MCCC1*	10.16	10.03	7.42x10^{-6}	10.16	10.03	5.25x10^{-6}	10.14	10.03	7.25x10^{-6}
	cg10343442	3	*TFG*	8.43	10.35	6.86x10^{-7}	8.44	10.35	7.24x10^{-7}	8.76	10.30	9.97x10^{-6}
	cg21910545	3	*HEG1*	8.12	10.46	3.15x10^{-6}	8.12	10.46	2.76x10^{-6}	8.37	10.40	5.31x10^{-6}
	cg00278494	4	*Intergenic*	10.16	10.03	3.91x10^{-6}	10.15	10.03	7.23x10^{-6}	10.12	10.03	6.62x10^{-5}
	cg10269431	4	*EGF*	10.13	10.03	6.06x10^{-6}	10.13	10.03	5.62x10^{-6}	10.08	10.03	1.21x10^{-3}
	cg25020279	5	*RICTOR*	7.09	10.80	7.43x10^{-6}	7.08	10.79	6.31x10^{-6}	7.74	10.69	1.30x10^{-4}
	cg18232548	**7**	***DDC***	**8.04**	**10.38**	**3.43x10^{-9}**	**7.95**	**10.37**	**3.88x10^{-10}**	**8.30**	**10.33**	**1.33x10^{-8}**
	cg08874430	8	*FDFT1*	7.91	10.55	1.19x10^{-5}	7.86	10.55	6.40x10^{-6}	8.30	10.48	6.46x10^{-5}
	cg17340655	8	*DDHD2*	6.86	10.82	1.65x10^{-6}	6.87	10.81	1.54x10^{-6}	7.56	10.71	4.98x10^{-5}
	cg19750657	**13**	***UFM1***	**10.26**	**10.04**	**1.36x10^{-8}**	**10.25**	**10.04**	**3.62x10^{-8}**	10.15	10.04	1.87x10^{-4}
	cg06192883	15	*MYO5C*	10.29	10.06	6.02x10^{-7}	10.28	10.06	1.05x10^{-6}	10.17	10.05	5.48x10^{-4}
	cg12533335	16	*TAF1C*	8.89	10.28	1.59x10^{-5}	8.85	10.27	6.54x10^{-6}	9.08	10.24	6.17x10^{-5}
	cg08857797	17	*VPS25*	10.14	10.03	6.75x10^{-6}	10.14	10.03	7.90x10^{-6}	10.06	10.03	4.32x10^{-2}
	cg01176028	21	*ABCG1*	10.21	10.05	9.17x10^{-6}	10.21	10.05	9.72x10^{-6}	10.12	10.04	4.07x10^{-3}
	cg03732014	21	*BACH1*	10.16	10.03	4.66x10^{-6}	10.15	10.03	7.78x10^{-6}	10.10	10.03	7.37x10^{-4}
	cg06500161	**21**	***ABCG1***	**10.25**	**10.04**	**1.29x10^{-9}**	**10.25**	**10.04**	**7.80x10^{-10}**	10.14	10.04	2.38x10^{-4}
	cg17648210	21	*Intergenic*	9.00	10.24	1.10x10^{-5}	8.97	10.24	6.23x10^{-6}	9.27	10.21	3.44x10^{-4}
	cg03625627	22	*PIK3IP1*	8.77	10.29	5.13x10^{-6}	8.79	10.29	6.81x10^{-6}	9.18	10.26	7.37x10^{-4}

Table 5-10(Continued)

				Model 1[a] (age, SVs, predicted cells)			Model 2 (age, SVs, predicted cells, smoking)			Model 3 (age, SVs, predicted cells, smoking, BMI)		
Phenotype	CpG	Chr	Gene	Beta	SE	P	Beta	SE	P	Beta	SE	P
HOMA-IR	cg04311473	1	*Intergenic*	10.30	10.07	$1.15x10^{-5}$	10.31	10.07	$5.57x10^{-6}$	10.24	10.06	$5.72x10^{-5}$
	cg10343442	3	*TFG*	8.43	10.35	$6.86x10^{-7}$	8.44	10.35	$7.24x10^{-7}$	8.76	10.30	$9.97x10^{-6}$
	cg21910545	3	*HEG1*	8.12	10.46	$3.15x10^{-6}$	8.12	10.46	$2.76x10^{-6}$	8.37	10.40	$5.31x10^{-6}$
	cg00278494	4	*Intergenic*	10.16	10.03	$3.91x10^{-6}$	10.15	10.03	$7.23x10^{-6}$	10.12	10.03	$6.62x10^{-5}$
	cg10269431	4	*EGF*	10.13	10.03	$6.06x10^{-6}$	10.13	10.03	$5.62x10^{-6}$	10.08	10.03	$1.21x10^{-3}$
	cg18232548	**7**	***DDC***	**8.04**	**10.38**	**$3.43x10^{-9}$**	**7.95**	**10.37**	**$3.88x10^{-10}$**	**8.30**	**10.33**	**$1.33x10^{-8}$**
	cg17340655	8	*DDHD2*	6.86	10.82	$1.65x10^{-6}$	6.87	10.81	$1.54x10^{-6}$	7.56	10.71	$4.98x10^{-5}$
	cg20078939	11	*TCP11L1*	9.02	10.24	$9.85x10^{-6}$	9.03	10.24	$1.21x10^{-5}$	9.32	10.21	$6.04x10^{-4}$
	cg15528501	12	*TM7SF3*	8.82	10.29	$1.42x10^{-5}$	8.81	10.29	$1.05x10^{-5}$	9.03	10.26	$5.62x10^{-5}$
	cg19750657	**13**	***UFM1***	10.26	10.04	$1.36x10^{-8}$	**10.25**	**10.04**	**$3.62x10^{-8}$**	10.15	10.04	$1.87x10^{-4}$
	cg06192883	15	*MYO5C*	10.29	10.06	$6.02x10^{-7}$	10.28	10.06	$1.05x10^{-6}$	10.17	10.05	$5.48x10^{-4}$
	cg08857797	17	*VPS25*	10.14	10.03	$6.75x10^{-6}$	10.14	10.03	$7.90x10^{-6}$	10.06	10.03	$4.32x10^{-2}$
	cg06500161	**21**	***ABCG1***	**10.25**	**10.04**	**$1.29x10^{-9}$**	**10.25**	**10.04**	**$7.80x10^{-10}$**	10.14	10.04	$2.38x10^{-4}$
HOMA-B	cg04311473	1	*Intergenic*	10.24	10.06	$1.89x10^{-5}$	10.25	10.05	$4.90x10^{-6}$	10.21	10.05	$3.27x10^{-5}$
	cg07198150	1	*COL24A1*	9.84	10.04	$1.49x10^{-5}$	9.83	10.04	$7.86x10^{-6}$	9.87	10.03	$1.77x10^{-4}$
	cg10343442	3	*TFG*	8.69	10.30	$2.52x10^{-6}$	8.73	10.30	$3.77x10^{-6}$	8.91	10.27	$1.91x10^{-5}$
	cg22875391	3	*TPRG1*	10.23	10.05	$1.70x10^{-6}$	10.22	10.05	$3.29x10^{-6}$	10.18	10.04	$4.40x10^{-5}$
	cg23436042	5	*Intergenic*	10.17	10.04	$8.50x10^{-6}$	10.17	10.04	$7.45x10^{-6}$	10.14	10.03	$8.26x10^{-5}$
	cg18232548	**7**	***DDC***	**8.50**	**10.30**	**$5.06x10^{-8}$**	**8.41**	**10.30**	**$2.82x10^{-9}$**	**8.60**	**10.28**	**$3.18x10^{-8}$**
	cg15553522	11	*TMEM132A*	10.45	10.10	$7.43x10^{-6}$	10.46	10.10	$3.31x10^{-6}$	10.37	10.09	$4.42x10^{-5}$
	cg06192883	15	*MYO5C*	10.22	10.05	$4.97x10^{-6}$	10.21	10.05	$6.33x10^{-6}$	10.14	10.04	$9.45x10^{-4}$
	cg06500161	**21**	***ABCG1***	**10.20**	**10.03**	**$3.40x10^{-9}$**	**10.20**	**10.03**	**$1.36x10^{-9}$**	10.12	10.03	$1.14x10^{-4}$

[a] Adjustment for sex was only considered in the EWAS of fasting glucose in ALSPAC, where females and males were included in the analysis. For the remaining EWAS, only females (ALSPAC), or males (SABRE), were included in the meta-EWAS. [b] Results are interpreted as the effect of 10% increase in methylation on a unit change in the outcome. Beta coefficients were transformed-back to the original units in which the outcome was measured using [exp(beta/100)] for log-transformed outcomes; P-value is the unadjusted-p. Associations were considered genome-wide significant at $p< 1.33x10^{-7}$ (Bonferroni corrected P for 376,415 tests) or FDR< 0.05.

5.6 Impact of adjusting for smoking and BMI in results of the meta-EWAS of glycaemic traits

By comparing the strength of the effect estimate between the basic model (adjusted for cells) and the models with further adjustment for smoking and BMI, it was evident that smoking and BMI were confounders of the main association by either reducing or increasing the strength of the effect estimate. For instance, adjustment for smoking on average decreased the strength of the effect estimate by 0.11% (range: 1.05% decrease to 0.26% increase) in fasting insulin, by 0.07% (range: 1.16% decrease to 0.18% increase) in HOMA-IR, and by 0.07% (range: 1.09% decrease to 0.42% increase) in HOMA-B, while adjustment for BMI on average increased the strength of the effect estimate by 1.84% (range: 1.11% decrease to 10.24% increase) in fasting insulin, by 0.66% (range: 1.16% decrease to 10.73% increase) in HOMA-IR, and by 0.08% (range: 0.77% increase to 2.54% increase) in HOMA-B. Opposite to this, in fasting glucose, it was identified that smoking tended to increase the strength of the effect estimate by 1.47% (range: 0.43 decrease to 4.04% increase), while BMI tented to decrease the strength of the effect estimate by 3.78% (range: 16.58% decrease to 8.14% increase). Lastly, in 2-h glucose, adjustment for smoking and BMI on average reduced the strength of the effect estimate by 0.11% (range: 2.95% decrease and 1.61 increase) and 5.03% (range: 2.51% to 9.09% decrease), respectively. Al together, results suggested that BMI and smoking were confounders of the main association, and the impact of these factors on a change in the effect estimate varied according to the trait evaluated.

5.7 Correlation across glycaemic traits based on effect estimates obtained in the meta-EWAS

As described earlier in section 5.3.3, a correlation analysis was conducted using unstandardized effect estimates obtained in the meta-EWAS of glycaemic traits for a subset of 20 CpG sites, which were originally identified in strong (FDR< 0.05, n=3 sites) or borderline association (FDR<0.10, n=17 sites) with fasting insulin. Methylation sites included in the correlation and in the cluster analysis (i.e. heatmap) were: *cg18232548* (*DDC*), cg06500161 (*ABCG1*), cg01176028 (*ABCG1*), cg19750657 (*UFM1*), cg10343442 (*TFG*), cg06192883 (*MYO5C*), cg17340655 (*DDHD2*), cg21910545 (*HEG1*), cg01212284 (*MCCC1*), cg10269431 (*EGF*), cg25020279 (*RICTOR*), cg08874430 (*FDFT1*), cg12533335 (*TAF1C*), cg03625627 (*PIK3IP1*), cg12994768 (*KRBA1*), cg03732014 (*BACH1*), cg08857797 (*VPS25*) and the intergenic CpG sites cg04311473, cg17648210, cg00278494 (Figure 5-9).

Overall, there was strong and positive correlation between traits at the specific CpG sites, with correlation p-values that ranged between $3.69x10^{-14}$ and $5.16x10^{-3}$. The strongest correlation was identified between fasting insulin and HOMA-IR (r= 1.00, p< 0.001), whilst the weakest correlation was identified between fasting glucose and HOMA-B (r=0.60, $p=5.16x10^{-3}$) (Figure 5-9, appendix Table S8-12). The average correlation value was r=0.88 for fasting insulin and HOMA-IR, r=0.86 for HOMA-B, r=0.81 for 2-h glucose, and r=0.64 for fasting glucose. The cluster analysis showed that there was more similarity in the effect estimates between fasting insulin, HOMA-IR and HOMA-B, than between these traits and 2-h glucose and fasting glucose. Looking at similarity between CpG sites across traits for three sites previously detected in strong association with fasting insulin, there was more similarity between the CpG sites in *ABCG1* and *UFM1*, than between them and the CpG in the *DDC* locus across the five traits compared.

Results of the cross-phenotype correlation analysis using effect estimates from the meta-EWAS corroborated the underlying characteristics of these glycaemic traits, even when only considering a subset of CpG sites for the analysis (i.e. n=20 sites). Fasting glucose was the trait most weakly correlated with any other phenotype, compared to the level of correlation observed for fasting insulin and the HOMA scores. Furthermore, the cluster analysis revealed that, even though the HOMA scores are equally dependent on fasting insulin and fasting glucose for their estimation, they were more closely related with fasting insulin than with fasting glucose. As expected, 2-h glucose and fasting glucose were grouped together within a same cluster, even though the correlation analysis showed that both traits were weakly correlated (r=0.63, Figure 5-9).

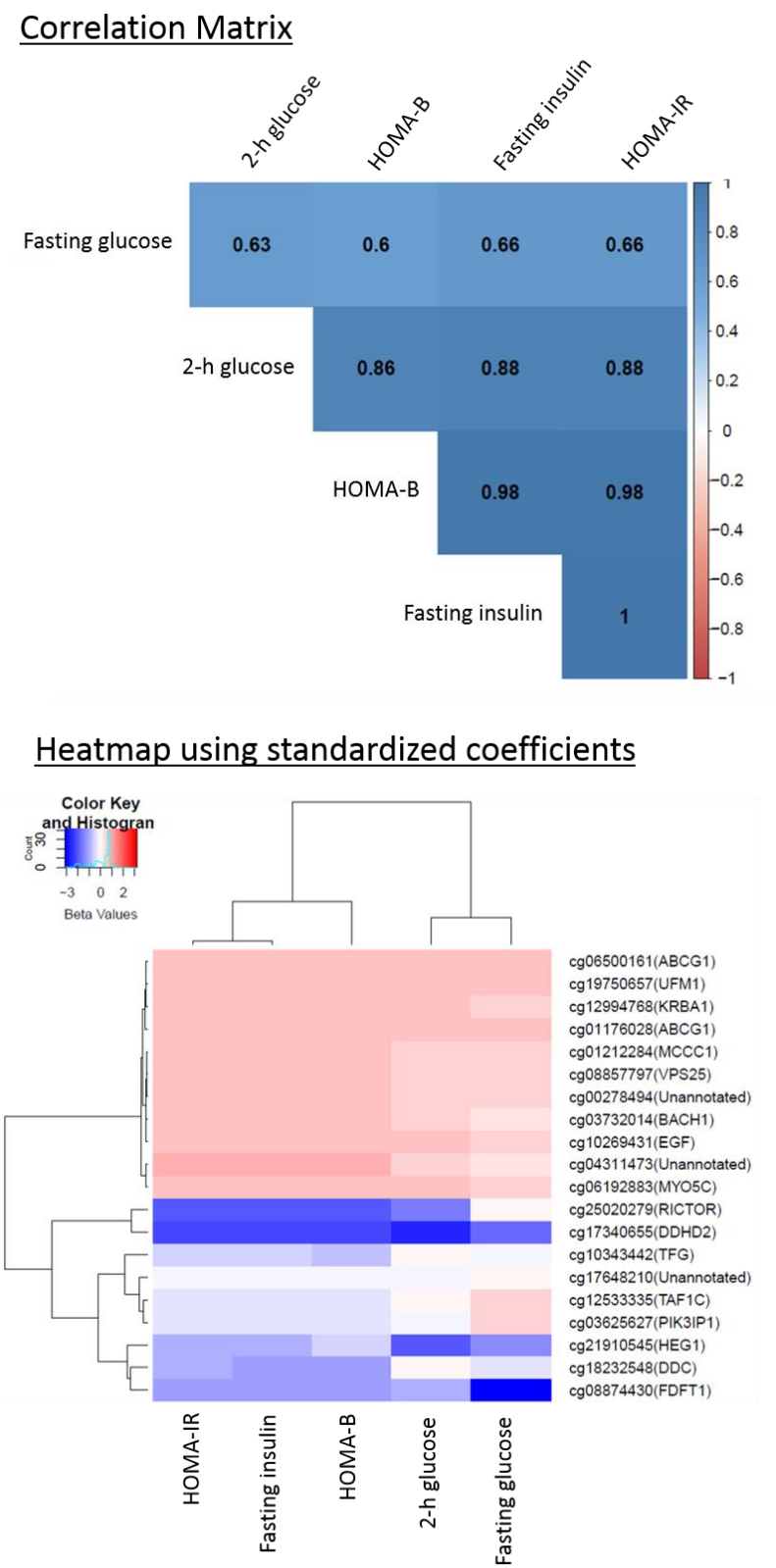


Figure 5-9 Correlogram and heatmap showing the level of correlation and similarity between effect estimates across traits for 20 CpG sites initially identified as top-ranking signals (smallest p-value) in the meta-EWAS of fasting insulin. Correlations were significant at $p < 0.01$; the strength of the correlation is indicated by the colour legend on the right. In the heatmap, phenotypes connected by a similar node in the dendrogram at the top are more similar between them, than with phenotypes connected by a more distant node.

5.8 Potential role of methylation as a mediator of the association between genetic variation and gene expression

To determine if the effect of genetic variation on gene expression could be mediated by methylation, the overlap between meQTL and eQTL at CpG sites of interest was investigated. A total of 1,346 meQTL were available at the antenatal and middle-age time-points for 5/9 sites previously identified in association with fasting insulin, HOMA-IR, HOMA-B and HbA1c. meQTL were compared to a list of eQTL reported by GTEx in different tissues (GTEx_Analysis_v7, analysis date:14-08-2018, https://www.gtexportal.org/home/datasets)[163]. A *cis* QTL was considered to overlap between datasets if a similar SNP was identified with $p<10^{-7}$ in the meQTL dataset, and with *q*-value < 0.05 in the eQTL dataset. An example of a *cis* QTL in overlap between datasets is illustrated in Figure 5-10.

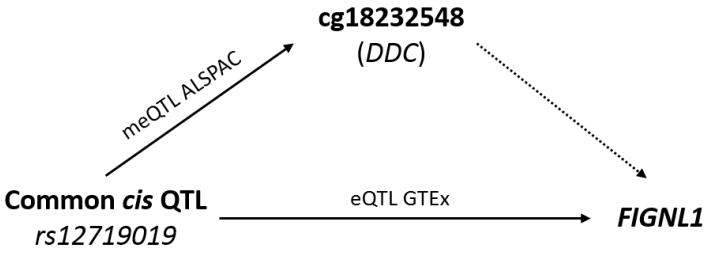


Figure 5-10 Identification of a cis QTL overlapping with an meQTL for cg18232548, and with an eQTL for the FIGNL1 gene. Because of the shared genetic variation between these two molecular markers, it is hypothesized that methylation at cg18232548 could be a mediator between the eQTL- FIGNL1 association.

If an meQTL overlapped with an eQTL, this suggested that DNA methylation was in the causal pathway between the genotype and gene expression, and that methylation was potentially a mediator of the observed SNP-gene expression association. Of the 1,346 meQTL identified, only one overlapped with a *cis* eQTL in blood. The *cis* QTL in common was the SNP rs12719019, which was associated with antenatal and middle-age methylation at the CpG cg18232548 in *DDC* (effect estimate=0.63, $p_{antenatal}=1.36x10^{-55}$ and $p_{middle\text{-}age}=1.15x10^{-56}$), and with gene expression of the *FIGNL1* gene (effect estimate=0.48, 95% CI= 0.36, 0.66, $q=2.70x10^{-8}$). Therefore, there was some evidence that the SNP-*FIGNL1* association could be mediated by methylation at cg18232548. For the *cis* QTL in rs12719019, the associations SNP-CpG and SNP-gene were in the same direction, indicating that the same effect allele was associated with higher methylation and higher gene expression. Figure 5-11 shows the genomic context of the CpG cg18232548 and the gene *FIGNL1*, both identified in association with the *cis* QTL SNP rs12719019. The CpG cg18232548 was downstream the region of the *FIGNL1* gene, and within a small CpG island located in an intron of the *DDC* gene. Interestingly,

the CpG cg18232548 was 142 kb upstream *GRB10*, a gene that has been previously reported in association with decreased insulin sensitivity and increased T2D risk[217, 218].

In summary, a *cis* QTL was found in common between an meQTL for the CpG cg18232548 in *DDC* (associated with fasting insulin and the HOMA scores), and an eQTL for the gene *FIGNL1*, and the same effect allele was associated with higher gene expression and higher methylation. Because methylation and gene expression are related outcomes, and because of the common genetic variation identified between cg18232548 and *FIGNL1*, it was hypothesized that methylation was possibly mediating the rs12719019-*FIGNL1* association. Additional *cis* QTL (n=6) were identified in association with methylation at the CpG in *DDC* (cg18232548), and with gene expression of *FIGNL1* and *DDC* across relevant tissues for T2D. Results in tissues different from blood showed that for some of these tissue-specific *cis* QTL, the effect allele was different between the meQTL and the eQTL (n=3 *cis* QTL), with SNP-methylation and SNP-gene expression associations going in opposite directions, especially when comparing meQTL obtained in blood with eQTL in omentum (i.e. visceral fat), skeletal muscle and liver. Because meQTL and eQTL are tissue-specific markers, it was less valid to assume that methylation was a mediator of the SNP-gene expression association when the eQTL and the meQTL were identified in different tissues.

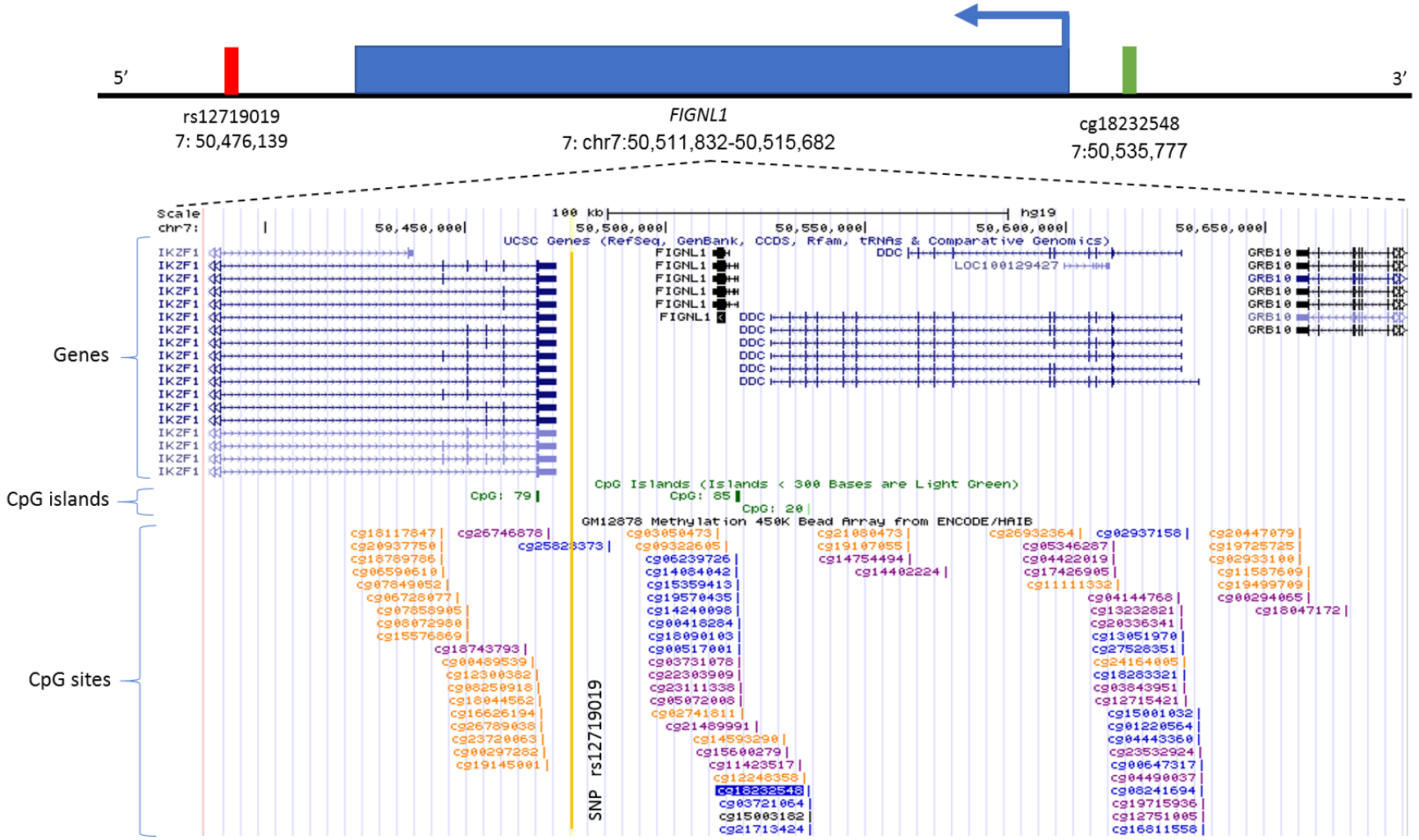


Figure 5-11 Genetic context of the CpG cg18232548 in DDC (highlighted in blue) and the FIGNL1 gene, both identified with a common genetic variant at the cis QTL rs12719019. The SNP rs12719019, which position in the figure is demarked by a vertical yellow line, was previously identified as a cis meQTL for cg18232548, and as an eQTL for the FIGNL1 gene (GTEx data); both markers were detected in peripheral blood samples. Plot adapted from the UCSC Genome Browser viewer tool.

5.9 Identifying eQTM for top-ranking CpG sites detected in the meta-EWAS of glycaemic traits

Previously, a lookup was conducted to determine the potential mediating role of methylation in the SNP-gene expression association by determining the overlap between meQTL and eQTL for top-ranking CpG sites of the meta-EWAS. In this second analysis, it was investigated if observationally there was evidence of an association between methylation at the CpG sites of interest, and gene expression of the nearby gene, in other words, to identify eQTM using the Bios QTL browser (https://genenetwork.nl/biosqtlbrowser/). From the nine sites of interest, methylation at the CpG cg06500161 was inversely associated with transcripts of the *ABCG1* gene (correlation coefficient=-0.32, p= 2.22×10^{-37}). For this site, no meQTL was available at the time-points of interest. Therefore, it was not possible to determine if methylation at cg06500161 was in the causal pathway between genetic variation and gene expression of the *ABCG1* gene.

Despite lack of an eQTM for the CpG in *DDC*, this was the best candidate site to suggest a mediating role of methylation in the SNP-gene expression association, and it is likely that methylation in *DDC* is in the causal pathway for gene expression of *FIGNL1* or *DDC* itself. Nonetheless, lack of an eQTM could respond to residual confounding still present in the observational association between methylation and gene expression for this CpG site. Although other alternative explanations may be possible.

5.10 Identifying shared genetics between methylation and the glycaemic traits

A comparison was made between meQTL identified for top-ranking CpG sites in fasting insulin, HOMA-IR, HOMA-B, and HbA1c, and GWAS variants for the same traits. This comparison allowed the investigation of genetic variation associated with the glycaemic traits, and with methylation. Detection of an overlap between an meQTL and GWAS SNPs indicated that methylation was a potential mediator of the SNP-outcome association. For this analysis, CpG sites taken forward were nine sites identified in strong (6/9 sites, FDR<0.05) or borderline (3/9 sites, FDR<0.10) association with fasting insulin, HOMA-IR, HOMA-B and HbA1c (see section 5.5 above). meQTL were looked-up in an online catalogue (mQTLdb, analysis date: 14-08-2018, http://www.mqtldb.org/)[109] using antenatal and middle-age DNA methylation. Genetic variation in fasting insulin, HOMA-IR, HOMA-B and HbA1c, was retrieved from the largest GWAS meta-analyses of glycaemic traits conducted by the MAGIC consortium (https://www.magicinvestigators.org/downloads/)[160-162].

In total, 1346 meQTL were associated with methylation at 5/9 sites of interest: 712/1346 meQTL were detected in association with antenatal methylation, 634/1346 were detected in association with middle-age methylation. All meQTL were retrieved with $p< 10^{-7}$. Based on their distance from the CpG site, 1335/1346 meQTL were in *cis* and 11/1346 meQTL were in *trans*. From the total number of meQTL identified, only a subset of them were found with nominal significance (GWAS p-value<0.05) in the GWAS of fasting insulin (n=37 *cis* meQTL), HOMA-IR (n=121 *cis* meQTL) and HOMA-B (n=7 *cis* meQTL) (Table 5-11). meQTL identified with nominal GWAS significance were associated with methylation at the CpG cg18232548 in *DDC*, and in cases where the effect allele was the same in the meQTL and the GWAS data, it was commonly observed that the associations SNP-methylation and SNP-outcome were in opposite directions. For instance, in 20/37 meQTL nominally associated with fasting insulin, the same effect allele was associated with a decrease in fasting insulin (effect estimate range -0.024 to -0.003) and an increase in methylation at the CpG in *DDC* (effect estimate range 0.426 to 0.632). Thus, results suggested that the association between methylation at cg18232548 (*DDC*) and fasting insulin and the HOMA scores was partially explained by common genetics, and that methylation was likely to be in the causal pathway in the SNP-outcome association. In contrast, none of the *trans* meQTL identified for two CpG sites associated with HbA1c, were detected with nominal significance in the GWAS of this trait (Table 5-11). Therefore, there was no evidence that the association between methylation at cg26316702 (*TEKT4*) and cg13583414 (*LZTS1*) and HbA1c was influenced by common genetics.

Table 5-11 Summary of meQTL identified for CpG sites associated with some of the glycaemic traits in the meta-EWAS, which were also found with nominal significance in the latest GWAS meta-analyses for the glycaemic traits.

Glycaemic trait	meQTL in GWAS	Nominal meQTL	Time-point	CpG	Same EA	Ratio[a]	Direction of effect	Ratio[b]
Fasting insulin	183	37/183 (cis)	Middle-age	*DDC* (cg18232548)	Yes No	20/37 17/37	Opposite NA	15/20 NA
HOMA-IR	266	121/266 (cis)	Middle-age	*DDC* (cg18232548)	Yes No	57/121 64/121	Opposite NA	56/57 NA
HOMA-B	266	7/266 (cis)	Middle-age	*DDC* (cg18232548)	Yes No	1/7 6/7	Opposite NA	1/7 NA
HbA1c	1	none	Antenatal	*TEKT4* (cg26316702)	NA	NA	NA	NA
	1	none	Antenatal	*LZTS1* (cg13583414)	NA	NA	NA	NA

meQTL in GWAS: number of meQTL identified in GWAS data; Nominal meQTL: subset of meQTL identified with GWAS p-value<0.05; Same EA: indicating if the meQTL-SNP and the GWAS-SNP had the same effect allele; [a] Ratio: proportion of nominal meQTL that had the same effect allele as the GWAS-SNP; direction of effect: indicating if associations were in opposite or similar direction. [b] Ratio: proportion of nominal meQTL with same effect allele as the GWAS-SNP, and with opposite/similar direction of effect.

5.11 Association between methylation at top-ranking CpG sites and established clinical risk factors

As a sensitivity analysis, the population was stratified by quintiles of methylation (even distribution of samples) for top-ranking CpG sites detected in the meta-analysis, and the association between methylation and clinical risk factors for T2D was investigated across the quintiles. CpG sites taken forward for this analysis were three sites identified with genome-wide significance in the meta-EWAS of fasting insulin, HOMA-IR and HOMA-B (i.e. cg06500161 in *ABCG1*, cg18232548 in *DDC* and cg19750657 in *UFM1*).

Significant difference in methylation at $p<0.05$ was identified between Q1 and Q5 for the three CpG sites of interest based on the different subsamples considered in the meta-analysis: at *ABCG1* difference in methylation ranged between 0.11 to 0.14, at *DDC* average difference in methylation was 0.01, and at *UFM1* difference in methylation ranged between 0.12 and 0.13. In relation to risk factors, strong associations were identified between quintiles of methylation at cg06500161 (*ABCG1*) and different anthropometric and metabolic variables: the strongest associations were detected with waist-circumference, triglycerides, HDL, systolic blood pressure and BMI, respectively (P for trend range: $1.10\text{x}10^{-15}$ to $1.57\text{x}10^{-8}$) (Table 5-12). Among glycaemic traits, strong association was identified between quintiles of *ABCG1* and fasting glucose, fasting insulin, HOMA-IR and HOMA-B (P for trend in the order of 10^{-6} and 10^{-4}). In addition, there was an overrepresentation of males compared to females in the upper quintiles of *ABCG1* (Table 5-12). In relation to the proportion of estimated cells (Houseman method), significant association was identified between granulocytes, monocytes, CD4T, NK-cells, B-cells and quintiles of *ABCG1* (P for trend range: $2.55\text{x}10^{-10}$ to $1.73\text{x}10^{-2}$). No association was observed between *ABCG1* and categories of glucose tolerance, HbA1c, 2-h glucose, total cholesterol and CD8T cells (Table 5-12).

For the cg18232548 in *DDC*, strong associations were identified with the glycaemic traits of fasting insulin, HOMA-IR and HOMA-B (P for trend range: $5.22\text{x}10^{-4}$ to $2.40\text{x}10^{-3}$). In addition, quintiles of *DDC* were associated with some metabolic and anthropometric risk factors: C-reactive protein, triglycerides, waist-circumference and HDL (P for trend range: 0.01 to 0.03). No association was identified between quintiles of *DDC* and fasting glucose, 2-h glucose, HbA1c, glucose tolerance status and other metabolic and anthropometric risk factors (see appendix Table S8-13). For the CpG cg19750657 in *UFM1*, strongest association was identified with the estimated cells CD4T, granulocytes, NK-cells, monocytes and B-cells (P for trend range: $1.10\text{x}10^{-15}$ to $2.34\text{x}10^{-3}$). Quintiles of *UFM1* were also associated with HOMA-IR, fasting insulin, waist-circumference and triglycerides (see

appendix Table S8-14), but not with HOMA-B, fasting glucose, 2-hours glucose, HbA1c, glucose tolerance status, or with LDL, HDL, total cholesterol and BMI (see appendix Table S8-14).

Interestingly, when measuring the association between continuous methylation at *ABCG1*, *DDC* and *UFM1* and the glycaemic traits using unadjusted regressions, these sites were strongly associated with fasting insulin, HOMA-IR and HOMA-B (p range: $1.75x10^{-5}$ to $2.98x10^{-2}$) (Table 5-13). This result was partially consistent with results of the meta-EWAS, except for the association between *UFM1* and HOMA-B, which was not detected before. In addition, *UFM1* was strongly associated with fasting glucose ($p=7.9x10^{-3}$) and 2-h glucose (p=0.04) (Table 5-13, Figure 5-12), and it was borderline associated with HbA1c (p=0.05) (Table 5-13, Figure 5-12). None of the top-three CpG sites analysed was associated with categories of glucose tolerance (p range between 0.33 to 0.75, Table 5-13).

Table 5-12 Association between quintiles of methylation at cg06500161 (ABCG1) and different clinical risk factors in a subsample of normoglycemic participants in ALSPAC (n=1002). Continuous variables were summarized using the mean and the standard deviation, while categorical variables were summarized using the proportion of samples per category on each quintile. P for trend represents the Bonferroni adjusted-p, with p< 0.05 indicating evidence of a linear trend in the methylation-outcome association.

	Quintile 1 (n=201)	Quintile 2 (n=200)	Quintile 3 (n=200)	Quintile 4 (n=200)	Quintile 5 (n=201)		
Continuous Phenotype	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P	P for trend (adjusted-p)
Age [years]	48.84(4.69)	49.81(5.12)	49.90(5.44)	50.40(5.36)	50.89(5.59)	2.03E-03	1.01E-02
BMI [kg/m2] [a]	25.11(3.87)	26.12(4.16)	26.32(4.38)	26.69(4.65)	27.96(4.96)	3.13E-09	1.57E-08
waist-circumference [cm]	82.02(10.17)	86.69(12.24)	88.31(11.94)	89.39(11.99)	95.14(14.59)	2.20E-16	1.10E-15
Fasting Glucose [mmol/l]	5.16(0.41)	5.24(0.47)	5.27(0.41)	5.29(0.49)	5.42(0.49)	1.15E-06	5.73E-06
2-hours Glucose [mmol/l] [b]	4.33(0.40)	4.28(0.40)	4.31(0.39)	4.24(0.38)	4.37(0.39)	7.95E-02	3.98E-01
HbA1c [%] [c]	5.52(0.31)	5.49(0.34)	5.56(0.28)	5.57(0.29)	5.52(0.33)	4.60E-01	2.30E+00
C-reactive Protein [mg/l] [a]	1.49(2.08)	1.8(2.45)	2.02(2.99)	2.08(2.82)	2.39(3.37)	3.85E-04	1.93E-03
fasting Insulin [µIU/ml] [a,b]	4.96(3.72)	4.60(2.93)	5.32(2.96)	5.71(3.64)	6.81(5.18)	2.13E-05	1.06E-04
HOMA-IR [a,b]	1.15(0.91)	1.07(0.74)	1.25(0.76)	1.32(0.89)	1.62(1.30)	3.02E-05	1.51E-04
HOMA-B [a,b]	63.85(71.14)	56.17(33)	63.97(32.84)	76.85(82.77)	77.59(59.11)	9.76E-05	4.88E-04
Cholesterol [mmol/l]	4.65(0.84)	4.79(0.9)	4.9(0.96)	4.89(0.97)	4.94(0.89)	1.42E-02	7.12E-02
Triglycerides [mmol/l] [a]	0.92(0.40)	1.10(0.64)	1.12(0.49)	1.27(0.76)	1.36(0.72)	3.79E-14	1.90E-13
HDL [mmol/l]	1.53(0.36)	1.49(0.35)	1.39(0.34)	1.37(0.32)	1.30(0.32)	9.00E-12	4.50E-11
LDL [mmol/l]	2.93(0.77)	3.0 (0.73)	3.20(0.89)	3.04(0.81)	3.12(0.80)	7.21E-03	3.61E-02
Systolic Blood Pressure [mmHg]	118.82(12.72)	121.67(13.67)	121.72(14.36)	123.23(12.6)	128.54(14.76)	5.53E-11	2.76E-10
Diastolic Blood Pressure [mmHg] [a]	72.82(10.99)	73.09(10.13)	72.95(9.53)	74.08(9.56)	76.78(11.26)	1.11E-04	5.56E-04
$CD8^+$ T cells	0.01(0.02)	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.02(0.03)	6.94E-02	3.47E-01
$CD4^+$ T cells	0.16(0.05)	0.16(0.05)	0.17(0.06)	0.18(0.05)	0.18(0.06)	8.68E-04	4.34E-03
Natural Killer Cells	0.19(0.05)	0.19(0.05)	0.20(0.05)	0.20(0.05)	0.21(0.05)	3.22E-03	1.61E-02
B cells	0.10(0.03)	0.09(0.03)	0.09(0.03)	0.09(0.03)	0.10(0.03)	3.47E-03	1.73E-02
Monocytes	0.06(0.03)	0.07(0.03)	0.07(0.03)	0.08(0.03)	0.08(0.03)	1.37E-04	6.84E-04
Granulocytes	0.53(0.08)	0.52(0.07)	0.51(0.09)	0.49(0.08)	0.48(0.08)	5.10E-11	2.55E-10
Categorical Phenotypes							
Sex [female/male]	172/29	136/64	135/65	107/93	72/129	2.20E-16	1.10E-15
Glucose tolerance [IFG/NGT] [b,d]	5/116	3/115	2/116	2/120	2/121	1.72E-01	8.60E-01
Glucose tolerance [IFG/IGT/NGT] [c]	7/1/68	6/0/70	1/2/71	8/0/68	9/2/65	3.62E-01	1.81E+00

[a] Variables log transformed to calculate the p-values. [b] Variables only available in a subset of 622 normoglycemic females in ALSPAC, distribution between quintiles (125/124/124/124/125). [c] Variable only available in 382 normoglycemic males in SABRE, distribution between quintiles (77/76/76/76/77). IFG: impaired fasting glucose. IGT: impaired glucose tolerance. NGT: normal glucose tolerance. [d] IGT was not considered in the subsample of 622 females in ALSPAC, as the maximum value of 2-h glucose < 5.0mmol/l. In contrast, IGT was reported in the subsample of males in SABRE.

Table 5-13 Summary of the association between continuous DNA methylation and glycaemic traits for top-three CpG sites identified in association with fasting insulin and the HOMA scores in the meta-EWAS. Coefficients are interpreted as the effect of 10% increase in methylation on a unit change in the outcome.

	cg06500161 (*ABCG1*)			cg18232548 (*DDC*)			cg19750657 (*UFM1*)		
	Coef.	SE	P	Coef.	SE	P	Coef.	SE	P
2-h glucose	0.23	0.33	0.49	-4.85	2.93	0.10	**0.70**	**0.33**	**0.04**
Fasting glucose	0.49	0.33	0.14	-2.75	2.96	0.35	**0.88**	**0.33**	**7.90E-03**
Fasting insulin [a]	**2.05**	**0.47**	**1.75E-05**	**-18.03**	**4.22**	**2.19E-05**	**1.58**	**0.48**	**1.03E-03**
HOMA-IR [a]	**2.14**	**0.50**	**2.32E-05**	**-18.56**	**4.48**	**3.95E-05**	**1.76**	**0.51**	**5.91E-04**
HOMA-B [a]	**1.74**	**0.44**	**8.91E-05**	**-16.02**	**3.92**	**5.02E-05**	**0.98**	**0.45**	**2.98E-02**
HbA1c [b]	0.26	0.39	0.51	-1.37	3.06	0.66	0.69	0.36	0.05
Glucose tolerance status [b, c]	Mean (SD)	Effect Size (%)	P [d]	Mean (SD)	Effect Size (%)	P	Mean (SD)	Effect Size (%)	P
NGT	0.54(0.04)	Ref	Ref	0.02(0.01)	Ref	Ref	0.76(0.04)	Ref	Ref
IFG	0.54(0.04)	0.36	0.64	0.02(0.00)	0.11	0.26	0.76(0.05)	0.83	0.32
IGT	0.55(0.04)	1.14	0.53	0.02(0.00)	0.21	0.36	0.73(0.05)	2.18	0.28
P for trend [e]			0.75			0.36			0.33

[a] Variable log-transformed to calculate the linear regression. [b] Variable only available in 382 males in SABRE. [c] Categorical variable described using the mean and standard deviation beta-values for the specific CpG sites. Effect size is interpreted as the absolute percentage change in mean beta-values between categories of glucose tolerance, considering NGT as the reference category. [d] Unadjusted p-value. [e] *P for trend* for the comparison across groups (adjusted p-value).

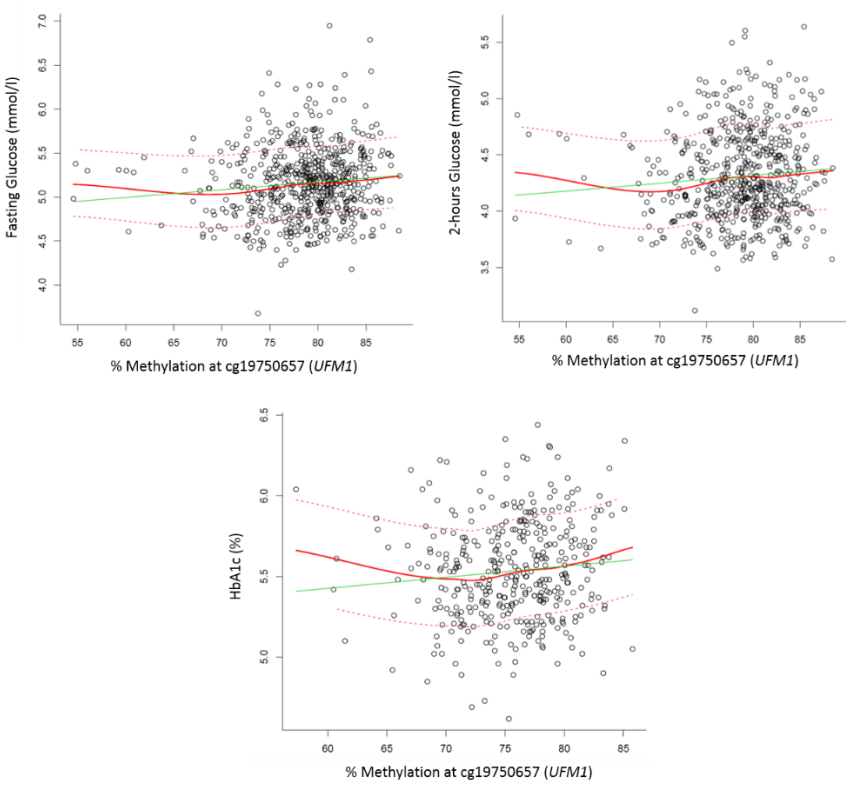


Figure 5-12 Scatterplots showing the correlation between methylation at cg19750657 (UFM1) and three glycaemic traits. Green line represents the fit line for the linear regression between methylation and the phenotype, while red-line is the mean fit smooth line for the non-parametric regression; dashed red-lines represent the upper and lower bounds of the smooth line. Methylation at the CpG in UFM1 was strongly associated with fasting glucose and 2-h glucose, but only borderline associated with HbA1c.

5.12 Enrichment of glycaemic traits-associated CpG sites for biological pathways

Biological pathway analysis was conducted independently for three glycaemic traits with the strongest evidence of association with methylation to identify processes enriched in genes annotated to the top 1000 CpG sites detected with the smallest p-value in the meta-EWAS of fasting insulin, HOMA-IR and HOMA-B. As mentioned earlier (methods Chapter 2), this arbitrary number of top sites was selected to improve the output of the enrichment analysis considering the small number of top-ranking sites (at $p<1.0x10^{-5}$) obtained in the meta-EWAS of these glycaemic traits (n= 9 to 20 sites). The enrichment analysis was conducted using the R package missMethyl with access to the Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG).

Methylation sites identified as top signals for fasting insulin, HOMA-IR and HOMA-B were spread-out around the genome, rather than being clustered in specific chromosomal regions. The 1000 CpG sites for fasting insulin, HOMA-IR and HOMA-B were near 760, 754 and 728 unique gene regions, respectively. In several cases, more than one CpG site mapped to the region of the same gene: *ABCG1* [4 sites], *NFIX* [2 sites], *AGMAT* [2 sites], and *CCNE2* [2 sites] in fasting insulin; *ABCG1* [4 sites], *NFIX* [4 sites], *HGS* [3 sites] and *HDAC4* [2 sites] in HOMA-IR, and finally, *RNF220* [4 sites], *SDCCAG8* [3 sites], *ABCG1* [3 sites] and *PHOSPHO1* [2 sites] in HOMA-B.

The list of genes annotated to CpG sites identified in fasting insulin and HOMA-IR was not enriched in gene ontology terms, but some of the genes annotated to CpG sites associated with HOMA-B were enriched in two overlapping terms, this after FDR correction for multiple testing. One GO term for CpG sites in HOMA-B was related with the *homophilic cell adhesion via plasma membrane adhesion molecules* (differentially methylated genes=26, FDR=0.04), and this term was nested within a second significantly enriched term related with *cell-cell adhesion via plasma-membrane adhesion molecules* (differentially methylated genes=26, FDR=0.04). These two terms made reference to "*the attachment of a plasma membrane adhesion molecule in one cell to an identical molecule in an adjacent cell*" according to the definition found in QuickGO (GO version: 2018-07-28, analysis date: 01-08-2018, https://www.ebi.ac.uk/QuickGO)[219]. In other words, enriched terms for HOMA-B were related to processes involved in intercellular communication. Despite identifying significant terms surpassing correction for multiple testing, in general, gene enrichment analyses were likely to be underpowered.

Likewise, none of the genes annotated to top CpG sites in fasting insulin, HOMA-IR and HOMA-B, showed enrichment for KEGG pathways after adjustment for multiple testing. Some of the top

pathways identified for fasting insulin were related with metabolic processes like c*arbohydrate digestion and absorption*, others with signalling pathways including *AMPK*, *neurotrophin* and *adipocytokine signalling pathways*, and additional pathways in fasting insulin were related with diseases including *insulin resistance, non-alcoholic fatty liver disease* and *Alzheimer's disease* (see appendix Table S8-15). Of interest, was the detection of the AMP-activated protein kinase signalling pathway, a molecular sensor of cellular energy that responds to stressful conditions like low glucose, hypoxia, ischemia, heat shock, among others (date created: April 2006, analysis date: 01-08-2018, https://www.cellsignal.co.uk)[220]. AMPK has been reported as a good therapeutic candidate for the treatment of T2D and obesity due to its role in the regulation of lipids and glucose metabolism[221]. Top 20 pathways identified for fasting insulin showed a 75% (15/20) overlap with top 20 pathways reported for HOMA-IR, but only a 5% (1/20) overlap with top 20 pathways reported for HOMA-B (see appendix Table S8-15). Pathways of interest identified in HOMA-B were the *notch signalling pathway*, *insulin resistance*, the *PPAR signalling pathway*, and the *ABC transporters pathway*, all of them related with the pathogenesis of T2D [222-224].

5.13 Comparison in the levels of methylation between blood and internal target tissues of relevance for T2D

To evaluate the relevance of peripheral blood as a source of methylation markers for glycaemic traits, the level of methylation between blood and six other internal tissues of relevance for T2D was compared using a publicly available dataset (GEO series GSE48472, https://www.ncbi.nlm.nih.gov/geo/) according to the study conducted by Slieker *et al.*[159]. Internal tissues included in the cross-tissue analysis were liver, skeletal muscle, pancreas, omentum, subcutaneous fat and spleen. CpG sites included in the analysis were top-ranking signals identified with borderline significance ($p<1.0x10^{-5}$) in the meta-EWAS of fasting insulin (n=20 sites), HOMA-IR (n=13 sites) and HOMA-B (n=9 sites).

Average correlation between methylation in blood and methylation in other tissues was always high and significant at the specific CpG sites (r range: 0.94 to 0.96, p-value range: 10^{-5} to 10^{-14}). The tissue with the weakest correlation with blood was muscle (r range: 0.89 to 0.91 and p-value range:10^{-4} to 10^{-8}), and tissues with the strongest correlation with blood were omentum (i.e. visceral fat) and spleen (r range: 0.98 to 0.99, p-value range: 10^{-7} to 10^{-14}). Methylation in blood was highly correlated with pancreatic tissue for CpG sites identified in association with HOMA-B, but this result could have been influenced by the small number of sites included in the cross-tissue comparison for this trait, relative to the number of sites included in the analysis for fasting insulin and HOMA-IR. Thus, by

increasing the number of sites compared across tissues, it also increases the chances of identifying differences between them, relative to an analysis including few sites for the comparison.

One of the limitations of this analysis was the overestimation of the correlation across tissues due to the presence of probes with very low and high average methylation, unbalancing the effect of probes with intermediate methylation. As observed in Figure 5-13 for CpG sites identified in association with fasting insulin, probes with intermediate methylation (i.e. located in the middle of the plot) deviated more from the line of complete linear correlation, compared to probes with low and high average methylation (i.e. located at the bottom-left and top-right corner of each quadrant).

High and positive correlation identified between methylation in blood and methylation in internal target tissues for T2D, suggested that methylation in blood positively resembled average methylation in more relevant but less accessible tissues for T2D and the glycaemic traits at the specific CpG sites. In addition, results suggested that blood could be a good source of biomarkers for methylation in association with fasting insulin, HOMA-IR and HOMA-B. However, these results do not suggest that methylation sites detected in blood will be successfully replicated in other tissues, and vice versa, since other tissue-specific factors are likely to play a role in the associations that can be detected, including cellular heterogeneity, batch effects, or the presence of comorbidities.

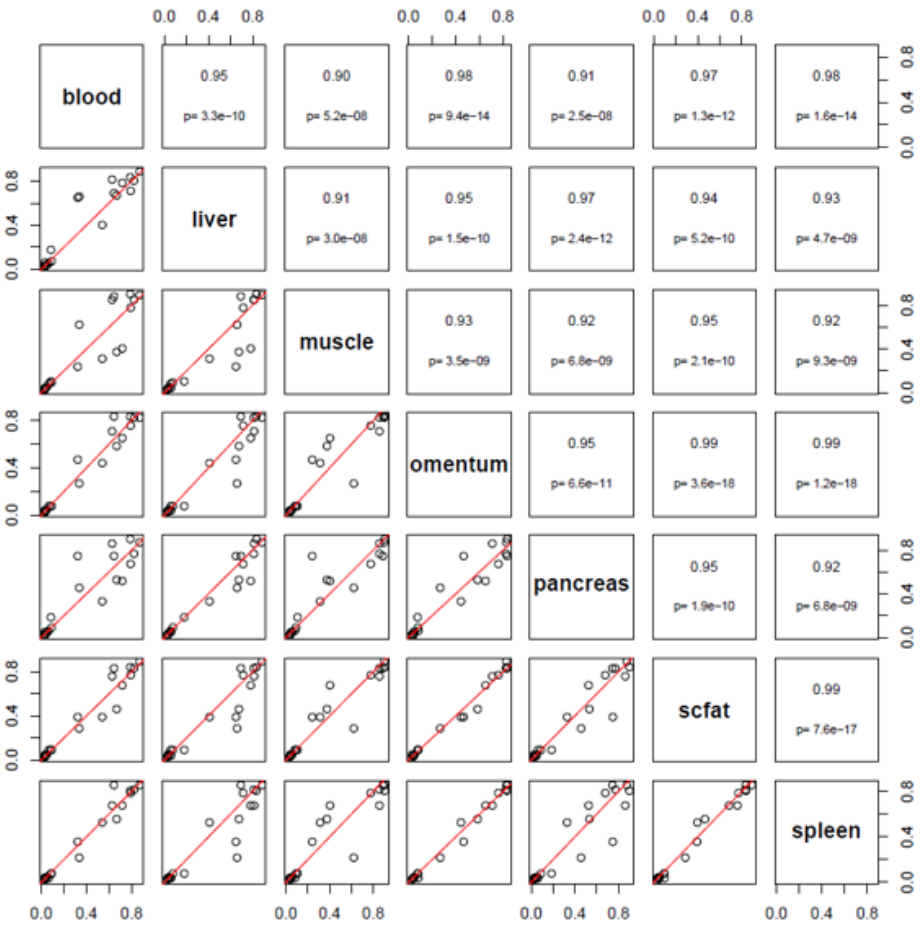


Figure 5-13 Cross-tissue comparison in the average levels of methylation for 20 CpG sites identified in association with fasting insulin in the meta-EWAS. Average methylation at the seven tissues compared was extracted from the GEO dataset GSE48472, based on the study published by Slieker et al.[159] X- and y-axis represent the average methylation calculated for CpG sites of interest in tissue one (x-axis) versus tissue two (y-axis). Scfat: subcutaneous fat; omentum: visceral fat.

5.14 Methylation score to determine the proportion of variance in the trait explained by top-ranking CpG sites identified in the meta-analysis

As referred to in the introductory section of this chapter, a methylation score was calculated using an effect size weighted linear combination of methylation values for stronger CpG sites (smallest p-value) identified in the meta-EWAS of fasting insulin, HOMA-IR and HOMA-B. Likewise, a score for HbA1c was generated using stronger markers identified in the EWAS of this trait in SABRE. The aim with constructing a score, was to determine the proportion of the variance in the trait that could be explained by stronger methylation markers identified in the meta-EWAS/EWAS, relative to the variance in the trait explained by established risk factors.

5.14.1 Methods in the assessment of the score

Methylation scores were calculated in the subsample of control females in ALSPAC (n=622) and replicated in the subsample of control males in SABRE (n=382), except for the score of HbA1c, which was only calculated in SABRE. Table 5-14 describes mean beta-values of methylation (%) for the CpG sites included in the different methylation scores, across the two studies.

Table 5-14 Comparison of the distribution of covariates and mean beta-values of methylation between ALSPAC (n=622 females) and SABRE (n=382 males). Difference in the covariates between studies was considered significant at p< 0.05.

	Units/Gene	ALSPAC (females)		SABRE (males)		
		Mean	SD	Mean	SD	P
Age	years	47.940	4.140	52.260	7.140	2.20E-16
Sex	% males	0.000	---	100.000	---	2.20E-16
Smoking	% smokers	8.840	---	72.965	---	2.20E-16
BMI	kg/m^2	25.970	4.770	26.040	3.650	6.56E-02
Fasting glucose	mmol/l	5.160	0.400	5.370	0.510	1.96E-11
Fasting insulin	µIU/ml	5.250	3.120	8.520	5.580	2.20E-16
2-h Glucose	mmol/l	4.310	0.390	4.910	1.210	2.20E-16
HOMA-IR	---	1.230	0.800	0.970	0.620	2.91E-08
HOMA-B	---	64.570	47.420	78.110	33.150	2.20E-16
cg06500161[a, b, c]	*ABCG1*	0.552	0.049	0.510	0.027	2.20E-16
cg18232548 [a, b, c]	*DDC*	0.019	0.005	0.532	0.053	2.20E-16
cg19750657[a, b]	*UFM1*	0.787	0.046	0.021	0.005	2.20E-16
cg10343442[a]	*TFG*	0.037	0.008	0.536	0.041	2.20E-16
cg17340655[a]	*DDHD2*	0.017	0.003	0.017	0.002	9.95E-01
cg06192883[a, b]	*MYO5C*	0.215	0.047	0.756	0.045	2.20E-16
cg12671247[d]	*RAD1*	0.567	0.088	0.035	0.005	2.20E-16
cg26316702 [d]	*TEKT4*	0.517	0.036	0.227	0.034	2.20E-16
cg13583414[d]	*LZTS1*	0.714	0.069	0.685	0.067	1.12E-10

CpG sites included in the methylation score for: [a] fasting insulin, [b] HOMA-IR, [c] HOMA-B, and [d] HbA1c (calculated only in SABRE). P-values were estimated using t-test for continuous parametric variables (age, bmi, fasting glucose and 2h-glucose), Mann-Whitney-Wilcoxon test for non-parametric variables (fasting insulin, HOMA-IR, HOMA-B), and Chi-square test for categorical binary variables (sex, smoking).

The strength of each score in predicting fasting insulin, HOMA-IR, HOMA-B and HbA1c, was assessed using different adjustment models (Table 5-15). A crude model without the score but including all the relevant risk factors, was considered as the reference model. Performance of models including the score versus the crude model was measured using the adjusted R^2, the root of the mean square error (RMSE), and significance of the Likelihood ratio test (P_{LRT}). Further detail on the interpretation of these parameters was provided in Chapter 2. Effect estimates of the regression between the score and the glycaemic trait were interpreted as a unit change in the outcome, per unit increase in the score (non-standardised score). Fasting insulin, HOMA-B and HOMA-IR were log-transformed before the analysis. Heterogeneity in the effect estimates of the score between studies was considered

significant at p<0.05 based on results of a Kruskal-Wallis test (i.e. assuming non-parametric distribution of means) using absolute effect estimates of the score for each study.

Table 5-15 Description of adjustment models implemented to assess independence of the methylation score from known risk factors associated with T2D-related outcomes.

Model	Covariates
Crude	Age, BMI, smoking, fasting glucose, fasting insulin[a] and 2-hours glucose[b].
M1	Score
M2	Score, age
M3	Score, age, smoking
M4	Score, age, smoking, BMI
M5	Score, age, smoking, BMI, fasting glucose, fasting insulin[a] and 2-hours glucose[b]

[a] Fasting insulin was included as a covariate in analyses for HOMA-IR and HOMA-B. [b] 2-h glucose was included as a covariate in the analysis for HbA1c.

5.14.1.1 Sensitivity analysis

Results of a basic score for fasting insulin and HOMA-IR were compared to those of an enriched score for the same traits after inclusion of additional markers identified in suggestive association (FDR ≤ 0.10). In addition, a sensitivity analysis was applied by stratifying the score into quartiles to compare the effect of the score on the outcome between quartiles (i.e. increase in the score from Q1 to Q4). Associations were regarded significant at p<0.05.

5.14.2 Methylation score for fasting insulin

Two scores were generated for fasting insulin, one containing three markers: cg06500161 (*ABCG1*), cg18232548 (*DDC*) and cg19750657 (*UFM1*), and a second score with six markers by adding three sites found in borderline association with fasting insulin (FDR > 0.05 and FDR ≤ 0.09): cg06192883 (*MYO5C*), cg10343442 (*TFG*) and cg17340655 (*DDHD2*). Descriptive statistics of the scores calculated in ALSPAC can be found in Table 5-16. In general, there was strong and positive correlation between the two scores, but there was no evidence of strong correlation between the two scores and levels of fasting insulin (rho range: -0.01 and 0.09, and p range: 0.83 and 0.02). The two scores for fasting insulin were replicated in SABRE, finding in this study similar characteristics with the scores in ALSPAC in terms of correlation between scores, and correlation between the scores and the outcome (see Table 5-16). However, differences in the mean of the scores were detected between the two studies (score1 $p_{ALSPAC\text{-}SABRE}= 4.28x10^{-10}$ and score2 $p_{ALSPAC\text{-}SABRE}=2.92x10^{-5}$). In addition, significant heterogeneity in the effect estimates of the score was detected between studies (p=0.03), where absolute effect estimates in SABRE were on average 1.54 units higher than in ALSPAC.

Table 5-16 Descriptive statistics of methylation scores calculated for fasting insulin, HOMA-IR and HOMA-B in samples in ALSPAC (n=622 females) and SABRE (n=382 males). The methylation score for HbA1c was uniquely calculated in SABRE.

Study	Outcome	Score	No. CpG	Mean median beta-values [a]	Mean absolute Effect [b]	Mean Weight	Mean (SD) Score	Median Score	Range Score	Correlation score-outcome (p-value) [c]	Correlation Score1-Score2 (p-value) [d]
ALSPAC	Fasting Insulin	1	3	0.45	9.27	1.00	0.40 (0.02)	0.40	(0.29, 0.53)	0.09 (0.02)	0.69 ($p<2.2x10^{-16}$)
		2	6	0.27	14.18	1.00	0.39 (0.02	0.39	(0.30, 0.72)	-0.01 (0.83)	
SABRE	Fasting Insulin	1	3	0.44	9.27	1.00	0.39 (0.02)	0.39	(0.32, 0.47)	0.08 (0.13)	0.77 ($p<2.2x10^{-16}$)
		2	6	0.27	14.18	1.00	0.39 (0.02)	0.39	(0.33, 0.46)	0.12 (0.02)	
ALSPAC	HOMA-IR	1	3	0.45	9.29	1.00	0.42 (0.02)	0.42	(0.30, 0.54)	0.11 (0.01)	0.86 ($p<2.2x10^{-16}$)
		2	4	0.39	7.69	1.00	0.58 (0.04)	0.58	(0.42, 0.76)	0.14 (0.001)	
SABRE	HOMA-IR	1	3	0.44	9.29	1.00	0.41 (0.02)	0.41	(0.33, 0.49)	0.08 (0.12)	0.89 ($p<2.2x10^{-16}$)
		2	4	0.39	7.69	1.00	0.58 (0.03)	0.57	(0.47, 0.69)	0.14 (0.01)	
ALSPAC	HOMA-B	1	2	0.29	9.69	1.00	0.15 (0.02)	0.15	(0.11, 0.24)	0.03 (0.43)	NA
SABRE	HOMA-B	1	2	0.28	9.69	1.00	0.15 (0.01)	0.15	(0.11, 0.19)	-0.01 (0.85)	NA
SABRE	HbA1c	1	3	0.58	2.45	1.00	1.67 (0.09)	1.67	(1.24, 1.86)	-0.11 (0.04)	NA

[a] Untransformed median beta-values of methylation. [b] Absolute effect estimates for CpG sites included in the score of fasting insulin, HOMA-IR and HOMA-B, are presented in natural-log units; effect estimates for these traits are interpreted as a log-unit change in the outcome per 10% increase in methylation. [c] Correlation calculated using the Spearman method for comparisons between the score and fasting insulin, HOMA-IR and HOMA-B, while Pearson correlation for comparisons between the score and HbA1c. [d] Correlation between two scores with different number of CpG sites, calculated for a same trait. For fasting insulin and HOMA-IR, correlation between scores was measured using the Pearson method, except for the scores of fasting insulin in ALSPAC, which were compared using the Spearman method.

Comparing between scores across studies, it was evident that the second score (n= 6 sites) had a better performance at estimating fasting insulin compared to the basic score (n= 3 sites), where no strong association was detected with fasting insulin. Thus, the second score was taken forward for further interpretation of results. Association between this score and fasting insulin was only detected after adjustment for BMI and fasting glucose (model 5) in ALSPAC, and in the unadjusted model (model 1) and the model adjusted for age (model 2) in SABRE (Table 5-17). Evidence suggested that per unit increase in the score was associated with an increase in 0.31 µIU/ml (95% CI=0.10-0.95, p=0.04) of fating insulin when the score was measured in ALSPAC, and this effect was independent of known risk factors for fasting insulin. In SABRE, per unit increase in the score was associated with an increase in 27.66 µIU/ml (95% CI=1.31-578.25, p=0.03), but this effect was likely overestimated and not independent of the effect of known risk factors for fasting insulin. Wide confidence intervals detected in results of model 2 in SABRE were indicative of large residual variation still present in this analysis, which was reduced after adjustment for BMI and fasting glucose, as indicated by a decrease in the value of RMSE from 0.559 in model 2, to 0.495 in model 5 (Table 5-17).

Maximum variation in fasting insulin explained by prediction models where the score was strongly associated with the outcome was 38.0% (ALSPAC model 5, p-model < $2.2x10^{-16}$) and 1.1% (SABRE model 2, p-model=$4.6x10^{-2}$). Looking at performance of the score in estimating the outcome, there was some suggestion that model 5 in ALSPAC performed slightly better than the crude model (RMSE model 5=0.453 versus RMSE crude model=0.454, P_{LRT}=0.04) (Table 5-17). Despite this finding, it was common to observe that the crude model (i.e. model without the score but including common risk factors) was a better predictor of fasting insulin, compared to models including the score (Table 5-17). Thus, results suggested that the score was not improving the prediction ability of common risk factors associated with fasting insulin.

Weakness of the score in predicting fasting insulin was an expected result based on the small number of CpG sites used to generate the score (n= 6 sites). Furthermore, any small variation explained by the score was likely to be confounded by some of the risk factors included as covariates, as it was demonstrated before that some of the CpG sites included in the score were associated with BMI (CpG site in *ABCG1* and *MYO5C*) and fasting glucose (CpG site in *ABCG1* and *UFM1*) (see section 5.11).

The sensitivity analysis using quartiles of the score suggested that there was an increase in 0.88 µIU/ml (95% CI=0.79-0.98, p=0.02) of fasting insulin for Q4 versus Q1 in the analysis in ALSPAC (Table 5-18), and this effect was independent of BMI and fasting glucose. In SABRE, there was an increase in 1.19 µIU/ml (95% CI=1.01-1.38, p=0.04) of fasting insulin for Q4 versus Q1 of the score, but this effect was not independent of BMI and fasting glucose (Table 5-18). In addition, differences in mean fasting insulin across quartiles were identified between Q2 and Q1 in SABRE (Figure 5-14).

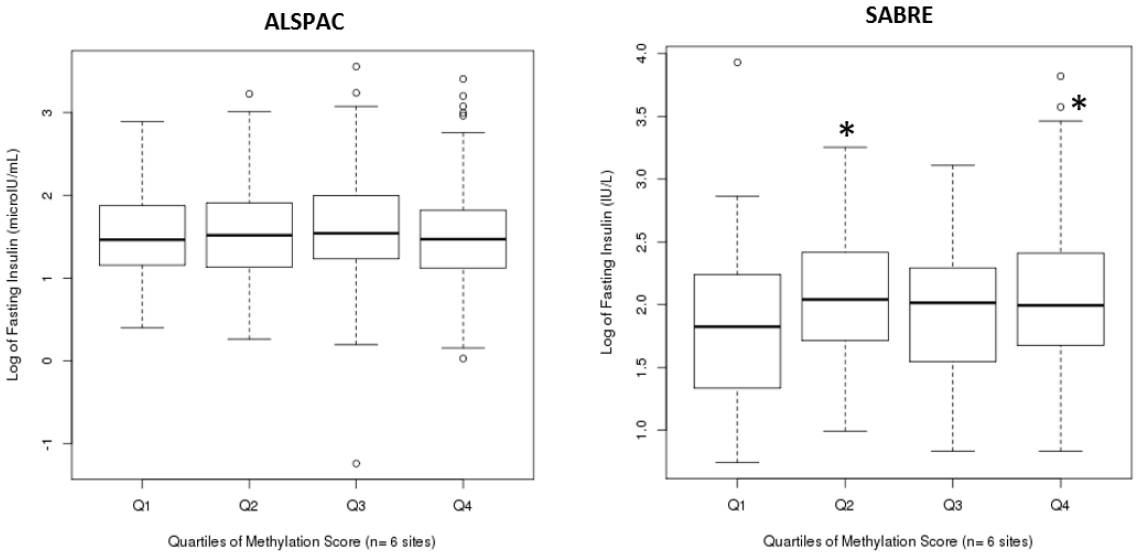


Figure 5-14 Distribution of fasting insulin per quartiles of methylation score in samples from ALSPAC and SABRE. Methylation score was generated using six CpG sites (extended score). In ALSPAC, there was no difference in mean levels fasting insulin between quartiles, while in SABRE differences in fasting insulin were detected for Q2 versus Q1 (beta=0.21, 95% CI=0.05-0.37, p=0.01), and for Q4 versus Q1 (beta=0.21, 95% CI=0.05-0.36, p=0.01). Star (*) indicates the quartile where significant differences were identified with Q1 at p < 0.05. Mean levels of fasting insulin per quartile of the score in ALSPAC (Q1=5.11, Q2=5.45, Q3=6.00, Q4=5.36 µIU/ml) and in SABRE (Q1=7.75, Q2=8.76, Q3=8.19, Q4=9.35 IU/L).

Table 5-17 Regression analysis between fasting insulin and the methylation score measured in ALSPAC (n=622 females), and replicated in SABRE (n=382 males). The score reported correspsonds to the second score for fasting insulin generated using six top CpG sites identified in the meta-EWAS of this trait. Results are interpreted as a log-unit change in fasting insulin (µIU/ml), per unit increase in the score. P-value is presented for the score, and for the different adjustment models. Associations were considered significant at p< 0.05.

Model [a]	ALSPAC						SABRE					
	Score parameters		Model parameters				Score parameters		Model Parameters			
	Effect	P	P	R^2 [b]	P (LRT) [c]	RMSE [d]	Effect	P	P	R^2	P (LRT)	RMSE
Crude	NA	NA	NA	0.38	NA	0.454	NA	NA	NA	0.21	NA	0.496
M1	-0.47	0.51	0.51	-9.12E-04	2.20E-16	0.577	**3.31**	**0.03**	**3.37E-02**	**0.01**	NA	0.560
M2	-0.50	0.48	0.17	2.43E-03	2.20E-16	0.576	**3.32**	**0.03**	**4.59E-02**	**0.01**	NA	0.559
M3	-0.49	0.49	0.05	0.01	2.20E-16	0.574	2.54	0.10	2.57E-04	0.05	NA	0.548
M4	-1.17	0.06	2.20E-16	0.27	2.20E-16	0.491	1.23	0.39	3.17E-16	0.19	NA	0.505
M5	**-1.16**	**0.04**	**2.20E-16**	**0.38**	**0.04**	**0.453**	1.11	0.43	2.20E-16	0.21	0.42	0.495

[a] Crude: model without the score but including as covariates age, smoking, BMI and fasting glucose. M1: score, M2: score and age, M3: score, age and smoking, M4: score, age, smoking and BMI, M5: score, age, smoking, BMI and fasting glucose. [b] Adjusted R^2 from the regression model measuring total variation in the outcome explained by the score alone (M1), or by the score and additional adjustment covariates included in the model (M2-M5). [c] Probability of the Likelihood ratio test to measure difference in the log-likelihood between models. P-LRT was significant at P<0.05. P-LRT was not calculated in SABRE due to the observed imbalance in the sample-size between models for this study. [d] RMSE is the root of the mean square error, and it's a measure of fitness of the model. A model with lower RMSE is a better predictor of the outcome than a model with higher RMSE. Highlighted in bold are associations were the effect of the score on the phenotype was significant at p<0.05.

Table 5-18 Association between quartiles of methylation score and fasting insulin in ALSPAC and SABRE. Quartiles were calculated for the score of fasting insulin including six top CpG sites (extended score). Associations were additively adjusted for age, smoking, BMI and fasting glucose, and were considered significant at p< 0.05.

Adjustment	ALSPAC						SABRE					
	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model
None	-0.03	0.07	(-0.15, 0.10)	0.70	-1.80E-03	0.60	**0.21**	**0.08**	**(0.05, 0.36)**	**0.01**	0.01	3.41E-02
Age	-0.03	0.07	(-0.16, 0.10)	0.69	1.43E-03	0.30	**0.20**	**0.08**	**(0.04, 0.36)**	**0.01**	0.02	4.31E-02
Smoking	-0.02	0.07	(-0.15, 0.11)	0.74	0.01	0.09	**0.17**	**0.08**	**(0.01, 0.32)**	**0.04**	0.05	2.57E-04
BMI	-0.11	0.06	(-0.22, -0.001)	0.05	0.27	< 2.2e-16	0.08	0.08	(-0.06, 0.23)	0.26	0.19	1.78E-15
Fasting Glucose	**-0.13**	**0.05**	**(-0.23, -0.02)**	**0.02**	0.38	< 2.2e-16	0.07	0.07	(-0.08, 0.21)	0.36	0.84	< 2.2e-16

Values of insulin were log-transformed before the analysis, and effect estimates should be interpreted as mean difference in fasting insulin for Q4 versus Q1 of methylation score after back-transformation of the coefficients using the exponential function [e^x]. The R^2 represents total variation in fasting insulin explained by the score alone, or by the score and additional adjustment covariates included in the model. P-values are reported for the score (P-score), and for the adjustment model (P-model). Sample size per quartile in ALSPAC (156/155/155/156) and in SABRE (96/95/95/96).

5.14.3 *Methylation score* for *HOMA-IR*

Two methylation scores for HOMA-IR were generated: a basic score using three sites which were also included in the score for fasting insulin (CpG sites in *ABCG1*, *DDC* and *UFM1*), and a second score where an additional site (cg06192883 in *MYO5C*) was included based on its borderline association with HOMA-IR (meta-EWAS FDR=0.08). Descriptive statistics of the scores calculated using samples in ALSPAC and SABRE can be found in Table 5-16. Briefly, across studies, there was high correlation between the two scores (r=0.86 to 0.89, $p<2.2x10^{-16}$), but they were weakly correlated with HOMA-IR. The level of correlation with the outcome was slightly higher for the second score (r=0.14) compared to the first score (r=0.11 and 0.08, see Table 5-16). Comparing mean values of the two scores and heterogeneity of effect estimates between studies, differences were observed in the mean of the scores (mean difference=0.01, $p_{score1}=9.93x10^{-11}$ and $p_{score2}=4.96x10^{-4}$), but not in the effect estimates, especially for the second score (difference effect estimate=-0.59, p=0.25), which showed the strongest correlation with the outcome.

From the regression analyses, the second score was strongly associated with HOMA-IR in the unadjusted model (model 1), and after adjustment for age (model 2) and smoking (model 3), but not after adjustment for BMI (model 4), fasting glucose and fasting insulin (model 5) (Table 5-19). This result suggests an overestimation of the effect of the score detected in less-adjusted models due to underlying confounding by BMI, fasting glucose and fasting insulin. No association was detected between the first score and HOMA-IR. Therefore, the second score was taken forward for further interpretation of results.

In the most adjusted model where the score was associated with HOMA-IR (model 3), evidence suggested that per unit increase in the score was associated with an increase in 7.03 (95% CI=1.77-27.66, p=0.01) and 14.59 (95% CI=2.18-97.51, p=0.01) units of HOMA-IR in the analysis in ALSPAC and SABRE, respectively (Table 5-19). Large confidence intervals in results of model 3 were indicative of large residual variation still present in these effect estimates. Maximum variation in HOMA-IR explained by the model where the score was significantly associated with HOMA-IR was 2.0% ($p=3.98x10^{-3}$, ALSPAC) and 6.0% ($p=3.8x10^{-5}$, SABRE), versus 84.6% (ALSPAC) and 82.6% (SABRE) variation in HOMA-IR explained by the crude model (without the score).

Evaluating performance of the score versus the crude model, there was no suggestion that the score outperformed the crude model in predicting variation in HOMA-IR knowing that: (1) there was no difference in RMSE between model 5 and the crude model, (2) variation reported by these two

models was equivalent (R^2= 0.83), and (3) the p-value of the likelihood ratio test indicated no significant difference in the likelihood between model 5 and the crude model (P_{LRT} =0.78 and 0.69 in ALSPAC and SABRE, respectively). Thus, evidence of the analysis in HOMA-IR suggested that the contribution of the methylation score to explain further variation in HOMA-IR was minimal or null, compared to the variation already captured by common risk factors (i.e. fasting glucose, fasting insulin and BMI) (Table 5-19).

The sensitivity analysis using quartiles of the score suggested that there was an increase in 1.19 (95% CI=1.05-1.38, p=0.01) and 1.21 (95% CI=1.03-1.42, p=0.02) units of HOMA-IR for Q4 versus Q1 of the score in the analysis in ALSPAC and SABRE, respectively. However, these effect estimates were not independent of BMI, fasting glucose and fasting insulin (Table 5-20). In addition, difference in mean HOMA-IR was also identified between Q3 and Q1 in ALSPAC (mean difference=1.26, SE=1.07, p=7.1x10^{-4}), but there was no evidence of a linear trend in the distribution of HOMA-IR across quartiles in this study (Figure 5-15).

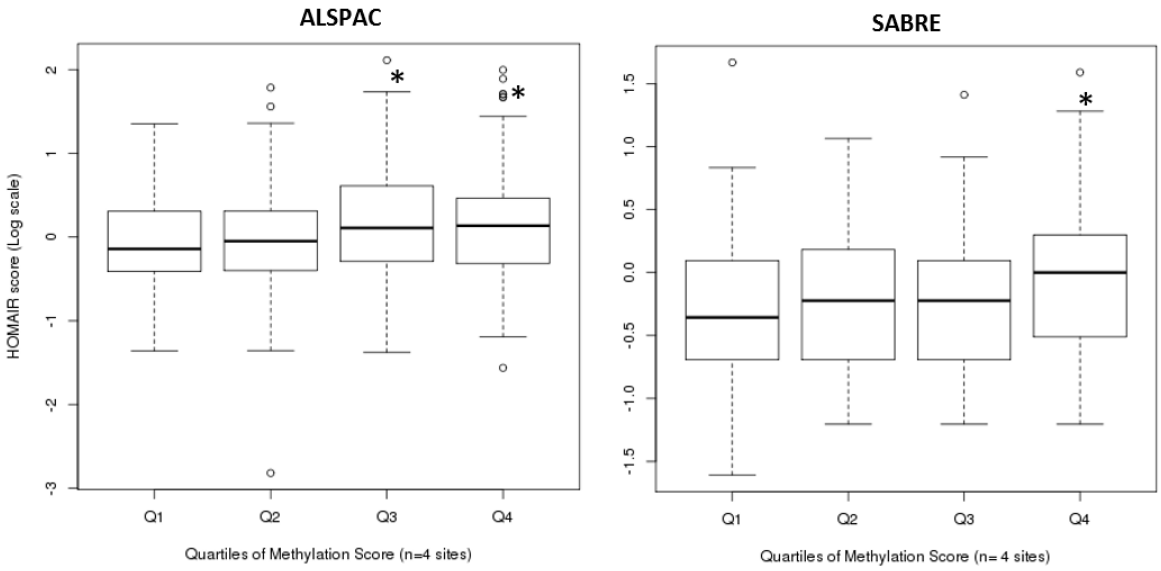


Figure 5-15 Distribution of HOMA-IR per quartiles of methylation score in samples from ALSPAC and SABRE. Methylation score was generated using four CpG sites (extended score) detected in the meta-analysis. Difference in mean of HOMA-IR was detected between Q4 and Q1 in ALSPAC and in SABRE. Star (*) indicates the quartile where significant differences in the mean of HOMA-IR were identified with the first quartile at p< 0.05. Mean values of HOMA-IR per quartile of the score in ALSPAC (Q1=1.10, Q2=1.13, Q3=1.46 and Q4=1.44) and in SABRE (Q1=0.90, Q2=0.91, Q3=0.97 and Q4=1.09).

Table 5-19 Regression analysis between HOMA-IR and the methylation score generated in ALSPAC (n=622 females) and replicated in SABRE (n=382 males). The score used for this analysis was the second score for HOMA-IR including four sites identified in the meta-EWAS of this trait. Results are interpreted as a log-unit change in HOMA-IR per unit increase in the score. P-value is reported for the score, and for the different adjustment models. Associations were considered significant at p < 0.05.

Model [a]	ALSPAC						SABRE					
	Score parameters		Model Parameters				Score parameters		Model parameters			
	Effect [b]	P	P	R^2	P (LRT)	RMSE	Effect	P	P	R^2	P (LRT) [c]	RMSE
Crude	NA	NA	NA	0.85	NA	0.240	NA	NA	NA	0.83	NA	0.233
M1	**2.00**	**4.49E-03**	**4.49E-03**	**0.01**	< 2.2e-16	0.610	**2.80**	**4.20E-03**	**4.20E-03**	**0.02**	NA	0.556
M2	**1.99**	**4.76E-03**	**0.01**	**0.01**	< 2.2e-16	0.609	**2.97**	**2.52E-03**	**5.59E-03**	**0.02**	NA	0.554
M3	**1.95**	**0.01**	**3.98E-03**	**0.02**	< 2.2e-16	0.607	**2.68**	**5.87E-03**	**3.80E-05**	**0.06**	NA	0.544
M4	0.76	0.22	< 2.2e-16	0.26	< 2.2e-16	0.525	1.11	0.23	< 2.2e-16	0.19	NA	0.503
M5	-0.08	0.78	< 2.2e-16	0.85	0.78	0.240	0.17	0.69	< 2.2e-16	0.83	0.69	0.233

[a] Crude: model without the score but including as covariates age, smoking, BMI and fasting glucose. M1: score, M2: score and age, M3: score, age and smoking, M4: score, age, smoking and BMI, M5: score, age, smoking, BMI and fasting glucose. [b] Effect estimates are in log-units since HOMA-IR was log-transformed before the regression analysis. In bold are results for the models where the score was significantly associated with HOMA-IR. [c] P-value of the likelihood ratio test was not calculated in SABRE as there was an imbalance in the number of samples between models for this study.

Table 5-20 Association between quartiles of methylation score and HOMA-IR in ALSPAC and SABRE. Quartiles were calculated for the score of HOMA-IR including four sites (extended score). Associations were additively adjusted for age, smoking, BMI, fasting glucose and fasting insulin, and were considered significant at p< 0.05.

Adjustment	ALSPAC						SABRE					
	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model
None	**0.18**	**0.07**	**(0.04, 0.31)**	**0.01**	0.02	2.25E-03	**0.203**	**0.08**	**(0.04, 0.36)**	**0.01**	0.01	7.92E-02
Age	**0.19**	**0.07**	**(0.05, 0.32)**	**0.01**	0.02	2.15E-03	**0.212**	**0.08**	**(0.05, 0.37)**	**0.01**	0.01	7.21E-02
Smoking	**0.19**	**0.07**	**(0.05, 0.32)**	**0.01**	0.03	9.50E-04	**0.189**	**0.08**	**(0.03, 0.35)**	**0.02**	0.05	5.25E-04
BMI	0.09	0.06	(-0.03, 0.21)	0.13	0.27	< 2.2e-16	0.081	0.08	(-0.07, 0.23)	0.28	0.18	3.36E-15
Fasting Glucose + Fasting Insulin	-0.01	0.03	(-0.06, 0.04)	0.72	0.85	< 2.2e-16	0.013	0.03	(-0.06, 0.08)	0.71	0.82	< 2.2e-16

Values of HOMA-IR were log-transformed before the analysis, and the effect estimates should be interpreted as mean difference in HOMA-IR for Q4 versus Q1 of methylation score after back-transformation of the coefficients using the exponential function [e^x]. Sample size per quartile in ALSPAC (156/155/155/156) and SABRE (96/95/95/96).

5.14.4 Methylation score for HOMA-B

Less informative than the score generated for fasting insulin and HOMA-IR, was the score generated for HOMA-B using two CpG sites mapping to the loci ABCG1 and DDC. General characteristics of the score can be found in Table 5-16. Briefly, this score was not correlated with HOMA-B (ALSPAC: rho=0.03 p=0.43, and SABRE: rho=-0.01 p=0.85), and results of the regression analysis showed no association between the score and HOMA-B in any of the adjustment models applied (p range: 0.18 to 1.00, appendix Table S8-16). Likewise, the sensitivity analysis using quartiles of the score indicated no significant difference in mean of HOMA-B between Q4 and Q1 of methylation score (appendix Table S8-17). Comparing the score between studies, no difference was detected in the mean of the score (mean difference=0.001, p=0.26), and there was no evidence of heterogeneity in effect estimates of the score between studies (difference effect estimate=-0.37, p=0.35). Taking together results in HOMA-B, the methylation score generated by combining two CpG sites did not capture any variation in HOMA-B and was not a better predictor of this outcome, compared to the combined effect of BMI, fasting glucose and fasting insulin (R^2=0.79 and 0.84, $p < 2.2x10^{-16}$, see appendix Table S8-16).

5.14.5 Methylation score for HbA1c

The score generated for HbA1c combined three CpG sites identified in strong association with this trait in the EWAS conducted in SABRE. CpG sites included were cg12671247 (*RAD1*), cg26316702 (*TEKT4*) and cg13583414 (*LZTS1*), all identified with FDR of 0.03 (see section 5.4.2). Descriptive statistics of the score can be found in Table 5-16. Briefly, the score was normally distributed, with mean value of 1.67 (SD=0.09), and range between 1.24 and 1.86. In addition, the score was weakly and inversely correlated with the percent of HbA1c (r=-0.104, p=0.04) (Figure 5-16).

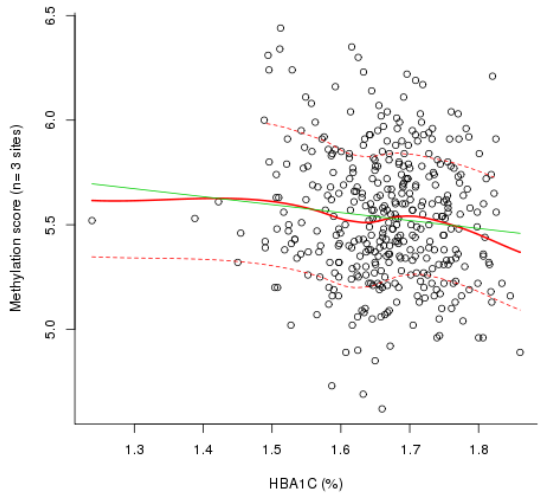


Figure 5-16 Correlation between percentage of HbA1c and methylation score. The score showed a weak negative correlation with HbA1c as represented by the fit line for the linear regression (shown in green) (r=-0.10, p=0.04), and by the smooth predictor line for the non-parametric regression (shown in red). Red-dashed lines are the confidence intervals for the smooth regression.

The regression analysis showed that the score was strongly associated with HbA1c, and this association surpassed adjustment for age, BMI, smoking, fasting glucose and 2-hours glucose (Table 5-21). In the most adjusted model, it was found that per unit increase in the score was associated with a 0.37% decrease in the percentage of HbA1c (95% CI=-0.70, -0.03, p=0.03), and the total variation in HbA1c explained by this model was 21.9% ($p < 2.2x10^{-16}$). A comparison between the crude model (i.e. without the score) and the most adjusted model with the score (model 5), suggested that inclusion of the score in model 5 improved the predictive performance of the crude model by 0.8%. This result was supported by a decrease in the value of the RMSE from 0.283 in the crude model to 0.272 in model 5, and by the significance of the p-value of the likelihood ratio test (P_{LRT}=0.03, Table 5-21), indicating strong difference in the likelihood between the two models. Despite this finding, the score alone did not explain as much variation in HbA1c (R^2=1.0%) as did common risk factors (R^2=21.1%).

Table 5-21 Regression analysis between HbA1c and a methylation score measured in samples from SABRE (n=382). The score was generated using three top CpG sites (in RAD1, TEKT4 and LZTS1) identified in the EWAS of HbA1c in SABRE. Results are interpreted as a percent increase in HbA1c, per unit increase in the score. P-value is presented for the score, and for the different adjustment models. Associations were considered significant at p< 0.05.

Model	Score parameters		Model Parameters			
	Effect	P	P	R2	P (LRT) [a]	RMSE [b]
Crude	NA	NA	NA	0.21	NA	0.283
M1	**-0.38**	**0.04**	0.04	0.01	NA	0.309
M2	**-0.37**	**0.04**	1.94E-08	0.08	NA	0.297
M3	**-0.41**	**0.02**	1.62E-09	0.11	NA	0.292
M4	**-0.43**	**0.02**	3.10E-09	0.11	NA	0.292
M5	**-0.37**	**0.03**	< 2.2e-16	0.22	0.03	0.272

[a] Probability of the likelihood ratio test to measure difference in the log-likelihood between models. P-LRT significant at p<0.05. P-LRT was not calculated in SABRE due to the observed imbalance in the sample-size between models for this study. [b] RMSE is the root of the mean square error, and it's a measure of fitness of the model. A model with lower RMSE is a better predictor of the outcome than a model with higher RMSE.

A sensitivity analysis using quartiles of the score revealed some evidence of a decrease in 0.08% (95% CI= -0.16, $-3.0x10^{-4}$, p=0.05) of HbA1c for Q4 versus Q1 of the score, and this association was likely to be independent of the known HbA1c risk factors of fasting glucose, 2-h glucose and BMI (Table 5-22). Despite suggestive differences in mean of HbA1c between Q4 and Q1 of the score, there was no evidence of a linear decrease in mean values of HbA1c across the quartiles (appendix Figure S8-16). In addition, a reduction in 0.09% (95% CI=-0.18, -0.003, p=0.04) of HbA1c for Q2 versus Q1 of methylation score was identified (see appendix Figure S8-16).

Table 5-22 Sensitivity analysis showing the association between quartiles of methylation score and HbA1c in SABRE. The score was generated using three top CpG sites identified in the EWAS of HbA1c in SABRE. Associations were additively adjusted for age, smoking, BMI, fasting glucose and 2-h glucose, and they were regarded significant at p < 0.05.

Adjustment	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model
None	**-0.09**	**0.04**	**(-0.18, -0.01)**	**0.04**	0.01	5.88E-02
Age	**-0.09**	**0.04**	**(-0.17, -4E-04)**	**0.04**	0.09	8.13E-08
Smoking	**-0.09**	**0.04**	**(-0.18, -0.01)**	**0.03**	0.11	6.73E-09
BMI	**-0.10**	**0.04**	**(-0.18, -0.01)**	**0.02**	0.11	1.02E-08
Fasting Glucose + 2h-Glucose	-0.08	0.04	(-0.16, -3E-04)	0.05	0.22	< 2.2e-16

Sample-size per quartile of the score in SABRE (96/95/95/96).

5.14.6 Summary of main findings in the methylation score analysis

Due to the small number of CpG sites included in the different scores (range between two and six sites), only modest variation in fasting insulin, HOMA-IR, and HbA1c was captured by the different scores (adjusted R^2 range: 0.3% to 1.9%). Based on results for fasting insulin and HOMA-IR, it was demonstrated that a score that includes more sites shows stronger association with the outcome and improves the predictive ability of the score, compared to a score with less CpG sites, even if some of the sites included in the score are not identified with epigenome-wide significance. In cases where the score was strongly associated with the outcome (i.e. score for fasting insulin, HOMA-IR and HbA1c, but not for HOMA-B), this association was not always independent of known risk factors. The score for HbA1c was the only one where the association with the outcome remained significant at $p < 0.05$ after adjustment for age, smoking, BMI, fasting glucose and 2-h glucose, while the score for fasting insulin in ALSPAC only reached significance after adjustment for BMI and fasting glucose.

Confounding of the score by clinical risk factors included as covariates in the model was an expected result, knowing that some of the CpG sites in the score were previously identified in association with BMI, fasting insulin, fasting glucose and 2-h glucose (see section 5.11). For instance, methylation by quintiles at the CpG in *ABCG1* was associated with BMI, fasting insulin, and fasting glucose. The sensitivity analysis using quartiles of the score confirmed previous evidence of confounding in the effect of the score by covariates, and the score for HbA1c was the only example where a modest effect of the score was still detected after complete adjustment for covariates. In summary, the methylation scores for fasting insulin, the HOMA scores and HbA1c, did not improve the predictive performance of known risk factors related with the outcome. Thus, the strength of the score should be enhanced further by using a more representative list of CpG sites obtained from well-powered EWAS or met-EWAS analyses.

Comparison in the mean of the score between studies showed that there was significant difference in the score for fasting insulin and HOMA-IR, but not for HOMA-B. In contrast, there was weak evidence of heterogeneity in effects estimates of the score between studies, with only modest heterogeneity detected in the score for fasting insulin. Even though no evidence of strong heterogeneity, it was common to observe that the absolute effect estimates of the score reported in SABRE were higher than in ALSPAC. Difference in the mean of the score between studies was an expected result, as it was previously demonstrated that mean methylation was different between studies for the CpG sites evaluated (see Table 5-14). One possible reason to explain the observed differences in methylation between studies is sex composition, which was different between the two

samples, and it is known that variation in methylation is also driven by sex[64]. Low heterogeneity in the absolute effect estimate of the score between studies suggested similar association between the score and the phenotypes, which was independent of observed differences in the distribution of fasting insulin, HOMA-IR and HOMA-B amongst studies (see Table 5-14).

5.15 Chapter summary

In this study, a meta-analysis of EWAS identified three CpG sites strongly associated with fasting insulin and HOMA-IR at *ABCG1* (cg06500161) *DDC* (cg18232548) and *UFM1* (cg19750657), two of them were also associated with HOMA-B (*ABCG1* and *DDC*). In addition, a borderline association was identified between methylation at CpG sites in *TFG* (cg10343442), *MYO5C* (cg06192883) and *DDHD2* (cg17340655), and fasting insulin; the CpG in *MYO5C* was also identified in borderline association with HOMA-IR. From the sites identified, the only one surpassing adjustment for BMI was the CpG in *DDC* associated with fasting insulin and the HOMA scores. It was demonstrated that BMI was not a confounder or a mediator of the *DDC*-HOMA's or *DDC*-fasting insulin associations, since no change in the effect estimate was identified at this site across models. Despite this result, effect estimates at the top signals identified across glycaemic traits suggested that smoking and BMI were confounders of the main association by either increasing or decreasing the effect estimate in a way that was dependent on the trait evaluated.

Univariate regressions between quintiles of methylation at CpG sites in *ABCG1*, *DDC*, *UFM1,* and risk factors for T2D, revealed strong association between these markers and the T2D metabolic and anthropometric risk factors of HDL, LDL, triglycerides, BMI and waist-circumference. Associations were also detected between quintiles of methylation and the proportion of predicted cell-counts, HOMA scores, fasting insulin, and fasting glucose for the CpG in *ABCG1*, but no association was detected between quintiles of methylation and HbA1c, 2-h glucose, or glucose tolerance status. Using continuous beta-values of methylation at the same CpG sites, *UFM1* showed strong association with fasting glucose and 2-h glucose, and borderline association with HbA1c.

Results of a cross-phenotype correlation and cluster analysis based on effect estimates obtained in the meta-analysis, showed that the level of correlation and similarity between the HOMA scores and fasting insulin was higher, than their correlation and similarity with fasting glucose and 2-h glucose. This evidence goes in line with known characteristics of these glycaemic traits. Fasting glucose was the trait with the weakest correlation with other outcomes, while fasting insulin and HOMA-IR were the traits with the highest correlation between them and with other outcomes.

Evidence of shared genetics between methylation and the glycaemic traits was identified at the CpG in *DDC*, where meQTL detected in association with this site based on middle-age methylation, were also found nominally associated with GWAS variants reported for fasting insulin, HOMA-IR and HOMA-B in recent GWAS meta-analyses for these traits. Evidence suggested that the effect of methylation at *DDC* on fasting insulin, HOMA-IR and HOMA-B was partially explained by genetic variants simultaneously influencing variation in methylation and in the glycaemic trait, meaning that methylation could be considered a mediator of the SNP-phenotype association.

In addition, an meQTL for the CpG in *DDC* was found in overlap with an eQTL for the *FIGNL1* gene identified in blood samples, and the associations meQTL-CpG and eQTL-gene were in the same direction, suggesting that methylation at the *DDC* locus was potentially mediating the eQTL-*FIGNL1* association. Despite this evidence, there was no suggestion of an association between methylation at the *DDC* locus and gene expression of the same gene, in other words, the CpG in *DDC* was not an eQTM. In contrast, an eQTM was identified at the CpG in *ABCG1*, but no meQTL was associated with this CpG site neither at the middle-age nor at the antenatal time-point. Thus, results at the *ABCG1* locus suggested no evidence of shared genetics between methylation and gene expression, and less evidence that methylation could be in the causal pathway between the genotype and gene expression.

The inclusion of 1000 of the most associated CpG sites identified for fasting insulin, HOMA-IR and HOMA-B, among which were the CpG sites in *ABCG1* and *PHOSPHO1,* two important methylation loci previously reported in association with T2D[62, 63, 87], in addition to newly reported loci in this study, revealed significant enrichment of sites in HOMA-B for a molecular process related with *intercellular communication via plasma membranes.* Furthermore, suggestive enrichment of HOMA-B-associated CpG sites was identified for the *notch signalling pathway*, a pathway for *insulin resistance*, the *PPAR signalling pathway*, and the *ABC transporters pathway*, all of them of importance in the study of T2D. A 75% overlap was detected in pathways identified for CpG sites in fasting insulin and HOMA-IR, none of them surpassing significance after multiple testing correction for enrichment. Pathways of interest detected in common between fasting insulin and HOMA-IR were the *AMPK* and the *neurotrophin signalling pathways*, and disease-related pathways for *insulin resistance* and *Alzheimer's disease.*

In a cross-tissue comparison in the levels of methylation between blood and internal target tissues for T2D using top CpG sites identified in association with fasting insulin, HOMA-IR and HOMA-B, high and strong correlation was identified between blood and other tissues of relevance for T2D. The

lowest correlation was detected between blood and muscle, and the highest correlation was detected between blood and visceral fat. High correlation between blood and pancreatic tissue was identified using HOMA-B-associated CpG sites, but this correlation was likely to be overestimated due to the reduced number of sites included in this cross-tissue comparison. Despite the observed high correlation in methylation across tissues, it is unlikely that associations identified in blood can be successfully replicated in other tissues and vice versa. Lack of replication of signals across tissues can be due to tissue-specific factors, including cellular composition, distribution of batch effects, among others.

Considering the reduced number of sites identified in the meta-EWAS of fasting insulin, HOMA-IR, HOMA-B, and in the EWAS of HbA1c, the methylation scores calculated for these traits only captured a small proportion of the total variation in the outcome (adjusted R^2 range: 0.3% to 1.9%), relative to the variation already explained by known risk factors (adjusted R^2 range: 21% to 85%). The only instance where adding the score to the predictive model including the risk factors resulted in an increase in the performance of the model, was in the score for fasting insulin in ALSPAC (n=6 sites), and in the score for HbA1c in SABRE (n=3 sites). It was also evident that the effect of the score on the outcome was not totally independent of the risk factors, and this was explained by the previously detected association between some of the CpG sites in the score, and the risk factors of BMI, fasting glucose, 2-h glucose and fasting insulin. In addition, it was demonstrated that the predictive strength of the score increased with an increase in the number of sites included in the analysis, but this also augmented the probability of adding noise (i.e. confounding) in the estimation of the phenotype. Maximum variation explained by a predictive model where the score was strongly associated with the outcome was 38.0% and 1.1% for fasting insulin, and 1.7% and 5.6% for HOMA-IR based on results in ALSPAC and SABRE, respectively. For HbA1c, maximum variation explained by models including the score in strong association was 21.9%. The score calculated for HOMA-B was not associated with the trait, therefore, it was not considered an informative predictor.

Despite important findings reported in this study, one of the limitations met was the use of two samples with different sex composition and different sample-size in the meta-analysis of EWAS, which could have limited the number of associations detected. It is known that sex can influence difference in mean methylation, and this could translate into differences in the associations detected within datasets and higher heterogeneity in results of the meta-analysis. In addition, including only two samples in a meta-analysis reduces the strength and reproducibility of signals detected, and it is more common that meta-analyses with few studies underestimate the value of

heterogeneity observed, as it was the case in the present analysis. Another limitation was the unequal representation of glycaemic traits of interest across studies, which prevented the opportunity to strengthen results identified in the individual EWAS of HbA1c, fasting proinsulin and 2-hours insulin. Limited number of signals identified in strong association in the meta-EWAS of fasting insulin, HOMA-IR and HOMA-B, implied that the methylation score for these traits was underpowered and unable to capture independent variation for these traits, as it was demonstrated here.

In conclusion, this study provided evidence of a strong novel signal identified at the CpG in *DDC* in association with multiple glycaemic outcomes, which effect was independent of BMI, and likely to be causally associated with gene expression of the *FIGNL1* gene, and with measures of fasting insulin, HOMA-IR and HOMA-B according to common genetics identified using meQTL, eQTL and GWAS data. Another signal newly reported was the CpG site in *UFM1*, with less evidence of independence from BMI in its association with fasting insulin and HOMA-IR, and with no observed overlap between meQTL for this site, and eQTL and GWAS data. This study also identified an association between multiple glycaemic traits and methylation at *ABCG1,* which has been widely reported to be associated with T2D and different glycaemic traits in previous studies. To strengthen evidence of epigenetics in glycaemic traits, future studies need to include more representative samples in the meta-EWAS, measured homogenously across sex, to obtain stronger signals that can provide better information of their functional role.

Similar to the methods and results presented in this chapter, Chapter 6 provides evidence of a meta-analysis of EWAS in T2D conducted across four independent European cohorts to strengthen current evidence of the influence of methylation on T2D, and to take forward for a causal analysis markers detected with the strongest evidence from the meta-analysis.

Chapter 6 Meta-analysis of EWAS in prevalent type 2 diabetes among Europeans

Introduction and aims of the chapter

In Chapter 4, I conducted an EWAS of prevalent T2D in a subsample of middle-age adults from ALSPAC using T2D as the exposure and DNAm as the outcome. From this analysis, I identified a strong association at the *NFYC* gene (cg15986668), and a second association with borderline significance at the *STARD10* gene (cg14045803). Knowing that there were sample size limitations in ALSPAC, I replicated the EWAS of T2D in another three European cohorts: KORA, LBC1936 and the Rotterdam Study (RSIII-1 and RS-Bios). Aiming at providing stronger evidence of differentially methylated loci associated with T2D, I conducted a meta-analysis of summary statistics from these five EWAS of T2D. Thus, the main aim in this chapter is to identify novel or ubiquitous loci associated with T2D based on results of the meta-analysis, and to investigate the functional implications of these loci in the aetiology of T2D.

In general, a meta-analysis is an statistical method that synthesises information from multiple independent studies, increasing sample size and power, and reducing the chances of false-positive findings[145]. Furthermore, a meta-analysis is based on summary data, which avoids the difficulties of requesting access to individual level data, and facilitates undertaking large-scale studies[145]. Meta-analysis has been widely used in large-scale GWAS of complex traits[145], but its implementation in epigenetic epidemiological studies is less common[225]. Even though most EWAS in T2D have included replication of selected CpG sites detected in the discovery stage[62, 64-67], few of them have implemented a meta-analysis to combine results across the discovery and the replication stage, except for Chambers *et al*.[62] and Al Muftah *et al*.[67]. However, there are two disadvantages in the approach for the meta-analysis followed by Chambers *et al*. and Al Muftah *et al*. The first disadvantage is that these studies focused on meta-analysing results only for CpG sites detected in the discovery stage, overlooking important associations that might have been identified by a meta-EWAS. The second disadvantage is that these studies used samples from different ethnicities in the discovery and the replication stage, which might have introduced further heterogeneity in the results.

In the present study I replicated the EWAS of T2D in five different studies, all of them including samples from European origin, and I summarized results from these EWAS via meta-analysis. By

undertaking this approach, I increased the sample size from previous EWAS in T2D, increased the power to detect differentially methylated loci associated with T2D, and I reduced the likelihood of introducing heterogeneity in the results by restricting the analysis to samples from similar ethnicity. All in all, results from this meta-analysis aim to provide the strongest evidence of differentially methylated loci available to date in association with prevalent T2D amongst Europeans.

The research described in this chapter aims to contribute to our understanding of DNAm in the risk of T2D by:

1. Providing evidence of stronger differentially methylated loci associated with T2D based on results from a meta-analysis of EWAS.
2. Evaluating the relevance of DNAm as an indicator of T2D by estimating the percentage of variance in T2D explained by top signals identified in the meta-analysis
3. Assessing the functional relevance of T2D-related methylation sites by implementing eQTM inspection and gene enrichment analysis.
4. Investigating the role of methylation as a potential mediator of the SNP-T2D and SNP-gene expression associations, by using publicly available data from GWAS meta-analyses of T2D, meQTL and eQTL data.
5. Interrogating possible mechanisms by which DNAm can influence T2D by estimating the correlation between DNAm and other clinical risk factors.
6. Assessing the relevance of blood as source of methylation markers for T2D by comparing levels of methylation across relevant tissues for T2D at CpG sites identified in the meta-EWAS.
7. Identifying differentially methylated regions associated with T2D using summary data from the meta-analysis and determining the functional relevance of these regions in the aetiology of T2D.

6.1 Baseline characteristics of participating cohorts

I meta-analysed results from five independent studies to test for the association between DNAm against prevalent T2D in middle-age adults from the ALSPAC, KORA, LBC1936 and the Rotterdam studies RSIII-1 and RS-Bios, using a cross-sectional case control study design. Each cohort conducted their own EWAS based on a common pre-specified analysis plan (see Chapter 2 for detail on establishing collaboration for the meta-analysis). Further description of the cohorts, definition of T2D, and cohort-specific methods implemented for pre-processing DNAm data, can be found in Chapter 2. A summary of EWAS results for each cohort were presented in Chapter 4.

The overall sample size included in the meta-analysis was 5,147 participants. From this, 496 (9.6%) were T2D cases, and 4,651 (90.4%) were controls. Table 6-1 summarizes baseline characteristics of each cohort. Briefly, the highest proportion of T2D cases was observed in the RS-Bios and the LBC1936 studies, while cases were underrepresented in ALSPAC, probably related with the age of these participants. Adults included in this analysis were in their early 50's and late 60's, there was similar proportion of sex in KORA and LBC1936, but there was overrepresentation of females in ALSPAC and in the Rotterdam studies. Mean levels of fasting glucose indicated average normoglycemia in participants across studies, but according to the definition of pre-diabetes by the WHO [1] and ADA [1], participants in RS-Bios were at higher risk of pre-diabetes (FG ≥ 5.7). Based on WHO categories for BMI (underweight if BMI < 18.5kg/m^2, normal weight if BMI 18.5-24.9kg/m^2, and obese if BMI ≥ 25.kg/m^2)[225], most participants included in this analysis were in the obese category.

In general, there was a small proportion of smokers in this sample, the highest percent was observed in RSIII-1, while the smallest percent was observed in ALSPAC. Underrepresentation of smokers in ALSPAC could have been a consequence of misclassification error by using a methylation score to predict self-reported smoking (see Chapter 4 for more on methylation score to predict smoking).

Table 6-1 Baseline characteristics of participants in each cohort included in the meta-analysis of EWAS in T2D. Continuous variables were described based on the mean and SD.

Cohort	N	T2D (% cases)	Age (yrs)	Sex (% male)	FG‡ (mmol/l)	BMI (kg/m^2)	Smoking (% current)	Ethnicity
ALSPAC	1,050	4.6	49.9 (5.4)	39.5	5.4 (1.1)	26.8 (4.7)	9.2	Europeans
KORA (F4)	1,719	9.0	61.0 (8.9)	48.9	5.6 (1.1)	28.1 (4.8)	12.5	Europeans
LBC1936	915	12.0	69.6 (0.8)	50.5	---	27.8 (4.4)	11.3	Europeans
RSIII-1*	728	10.2	59.7 (8.1)	45.4	5.5 (1.1)	27.5 (4.7)	26.9	Europeans
RS-Bios*	735	14.7	67.6 (6.0)	42.0	5.7 (1.1)	27.7 (4.1)	10.6	Europeans
Meta-analysis†	**5,147**							

*Sub-cohorts from the Rotterdam study. †From the total meta-analysis sample, 496 were T2D cases and the rest were controls. ‡Mean values of fasting glucose were not available in LC1936. Instead, mean value of HbA1c was reposted at 5.92 (SD=0.71).

Comparing baseline characteristics of participants by T2D status (see appendix Table S 8-18 to Table S8-20), differences were commonly observed in BMI, measures of blood glucose, in at least one lipid marker, in some of the predicted white blood cells (i.e. CD8T, lymphocytes or granulocytes), and in diastolic BP. Conversely, no differences were observed by T2D status in age, sex, smoking, SES and physical activity. KORA was the only cohort where an association was observed between age, sex and T2D, indicating that participants who were older or males were at higher risk of T2D compared to younger or female participants.

6.2 Quality control before meta-analysis

Results of the EWAS in T2D were passed through a QC pipeline (see Chapter 2 for a detailed description) before including them in the meta-analysis. The number of normalized autosomal probes included in the EWAS ranged between 357,018 and 385,048 CpG sites, and after QC of the datasets, the number of probes available for the meta-analysis ranged between 349,413 and 376,820 CpG sites. The QC reports (summarized in Table 6-2) showed that results of the EWAS in RS-Bios were more precise than those of the EWAS in ALSPAC, with minor variations in the precision of results between studies across adjustment models. Higher precision in RS-Bios compared to ALSPAC, can be attributed to the larger proportion of T2D cases versus controls in the former dataset, and to the smaller median SE observed in RS-Bios.

Table 6-2 QC report of results of the EWASs in T2D for five cohorts included in the meta-analysis and based on different EWAS models. Beta refers to the regression coefficient (i.e. difference in methylation β-values in cases versus controls), Min-P is the smallest P-value detected for each EWAS.

Model†	Study	Probes after QC*	Min-Beta	Max-Beta	Median SE	Min-P	Lambda
1	ALSPAC	374,901	-0.13	0.13	3.90E-03	6.24E-08	1.43
	KORA	376,820	-0.10	0.09	1.77E-03	5.84E-07	1.00
	LBC1936	370,910	-0.10	0.08	2.74E-03	1.94E-08	1.30
	RSIII-1	369,000	-0.09	0.12	2.43E-03	8.30E-07	1.40
	RS-BIOS	349,413	-0.07	0.08	1.65E-03	1.69E-06	1.04
2	ALSPAC	374,901	-0.14	0.14	3.89E-03	2.70E-07	1.48
	KORA	376,820	-0.13	0.10	1.69E-03	4.79E-07	1.02
	LBC1936	370,910	-0.10	0.08	2.65E-03	5.89E-08	1.07
	RSIII-1	369,000	-0.09	0.12	2.23E-03	2.71E-07	1.15
	RS-BIOS	349,413	-0.07	0.07	1.67E-03	1.95E-06	1.04
3	ALSPAC	374,901	-0.13	0.14	3.89E-03	3.07E-07	1.46
	KORA	376,820	-0.11	0.11	1.69E-03	4.52E-07	1.02
	LBC1936	370,910	-0.10	0.08	2.65E-03	5.18E-08	1.07
	RSIII-1	369,000	-0.09	0.12	2.23E-03	2.75E-07	1.15
	RS-BIOS	349,413	-0.07	0.07	2.36E-03	1.85E-07	0.99
4	ALSPAC	374,901	-0.12	0.14	3.94E-03	5.48E-08	1.44
	KORA	376,820	-0.10	0.10	1.71E-03	6.01E-07	1.02
	LBC1936	370,910	-0.09	0.08	2.74E-03	3.47E-07	1.08
	RSIII-1	369,000	-0.11	0.11	2.30E-03	6.63E-07	1.07
	RS-BIOS	349,413	-0.07	0.07	1.69E-03	3.32E-06	1.03

*Final number of probes after excluding probes in allosomes, and those with incomplete estimates from the EWAS. †Model 1: basic model adjusted for age, sex and SVs; Model 2: EWAS additionally adjusted for 6 Houseman cells; Model 3: EWAS additionally adjusted for smoking. Categories of smoking were non-smoker/smoker in ALSPAC, and non-smoker/former smoker/current smoker in the remaining cohorts; Model 4: EWAS additionally adjusted for BMI.

6.3 Results of the meta-analysis of EWAS in prevalent Type 2 Diabetes

A fixed-effect inverse-variance weighted meta-analysis was run to model the association between T2D as the exposure against DNAm as the outcome. The number of autosomal probes included in the analysis was 376,820 CpG sites. Overall, effect estimates detected were small: considering top CpG sites with $p<1.0x10^{-5}$, effect estimates ranged from a 0.1% decrease to a 1.4% increase in the levels of methylation in T2D cases versus controls, and median absolute effect estimate was 0.6% (see appendix Table S 8-21 to Table S8-23). The strongest signal (at Bonferroni correction for 376,820 tests $p<1.33x10^{-7}$) in association with T2D was detected at the CpG site cg19693031 in *TXNIP*, and significance of this association survived adjustment for common covariates (Figure 6-1). For the association at *TXNIP*, BMI was a confounder since adjustment for BMI reduced in 7.7% the absolute effect estimate, and the level of significance of the association (see Table 6-4). In contrast, no confounding was observed in the association at *TXNIP* with respect to predicted cell-counts and smoking (Figure 6-1). Borderline association with T2D (at $p<1.08x10^{-6}$) was detected at the intergenic

CpG cg13826139 only after adjustment for cells, smoking and BMI. As with *TXNIP*, BMI was also a confounder of the association at cg13826139 (see Table 6-4).

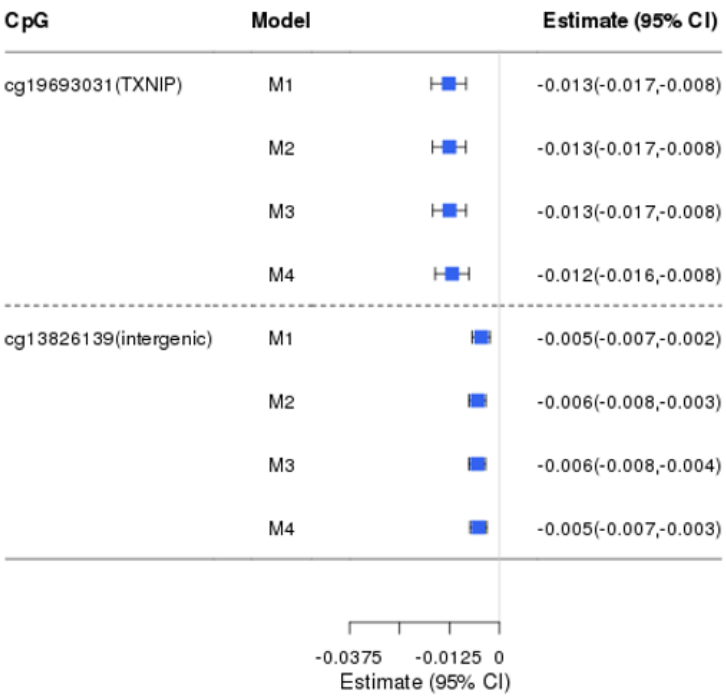


Figure 6-1 Forest-plot showing the distribution of effect estimates across adjustment models for the top two CpG sites identified in association with T2D in the meta-analysis. In brackets, the nearest gene annotated to the CpG site. Results based on models adjusted for: M1: age, sex and SVs, M2: additionally, adjusted for predicted cell-counts, M3: additionally, adjusted for smoking, M4: additionally, adjusted for BMI.

6.3.1 Effect of adjusting for smoking and BMI in results of the meta-analysis

Comparing across meta-EWAS models, most associations with borderline significance at $p<1.0x10^{-5}$ were detected in the model adjusted for smoking (n=25 CpG sites) but not for BMI (n=7 CpG sites) (Table 6-3, Figure 6-2). There was some evidence of inflation in results of the meta-EWAS based on an average lambda of 1.27. Of the associations surpassing borderline significance in the model adjusted for cells (n=22 CpG sites), 17 CpG sites were found in common between this and the model with further adjustment for smoking. At these 17 CpG sites, adjustment for smoking was shown to have little effect: median percentage difference in effect size before and after adjustment for smoking was 0.97% and only 4/17 sites changed by 2% or more. At more than half of the top 17 sites, adjustment for smoking reduced the effect size towards the null. At the remaining 7/17 sites the effect size increased after adjustment for smoking. Adjustment for smoking increased the standard error (i.e. decreased precision) at 6/17 sites.

Table 6-3 Summary of results of the meta-EWAS in T2D across different adjustment models. Effect size measured as mean difference in untransformed β-values between T2D cases and controls.

Model	Lambda	Min effect size	Max effect size	Lowest P	Top CpG	Gene	Bonferroni†	Borderline‡
1	1.19	-0.013	0.010	1.13E-08	cg19693031	*TXNIP*	1	10
2	1.29	-0.013	0.014	3.80E-09	cg19693031	*TXNIP*	1	22
3*	1.33	-0.013	0.014	4.26E-09	cg19693031	*TXNIP*	1	25
4	1.25	-0.012	0.013	4.99E-08	cg19693031	*TXNIP*	1	7

*A borderline association with T2D was detected at the intergenic CpG cg13826139 in this model (adjusted-p=0.048).
†Associations identified at $p<1.33x10^{-7}$. ‡ Associations identified at $p<1.0x10^{-5}$.

Of the associations surpassing borderline significance in the model additionally adjusted for BMI (n=7 CpG sites), 5 CpG sites were found in common between this and the model adjusted for cells. At these 5 CpG sites, adjustment for BMI was shown to have moderate effect: median percentage difference in effect size before and after adjustment for BMI was 4.65% and the maximum percentage change in effect size was 7.44%. At almost half of the top sites, adjustment for BMI reduced the effect size towards the null. At the remaining 3/5 sites the effect size increased after adjustment for BMI. Adjustment for BMI decreased the standard error (i.e. increased precision) at 5 CpG sites in common across models. Taken together results, it was evident that some residual confounding remained in the main association due to smoking and BMI. Because adjustment for BMI drastically increased the p-value of most associations detected after adjustment for cells and smoking, it is possible that much of the association between T2D and DNAm is due to differences in BMI.

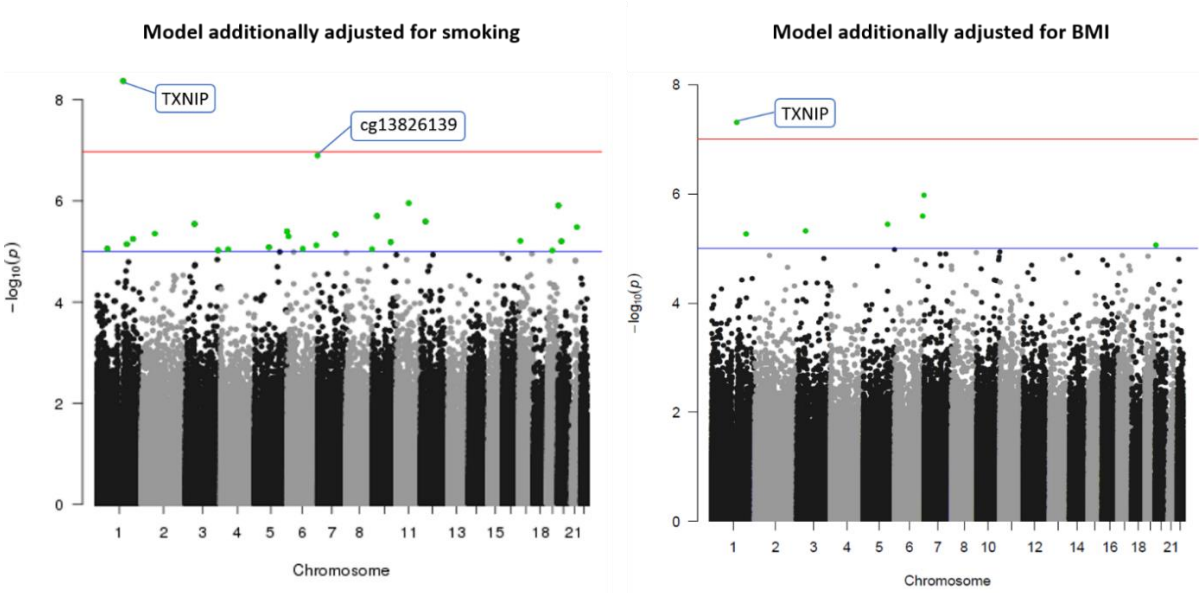


Figure 6-2 Manhattan plot for the meta-analysis of associations between middle-age DNAm and prevalent T2D after adjustment for common covariates (age, sex, SVs, Houseman cells), smoking (left plot) and BMI (right plot). Red line is the Bonferroni threshold for multiple testing at $p<1.33x10^{-7}$. Blue line represents the borderline threshold of significance at $p<1.0x10^{-5}$. Highlighted in green are CpG sites that surpassed borderline significance.

Top associations identified in the model adjusted for smoking were selected for further analysis based on their similarity with top signals detected across adjustment models. Furthermore, it was after adjustment for smoking that the highest number of associations with T2D were detected at Bonferroni significance or borderline significance. At the top 25 CpG sites, median absolute effect was a difference in methylation β-value of 0.0054 in association with T2D (i.e. 0.54% absolute change, range: 1.25% decrease to 1.36% increase) (see Table 6-4). For most of the top CpG sites, the effect of T2D was associated with a decrease in peripheral blood DNAm in middle-age adults (Figure 6-3, Table 6-4). An enrichment analysis for CpG islands and genomic features was done using Fisher's tests. However, because of the small number of probes used in the comparison (n=25 top sites vs 376,820 sites in the whole array), the analysis was underpowered to demonstrate significant enrichment (p>0.05). Overall, there was some suggestion that top CpG sites were enriched at first exons (12.0% of the top 25 sites vs 5.3% in the whole array: p=0.28), but under-represented in distant promoters (0.0% vs 14.0%; p=0.08).

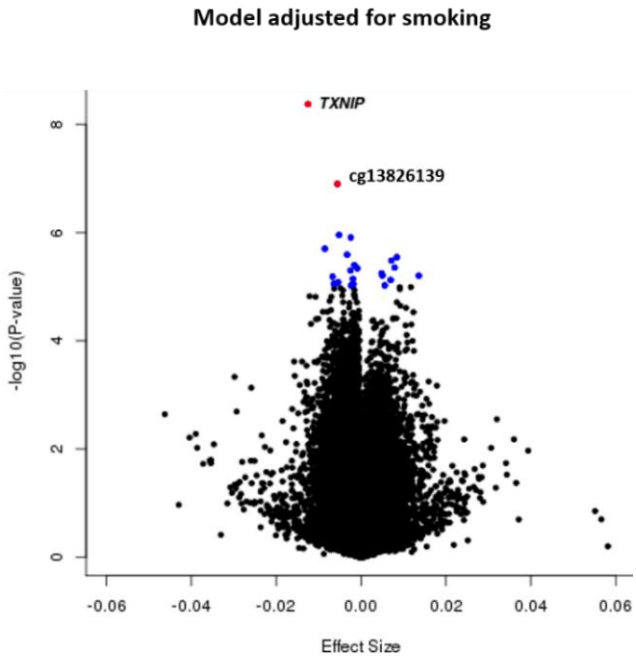


Figure 6-3 Volcano plot showing the enrichment for negative effects among the top CpG sites associated with T2D based on the model adjusted for smoking. Strongest CpG sites at Bonferroni significance are highlighted in red, while those with borderline significance (at $p<1.0x10^{-5}$) are highlighted in blue. Effect size is the difference in DNAm between cases and controls.

In a sensitivity analysis, the level of inter-study heterogeneity and influence of individual studies in results of the meta-analysis was assessed. Heterogeneity was detected at a minority of the top 25 CpG sites: 6/25 sites had I^2> 40 (range 46.3% to 89.4%), and in 3 of these 6 sites, the heterogeneity p-value was <0.05 (CpG sites mapping to *TXNIP*, *ABCG1* and *CPT1A*) (Table 6-4). For the remaining

19/25 sites, median value of heterogeneity was 12.5% (range 0.0 to 31.3%). High heterogeneity for some of the top sites was potentially influenced by effect estimates in KORA, which were very small and close to cero compared to estimates in other studies (see Figure 6-4). Despite this finding, KORA did not have large influence in results of the meta-analysis at the top 25 CpG sites.

6.3.2 Sensitivity analyses

6.3.2.1 Fixed-effect versus random-effect meta-analysis

Considering the possibility of inter-study variation in the effect estimates, the meta-analysis was repeated using a random effect model for the 25 top CpG sites. Comparing coefficients between the fixed and the random effect model, the percentage change in estimates was 0% for 10/25 sites, and it was <10% for another 10/25 sites (median percentage change in estimates: 1.2). For the remaining 5/25 sites, percentage change in estimates was >10% (maximum percentage change: 65.3). High percentage change in estimates across models was characteristic of CpG sites with high heterogeneity (I^2> 40%) (see Figure 6-4). At almost half of the top CpG sites, the random effect model reduced estimates towards the null. In the random effect model, the largest p-value at the 25 top CpG sites was 0.018, and only one site had $p<1.33x10^{-7}$ (see appendix Table S8-24).

Table 6-4 Top associations between middle-age DNAm and prevalent T2D identified in the meta-analysis using a model adjusted for age, sex, SVs, 6-Houseman cells and smoking. Beta is the effect estimate showing the difference in untransformed β-values of methylation in association with T2D. Direction: direction of effect in the individual EWAS, and they appear in the order in which studies are displayed in the table. I^2 is the heterogeneity test and the P-value for heterogeneity. Evidence of heterogeneity if p<0.05. Association in the meta-analysis were regarded significant at p<1.33x10^{-7}.

CpG	Nearest Gene	ALSPAC (N=1050)		KORA (N=1719)		LBC1936 (N=915)		RSIII-1 (N=728)		RS-Bios (N=735)		Meta-analysis (N=5147)					
		Beta	P	Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Direction	I^2	P
cg19693031	*TXNIP*	-0.022	5.93E-03	0.002	6.47E-01	-0.026	3.88E-06	-0.019	1.86E-05	-0.015	1.27E-03	-0.013	2.13E-03	4.26E-09	-+---	83.30	8.18E-05
cg13826139	*Intergenic*	-0.004	4.30E-01	-0.005	5.81E-04	-0.005	4.68E-02	-0.007	5.78E-04	NA	NA	-0.006	1.06E-03	1.27E-07	----?	0.00	9.17E-01
cg00574958	*CPT1A*	-0.005	3.60E-02	5.14E-05	9.80E-01	-0.008	6.68E-05	-0.018	6.19E-06	-0.004	1.51E-01	-0.005	1.07E-03	1.11E-06	-+---	79.50	6.16E-04
cg14275576	*Intergenic*	-0.003	2.55E-02	-0.002	1.38E-02	-0.002	5.52E-03	-0.004	2.36E-02	-0.002	5.37E-01	-0.002	5.05E-04	1.24E-06	-----	0.00	9.36E-01
cg27237541	*MYO3A*	-0.019	1.10E-02	-0.010	2.31E-03	-0.011	3.33E-02	-0.008	6.55E-02	-0.005	1.13E-01	-0.009	1.80E-03	1.99E-06	-----	0.00	4.52E-01
cg19611616	*STK38L*	-0.006	1.45E-03	-0.004	3.59E-03	-0.002	5.48E-02	-0.003	5.06E-02	-0.002	8.29E-01	-0.003	7.03E-04	2.56E-06	-----	3.90	3.85E-01
cg00082384	*NISCH*	0.014	2.60E-02	0.008	1.18E-02	0.015	1.04E-03	0.006	1.87E-01	0.005	1.31E-01	0.008	1.79E-03	2.86E-06	+++++	6.30	3.71E-01
cg06500161	*ABCG1*	0.026	2.68E-04	-0.003	2.50E-01	0.024	5.18E-08	0.010	5.06E-03	0.008	9.90E-03	0.007	1.53E-03	3.30E-06	+-+++	89.40	1.22E-07
cg14186584	*Intergenic*	-0.002	7.62E-03	-0.001	8.03E-02	-0.002	7.68E-03	-0.003	3.45E-02	-0.002	6.28E-01	-0.002	3.47E-04	4.01E-06	-----	0.00	9.37E-01
cg25741837	*SMYD5*	0.022	4.86E-03	0.004	1.88E-01	0.007	9.85E-02	0.009	2.33E-02	0.009	4.74E-03	0.008	1.71E-03	4.44E-06	+++++	12.50	3.34E-01
cg15560632	*LRCH4*	-0.001	7.24E-03	-2.86E-04	7.72E-01	-0.001	3.40E-03	-0.002	1.02E-02	-0.001	1.62E-01	-0.001	2.00E-04	4.58E-06	-----	0.00	4.72E-01
cg07400328	*MUTED*	-0.003	1.66E-02	-0.003	3.55E-03	-0.002	5.62E-03	-0.002	3.38E-01	0.001	6.99E-01	-0.003	5.48E-04	5.03E-06	----+	0.00	6.94E-01
cg22628512	*Intergenic*	0.012	2.49E-04	0.005	5.62E-02	0.004	1.41E-01	0.003	2.35E-01	0.004	1.90E-02	0.005	1.06E-03	5.66E-06	+++++	29.60	2.24E-01
cg06468695	*CCDC42*	0.003	6.33E-01	0.006	1.02E-03	0.012	1.61E-03	0.001	5.71E-01	0.004	2.77E-02	0.005	1.11E-03	6.19E-06	+++++	31.30	2.13E-01
cg06039489	*C20orf26*	0.008	3.03E-01	0.005	4.93E-01	0.028	2.22E-03	0.010	1.39E-01	0.018	3.43E-04	0.014	3.01E-03	6.27E-06	+++++	29.60	2.24E-01
cg27374726	*Intergenic*	-0.006	2.35E-01	-0.003	2.62E-01	-0.015	5.06E-04	-0.004	2.66E-01	-0.011	3.93E-04	-0.007	1.49E-03	6.52E-06	-----	52.70	7.63E-02
cg01009875	*TMCO1*	-0.001	5.65E-02	-0.002	6.95E-02	-0.002	7.90E-04	-0.002	1.50E-01	-0.008	1.81E-01	-0.002	4.31E-04	7.17E-06	-----	0.00	7.35E-01
cg17566334	*PACRG*	-0.004	6.89E-01	0.006	5.88E-03	-0.001	8.81E-01	0.014	7.53E-04	0.007	1.25E-02	0.007	1.55E-03	7.52E-06	-+-++	26.80	2.43E-01
cg07184465	*SPZ1*	-0.012	3.83E-02	-0.003	1.27E-01	-0.006	6.66E-02	-0.011	1.40E-04	-0.004	1.03E-01	-0.005	1.21E-03	8.27E-06	-----	46.30	1.14E-01
cg11851382	*PPAP2B*	-0.012	2.34E-02	-0.003	1.94E-01	-0.012	3.79E-03	-0.005	2.13E-01	-0.006	5.87E-03	-0.006	1.38E-03	8.81E-06	-----	15.10	3.18E-01
cg08273233	*HTR1E*	-0.005	4.68E-01	-0.004	7.61E-02	-0.012	6.96E-03	-0.005	1.97E-01	-0.009	9.90E-04	-0.006	1.45E-03	8.85E-06	-----	0.00	4.27E-01
cg20154947	*PLEC1*	-0.002	3.91E-04	0.001	6.47E-01	-0.002	1.49E-03	-0.002	1.48E-01	0.000	7.36E-01	-0.002	4.02E-04	9.02E-06	-+---	27.70	2.37E-01
cg13927560	*TMEM33*	-0.002	4.61E-02	-0.002	5.27E-02	-0.002	1.20E-02	-0.004	9.56E-03	-0.002	4.35E-01	-0.002	4.64E-04	9.05E-06	-----	0.00	7.82E-01
cg01317029	*FAM131A*	0.014	2.11E-03	0.004	7.64E-02	0.009	1.79E-03	0.007	5.05E-02	0.002	3.68E-01	0.006	1.26E-03	9.48E-06	+++++	46.60	1.12E-01
cg17155612	*LOC148189*	-0.002	1.02E-03	-0.003	7.01E-02	-0.001	2.66E-01	-0.005	4.71E-03	-0.001	9.15E-01	-0.002	5.33E-04	9.55E-06	-----	0.00	4.29E-01

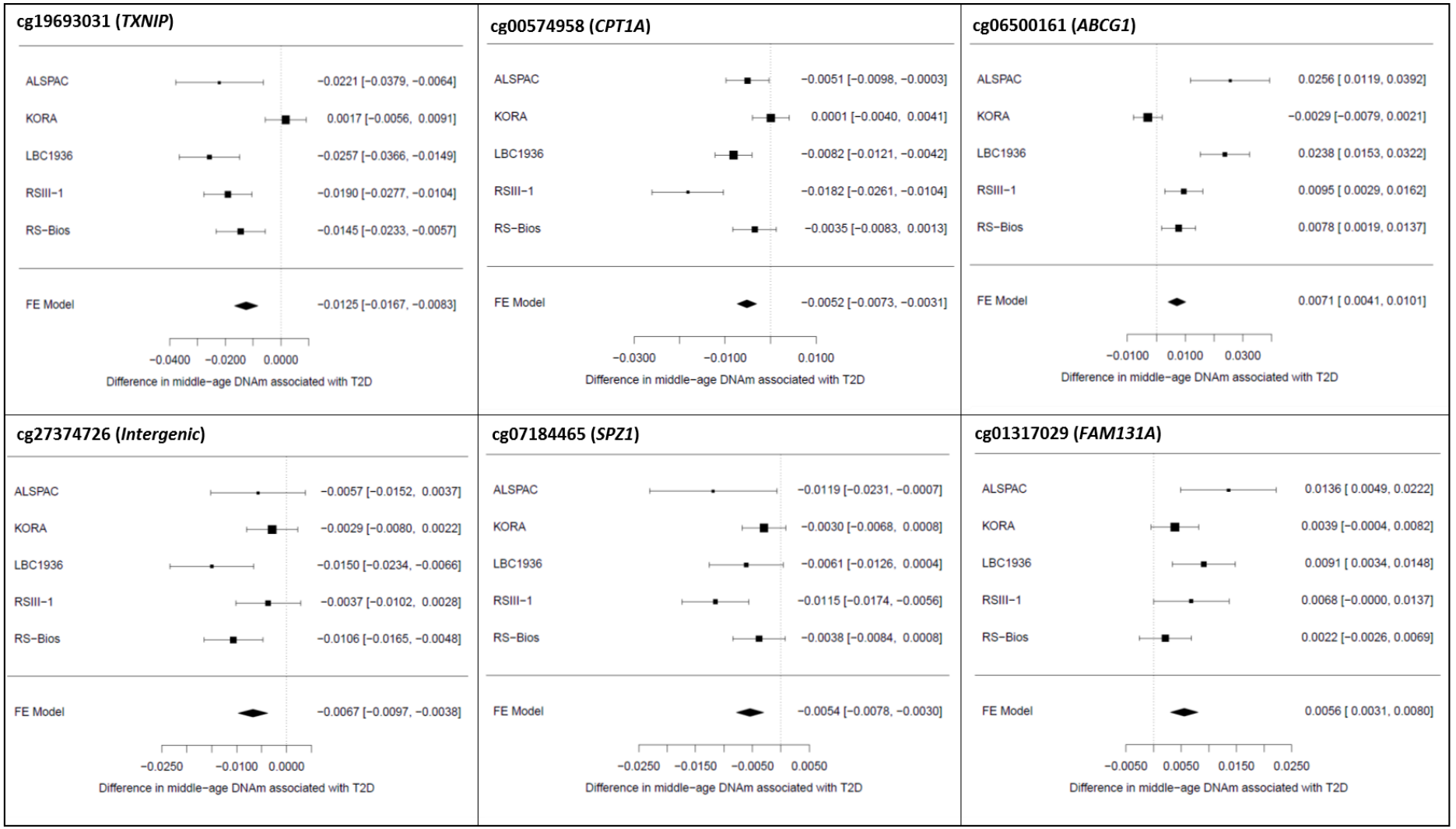


Figure 6-4 Forest-plot showing results of the EWAS in T2D for each cohort, and the combined result using a fixed-effect meta-analysis for six top CpG sites identified with the highest inter-study heterogeneity (I^2>40%). Results based on the meta-analysis adjusted for age, sex, SVs, 6-Houseman cells and smoking.

6.3.2.2 *Leave-one-out sensitivity analysis*

Forest plots showed that most cohorts agreed in the direction of the effect at the 25 top CpG sites (appendix Figure S 8-17), but one of the cohorts consistently had a disproportionately large influence on the meta-analysis at 3/25 sites (Figure 6-5, appendix Figure S 8-17). Repeating the meta-analysis excluding one study at a time showed that estimates from the EWAS in KORA were biasing results of the meta-analysis towards the null at the CpG sites cg19693031 (*TXNIP*), cg06500161 (*ABCG1*) and cg00574958 (*CPT1A*) (Figure 6-5). One possible reason why KORA had a larger influence on the meta-analysis at these sites, was because of the smaller effect estimates observed in this study, which could have been due to an over-adjustment of the EWAS by removing important variation in methylation when correcting for non-independent PCs in KORA. For the CpG site cg13826139 (intergenic), there were no results available in the RS-Bios cohort.

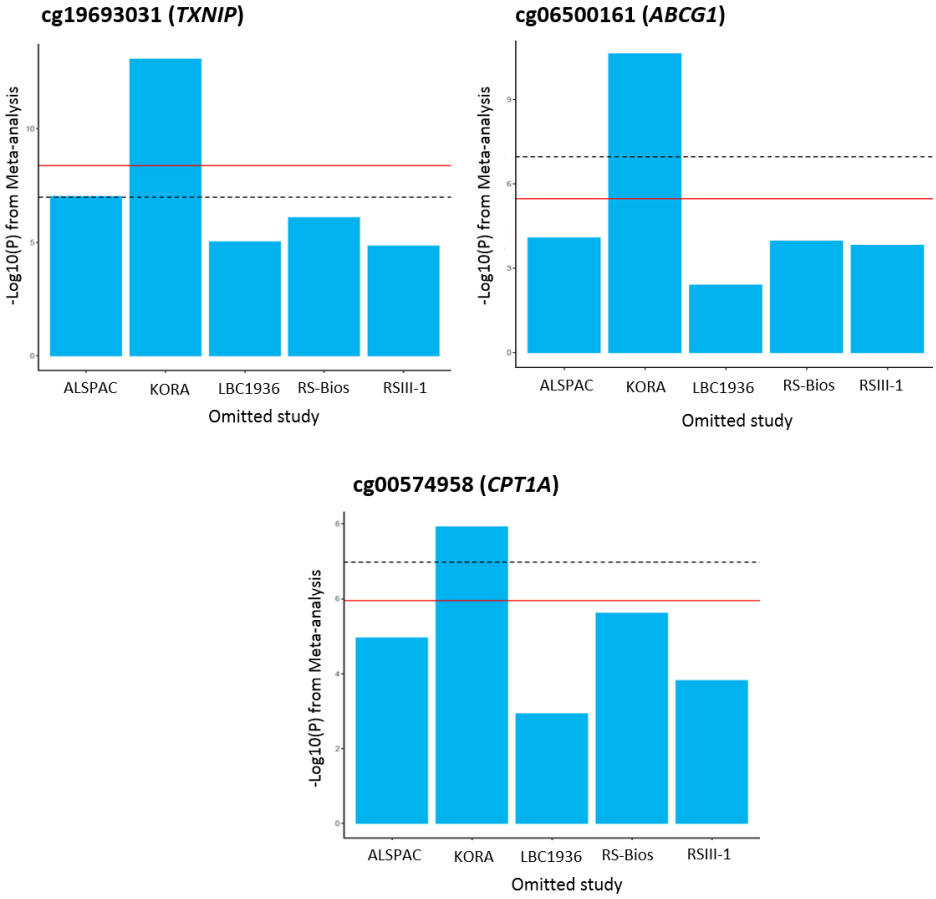


Figure 6-5 Leave-one-out sensitivity analysis showing the effect of removing one study at a time in results of the meta-analysis. Top sites shown here are three CpG sites for which the leave-one-out analysis showed that removing one study (i.e. KORA) had a large influence on the meta-analysis. Red line is the P-value of the meta-analysis, and dashed black line is the Bonferroni corrected P-value. The study showed at the bottom of the plot, is the one removed from the meta-analysis in the sensitivity analysis. From left to right: ALSPAC, KORA, LBC1936, RS-Bios and RSIII-1.

6.3.2.3 *Meta-analysis excluding KORA*

Another sensitivity analysis was performed by including in the meta-analysis 4/5 cohorts which used the same protocol to adjust for batch effects in the EWAS. In this sense, KORA was excluded from the analysis as instead of using independent SVs to correct for batch effects, this cohort used the first 10 PCs derived from control probes in the methylation array to correct for latent variation in methylation. Apart from this methodological difference, there was no other reason to exclude KORA from the meta-analysis as it was demonstrated that effect estimates of the EWAS in KORA surpassed QC inspection and were among the most precise effect estimates (i.e. smaller standard error) (see section 6.2).

Results of the sensitivity analysis showed that there was more power to detect stronger associations between DNAm and T2D in the analysis excluding KORA, compared to the main meta-analysis (Table 6-5, Figure 6-6). Furthermore, there was no consistency in results between the main meta-analysis and the sensitivity analysis: The Spearman's correlation was 0.07 among regression coefficients, and the percentage change in estimates was >10% for 12/25 top-ranking sites (median percentage change in estimates: 8.7%). For most of these 25 top sites, the effect of excluding KORA was a decrease in the effect estimate towards the null. Overall, a 40% overlap was identified between top CpG sites of the main and the sensitivity analysis (Table 6-5).

Table 6-5 Comparison of results between the meta-analysis of T2D including five cohorts, and the sensitivity analysis excluding results from the EWAS in KORA.

Model	Main analysis			Sensitivity analysis (No KORA)			Comparison	
	Top hits†	Bonferroni hits	Lambda	Top hits	Bonferroni hits	Lambda	% Overlap top hits	*rho*‡
1	10	1	1.19	56	5	1.27	60.0	0.05
2	22	1	1.29	58	6	1.36	50.0	0.07
3	25	1	1.33	58	6	1.40	40.0	0.07
4	7	1	1.25	24	2	1.30	14.3	0.06

†Associations surpassing $p<1.0x10^{-5}$. ‡Spearman's correlation coefficient calculated for regression estimates across analyses.

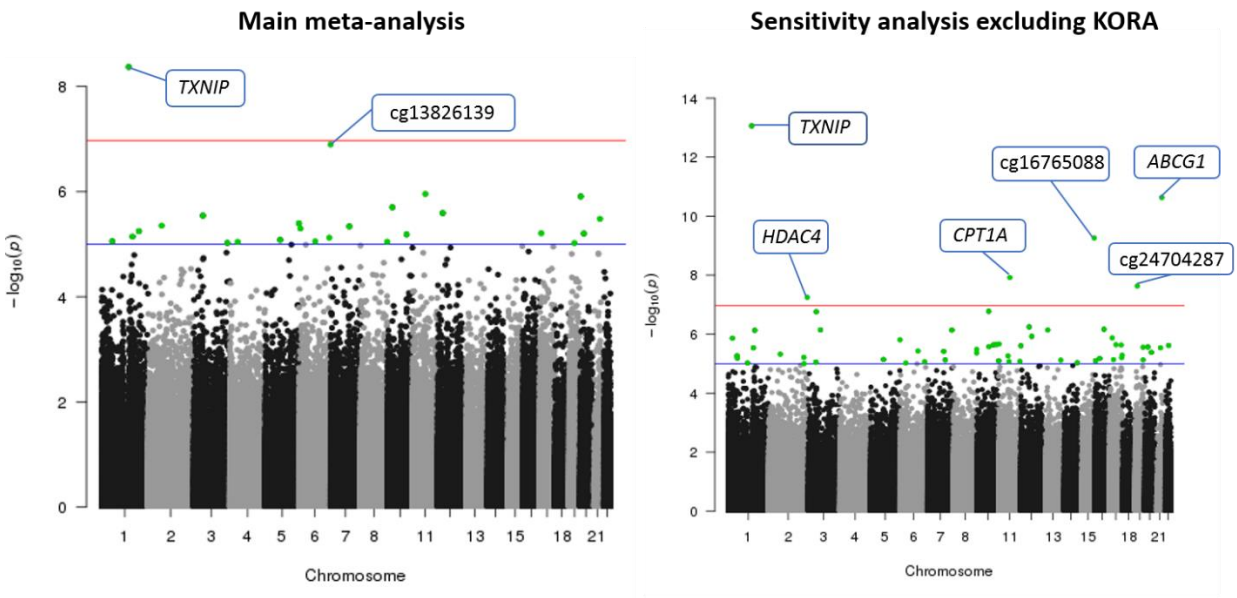


Figure 6-6 Manhattan plot comparing associations obtained between the main meta-analysis (5 cohorts), and the sensitivity analysis excluding KORA. Results shown are based on the meta-analysis adjusted for age, sex, SVs, 6-Houseman cells and smoking. Compared to the main meta-analysis, the sensitivity analysis was able to detect stronger associations with T2D surpassing epigenome-wide significance at $p<1.33x10^{-7}$ (red line), and borderline significance at $p<1.0x10^{-5}$ (blue line).

Looking at results of the sensitivity analysis, strongest associations surpassing Bonferroni correction were identified at six CpG sites mapping to the genes *TXNIP* (cg19693031), *ABCG1* (cg06500161), *CPT1A* (cg00574958), *HDAC4* (cg00144180), and at the intergenic CpG sites cg16765088 and cg24704287 (Table 6-6). The effect of T2D on difference in DNA methylation was small: at the six strongest CpG sites, T2D was associated with a median absolute change in β-values of methylation of 1.2% (range: 1.9% decrease to 1.3% increase). At most of the Bonferroni significant CpG sites, T2D was associated with a decrease in mean values of DNA methylation. Three of the six top sites identified in the sensitivity analysis have been previously reported in the literature in association with T2D, or with metabolic traits related with T2D (i.e. fasting glucose, BMI, HDL, triglycerides). There was some evidence of heterogeneity among top associations of the sensitivity analysis: 2/6 sites had I^2>40, and for the same two sites, the p-value for heterogeneity was <0.05 (Table 6-6).

Table 6-6 Strongest associations detected in a sensitivity meta-analysis excluding results from KORA, and using a model adjusted for age, sex, SVs, 6-Houseman cells and smoking. Adj. P is the Bonferroni corrected P-value of the meta-analysis, I^2 is the heterogeneity estimate, and the P value for heterogeneity.

CpG site	Chr	Gene	Meta-analysis (N=3428)						
			Beta	SE	P	Adj. P	Direction	I^2	P
cg19693031	1	*TXNIP*	-0.019	2.59E-03	8.75E-14	3.29E-08	----	0.00	0.455
cg06500161	21	*ABCG1*	0.013	1.92E-03	2.34E-11	8.81E-06	++++	77.70	0.004
cg16765088	15	*Intergenic*	-0.011	1.75E-03	5.50E-10	2.07E-04	----	0.00	0.429
cg00574958	11	*CPT1A*	-0.007	1.25E-03	1.20E-08	4.53E-03	----	72.20	0.013
cg24704287	19	*Intergenic*	-0.011	1.97E-03	2.34E-08	8.79E-03	----	0.00	0.968
cg00144180	2	*HDAC4*	0.012	2.23E-03	5.64E-08	2.12E-02	++++	4.70	0.369

At the top 58 CpG sites detected with $p<1.0x10^{-5}$ in the sensitivity analysis, median absolute effect was a difference in methylation β-value of 0.008 in association with T2D (i.e. 0.8% absolute change, range: 1.9% decrease to 1.6% increase) (see appendix Table S 8-25). Similarly to what was observed in the main analysis, for most of the top CpG sites in the sensitivity analysis the effect of T2D was associated with a decrease in peripheral blood DNAm (Figure 6-7). Heterogeneity was detected at some of the top 58 CpG sites: 16/58 sites had I^2> 40 (range 42.4% to 77.7%), and the p-value for heterogeneity was <0.05 at six of the 16 sites with high heterogeneity (see appendix Table S 8-25). The enrichment analysis for CpG islands and genomic features showed that top sites of the sensitivity analysis were not significantly enriched at any genomic feature. With respect to CpG islands, there was strong evidence that top CpG sites were enriched in shelf regions (17.2% of the top 58 sites vs 8.4% in the whole array: p=0.03) and in the open sea (58.6 vs 33.5; p<0.0001), but they were under-represented at CpG islands (8.6% vs 35.6%; p<0.0001).

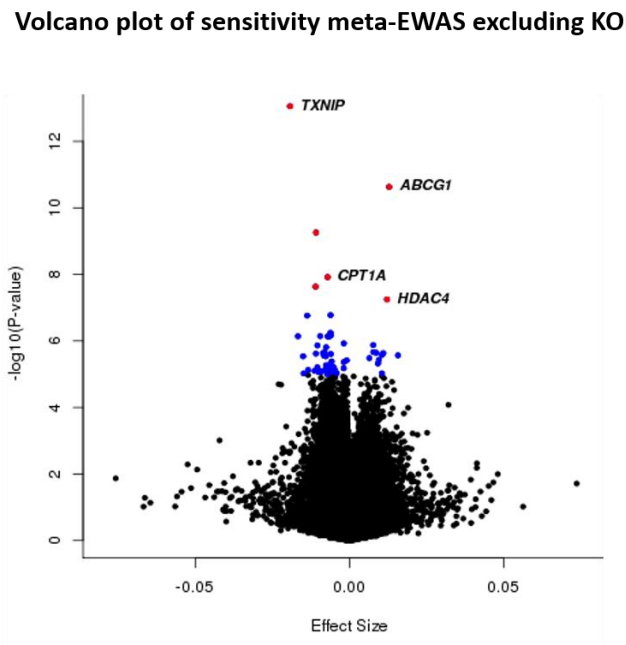


Figure 6-7 Volcano plot showing the enrichment for negative effects among top CpG sites identified in the sensitivity meta-analysis using a model adjusted for age, sex, SVs, 6-Houseman cells and smoking. Sensitivity analysis after excluding KORA from the meta-analysis. Strongest CpG sites at Bonferroni significance are highlighted in red, while those with borderline significance (at $p<1.0x10^{-5}$) are highlighted in blue. Effect size is the difference in DNAm between T2D cases and controls.

6.3.3 Risk of T2D and proportion of variance in the trait explained by DNA methylation

A logistic regression was used to test for the association between DNAm at seven top CpG sites (exposure), and T2D (outcome) using a model adjusted for common covariates and BMI. This association analysis was restricted to samples in the ALSPAC dataset. Overall, the regression analysis showed that the odds for T2D were the highest for the CpG site cg06500161 (*ABCG1*) [OR=1.13, 95%CI=1.06, 1.21] (Table 6-7). At the CpG in *ABCG1*, there was an attenuation by 2.65% in the effect of DNA methylation on T2D after adjustment for BMI, but the association remained significant. For most of the seven top CpG sites in the meta-analysis, DNA methylation was associated with a protective effect on T2D: at 5/7 sites the median odds of T2D was 0.93 per 1% increase in methylation, and the association p-value was <0.05 in 4/7 sites.

Small variance in T2D was explained by the seven top CpG sites in the meta-analysis according to estimates from the Nagelkerke's R^2 statistic using unadjusted regressions: at the seven CpG sites, the median variance explained in T2D was 2.0% (range 5.7% to 1.3%) (Table 6-7). The highest variance was captured by the CpG in *ABCG1* (5.7%). In total, the combined seven sites explained 11.2% of the total variance in T2D. In general, total variance explained by the CpG sites was lower than the variance explained by established risk factors (i.e. age, sex, smoking and BMI). For instance, for the CpG in *ABCG1*, the model including the risk factors explained 18.5% of the variance in T2D compared

to 5.7% explained by the CpG site alone, independent of the risk factors. In addition, by including in the model all the seven top CpG sites identified in the meta-analysis and the risk factors, the total variance explained in T2D was 21.55% compared to 11.2% for the model including only the CpG sites.

Table 6-7 Summary estimates of the association between DNAm and T2D at seven CpG sites detected with Bonferroni significance in the meta-analysis and sensitivity analysis. Basic model adjusted for age, sex, SVs, 6-Houseman cells and smoking; second model additionally adjusted for BMI. Results are interpreted as the odds of T2D per 1% increase in methylation β-values. Highlighted in bold are associations surpassing significance at p<0.05.

CpG	Chr	Gene	Feature‡	Basic model			Adjusted for BMI			Variance in T2D (%) †
				OR	95% CI	P	OR	95% CI	P	
cg19693031[a,b]	1	*TXNIP*	3'UTR	0.93	(0.89,0.98)	**1.10E-02**	0.94	(0.89,0.99)	**0.03**	2.0
cg00144180[b]	2	*HDAC4*	5'UTR	1.08	(1.01,1.16)	**2.40E-02**	1.08	(1.00,1.17)	**0.04**	2.9
cg13826139[a]	6	*Intergenic*	Intergenic	0.98	(0.90,1.06)	5.59E-01	0.98	(0.89,1.06)	0.57	1.3
cg00574958[b]	11	*CPT1A*	5'UTR	0.79	(0.62,1.00)	5.13E-02	0.83	(0.65,1.04)	0.11	1.3
cg16765088[b]	15	*Intergenic*	Intergenic	0.93	(0.88,0.99)	**1.64E-02**	0.93	(0.88,0.99)	**0.02**	3.3
cg24704287[b]	19	*Intergenic*	Intergenic	0.95	(0.89,1.02)	1.57E-01	0.96	(0.89,1.03)	0.22	1.3
cg06500161[b]	21	*ABCG1*	Body	1.13	(1.06,1.21)	**3.77E-04**	1.10	(1.03,1.18)	**0.01**	5.7

[a] CpG sites detected in the main meta-analysis using a fixed effect model and adjusting for age, sex, SVs, 6-Houseman cells and smoking. [b] CpG sites detected in a sensitivity analysis after excluding KORA from the main meta-analysis. †Variance calculated using the Nagelkerke's R^2 statistic derived from a completely unadjusted logistic regression. ‡Based on annotation data from the Illumina manifest.

6.3.4 Risk of T2D by quartiles of DNA methylation

Repeating the association with T2D using quartiles of DNA methylation, it was identified that most of the seven top CpG sites had a protective effect on T2D when comparing Q4 versus Q1 of methylation (Table 6-8). At 5/7 CpG sites, the upper quartile (higher methylation) was associated with a decreased risk of T2D compared to the lower quartile, and in 4/5 of these associations the p-value was <0.05. Increased risk of T2D in Q4 versus Q1 was identified for the CpG in *HDAC4*, and for the intergenic CpG cg13826139 (Table 6-8). Except for the association at *TXNIP*, most of the other associations showed an attenuation in the effect estimate towards the null, or an increase in the p-value, after adjustment for common covariates and BMI (Table 6-8). The CpG in *TXNIP* had the strongest protective effect on T2D, while the CpG in *HDAC4* had the strongest effect towards increasing the risk of T2D when comparing Q4 versus Q1 of methylation (Table 6-8).

Table 6-8 Summary of the association between quartiles of methylation at the seven top CpG sites in the meta-analysis, and risk of T2D. Associations were interpreted as the difference in the risk of T2D between Q4 and Q1 of methylation.

			Unadjusted†			Basic model†			Adjusted for BMI†		
CpG	Chr	Gene	OR	95%CI	P	OR	95%CI	P	OR	95%CI	P
cg19693031	1	*TXNIP*	0.38	(0.16,0.88)	0.02	0.32	(0.12,0.82)	0.02	0.33	(0.12,0.88)	0.03
cg00144180	2	*HDAC4*	5.94	(2.02,17.47)	1.22E-03	4.68	(1.48,14.81)	0.01	3.05	(0.94,9.86)	0.06
cg13826139	6	*Intergenic*	3.79	(1.39,10.37)	0.01	3.73	(1.06,13.16)	0.04	3.57	(0.99,12.80)	0.05
cg00574958	11	*CPT1A*	0.40	(0.16,0.97)	0.04	0.42	(0.15,1.15)	0.09	0.45	(0.16,1.28)	0.13
cg16765088	15	*Intergenic*	0.32	(0.11,0.89)	0.03	0.32	(0.11,0.97)	0.04	0.41	(0.13,1.26)	0.12
cg24704287	19	*Intergenic*	0.21	(0.06,0.73)	0.01	0.24	(0.06,0.98)	0.05	0.28	(0.06,1.2)	0.09
cg06500161	21	*ABCG1*	0.48	(0.21,1.09)	0.08	0.76	(0.3,1.91)	0.56	0.78	(0.31,2)	0.61

†Unadjusted model was a univariate T2D~ CpG regression; Basic model: adjusted for age, sex, SVs and smoking; BMI model: including covariates of basic model in addition to BMI. Proportion of T2D cases and controls by quartiles of: *TXNIP* (Q1: 20/242 Q2: 7/255 Q3: 11/250 Q4: 8/254); cg13826139 (Q1: 5/258 Q2: 9/253 Q3: 16/246 Q4: 18/245); *ABCG1* (Q1: 18/244 Q2: 11/251 Q3: 10/251 Q4: 9/253); cg16765088 (Q1: 15/244 Q2: 13/246 Q3: 14/245 Q4: 5/254); *CPT1A* (Q1: 17/245 Q2: 14/248 Q3: 10/252 Q4: 7/255); cg24704287 (Q1: 14/249 Q2: 13/249 Q3: 17/245 Q4: 3/259); *HDAC4* (Q1: 4/259 Q2: 10/252 Q3: 12/250 Q4: 22/240).

6.3.5 Methylation against categories of glucose tolerance

Based on the classification of prediabetes by the American Diabetes Association (ADA), participants with an FG above or equal to 5.6mmol/l and lower than 7.0mmol/l, are considered prediabetic. In ALSPAC, 235 participants were prediabetic, 36 were cases of T2D, and the remaining 779 were controls. Of the seven CpG sites previously detected in strong association with T2D, six sites were also identified in strong association with categories of glucose tolerance (Figure 6-8). At most of these sites, difference in methylation was observed between controls and prediabetics, and controls and T2D cases (Figure 6-8). At the CpG in *ABCG1,* difference in methylation was additionally detected between prediabetics and T2D cases. For the intergenic CpG cg16765088, difference in methylation was solely identified between prediabetics and T2D cases, and for the CpG in *CPT1A,* difference in methylation was only detected between controls and prediabetics (Figure 6-8).

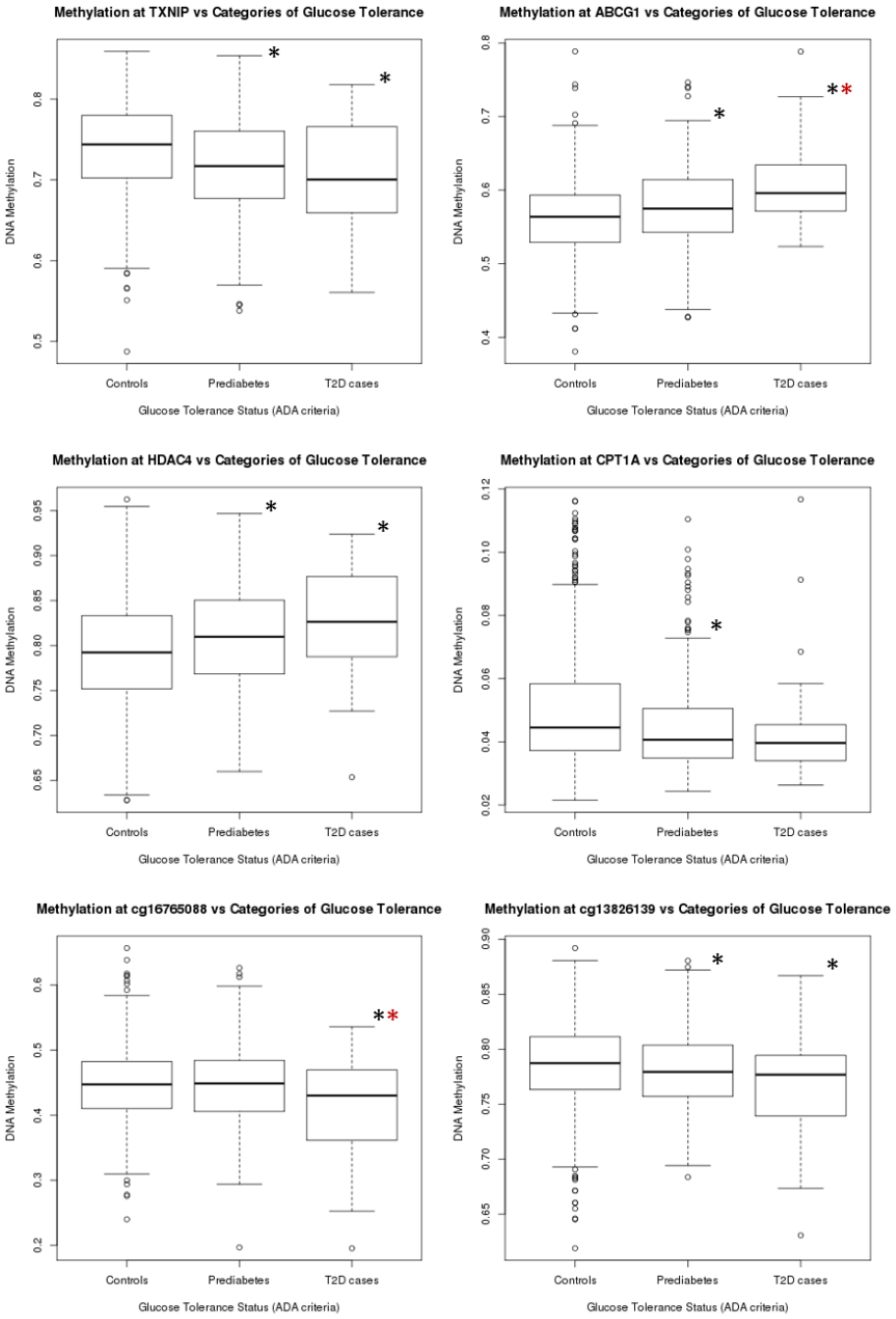
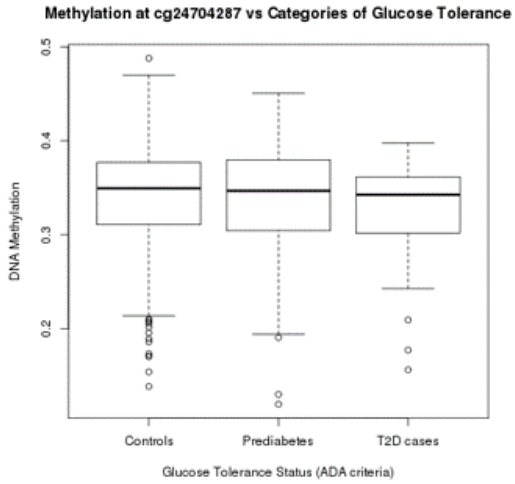


Figure 6-8 Difference in DNA methylation across categories of glucose tolerance defined by the American Diabetes Association (ADA), for some of the strongest CpG sites identified in the meta-analysis of T2D. Prediabetes was defined based on FG levels equal or above 5.6mmol/l, and lower than 7.0mmol/l . Black stars represent the category where a significant difference was identified with disease-free controls at p<0.05. Red stars for comparison where difference in methylation was identified between prediabetes and T2D cases categories.

Figure 6-8 (Continuation). No difference in methylation across categories of glucose tolerance was detected at the intergenic CpG cg24704287.



6.4 Functional interrogation of main findings

6.4.1 Identification of eQTM at the top CpG sites associated with T2D

Using the Bios QTL browser, a lookup was performed for CpG sites associated with levels of expression of the nearest gene. At 3/7 top CpG sites identified in the meta-analysis, an association was identified between DNA methylation and the expression of transcripts for the genes ABCG1, TXNIP and CPT1A (Table 6-9). In all cases, DNAm was inversely correlated with gene expression (Table 6-9). For the CpG sites in *ABCG1* and *TXNIP*, an association between DNA methylation and differential gene expression was previously reported in skeletal muscle samples from T2D donors[63]. According to previous evidence, DNA methylation was associated with reduced expression of *ABCG1* and increased expression of *TXNIP* in skeletal muscle samples from subjects with T2D[63]. Functional information of the genes annotated to the strongest CpG sites in the meta-analysis, can be found in the appendix Table S 8-26.

Table 6-9 Summary of top CpG sites detected in the meta-analysis that were identified as an eQTM for their association with gene expression in cis. eQTM were extracted from the Bios QTL browser. Associations were considered significant at p<0.05. r is the reported correlation coefficient.

CpG	Chr	Gene	Transcript	r	P
cg19693031	1	*TXNIP*	ENSG00000117289	-0.12	7.14E-08
cg00574958	11	*CPT1A*	ENSG00000110090	-0.17	3.05E-20
cg06500161	21	*ABCG1*	ENSG00000160179	-0.32	2.22E-37

6.4.2 Shared genetics between DNA methylation and gene expression

To determine if DNAm was potentially mediating the genetic association with gene expression (i.e. SNP-gene expression or eQTL), an overlap between an eQTL and an meQTL SNP for top seven CpG sites identified in the meta-analysis was investigated. meQTL were extracted from the GoDMC consortium based on associations detected in middle-age DNA methylation from peripheral blood samples. eQTL were retrieved from the GTEx portal using associations detected across different tissues (further detail on Chapter 2). A *cis* QTL was considered to overlap between datasets if a similar SNP was identified at $p<1.0x10^{-5}$ in the meQTL dataset, and at *q-value* < 0.05 in the eQTL dataset. In total, six meQTL were identified for 5/7 top CpG sites in the meta-analysis mapping to the genes *HDAC4*, *ABCG1*, *TXNIP*, and to the intergenic CpG sites cg16765088 and cg24704287. Most of these six meQTL were in *cis,* except for the meQTL detected in *trans* for the SNP-CpG pair rs6657798: cg19693031 (*SLC2A1*: *TXNIP*). No overlap was detected between six meQTL for top CpG sites, and eQTL reported in the GTEx dataset using a *q-value* < 0.05. This mismatch across datasets persisted even after relaxing the threshold to include eQTL identified with borderline significance (i.e. *q-value* >0.05).

6.4.3 Shared genetics between DNA methylation and T2D and glycaemic traits

Similar to the analysis performed previously, the overlap between meQTL for top seven CpG sites in the meta-analysis, and GWAS variants for T2D and some glycaemic traits, including fasting insulin, HOMA-IR, HOMA-B, 2-h glucose and HbA1c was investigated. GWAS variants for T2D were retrieved from a trans-ethnic GWAS meta-analysis conducted by Mahajan *et al*.[30], and from a list of 75 SNPs previously selected from different DIAGRAM studies to construct a polygenic risk score for T2D using ALSPAC samples (see Chapter 3 and Chapter 7 for more on polygenic risk score for T2D). GWAS variants for the glycaemic traits were retrieved from the most recent GWAS meta-analyses reported in the MAGIC consortium (see Chapter 2 for more on GWAS datasets included in the meQTL lookup).

Of the six meQTL previously identified for 5/7 top CpG sites in the meta-analysis, one was found in a recent GWAS of fasting insulin, but this association did not surpass nominal significance (p=0.22). Similarly, 4/6 meQTL were found in recent GWASs meta-analyses of HOMA-IR, HOMA-B, HbA1c and 2-h glucose (GWAS p-value range 0.01 to 0.89), but only 2/4 meQTL were nominally associated with HOMA-B and with 2-h glucose. The meQTL nominally associated with HOMA-B was the SNP rs220182, a *cis* meQTL for the CpG in *ABCG1*. At the SNP rs220182, the effect allele was the same across datasets, and it was associated with an increase in DNA methylation (coefficient=0.061, $p=3.5x10^{-202}$), but a decrease in the HOMA-B score (coefficient=-0.008, p=0.02). The meQTL

nominally associated with 2-h glucose was the SNP rs6657798, a *trans* meQTL for the CpG in *TXNIP*. At the SNP rs6657798, the effect allele was the same across datasets, and it was associated with an increase in 2-h glucose (coefficient=0.059, p=0.01), but a decrease in DNA methylation (coefficient=-0.459, $p=3.5x10^{-202}$). None of the six meQTL associated with methylation at 5/7 top CpG sites, were found in a GWAS meta-analysis of T2D conducted by Mahajan *et al*.[30], or within the list of 75 T2D SNPs previously selected to construct a polygenic risk score for T2D in ALSPAC.

6.4.4 Association between DNA methylation and phenotypes related with T2D

To determine possible mechanisms by which DNA methylation is associated with T2D at the top seven CpG sites detected in the meta-analysis and the sensitivity analysis, the association between methylation and important phenotypes for T2D was investigated. In further analysis, the population was stratified by quartiles of methylation for the seven CpG sites to establish if the association between DNA methylation and the trait differed across the quartiles (i.e. evidence of a linear trend). This risk factor analysis was restricted to the dataset of ALSPAC samples.

In the analysis with DNA methylation as a continuous exposure, strong associations were identified between methylation and some glycaemic traits (Table 6-10, appendix Table S 8-29). Methylation at *TXNIP* and the two intergenic CpG sites cg13826139 and cg16765088, was inversely associated with levels of 2-h glucose and FG, whereas methylation at *ABCG1* was positively associated with levels of 2-h glucose, FG, fasting insulin and the HOMA scores. Per 1% increase in methylation at *HDAC4* was positively associated with FG, fasting insulin and HOMA-IR. Similarly, methylation at *CPT1A* was positively associated with fasting insulin and the HOMA scores, but inversely associated with FG (Table 6-10). No association was identified between methylation at the intergenic CpG cg24704287, and any of the glycaemic traits (appendix Table S 8-29). At 6/7 top CpG sites, an association was identified with categories of glucose tolerance (normoglycemic, prediabetes and diabetes state). For 3/6 of these sites, an increase in methylation was associated with an increase in glucose tolerance (i.e. higher proportion of normoglycemic), while for 2/6 sites the association was the inverse. At the remaining CpG site in *CPT1A*, reduced methylation was specifically associated with the prediabetic state (Table 6-10). Overall, this analysis showed potential mechanisms by which methylation at the target CpG sites can influence the risk of T2D.

Mean β-values of methylation were significantly different across quartiles at *TXNIP*, and at the intergenic CpG cg24704287 (appendix Table S8-30). For most of the seven top CpG sites, sociodemographic, metabolic and anthropometric factors were differentially distributed across

quartiles of methylation. Generally, an increase in methylation in the upper quartile was inversely associated with age, FG, BMI, waist circumference, levels of triglycerides, total cholesterol, and measures of blood pressure (Table 6-11, appendix Table S8-31). Conversely, increased methylation in the upper quartile was generally positively associated with sex (i.e. proportion of females > males), HDL and glucose tolerance (i.e. proportion of controls > prediabetes > T2D cases). All the top CpG sites in the meta-analysis showed differential distribution of white cells across quartiles of methylation.

Fewer associations were identified between methylation by quartiles and 2-h glucose (only at *TXNIP*), C-reactive protein (CRP) (in 3/7 top sites), fasting insulin and the HOMA scores (in 3/7 top sites) (appendix Table S8-31). At *TXNIP*, increased methylation in the upper quartile was inversely associated with 2-h glucose (Table 6-11), while levels of CRP were higher in the upper quartile of *HDAC4* and the intergenic CpG cg13826139. Levels of fasting insulin and the HOMA scores were also higher in the upper quartile for *ABCG1* and *HDAC4*. The CpG in *CPT1A* and the intergenic CpG cg24704287, showed no association with anthropometric or lipid measures. At *CPT1A*, quartiles of methylation were inversely associated with age, fasting insulin, the HOMA scores, and with various white-cell types. The intergenic CpG cg24704287 was only associated with white-cell types (appendix Table S8-31).

Table 6-10 Summary of associations between DNA methylation at four of the top seven CpG sites identified in the meta-analysis, and phenotypes related with T2D. Results are based on unadjusted regressions between β-values of methylation as the exposure, against the phenotype as the outcome. Analysis restricted to samples in ALSPAC (n=1050). In bold are associations surpassing significance at p<0.05. Results for the remaining top three CpG sites can be found in the appendix Table S 8-29.

	TXNIP (cg19693031)			*ABCG1* (cg06500161)			*CPT1A* (cg00574958)			*HDAC4* (cg00144180)		
	Estimate†	SE	P	Estimate	SE	P	Estimate	SE	P	Estimate	SE	P
2-hours Glucose	-1.15	0.43	**7.22E-03**	1.38	0.50	**6.20E-03**	-2.49	1.49	0.096	0.65	0.43	0.130
Fasting glucose	-2.75	0.50	**5.00E-08**	3.40	0.60	**1.77E-08**	-5.77	1.80	**1.42E-03**	1.98	0.52	**1.34E-04**
Fasting insulin*	-0.56	0.47	2.29E-01	2.20	0.49	**7.39E-06**	-3.65	1.35	**6.89E-03**	0.97	0.41	**1.72E-02**
HOMA-IR*	-0.80	0.51	1.19E-01	2.45	0.53	**5.27E-06**	-3.87	1.46	**8.41E-03**	1.08	0.44	**1.55E-02**
HOMA-B*	0.05	0.43	9.05E-01	1.58	0.45	**4.52E-04**	-3.23	1.24	**9.24E-03**	0.64	0.37	8.73E-02
Glucose Tolerance	Mean (SD)	Effect size (%) ‡	P**	Mean (SD)	Effect size (%)	P	Mean (SD)	Effect size (%)	P	Mean (SD)	Effect size (%)	P
Normoglycemic	0.74(0.06)	Ref	Ref	0.56(0.05)	Ref	Ref	0.05(0.02)	Ref	Ref	0.79(0.06)	Ref	Ref
Prediabetes	0.72(0.06)	2.13	2.90E-06	0.58(0.06)	1.51	2.29E-04	0.04(0.02)	0.47	6.15E-04	0.81(0.06)	1.83	1.37E-04
T2D cases	0.70(0.07)	3.29	3.51E-03	0.61(0.05)	4.41	1.60E-06	0.04(0.02)	0.58	1.21E-01	0.83(0.06)	3.51	1.80E-03
P for trend‡‡			**1.45E-07**			**9.81E-09**			**2.93E-04**			**2.73E-06**

†Estimates are interpreted as the effect of 1% increase in methylation on a unit change in the phenotype. ‡ Percentage of the absolute difference in DNA methylation between the reference category (normoglycemic) and the comparison categories (prediabetes, diabetes). *Variables log-transformed before conducting the regression analysis, and available only for a subset of 645 females in ALSPAC. **Unadjusted p-value for the paired comparison between categories of glucose tolerance. ‡‡ Adjusted P-value for the comparison across categories. P< 0.05 indicates evidence of a linear trend in the association between methylation and glucose tolerance categories.

Table 6-11 Summary of associations between quartiles of methylation at the CpG in TXNIP (cg19693031), and different sociodemographic, anthropometric and metabolic factors of relevance in T2D. Sensitivity analysis restricted to samples in the ALSPAC dataset (n=1,050). Continuous variables were summarized using the mean and the standard deviation, while for categorical variables the proportion of samples per category is shown. Results are interpreted as the change in the trait between Q1 (lower methylation) and Q4 (higher methylation) of methylation at TXNIP.

	Quartile 1 (n=262)	Quartile 2 (n=262)	Quartile 3 (n=261)	Quartile 4 (n=262)	
Phenotype	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	51.94(5.74)	49.94(5.38)	49.17(4.94)	49.10(4.96)	**1.64E-10**
BMI [kg/m2]	27.31(4.5)	26.37(4.39)	26.30(4.36)	26.46(5.12)	**4.08E-02**
waist-circumference [cm]	92.89(13.24)	88.63(13.08)	87.13(12.43)	86.65(13.54)	**7.17E-08**
Fasting Glucose [mmol/l] *	5.67(1.28)	5.38(0.97)	5.31(0.79)	5.27(0.68)	**5.61E-08**
2-hours Glucose [mmol/l]	4.49(1.04)	4.33(0.82)	4.34(0.66)	4.29(0.60)	**3.02E-02**
C-reactive Protein [mg/l] *	2.05(2.65)	2.05(2.91)	2.01(3.18)	1.79(2.41)	3.55E-01
Fasting Insulin [µIU/ml] *[a]	6.18(8.08)	5.63(3.86)	6.09(4.66)	5.46(4.24)	3.43E-01
HOMA-IR *[a]	1.67(3.15)	1.33(0.99)	1.47(1.33)	1.30(1.15)	3.25E-01
HOMA-B *[a]	64.79(44.97)	66.34(40.55)	78.17(97.94)	63.55(37.13)	3.34E-01
Cholesterol [mmol/l]	4.96(0.97)	4.75(0.96)	4.79(0.84)	4.74(0.91)	**2.98E-02**
Triglycerides [mmol/l] *	1.36(0.77)	1.17(0.66)	1.1(0.56)	1.06(0.51)	**8.70E-08**
HDL [mmol/l]	1.35(0.31)	1.41(0.34)	1.42(0.35)	1.45(0.38)	**1.00E-02**
LDL [mmol/l]	3.09(0.87)	2.99(0.82)	3.06(0.77)	3.01(0.81)	5.06E-01
Systolic Blood Pressure [mmHg]	127.37(14.33)	122.89(12.96)	122.77(14.32)	119.35(13.86)	**2.65E-09**
Diastolic Blood Pressure [mmHg] *	75.53(10.74)	74.13(9.51)	74.35(12.17)	72.31(9.28)	**3.34E-03**
$CD8^+$ T cells	0.02(0.03)	0.02(0.03)	0.01(0.03)	0.02(0.03)	1.70E-01
$CD4^+$ T cells	0.18(0.06)	0.17(0.06)	0.17(0.05)	0.16(0.05)	9.09E-02
Natural Killer Cells	0.21(0.06)	0.2(0.05)	0.20(0.05)	0.20(0.05)	2.80E-01
B cells	0.1(0.03)	0.09(0.03)	0.10(0.03)	0.10(0.03)	6.01E-01
Monocytes	0.08(0.03)	0.08(0.03)	0.07(0.03)	0.07(0.03)	**2.67E-04**
Granulocytes	0.49(0.08)	0.51(0.09)	0.52(0.08)	0.52(0.08)	**7.32E-05**
Categorical Phenotypes					
Sex [female/male]	101/161	159/103	185/76	199/63	**<0.001**
Glucose tolerance [Normoglycaemic/Prediabetes/Diab] †	165/80/17	191/69/2	210/42/9	212/44/6	**7.18E-07**

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females in ALSPAC, distribution between quartiles (161/161/161/161).
†categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and diabetes if FG≥7.0mmol/l).

6.4.5 Cross-tissue comparison in the levels of DNA methylation

Apart from blood, differential DNA methylation at some of the top CpG sites identified in the meta-analysis has been previously reported in primary target tissues for T2D in other studies. For instance, Dayeh *et al*.[63] reported dysregulation of methylation at the CpG cg19693031 (*TXNIP*) in skeletal muscle and pancreatic islets from T2D cases. In addition, Dayeh *et al*.[63] and Nilsson *et al.*[71] identified that methylation of the CpG cg06500161 in *ABCG1* was significantly increased in adipose tissue of subjects with T2D compared to controls. However, less is known about differential methylation in target primary tissues for other CpG sites detected in the meta-analysis.

Top CpG sites in the meta-analysis of T2D identified at $p<1.0x10^{-5}$ (n=25 sites), were used to estimate the correlation between DNA methylation in blood, and methylation in six internal tissues of relevance for T2D. This comparison was made *in silico* using a publicly available dataset (GEO series GSE48472, https://www.ncbi.nlm.nih.gov/geo/) that has been implemented throughout this thesis for the cross-tissue comparison analysis. At the top 25 CpG sites, high and strong correlation was consistently identified between DNA methylation in blood, and methylation in liver, skeletal muscle, visceral fat (i.e. omentum), pancreas, and subcutaneous fat (r range 0.87 to 0.98) (Figure 6-9). The lowest correlation was detected between blood and pancreas (r=0.87). Although overall correlation is high, there are specific CpG sites where correlation among tissues is low.

The same cross-tissue comparison was made for top 58 CpG sites detected in a sensitivity analysis (i.e. meta-analysis excluding KORA, see section 6.3.2.3). As before, high and strong correlation was observed between blood and other internal tissues (r range 0.65 to 0.94), and the lowest correlation was observed between blood and pancreas (r=0.65) (Figure 6-9). Overall, this *in silico* comparison gives an idea of the average methylation expected in relevant tissues for T2D based on CpG sites detected in blood, but profiles of methylation used in this comparison were not characteristic of T2D status.

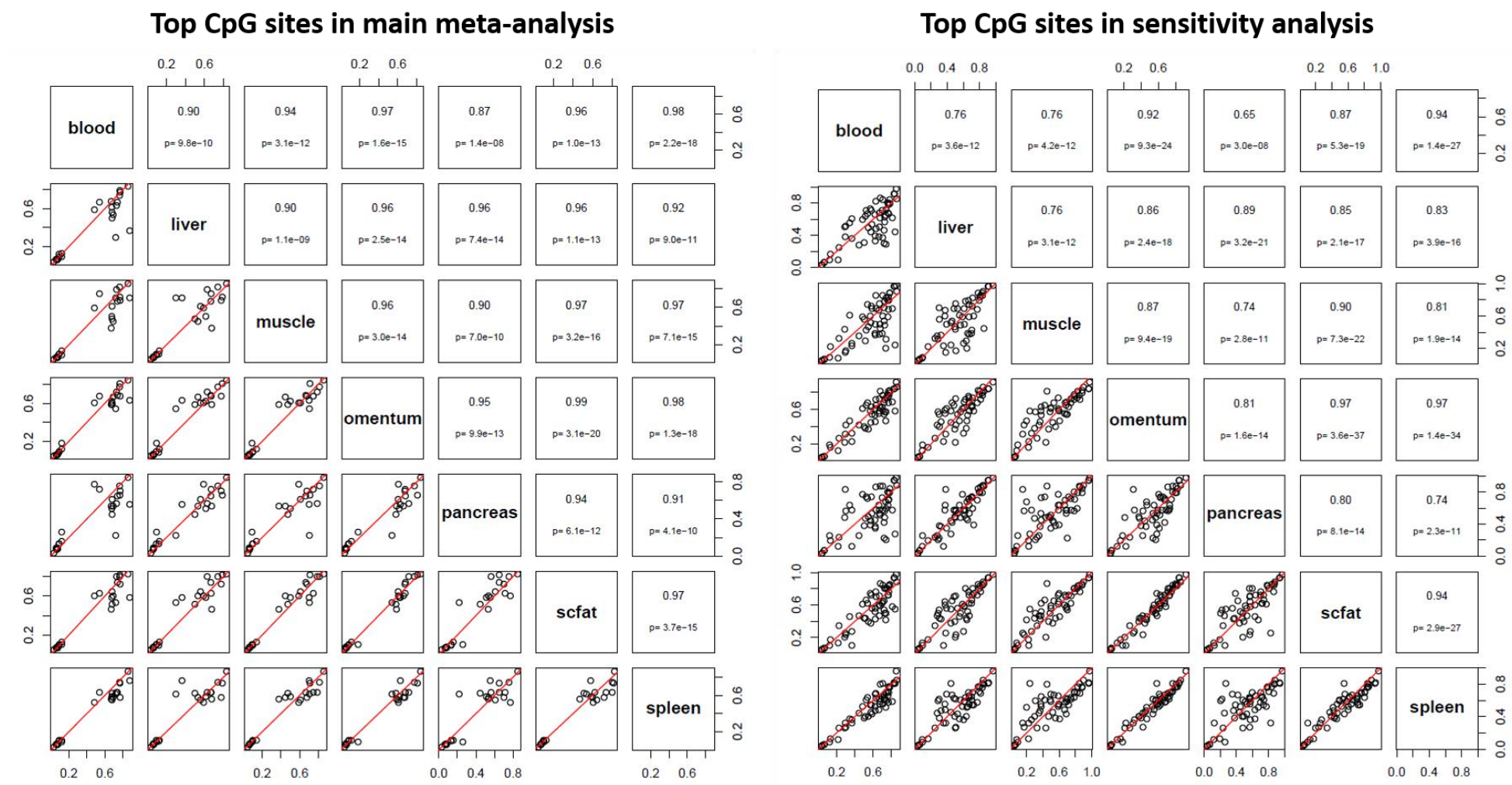


Figure 6-9 Cross-tissue comparison of DNA methylation at top CpG sites (at $p<1.0x10^{-5}$) identified in the meta-analysis of T2D, and in a sensitivity analysis excluding results from the KORA study. Methylation profile across tissues was extracted from the GEO dataset GSE48472. Scfat: subcutaneous fat.

6.4.6 Enrichment analysis for biological processes

The GO and KEGG libraries were used to identify potential enrichment for molecular, cellular or biological processes, or for metabolic pathways related with T2D, among top 25 CpG sites detected in the meta-analysis. T2D-associated methylation sites were not restricted to certain chromosomal region, as they were found throughout the genome. Top 25 T2D-associated methylation sites are near 20 unique gene regions (5 sites were intergenic), and there was no enrichment for GO terms, or for metabolic pathways in KEGG after adjustment for multiple testing (see appendix Table S8-27 and Table S 8-28). Despite no evidence of enrichment, some of the pathways detected in KEGG were related with the pathogenesis of T2D: "fat digestion and absorption", "ABC transporters", "PPAR signalling pathway", and the "insulin resistance" pathway.

The enrichment analysis was repeated for top 58 CpG sites detected in the sensitivity analysis. These CpG sites were near 41 unique gene regions across the genome: 2 sites were annotated to the same gene (*RPL13AP5*), and 16 sites were intergenic. The list of 58 CpG sites was also not enriched for GO terms or KEGG pathways (see appendix Table S8-27 and Table S 8-28). Even though there was no enrichment, some of the top terms reported in GO related with the positive regulation of cholesterol, sterols, and biosynthetic processes for lipids, which are important metabolic processes in T2D. Top pathways reported in KEGG in relation with T2D were: "insulin resistance", "adipocytokine signalling pathway", "TNF signalling pathway", "steroid biosynthesis", "insulin signalling pathway", "fat digestion and absorption", and "ABC transporters" (appendix Table S 8-28).

6.5 Regional analysis of differential methylation associated with T2D

6.5.1 Top DMRs associated with T2D

CpG sites in a region tend to share similar levels of methylation, and regions of differential methylation are more likely to have biological implications in gene function and regulation than isolated CpG sites. Thus, a DMR analysis was conducted using *Comb-p* and summary estimates for the individual CpG sites obtained from the meta-analysis. Further detail of the method used for DMR identification can be found in Chapter 2.

The region-based analysis accounting for the spatial correlation of p-values, allowed identification of 33 DMRs associated with T2D (Sidak corrected p-value<0.05). DMRs were identified at 33 unique genomic regions spread out throughout the genome (Table 6-12). At the 33 T2D DMRs, median DMR size was 268bp (size range 64bp to 388bp), median CpG count per DMR was 4 sites (CpG range 2 to 9 sites), and median of the absolute percentage change in methylation between T2D cases and

controls was 0.35 (change range 0.06 to 1.45). For more than half of the DMRs, T2D was associated with hypomethylation of the region (Table 6-12). The strongest DMR (estimate=0.01; Sidak p-value=$1.04x10^{-6}$) was annotated to the *ADCY7* gene. This DMR was hypermethylated in T2D cases compared to controls. Three gene regions were found in common between the DMR analysis and the individual CpG site meta-analysis. The overlapped regions were annotated to the *CPT1A*, *TXNIP* and *PLEC1* genes, finding within these gene regions 3/33 DMRs, and 3/25 top CpG sites in the meta-analysis.

Repeating the DMR analysis using CpG site estimates obtained in the sensitivity meta-analysis (i.e. excluding KORA), 77 DMRs were identified in association with T2D. From these, 12/77 DMRs overlapped with DMRs identified in the main analysis (Table 6-12, appendix Table S 8-32). DMRs were located across the genome, and most of them were hypomethylated in T2D cases compared to controls. At the 77 DMRs, median DMR size was 262bp (size range 50bp to 604bp), median CpG count per DMR was 4 sites (CpG range 1 to 10 sites), and median of the absolute percentage change in methylation between T2D cases and controls was 0.59 (percentage change range 0.06 to 1.93). The strongest DMR was annotated to the 5'UTR region of the *CPT1A* gene (estimate=-0.01; Sidak p-value=$1.11x10^{-9}$) (appendix Table S 8-32). In addition, an overlap was identified at eight gene regions between the DMR analysis and the sensitivity meta-analysis. Gene regions in common across analyses were annotated to the genes *HDAC4*, *CPT1A*, *ABCG1*, *RPL13AP5*, *PHGDH*, *TXNIP*, *PLEC1* and *HCCA2*. Interestingly in this secondary analysis, a DMR annotated to the region of the *SLC1A5* gene, contained the CpG site cg21766592, which has been previously reported in association with T2D in peripheral blood DNA methylation[64].

Table 6-12 Regions identified using Comb-p as differentially methylated in association with T2D. Results obtained using CpG site summary statistics from the meta-analysis. Effect size measured as %Meth: median of the absolute percentage change in methylation between T2D cases and controls, calculated for all the DMPs within a region. Strongest DMR identified with the smallest Sidak p-value is highlighted in bold.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Index CpG	Lowest P	P_{region}	Sidak
1	Chr1:108023249-108023483*	*NTNG1*	234	5	1.19	↑	cg20016673	7.39E-03	7.45E-07	1.20E-03
1	Chr1:145441552-145441553*	*TXNIP*†**	1	1	1.25	↓	cg19693031	4.77E-04	4.26E-09	1.60E-03
1	Chr1:36023134-36023415	*NCDN*	281	6	0.12	↓	cg10905247	2.09E-02	4.37E-06	5.84E-03
1	Chr1:871308-871547	*SAMD11*	239	3	1.45	↑	cg02439789	2.11E-02	7.68E-06	1.20E-02
1	Chr1:159825552-159825762	*VSIG8*	210	4	0.29	↑	cg17986992	2.37E-02	9.04E-06	1.61E-02
1	Chr1:201924337-201924584	*TIMM17A*	247	7	0.12	↓	cg24396741	2.61E-02	1.28E-05	1.93E-02
2	Chr2:113992762-113993143	*PAX8*	381	6	1.39	↑	cg21482265	1.88E-03	6.37E-06	6.28E-03
2	Chr2:231692812-231693071*	*Intergenic*	259	3	0.63	↑	cg19184455	2.37E-02	8.44E-06	1.22E-02
2	Chr2:233924789-233925031	*INPP5D*	242	8	0.08	↓	cg06272010	1.38E-02	1.07E-05	1.66E-02
3	Chr3:119217133-119217447	*C3orf1*	314	8	0.12	↑	cg19231082	8.44E-03	3.37E-06	4.04E-03
3	Chr3:4534791-4535155*	*ITPR1*	364	9	0.08	↓	cg02808075	2.15E-02	6.97E-06	7.19E-03
3	Chr3:128968351-128968635	*COPG*	284	4	0.06	↓	cg12821290	3.88E-02	2.89E-05	3.77E-02
3	Chr3:15311021-15311270	*SH3BP5*	249	4	0.35	↓	cg07078958	3.88E-02	2.66E-05	3.94E-02
6	Chr6:5261291-5261561	*LYRM4*	270	9	0.10	↓	cg17107193	7.39E-03	5.44E-06	7.56E-03
7	Chr7:94953770-94954060	*PON1*	290	4	1.05	↑	cg01874867	2.61E-02	3.29E-05	4.19E-02
8	Chr8:41583136-41583524	*ANK1*	388	4	0.90	↑	cg19537719	6.45E-03	5.70E-07	5.54E-04
8	Chr8:145018010-145018301*	*PLEC1*†	291	5	0.10	↓	cg20154947	2.39E-02	9.82E-06	1.26E-02
10	Chr10:121356513-121356866*	*TIAL1*	353	9	0.08	↓	cg15856091	1.88E-03	2.58E-06	2.75E-03
10	Chr10:6214016-6214080*	*PFKFB3*	64	3	0.64	↓	cg05014727	1.31E-02	2.01E-06	1.18E-02
11	Chr11:68607622-68607738	*CPT1A*†	116	3	0.43	↓	cg00574958	1.70E-03	6.77E-08	2.20E-04
11	Chr11:44642868-44642933	*Intergenic*	65	3	1.21	↑	cg00233028	1.70E-03	4.71E-08	2.73E-04
11	Chr11:1769152-1769523	*HCCA2*	371	8	0.81	↑	cg03300078	8.41E-03	1.24E-06	1.26E-03
11	Chr11:63998250-63998439	*DNAJC4*	189	2	0.20	↓	cg11468835	2.25E-02	7.33E-06	1.45E-02
12	Chr12:49463725-49464042*	*RHEBL1*	317	9	0.07	↓	cg14906565	1.30E-02	6.50E-06	7.69E-03
16	**Chr16:50321818-50322157**	***ADCY7***	**339**	**4**	**0.53**	**↑**	**cg06897661**	**2.54E-04**	**9.35E-10**	**1.04E-06**

Continuation Table 6-12.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Index DMP	Lowest P	P_{region}	Sidak
16	Chr16:29296614-29296798	*Intergenic*	184	4	1.00	↓	cg23515125	1.70E-03	5.64E-08	1.16E-04
16	Chr16:75150456-75150745	*LDHD*	289	3	0.78	↓	cg03991512	1.97E-02	2.15E-05	2.77E-02
19	Chr19:38806746-38806875*	*YIF1B*	129	5	0.12	↓	cg11436475	3.46E-03	1.93E-07	5.63E-04
19	Chr19:55549590-55549843*	*GP6*	253	6	0.93	↓	cg18355337	6.87E-03	6.03E-07	8.98E-04
19	Chr19:39389915-39390199*	*SIRT2*	284	2	0.10	↓	cg11396509	5.84E-03	9.30E-06	1.23E-02
20	Chr20:17549599-17549866	*DSTN*	267	4	0.35	↓	cg14158573	2.58E-02	1.18E-05	1.66E-02
21	Chr21:35320596-35320668*	*Intergenic*	72	2	1.14	↓	cg27037013	4.77E-04	5.06E-09	2.65E-05
22	Chr22:41864805-41865101	*ACO2*	296	7	0.07	↓	cg05365887	8.03E-03	2.20E-05	2.76E-02

*DMRs also identified by c*omb-p* using estimates of the sensitivity meta-analysis (i.e. excluding KORA). †Gene regions in overlap between the single-site analysis (i.e. meta-analysis) and the DMR analysis. **DMR less informative as only 1 CpG site was identified within the region. CpG count: number of DMPs detected within a DMR. Direction: relative effect observed across CpG sites within a DMR. Index CpG: CpG site with the lowest P from the meta-analysis. lowest P: smallest P-value identified within a region based on the meta-analysis. Pregion: P-value of the region calculated by *comb-p* using the Stouffer-Liptak correction. Sidak: significance of the DMR after multiple-testing correction. DMRs were considered significant at Sidak < 0.05

6.5.2 Genomic context of DMRs associated with T2D

Regulatory annotation of DMRs associated with T2D. The enrichment analysis for genomic regions conducted in LOLAweb (http://lolaweb.databio.org/), demonstrated that hypomethylated DMRs in T2D are located within core promoters and exons, and to a less extent, within introns and proximal promoters. With respect to transcription start sites (TSS), hypomethylated DMRs are mostly located 1Mb downstream the nearest TSS, and they are less represented near the TSS (within 0-1kb distance), either upstream or downstream (appendix Figure S8-20). Binding sites for the transcription factors *Pol2*, *Max* and *c-Myc*, are among the most commonly identified within hypomethylated DMRs according to ENCODE data[150]. Genomic annotation for hypermethylated DMRs in T2D demonstrated that these DMRs are located within intergenic regions, exons and introns, and to a less extent within core promoters and proximal promoters (appendix Figure S8-20). In relation to TSS, hypermethylated DMRs are also located away from the TSS (1Kb-1Mb distant), and they could be found either upstream or downstream the TSS. The most commonly observed transcription factor binding sites in hypermethylated DMRs were for *NRSF* and *GATA-2* based on ENCODE data (appendix Figure S8-20).

Additional annotation of DMRs using EpiExplorer (https://epiexplorer.mpi-inf.mpg.de/) and lymphoblastoid cells to proxy peripheral blood cells, demonstrated that hypomethylated DMRs in T2D overlapped with regions for DNase I hypersensitive sites (DHS), and histone marks associated with transcriptional active regions (H3K4m2, H3K4m3, H3K9ac). In addition, hypomethylated DMRs were primarily located within genes, gene promoters (-5kb to 1kb), CpG islands (63.6%) (Figure 6-10), and they overlapped with binding sites for *Pol2* (68.2%). In comparison, hypermethylated DMRs in T2D were depleted of DHS regions (27.3%), strong enhancers (9.1%) and gene promoters (18.2%). Furthermore, hypomethylated DMRs were enriched in histone marks associated with transcription repression (H3K27m3), conserved regions, exons (81.8%) and CpG islands (54.5%) (Figure 6-10). Binding sites for the transcription factors *Pol2* (36.4%) and *CTCF* (27.2%) were also common among hypermethylated DMRs based on ENCODE data.

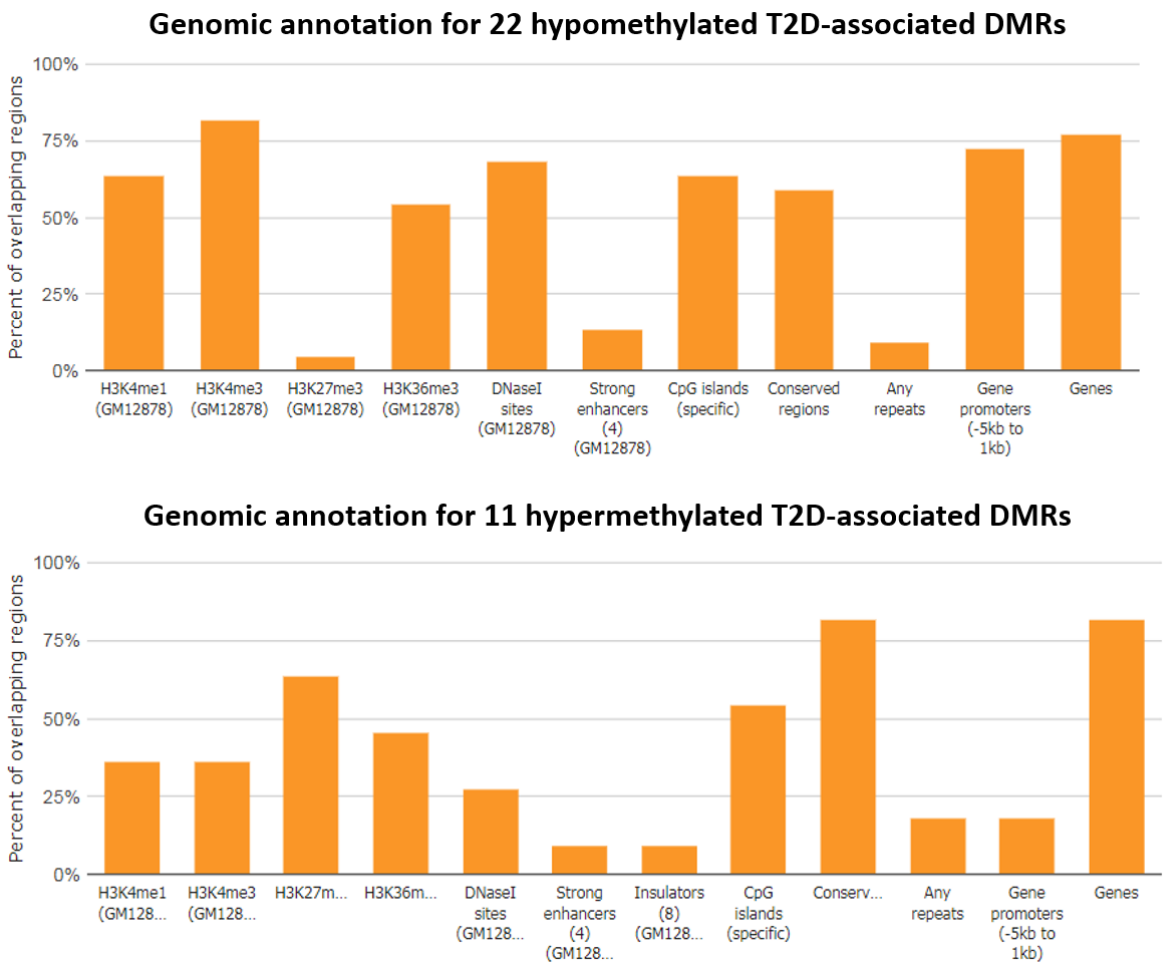


Figure 6-10 Annotation of T2D-associated DMRs across multiple regulatory regions using EpiExplorer (https://epiexplorer.mpi-inf.mpg.de/). Genomic annotation based on data from lymphoblastoid cells (GM12878).

Cross-tissue identification of enrichment for DNaseI hypersensitive sites (DHS) and histone marks among CpG sites within T2D-associated DMRs. Cell-type and tissue-specific enrichment for DHS was conducted in eFORGE v1.2, revealing that CpG sites within T2D-associated DMRs were not enriched for DHS in any tissue or cell type assayed included in the consolidated ROADMAP project, ENCODE and BLUEPRINT datasets. Thus, this lookup suggested that CpG sites identified in blood are less likely to reside in transcriptionally active regulatory regions in other important metabolic tissues. The enrichment lookup for H3 methylation marks indicated that CpG sites within T2D-associated DMRs predominantly overlapped with regions of mono-methylation, and to a lesser extent, with regions of tri-methylation of lysine 4 of H3 (H3K4m1 and H3K4m3). Enrichment for the H3K4m1 histone mark was observed in peripheral blood, placenta and skin tissue based on data from the consolidated ROADMAP project (appendix Figure S 8-18). Because H3K4m1 is a mark of transcriptional activation in enhancers[90], this finding suggests that CpG sites within T2D-ssociated DMRs were likely to reside in regions of active enhancers for specific tissues, some of them less related with T2D pathogenesis.

The enrichment lookup was also conducted for CpG sites detected in a secondary DMR analysis (i.e. 358 sites within 77 T2D-associated DMRs), obtaining similar results as with the main analysis. This second lookup confirmed that T2D-associated DMRs were depleted of DHS and most histone marks, except for H3K4m1 and H3K4m3. Strong signals of enrichment for H3K4m1 and H3K4m3 were identified across different tissues and cell types based on data from the consolidated ROADMAP project (appendix Figure S 8-19). Results of this second lookup suggested that CpG sites within T2D-associated DMRs were likely to reside within active enhancers and promoters, not only in blood, but also in other important tissues for T2D, including muscle and small intestine.

6.5.3 Percentage of the variance in T2D captured by CpG sites within top DMRs

As an additional analysis, the percentage of the variance in T2D explained by CpG sites within the DMRs in *ADCY7* (Chr16:50321818-50322157) and *CPT1A* (Chr11:68607622-68608226) was estimated. These two DMRs were the strongest regions associated with T2D based on two DMR analyses previously conducted (see section 6.5.1). The average methylation at the CpG sites within each DMR was used in a logistic regression to compute the Nagelkerke's R^2 statistic. Regressions were adjusted for common covariates and BMI. Strongest variance in T2D was explained by the DMR in *CPT1A* (1.1%) compared to the DMR in *ADCY7* (0.5%). Results of the regression analysis suggested that per 1% increase in methylation at the DMR in *CPT1A*, was associated with a 35% (95%CI 0.45-0.92) reduced risk of T2D only after adjustment for common covariates, but not for BMI. No association was identified between T2D and methylation at the DMR in *ADCY7*.

Table 6-13 Summary of the association between methylation and T2D at two top DMRs derived from estimates of the meta-analysis. The regression was calculated using the average methylation across CpG sites within a DMR. The variance in T2D was estimated using the Nagelkerke's statistic from the logistic regression.

				Basic model*			Adjusted for BMI**			Variance in T2D (%)
DMR	Chr	Position	Gene	OR	95% CI	P	OR	95% CI	P	
Main Analysis	16	50321818-50322157	*ADCY7*	1.00	(0.91,1.09)	9.54E-01	0.99	(0.9,1.08)	0.76	0.5
Secondary Analysis	11	68607622-68608226	*CPT1A*	0.65	(0.45,0.92)	1.42E-02	0.70	(0.48, 1.00)	0.05	1.1

*Basic model adjusted for age, sex, SVs, 6-Houseman cells and smoking. **Model additionally adjusted for BMI.

6.5.4 Enrichment for biological processes and metabolic pathways for CpG sites within T2D-associated DMRs

To determine if CpG sites within DMRs associated with T2D were enriched for specific biological processes or metabolic pathways, a gene enrichment analysis was conducted using genes annotated to these CpG sites. Background probes used for the analysis were all the CpG sites included in DMRs initially detected by *Comb-p* before applying multiple testing correction. The 166 CpG sites identified within 33 DMRs were near 29 unique gene regions (3 DMRs and 12 sites were intergenic), and there was no enrichment for GO terms or KEGG pathways at these CpG sites after adjustment for multiple testing (see appendix Table S 8-33 and Table S 8-34). Some of the top GO terms identified were related with cellular components of the mitochondrion, biological processes of "cellular response to caloric restriction" and "regulation of thyroid-stimulating hormone secretion" (Figure 6-11), and molecular functions of "deacetylation of histones" and "D-lactate dehydrogenase activity". Among top KEGG pathways were the "thyroid hormone synthesis", "pancreatic secretion" and "cortisol synthesis and secretion" (see appendix Table S 8-34). Repeating the enrichment analysis for 358 CpG sites included within 77 DMRs in T2D identified in a secondary analysis, there was no enrichment for GO terms or KEGG pathways after multiple testing correction. Top terms and pathways identified across analyses were highly similar. Of interest in this secondary analysis was the detection of the *PPAR* signalling pathway, which is related with the metabolism of lipids and T2D pathogenesis.

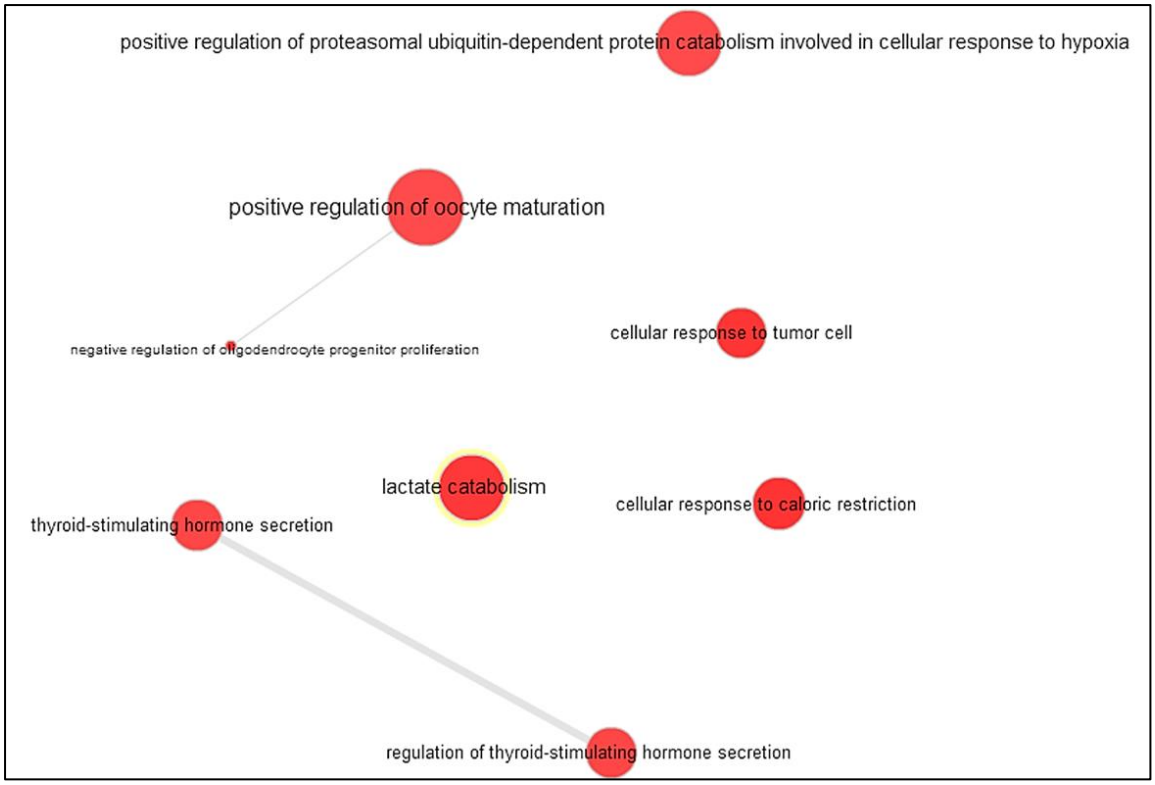


Figure 6-11 Network plot of top GO terms related with biological processes enriched for genes annotated to CpG sites within DMRs in T2D. Plot generated using REVIGO (http://revigo.irb.hr/revigo.jsp), illustrating the link between terms to facilitate the interpretation of results from GO.

6.6 Chapter summary

This study was able to provide evidence of differential methylation in association with prevalent T2D using a meta-analysis of EWAS and five European cohorts. Top association identified in the meta-analysis was at the well-known CpG site in *TXNIP* (cg19693031), which has been reported as an ubiquitous methylation marker for T2D not only in Europeans, but also in African and Mexican-American populations[64, 68]. A second signal was detected with epigenome-wide significance at the intergenic CpG site cg13826139, which to my knowledge has not been yet reported in association with T2D among Europeans. From a second meta-analysis restricted to studies that used a similar method to adjust for batch effects in the EWAS, it was possible to identify another five methylation sites associated with T2D, in addition to the signal at *TXNIP*. Methylation sites detected with epigenome-wide significance in the second meta-analysis mapped to the genes *ABCG1* (cg06500161), *CPT1A* (cg00574958) and *HDAC4* (cg00144180), and to the intergenic CpG sites cg16765088 and cg2470428. Apart from the associations at *ABCG1* and *CPT1A*, the remaining three sites are to my knowledge novel associations for prevalent T2D among Europeans.

Overall, it was identified that few associations surpassed epigenome-wide significance after adjustment for BMI, except for the association at the CpG in *TXNIP* and the intergenic CpG in cg16765088. This indicates that for most of the top signals detected in the meta-analysis, BMI was confounding the association between T2D and DNA methylation. Because the associations at *TXNIP* and cg16765088 were not influenced by BMI, this suggested that T2D might be associated with methylation at these sites via mechanisms independent of obesity.

Furthermore, it was evident that in the sensitivity meta-EWAS there was more power to detect stronger associations compared to the main meta-EWAS, and this was partially attributed to the higher homogeneity observed in effect estimates across studies once results from KORA were excluded. Generally, effect estimates in KORA were biased towards the null at the top-ranking CpG sites with higher interstudy heterogeneity. One reason to explain why effect estimates in KORA were considerably smaller than in other studies, was because of over-adjustment of the EWAS, which might have been attributed to using PCs, rather than SVs, to correct for unwanted variation in methylation. A PC analysis is an unsupervised method to reduce the complexity of high-dimensional data without considering particular characteristics of the samples compared[226]. The aim of this analysis is to find components that capture the largest variance in the data using the minimum dimensions possible to represent the original complexity of the data[226]. The disadvantage of using this method to correct for batch effects in methylation data, is that PCs can also capture biological variation that is important for the comparison between groups (T2D cases and controls). As a result, the analysis including non-independent PCs is over-adjusted and removes variation between samples that is indicative of different patterns of methylation between groups. In contrast, an SV analysis has the advantage of generating components (SVs) of variation in methylation that are independent of the variable of interest, and of selected covariates for the analysis[137]. Thus, the SV analysis provides more certainty that the variation in methylation removed is due to technical artefacts, rather than to consistent differences between groups[137].

Results of the meta-EWAS also indicated that T2D-associated methylation sites were in general hypomethylated in T2D cases compared to controls, and that the overall effect of T2D on difference in methylation at these sites was small. Furthermore, regression analyses conducted in the ALSPAC dataset showed that increased methylation at top sites in the meta-analysis was generally associated with a protective effect on T2D. This protective effect of hypermethylation on T2D was also identified in a sensitivity analysis using quartiles of methylation for the top CpG sites. In addition, it was demonstrated that top sites in the meta-analysis were able to capture a small but

significant variance in T2D, either individually or in combination, but this variation was less than the variation captured by established risk factors.

A risk factor analysis showed that strongest CpG sites were associated with important clinical risk factors for T2D, providing further evidence of potential mechanisms by which methylation can influence the risk of T2D. Most of the top CpG sites were also associated with categories of glucose tolerance, and for the CpG in *CPT1A*, the association was only identified with the prediabetic state. Functional lookup of the strongest CpG sites revealed that methylation at *TXNIP*, *CPT1A* and *ABCG1*, was inversely associated with gene expression of the same genes. Additionally, there was some evidence of genetic interaction between DNA methylation and the glycaemic traits of HOMA-B and 2-h glucose, for the CpG sites in *ABCG1* and *TXNIP*, respectively. In both cases, the genetic interaction was in opposite directions for DNA methylation, and for levels of the glycaemic trait. No genetic interaction was identified between DNA methylation and T2D. A gene enrichment analysis did not identify pathways or biological processes strongly associated with top methylation sites in the meta-analysis. However, some of the top pathways obtained were related with the metabolism of lipids, insulin resistance, and the tumour necrosis factor (TNF) signalling pathway.

A regional analysis of differential methylation revealed that most of the DMRs associated with T2D were hypomethylated in T2D cases compared to controls. In addition, hypomethylated DMRs are most commonly identified in overlap with gene promoters and genes, rather than with intergenic regions and introns as it was observed for hypermethylated DMRs. The two strongest DMRs were detected at the *ADCY7* and *CPT1A* genes, one was hypermethylated and the other one was hypomethylated in T2D cases compared to controls, respectively. Using the ALSPAC dataset, it was possible to replicate the association between the DMR in *CPT1A* and T2D, but not the association at the DMR in *ADCY7*. For the DMR in *CPT1A,* it was confirmed that hypermethylation of this region is associated with a protective effect on T2D. The proportion of variance in T2D explained by CpG sites within the DMRs in *ADCY7* and *CPT1A* was considerably lower than the variance in T2D captured by top CpG sites identified in the meta-analysis. As with the single CpG site analysis, no enrichment was identified for biological processes or metabolic pathways when using CpG sites identified within top DMRs.

Because findings from the meta-analysis come from an observational analysis where associations can be true, or they can be confounded by unmeasured factors, top results from the meta-analysis in T2D were followed up using Mendelian randomization methods for causal inference analysis.

Furthermore, due to the interaction between DNA methylation and environmental factors, the association with T2D can be either from methylation to T2D (predictive scenario), or from T2D to DNA methylation (consequential scenario). Therefore, it was necessary to incorporate a bidirectional MR to determine true direction of causality in this association. Causal inference analysis for top signals identified in the meta-analysis is the topic of Chapter 7.

Chapter 7 Exploring causality in DNA methylation and type 2 diabetes

In previous chapters the association between T2D and methylation from an observational perspective has been addressed. This has included evidence obtained using samples from ALSPAC alone (see Chapter 4) or combining results across five different cohorts in a meta-analysis (see Chapter 6) to increase the statistical power of findings. In this chapter, top methylation markers identified in the observational analysis were investigated for causality and direction using a bidirectional Mendelian randomization (MR).

Methods to assess causality include RCTs, negative control analysis, triangulation, parental comparison and cross-cohort comparison, among others (refer to Richmond *et al*.[1]), but the one implemented was Mendelian randomization (see Chapter 3). Briefly, an MR is analogous to an RCT[1-4], but instead of randomly allocating participants to treatments, it exploits the Mendel's laws of independent assortment of the genotype during conception, to postulate the use genetic variants strongly associated with a modifiable exposure as unbiased instruments[3, 5, 6]. The genetically predicted exposure is used to determine causality in the exposure-outcome association. Because the genotype is established from conception, is unaffected by environmental factors (not susceptible to reverse causation), and is fixed over the life course (unaffected by regression dilution bias), causal estimates obtained in an MR are a combination of the effect of long-term exposure and developmental compensation[7].

Limitations associated with MR have been mentioned elsewhere [4, 8], including low statistical power due to the small amount of variation in the exposure explained by the genetic variant(s), bias in the predicted estimate introduced by population stratification, pleiotropic effects due to the association of the genetic instrument with the outcome through pathways independent of the exposure of interest, and canalization or developmental compensation[3, 4]. Canalization occurs when the phenotypic penetrance of a variant is reduced during the life-time due to physiological changes that counterbalance the detrimental effect of the genotype.

The application of MR methods to address causality using -omics data is a growing field[7, 9] that benefits from the stronger effect that the genotype may have on these intermediate molecular traits, compared to the effect in more common phenotypes like BMI[9]. As any other phenotypic exposure and outcome, DNA methylation is susceptible to confounding as it is responsive to environmental triggers associated with disease risk[10], and it is also susceptible to reverse causation, as changes in the epigenotype can occur as a consequence of disease[9]. Thus, an MR represents a

valid strategy to infer causality in epigenetic studies. Genetic instruments associated with variation in methylation are known as methylation QTL, and this association is time- and tissue-specific[11, 12]. By elucidating direction of causality in epigenetics of common diseases such as T2D, it is possible to use epigenetic markers to intervene in disease prevention and treatment, rather than limiting their use as biomarkers of disease risk.

Despite extensive literature supporting the role of DNA methylation in the aetiology of T2D, there is limited evidence supporting causality of these associations, particularly for those identified in cross-sectional studies, where it is difficult to discern true direction of effect. There are examples of studies addressing causality and directionality in the association between DNA methylation and BMI[13-15], an anthropometric trait related with T2D. For instance, Wahl *et al*.[13] replicated 187 loci associated with BMI, and results of the bidirectional MR suggested that for most of these sites variation in methylation occurred as a consequence of BMI rather than vice versa[13]. Directly related with T2D, Richardson *et al*.[16] conducted an MR analysis to study the genetic liability to T2D based on DNA methylation measured in early-life. In this study, variation in methylation at CpG sites in *HNF1B*, *KCNJ11*, *IGF2BP2* and *WFS1,* was identified in association with future risk of T2D. Another example of using genetics to explain direction of causality in the association between DNA methylation and T2D was provided by Elliott *et al.*[17]. In this study, well-known genetic variants for T2D were used as causal anchors to investigate the mediating role of DNA methylation in future risk of T2D; methylation was measured in disease-free participants. Even though evidence provided by Elliott *et al*. and Richardson *et al*., gap in the knowledge remains for observational associations detected in the context of cross-sectional epigenetic studies in T2D, and this was the focus of the present Chapter.

Aims:

1. Determine causality for observational associations detected between T2D and DNA methylation at middle-age using Mendelian Randomization methods.
2. Establish direction of causality between T2D and DNA methylation using a bidirectional two-sample MR analysis.
3. Determine the functional implications of methylation markers detected in the causal analysis by implementing a gene enrichment analysis, and a lookup for evidence of potential associations between methylation, gene expression and other outcomes related with T2D.

7.1 Study Population

As mentioned in the methods section for the causal analysis (see Chapter 3), a subset of 1,252 participants were successfully genotyped for T2D SNPs. Of the initial 148 SNPs selected for genotype extraction, 142 SNPs were successfully typed across middle-age females and males in ALSPAC. Baseline characteristics of the subsample included in the genetic analysis are summarized in Table 7-1. Single-SNP regressions and a polygenic risk score for T2D were conducted in a subsample composed of 36 T2D cases and 804 controls.

7.2 Proxies for T2D and their relationship with other glycaemic outcomes

Of the 142 T2D SNPs successfully typed in participants in ALSPAC, 126 SNPs remained in the dataset after additional pruning for MAF, Hardy-Weinberg equilibrium, and missing genotyping rate (refer to Chapter 3 for more on QC of the genetic data). Of these 126 SNPs, 75 were selected as independent SNPs based on an LD threshold < 0.2. Looking at the overlap between 75 SNPs in T2D and SNPs identified in recent GWAS meta-analyses for five glycaemic traits (i.e. fasting glucose, fasting insulin, HOMA-IR, HOMA-B and HbA1c), 41.2% SNPs uniquely matched with one glycaemic trait, while the remaining 58.8% were identified in overlap with more than one trait. For the SNPs that uniquely matched to one trait, most of them were nominally associated (at $p < 0.05$) with fasting glucose, while for those matching to more than one trait, they were overrepresented in HOMA-B and HbA1c levels (Figure 7-1).

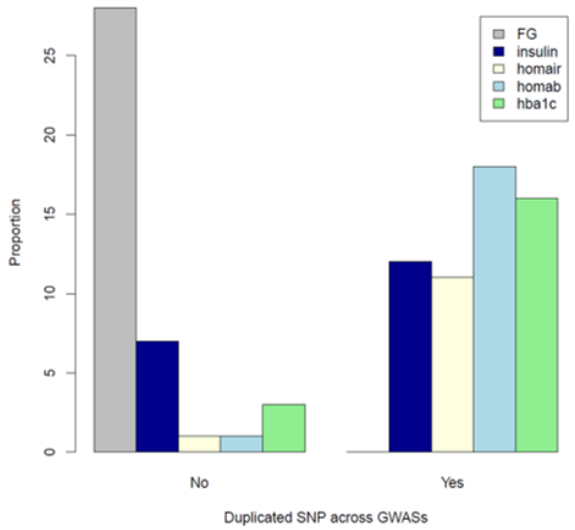


Figure 7-1 Overlap between 75 independent SNPs for T2D, and variants identified in GWAS meta-analyses of glycaemic traits. SNPs in T2D were in overlap with a glycaemic trait SNPs if GWAS p<0.05. GWAS data for the glycaemic traits was extracted from the MAGIC consortium (https://www.magicinvestigators.org).

Table 7-1 Baseline characteristics of the subsample of 1,252 females and males in ALSPAC genotyped for 142 T2D SNPs, and with availability of DNA methylation at the middle-age time-point. Continuous variables were described using mean and SD, while categorical variables were described using the proportion and sample-size per category. N is the total number of samples with complete data for the covariates, and samples with unknown data were reported as missing.

N=1252	**Non-missing**	**Females** (n=867)	**Males** (n=385)	**Missing**	**P** [b]
Age [a]	1,212	46.91 (4.66)	53.82 (5.27)	40	<0.01
T2D status [c]	840			412	
Cases	36	2.40 (21)	3.80 (15)		0.22
Controls	804	63.70 (552)	65.50 (252)		
Metabolic/anthropometric					
Fasting Glucose (mmol/L)	1,013	5.28 (1.13)	5.65 (1.09)	239	<0.01
BMI (kg/m^2)	1,022	26.26 (4.85)	27.06 (3.81)	230	<0.01
Waist circumference (cm)	1,025	83.87(11.79)	96.31 (10.44)	227	<0.01
Cholesterol (mmol/L)	1,023	4.61 (0.86)	5.12 (0.93)	229	<0.01
Triglycerides (mmol/L)	1,013	1.02 (0.59)	1.40 (0.66)	239	<0.01
HDL (mmol/L)	1,013	1.49 (0.38)	1.29 (0.25)	239	<0.01
LDL (mmol/L)	1,013	2.95 (0.81)	3.17 (0.81)	239	<0.01
Smoking	828			424	
Never smoked		55.67 (275)	50.00 (167)		0.17
Ever smoked		36.64 (181)	43.11 (144)		
Current smoker		7.69 (38)	6.89 (23)		
Physical activity	1,031			221	
Less than or 4h/week		60.15 (412)	45.95 (159)		<0.01
More than 4h/week		39.85 (273)	54.05 (187)		
Socioeconomic status	905			347	
Low income		13.34 (91)	8.97 (20)		0.01
Middle income		33.28 (227)	25.56 (57)		
High income		53.37 (364)	65.47 (146)		
Principal components [d]					
PC1	1,174	8.90×10^{-4} (0.01)	4.80×10^{-4} (0.02)	78	0.68
PC2	1,174	1.10×10^{-4} (0.01)	-5.10×10^{-4} (0.02)	78	0.51
PC3	1,174	-4.30×10^{-4} (0.01)	1.60×10^{-4} (0.02)	78	0.52
PC4	1,174	-1.10×10^{-4} (0.01)	3.10×10^{-4} (0.02)	78	0.66
PC5	1,174	1.00×10^{-5} (0.01)	-5.10×10^{-5} (0.02)	78	0.95
PC6	1,174	2.00×10^{-4} (0.01)	-5.40×10^{-4} (0.02)	78	0.45
PC7	1,174	6.10×10^{-4} (0.01)	3.50×10^{-4} (0.02)	78	0.79
PC8	1,174	4.80×10^{-4} (0.01)	-2.30×10^{-3} (0.02)	78	<0.01
PC9	1,174	-7.00×10^{-5} (0.01)	7.90×10^{-4} (0.02)	78	0.36
PC10	1,174	3.50×10^{-4} (0.01)	-6.40×10^{-4} (0.02)	78	0.31

[a] Age was missing in 40 male samples. [b] P is the p-trend in categorical ordinal variables. [c] Twelve samples considered as missing for T2D status in the genetic dataset, were recategorized as controls in an updated version of the phenotypic dataset in ALSPAC (Chapter 4). [d] Principal components were previously calculated in the genetic dataset of females and males in ALSPAC. Missing values for the first ten PCs were detected in 78 female samples.

7.3 Genetic proxies versus T2D and confounders in ALSPAC

Proxies versus T2D

An additive genetic model was applied to investigate the association between 75 independent T2D SNPs and T2D and nine potential confounders. Regressions were adjusted for the first ten genetic PCs and a batch variable. Associations were considered significant at $p< 6.67x10^{-4}$ after correction for multiple testing using Bonferroni (α=0.05/75 SNPs). Further detail of the method used to conduct these genetic regressions was described previously in Chapter 3.

Table 7-2 Main results of the regression between 75 independent SNPs and T2D and potential confounders using a subsample of 1,252 participants in ALSPAC. Results were adjusted for 10 genetic PCs and a batch effect variable. Summary statistics are presented for the SNP with the smallest P-value of association with the outcome of interest. BMI and HDL were log-transformed before the analysis. Associations were regarded significant at $p < 6.67x10^{-4}$ (i.e. α=0.05/75 SNPs).

	N†	SNPs+	Top SNP	Chr	Estimate (95%CI)	P
T2D	32/768/452	75	rs1496653	3	0.54 (0.32, 0.93)	0.03
Sex	789/385/78	75	rs2812533	10	0.68 (0.51, 0.89)	0.01
Age	1,212/40	75	rs1470579	3	-0.64 (-1.06, -0.21)	$4.0x10^{-3}$
BMI (Ln)	1,022/230	75	rs7202877	16	0.04 (0.01, 0.06)	$2.0x10^{-3}$
HDL (Ln)	1,013/239	75	rs13389219	2	-0.04 (-0.06, -0.02)	$1.3x10^{-4}$
LDL	1,013/239	75	rs319598	5	-0.09 (-0.17, -0.02)	0.01
Cholesterol	1,023/229	75	rs7795991	7	0.11 (0.03, 0.19)	0.01
PC1	1,174/78	75	rs7163757	15	$2.0x10^{-3}$ ($5.0x10^{-4}$, $3.0x10^{-3}$)	0.01
PC2	1,174/78	75	rs7178572	15	$-2.0x10^{-3}$ ($-3.5x10^{-3}$, $-1.0x10^{-3}$)	$1.0x10^{-3}$
PC3	1,174/78	75	rs16861329	3	$2.0x10^{-3}$ ($5.3x10^{-4}$, $4.0x10^{-3}$)	0.01
PC4	1,174/78	75	rs7163757	15	$2.0x10^{-3}$ ($6.0x10^{-4}$, $3.1x10^{-3}$)	$4.0x10^{-3}$
PC5	1,174/78	75	rs12970134	18	$-2.0x10^{-3}$ ($-3.0x10^{-3}$, $-3.0x10^{-4}$)	0.02
PC6	1,174/78	75	rs516946	8	$2.0x10^{-3}$ (4.0x10-4, $3.04x10^{-3}$)	0.01
PC7	1,174/78	75	rs7163757	15	$-1.6x10^{-3}$ ($-2.8x10^{-3}$, $-3.8x10^{-4}$)	0.01
PC8	1,174/78	75	rs1801282	3	$-2.0x10^{-3}$ ($-4.0x10^{-3}$, $-1.2x10^{-4}$)	0.04
PC9	1,174/78	75	rs516946	8	$3.0x10^{-3}$ ($1.2x10^{-3}$, $4.0x10^{-3}$)	$2.9x10^{-4}$
PC10	1,174/78	75	rs4812829	20	$-2.2x10^{-3}$ ($-4.1x10^{-3}$, $-5.0x10^{-4}$)	0.01
Smoking‡	374/424/454	52	rs1169288	12	0.71 (0.56, 0.88)	$2.0x10^{-3}$
SES‡	748/100/404	72	rs2972156	2	0.66 (0.47, 0.91)	0.01
Physical Activity‡	544/429/279	56	rs2820446	1	1.36 (1.11, 1.66)	$3.0x10^{-3}$

†N is the total sample included in the regression, divided into cases/controls/missing samples for categorical variables, and into complete-data/missing-data for continuous variables. In all the regressions, 78 samples with missing records for the first ten genetic PCs were all females. SNPs are the total number of variants included in the regression analysis. Top SNP is the variant where the smallest P-value was detected in association with the outcome. ‡ Due to the small proportion of samples observed per category of the genotype for these traits, only a subset of the total SNPs was included in the regression analysis.

Overall, none of the SNPs extracted from DIAGRAM were strongly associated with T2D in ALSPAC, and the power to detect an effect of 0.54 (T2D~SNP rs1496653 in *UBE2E2*) at p<0.05, was between 18% and 24%. Power was calculated using the *Power Calculator for Mendelian Randomization* tool

(http://cnsgenomics.com/shiny/mRnd/) and including as input parameters regression statistics for the SNP identified with the strongest association with T2D in ALSPAC (rs1496653 in *UBE2E2*, 95%CI=0.32-0.93, p=0.03). The proportion of the genetic variation in T2D used was 10%-15% based on previous GWAS of T2D[18, 19], sample-size was 1,252 participants in ALSPAC, and the proportion of cases used was 0.04. In contrast, it was estimated that the minimum sample size required to confidently detect an effect of 0.54 (T2D~SNP rs1496653 in *UBE2E2*) with 80% power at p<0.05, was 169,390 samples. Association statistics obtained in the single SNP regression against T2D, are presented in the appendix Table S8-35.

Proxies versus confounders

No association was detected between the genotype and sex, smoking, physical activity and socioeconomic status, and the average power to detect an effect between 0.66 and 1.36 at p<0.05 was 9% (β range between 8%-10%). Power was calculated based on a proportion of variance in the outcome explained by the genotype of 0.01, and on the effect estimate obtained for the SNP with the smallest p-value of association with the categorical confounder. Summary of main results obtained in these genetic regressions is presented in Table 7-2. In the regression between the genotype and continuous confounders, an association was identified between SNP rs13389219 and HDL, and between SNP rs516946 and PC9. Results suggested that per extra risk allele C in rs13389219 was associated with an average increase in 0.95 mmol/l of HDL (95%CI= 0.93-0.98, $p=1.3x10^{-4}$), and per extra risk allele C in SNP rs516946 was associated with an average increase in $3.0x10^{-3}$ in the genetic variation explained by PC9 (95%CI=$1.0x10^{-3}$- $4.0x10^{-3}$, $p=2.9x10^{-4}$) (see appendix Figure S8-21).

The estimated power to detect the single-SNP association with HDL and PC9 was 6% and 10% , respectively, using as parameters for this estimation: a sample size of 1,013 and 1,174, an observed unadjusted effect (β_{OLS}) of -0.06 and 0.003, a hypothetical causal effect (β_{yx}) of -0.04 and $3.0x10^{-3}$ (i.e. adjusted effect for PCs and batches), a proportion of variance in the outcome explained by the genotype of 0.01 (i.e. R^2 reported by the linear regression), a variance in the exposure (σ^2_x) of 0.58 and 0.78, (equivalent to the frequency of the risk allele for the SNP with the strongest association), and a variance in the outcome (σ^2_y) of 0.12 and $2.0x10^{-4}$ for HDL and PC9. No association was detected between the genotype and BMI, age, LDL, total cholesterol and PC1-PC8

7.4 Polygenic Risk score for T2D

It is well known that weak instruments can bias results of the causal analysis towards the observational association in the context of a single sample MR. Because only null associations were identified in the single SNP analysis against T2D in ALSPAC, power limitations prevented the use of these proxies individually to conduct the single sample MR. To increase the strength of the instruments and the power to identify a causal effect, genetic proxies were combined into a polygenic risk score (PRS) to conduct the single sample MR via 2SLS-IV regression.

Further information of 75 index SNPs included in two polygenic scores for T2D can be found in the appendix Table S8-36, while the method used to select these SNPs was described previously in Chapter 3. The two genetic scores were generated based on two p-value thresholds, one at $p \leq 5.0x10^{-8}$ (PRS1=58 SNPs), and the second at $p \leq 9.0x10^{-6}$ (PRS2=75 SNPs). Descriptive statistics of the scores are presented in Table 7-3, showing a difference in the mean value between the scores ($p<0.001$), and a positive strong correlation among them according to the Pearson estimate ($r=0.93$). Scores were normally distributed, with values ranging between 0.04 to 0.07 for PRS1, and between 0.04 to 0.06 for PRS2 (Figure 7-2).

Table 7-3 Characteristics of two polygenic risk scores (PRS) for T2D validated in a sub-sample of adults in ALSPAC (n=1,252). P-range is threshold of significance considered to include SNPs in either of the two PRS, and P evaluates significance in the mean difference between the scores.

Score	P-range	SNPs	N	Mean	SD	Range	P
PRS1	$5.0x10^{-8}$	56	1,252	0.053	0.004	0.041 - 0.067	<0.001
PRS2	$9.0x10^{-6}$	75	1,252	0.048	0.003	0.039 - 0.060	

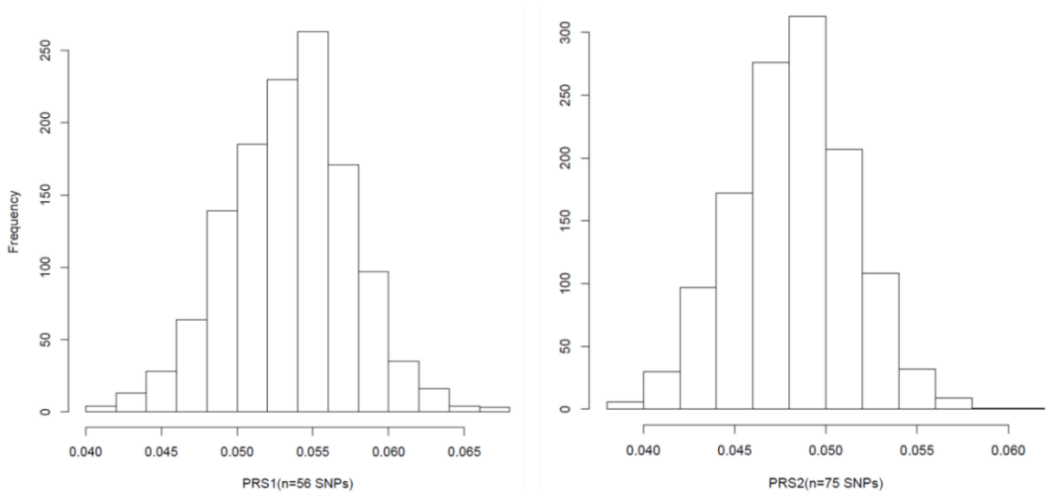


Figure 7-2 Histogram representing the distribution of two polygenic scores for T2D within adults in ALSPAC. PRS1 was composed of 56 SNPs and PRS2 was composed of 75 SNPs. Values correspond to unstandardized units of the scores.

7.4.1 Polygenic scores versus T2D

Mean values of PRS1 and PRS2 were on average 0.16% (p=0.01) and 0.10% (p=0.02) higher in cases compared to controls, respectively (Figure 7-3). Using standardizes values of scores, it was identified that these two instruments were strongly associated with an increased risk of T2D in the adjusted regression (covariates: sex and the first ten genetic PCs), but only PRS1 was also associated with T2D in the unadjusted regression. Results suggested that per SD increase in PRS1 was associated with 1.52 (95%CI=1.08, 2.14, p=0.02) and 1.59 (95%CI=1.12, 2.26, p=0.01) increased risk of T2D based on results of the unadjusted and the adjusted model, respectively, while per SD increase in PRS2 was associated with 1.42 (95%CI=1.01,2.01, p=0.04) increased risk of T2D only in the adjusted model.

Total variation in T2D explained by PRS1 alone was 1.9%, and by PRS1 and additional covariates was 6.6%. Variation in T2D explained by PRS2 alone was 1.5%, and by PRS2 and additional covariates was 5.5%. Using an ANOVA test to compare fitness between the unadjusted and the adjusted model including the polygenic risk score, revealed no significant difference between models at explaining variation in T2D at p<0.05 (PRS1 P_{ANOVA}=0.47, and PRS2 P_{ANOVA}=0.51). Even though the two scores captured a similar proportion of variation in T2D, only that attributed to PRS1 was statistically significant (p=0.01), and borderline significant for PRS2 (p=0.06). Summary statistics for the association between the two polygenic scores and T2D, are shown in Table 7-4 and Table 7-5**.** Associations were visually represented in the appendix Figure S8-22.

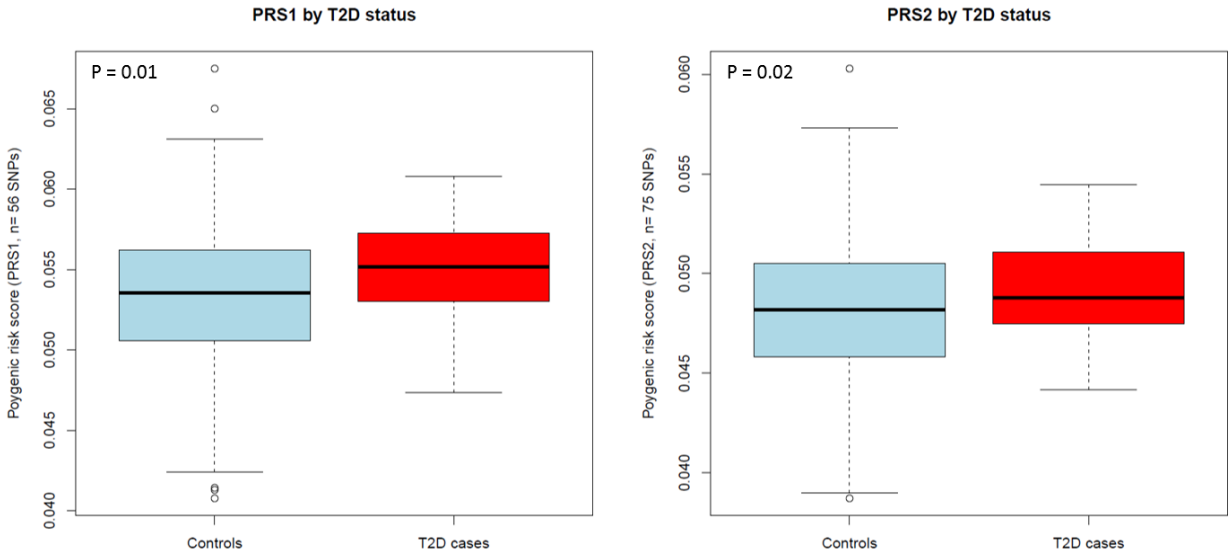


Figure 7-3 Distribution of two polygenic scores for T2D across disease groups. T2D cases had on average higher mean values of the scores compared to controls.

7.4.2 Scores versus confounders

Summary statistics of the association between the two polygenic scores and potential confounders, are described in Table 7-4 and Table 7-5. According to results, the two scores were not associated with BMI, age, sex, smoking and predicted cell-counts, which were covariates used to adjust the observational exposure-outcome association. Despite this, the polygenic scores were associated with factors related with T2D such as log-transformed levels of fasting glucose (p range 0.01 to 0.04), and log-transformed levels of HDL only for PRS2 (p=0.04). Considering that PRS1 strongly explained higher variation in T2D compared to PRS2, without being associated with direct confounders of the main association, this score was taken forward to estimate the association with methylation at selected DMPs (i.e. observed IV-outcome association) (see section 7.4.4), and also to conduct the single sample MR (see section 7.5).

Table 7-4 Association statistics between the polygenic score with 56 SNPs (PRS1), and T2D and potential confounders.

		PRS1 T2D (n= 56 SNPs) Unadjusted				PRS1 T2D (n= 56 SNPs) Adjusted†				
	N	Effect estimate (95%CI)	P	R^2	$P_{(F)}$ ††	Effect estimate (95%CI)	P	R^2	$P_{(F)}$	$P_{(ANOVA)}$**
T2D‡	**840**	**1.52 (1.08, 2.14)**	**0.02**	**2.00E-02**	**1.00E-02**	**1.59 (1.12, 2.26)**	**0.01**	**0.07**	**1.00E-02**	**4.70E-01**
Age	1,212	-2.64E-04 (-0.32, 0.32)	1.00	2.10E-09	9.99E-01	-0.05 (-0.33, 0.22)	0.71	0.30	2.20E-16	2.20E-16
Sex‡	1,252	0.98 (0.87, 1.11)	0.78	8.50E-05	7.80E-01	0.99 (0.87, 1.12)	0.85	0.08	7.80E-01	3.52E-11
Smoking‡	828	1.00 (0.87, 1.14)	0.98	1.00E-06	9.80E-01	1.01 (0.88, 1.16)	0.86	0.01	9.80E-01	5.90E-01
BMI (log)	1,022	0.01 (-0.26, 0.27)	0.96	1.66E-06	9.60E-01	-0.02 (-0.29, 0.24)	0.87	0.02	1.00E-02	3.26E-03
Cholesterol	1,023	-0.03 (-0.08, 0.03)	0.33	9.30E-04	3.30E-01	-0.03 (-0.08, 0.03)	0.34	0.07	6.80E-12	3.76E-12
HDL (log)	1,013	-0.01 (-2.79E-02, 1.59E-03)	0.08	3.02E-03	8.07E-02	-0.01 (-2.74E-02, 1.29E-03)	0.07	0.08	3.60E-13	4.99E-13
LDL	1,013	-0.01 (-0.06, 0.04)	0.79	6.90E-05	7.92E-01	-0.01 (-0.06, 0.04)	0.79	0.03	1.00E-02	1.02E-02
Waist-circumference	1,059	0.20 (-0.58, 0.98)	0.61	2.41E-04	6.14E-01	0.22 (-0.49, 0.93)	0.54	0.19	2.20E-16	2.20E-16
FG (log)	**1,030**	**0.01 (3.63E-04, 0.02)**	**0.04**	**4.05E-03**	**4.12E-02**	**0.01 (1.44E-03, 0.02)**	**0.02**	**0.06**	**2.83E-09**	**6.76E-09**
C-reactive protein (log)	1,021	-0.02 (-0.08, 0.05)	0.62	2.45E-04	6.17E-01	-0.01 (0.08, 0.05)	0.68	0.01	3.50E-01	2.93E-01
SBP*	1,052	0.21 (-0.63, 1.06)	0.63	2.27E-04	6.25E-01	0.20 (-0.55, 0.94)	0.60	0.24	2.20E-16	2.20E-16
DBP*(log)	1,052	-1.08E-03 (-8.56E-03, 6.40E-03)	0.78	7.66E-05	7.77E-01	-8.11E-04 (-0.01, 0.01)	0.83	0.08	8.60E-13	3.18E-13
Predicted cell-counts										
CD4T	1,065	6.62E-04 (-2.61E-03, 3.94E-03)	0.69	1.48E-04	6.90E-01	6.49E-04 (-2.64E-03, 3.94E-03)	0.70	0.02	6.00E-02	4.00E-02
CD8T	1,065	-6.02E-04 (-2.36E-03, 1.15E-03)	0.50	4.25E-04	5.00E-01	-6.86E-04 (-2.44E-03, 1.07E-03)	0.44	0.03	9.62E-04	6.21E-04
B-cells	1,065	6.38E-05 (-1.77E-03, 1.90E-03)	0.95	4.35E-06	9.50E-01	2.91E-05 (-1.83E-03, 1.88E-03)	0.98	0.01	8.90E-01	8.40E-01
NK	1,065	9.42E-04 (-2.21E-03, 4.09E-03)	0.56	3.23E-04	5.60E-01	1.02E-03 (-2.13E-03, 4.18E-03)	0.53	0.02	4.00E-02	3.00E-02
Monocytes	1,065	-5.81E-04 (-2.31E-03, 1.15E-03)	0.51	4.07E-04	5.10E-01	-6.76E-04 (-2.33E-03, 9.81E-04)	0.42	0.11	2.20E-16	2.20E-16
Granulocytes	1,065	-4.71E-04 (-5.95E-03, 0.01)	0.87	2.67E-05	8.70E-01	-2.89E-04(-5.78E-03, 5.20E-03)	0.92	0.02	4.00E-02	2.00E-02

† Adjusted regression for sex and the first ten genetic PCs. ‡ Estimates for these variables are in odds ratios (OR).*Systolic and diastolic blood pressure. DBP was log-transformed before analyses. †† $P_{(F)}$ P value of the regression model.** P value calculated using an ANOVA test to estimate the difference between regression coefficients across adjustment models. $P < 0.05$ was the threshold of significance.

Table 7-5 Association statistics between the polygenic score with 75 SNPs (PRS2), and T2D and potential confounders.

		PRS2 T2D (n= 75 SNPs) Unadjusted				PRS2 T2D (n= 75 SNPs) Adjusted†				$P_{(ANOVA)}$**
	N	Effect estimate (95%CI)	P	R^2	$P_{(F)}$††	Effect estimate (95%CI)	P	R^2	$P_{(F)}$	
T2D‡	**840**	**1.38 (0.99, 1.93)**	**0.06**	**1.00E-02**	**0.06**	**1.42 (1.01, 2.01)**	**0.04**	**0.06**	**6.00E-02**	**5.10E-01**
Age	1,212	0.04 (-0.29, 0.36)	0.82	4.13E-05	0.82	-0.02 (-0.29, 0.26)	0.90	0.30	2.20E-16	2.20E-16
Sex‡	1,252	0.98 (0.87, 1.11)	0.78	8.77E-05	0.78	0.98 (0.87, 1.11)	0.77	0.08	1.34E-14	3.44E-11
Smoking ‡	828	0.98 (0.86, 1.12)	0.74	1.70E-04	0.74	0.99 (0.87, 1.14)	0.92	0.01	7.40E-01	6.00E-01
BMI	1,022	0.03 (-0.23, 0.29)	0.82	4.29E-05	0.82	-5.60E-04 (-0.26, 0.26)	1.00	0.02	1.00E-02	3.36E-03
Cholesterol	1,023	-0.02 (-0.07, 0.04)	0.50	4.44E-04	0.50	-0.02 (-0.07, 0.04)	0.53	0.07	8.50E-12	3.82E-12
HDL	1,013	-0.01 (-2.97E-02, -2.70E-04)	0.05	3.93E-03	0.05	**-0.01 (-2.94E-02, -6.32E-04)**	**0.04**	**0.08**	**2.30E-13**	**4.79E-13**
LDL	1,013	4.6E-04 (-0.05, 0.05)	0.99	3.30E-07	0.99	3.08E-05 (-0.05, 0.05)	1.00	0.03	1.00E-02	1.00E-02
Waist-circumference	1,059	0.19 (-0.58, 0.97)	0.63	2.26E-04	0.63	0.24 (-0.47, 0.96)	0.50	0.19	2.20E-16	2.20E-16
FG (log)	**1,030**	**0.01 (0.00, 0.02)**	**0.04**	**4.28E-03**	**0.04**	**0.01 (2.06E-03, 0.02)**	**0.01**	**0.06**	**2.12E-09**	**5.58E-09**
C-reactive protein	1,021	-0.02 (-0.08, 0.04)	0.54	3.62E-04	0.54	-0.02 (-0.08, 0.05)	0.61	0.01	3.43E-01	2.94E-01
SBP*	1,052	0.09 (-0.76, 0.93)	0.84	4.09E-05	0.84	0.09 (-0.66, 0.84)	0.81	0.24	2.20E-16	2.20E-16
DBP*(log)	1,052	-1.14E-03 (-0.01, 0.01)	0.77	8.51E-05	0.77	-7.08E-04 (-0.01, 0.01)	0.85	0.08	8.64E-13	3.20E-13
Predicted cell-counts										
CD4T	1,065	4.17E-04 (-2.85E-03, 3.68E-03)	0.80	5.90E-05	0.80	3.30E-04 (-2.96E-03, 3.62E-03)	0.84	0.02	6.49E-02	4.47E-02
CD8T	1,065	-8.84E-04 (-2.63E-03, 8.65E-04)	0.32	9.22E-04	0.32	-9.73E-04 (-2.72E-03, 7.79E-04)	0.28	0.03	7.75E-04	6.05E-04
B-cells	1,065	2.73E-04 (-1.56E-03, 2.10E-03)	0.77	8.02E-05	0.77	2.23E-04 (-1.63E-03, 2.08E-03)	0.81	0.01	8.88E-01	8.44E-01
NK	1,065	1.87E-03 (-1.26E-03, 5.01E-03)	0.24	1.29E-03	0.24	2.03E-03 (-1.13E-03, 5.18E-03)	0.21	0.02	2.72E-02	2.69E-02
Monocytes	1,065	-8.18E-04 (-2.54E-03, 9.06E-04)	0.35	8.13E-04	0.35	-9.00E-04 (-2.56E-03, 7.56E-04)	0.29	0.11	2.20E-16	2.20E-16
Granulocytes	1,065	-7.59E-04 (-6.22E-03, 4.70E-03)	0.79	6.98E-05	0.79	-5.72E-04 (-6.06E-03, 4.92E-03)	0.84	0.02	3.64E-02	2.43E-02

† Adjusted regression for sex and the first ten genetic PCs. ‡ Estimates for these variables are in odds ratios (OR).*Systolic and diastolic blood pressure. DBP was log-transformed before analyses. †† $P_{(F)}$ P value of the regression model.** P value calculated using an ANOVA test to estimate the difference between regression coefficients across models. $P < 0.05$ was the threshold of significance.

7.4.3 Methylomic variation associated with PRS

An EWAS with the polygenic score (PRS EWAS) was conducted in a subset of 1,078 middle-age participants in ALSPAC, showing no strong association between the score and DNA methylation after Bonferroni correction for multiple testing (at $p < 1.31x10^{-7}$). Null associations were consistent across the two adjustment models used for this analysis, and there was no suggestion of genomic inflation in this EWAS based on a Lambda between 0.99 and 1.01. Top-ranked DMPs identified in borderline association with the PRS (at $p < 1.0x10^{-5}$), are described in Table 7-6. Q-Q plot and volcano plot summarizing results of the EWAS with the PRS are presented in the appendix Figure S8-23.

Table 7-6 Summary statistics for top-ranked DMPs identified in the EWAS with the polygenic risk score for T2D. Two models were implemented, a basic model adjusted for age, sex and the first-ten genetic PCs, and a second model additionally adjusted for 6 Houseman cells.

CpG	Chr	Position	Gene	Basic model			Adjusted for Cells		
				Effect	SE	P	Effect	SE	P
cg26799188	12	72629575	*Unannotated*	0.010	0.002	3.79E-07	0.010	0.002	3.59E-07
cg01554963	17	773023	*NXN*	-0.008	0.002	8.49E-07	-0.008	0.002	7.19E-07
cg03676624	11	77313686	*AQP11*	0.005	0.001	2.58E-06	0.005	0.001	3.42E-06
cg11362770	3	98451718	*ST3GAL6*	-0.003	0.001	6.95E-06	-0.003	0.001	1.28E-05

Comparing results of the PRS EWAS with those obtained in the EWAS of T2D (see Chapter 6), there was no overlap in top-ranked signals obtained across analyses. In addition, the PRS EWAS was more underpowered than the case control EWAS to identify differences in methylation across groups, even though the sample sizes were similar (PRS EWAS n=1,078, and EWAS of T2D n=1,050). Lack of power of the polygenic score to detect differences in methylation across disease groups, can be attributed to the relatively small variation in T2D explained by the score (R^2=2.0%). Thus, to strengthen findings from the PRS EWAS, it will be necessary to increase the sample studied in ALSPAC, or to meta-analyse different EWAS from additional cohorts, providing they have availability of genetic data to calculate the PRS. Another method that can be implemented to strengthen results of the PRS EWAS is to conduct a DMR analysis based on EWAS results. Main findings obtained in the DMR analysis are briefly described in section 7.4.6.

The next section shows results of the association between the PRS and methylation (i.e. observed IV-outcome association) for top-ranked DMPs ($p<1.0x10^{-5}$) detected in the observational analysis. To emphasize at this point is that throughout this chapter, results obtained for DMPs identified in the meta-EWAS of T2D are regarded as main evidence, while results for DMPs identified in a sensitivity meta-EWAS of T2D (i.e. excluding KORA samples), or in the EWAS of T2D in ALSPAC, are considered supplementary evidence.

7.4.4 Observational IV-outcome association

Results suggested weak association between the polygenic score (PRS1) and top 25 DMPs identified in the meta-EWAS of T2D (adjusted-p range 0.27 to 1.00). The association with the smallest P-value was identified at the DMP cg11851382, mapping to the *PPAP2B* gene (p=0.01). Additional associations identified with nominal significance (p<0.05) were between the polygenic score and the DMPs cg27374726 (p=0.04), cg14275576 (p=0.07) and cg07184465 in *SPZ1* (p=0.09). Direction of effect between the meta-EWAS estimate and the IV-outcome estimate was consistent for 14/25 DMPs. The magnitude of the absolute effect was on average 0.01 higher (difference range 0.001 to 0.02) in the meta-EWAS compared to the observed IV-outcome regression, and the correlation between effect estimates was small and non-significant (rho=0.21, p=0.49). Comparison of association statistics between the meta-EWAS and the IV-outcome regression are presented in Table 7-7.

Results of the IV-outcome regression for additional observational datasets

The polygenic score was also weakly associated with top DMPs detected in the sensitivity meta-EWAS of T2D (n=58 DMPs, adjusted-p range 0.62 to 1.00), and in the EWAS of T2D in ALSPAC (n=11 DMPs, adjusted-p range 0.46 to 1.00), this based on a p-threshold for significance below 1.31×10^{-7}. For DMPs identified in the sensitivity meta-EWAS, the smallest unadjusted p-value was detected at the DMP in *PPAP2B* (p=0.01), while other associations with nominal significance (p range 0.03 to 0.09) were identified at the DMPs cg20812370 in *PBX1*, cg24686009 in *RAP1B*, cg27374726, cg13178597 in *RGS17*, cg20231084, cg07184465 in *SPZ1* and cg01577083. For top DMPs identified in the EWAS in ALSPAC, the DMP cg04656330 in *PNKD* was the only one detected in borderline association with PRS1 (unadjusted p=0.04). Consistency in the direction of effect between estimates of the IV-outcome regression and the observational analysis was seen for 38/58 DMPs in the sensitivity meta-EWAS, and for 7/11 DMPs in the EWAS in ALSPAC. Magnitude of the absolute effect was always higher in the observational compared to the IV-outcome regression, and only weak correlation was identified between these estimates (sensitivity meta-EWAS r=0.16, p=0.23 and EWAS in ALSPAC rho=0.47, p=0.15). Association statistics of the IV-outcome regression for DMPs detected in additional observational datasets, are presented in the appendix Table S8-37.

Table 7-7 Comparison of association statistics between the case control analysis (T2D~Meth) and a polygenic risk score analysis (observed IV~Meth) for top 25 DMPs identified in the meta-EWAS of T2D at p < 1.0x10^{-5}. Meta-EWAS results were obtained using a model adjusted for age, sex, SVs, smoking and 6 Houseman cells. Results of the PRS~Meth regression were adjusted for age, sex, 10 genetic PCs and Houseman cells. Highlighted in bold are associations with opposite direction of effect across analyses.

CpG	Chr	Gene	T2D vs Meth†				PRS vs Meth‡			
			Estimate	SE	P	Bonf*	Estimate	SE	P	Bonf*
cg19693031	1	*TXNIP*	-0.013	2.13E-03	4.26E-09	0.002	-1.85E-03	1.43E-03	0.20	1.00
cg13826139	6	*Unannotated*	-0.006	1.06E-03	1.27E-07	0.05	-1.43E-03	9.62E-04	0.14	1.00
cg00574958	11	*CPT1A*	-0.005	1.07E-03	1.11E-06	0.42	-4.79E-04	3.93E-04	0.22	1.00
cg14275576	**20**	***Unannotated***	**-0.002**	5.05E-04	1.24E-06	0.47	**3.61E-04**	2.02E-04	0.07	1.00
cg27237541	**10**	***MYO3A***	**-0.009**	1.80E-03	1.99E-06	0.75	**3.85E-04**	1.24E-03	0.76	1.00
cg19611616	12	*STK38L*	-0.003	7.03E-04	2.56E-06	0.97	-3.97E-04	3.33E-04	0.23	1.00
cg00082384	**3**	***NISCH***	**0.008**	1.79E-03	2.86E-06	1.00	**-1.18E-03**	1.32E-03	0.37	1.00
cg06500161	21	*ABCG1*	0.007	1.53E-03	3.30E-06	1.00	9.34E-04	1.40E-03	0.50	1.00
cg14186584	5	*Unannotated*	-0.002	3.47E-04	4.01E-06	1.00	-1.19E-05	1.19E-04	0.92	1.00
cg25741837	**2**	***SMYD5***	**0.008**	1.71E-03	4.44E-06	1.00	**-1.46E-05**	1.45E-03	0.99	1.00
cg15560632	7	*LRCH4*	-0.001	2.00E-04	4.58E-06	1.00	-5.95E-05	5.24E-05	0.26	1.00
cg07400328	6	*MUTED*	-0.003	5.48E-04	5.03E-06	1.00	-2.23E-04	2.57E-04	0.39	1.00
cg22628512	1	*Unannotated*	0.005	1.06E-03	5.66E-06	1.00	3.28E-04	5.78E-04	0.57	1.00
cg06468695	17	*CCDC42*	0.005	1.11E-03	6.19E-06	1.00	4.29E-04	1.09E-03	0.69	1.00
cg06039489	**20**	***C20orf26***	**0.014**	3.01E-03	6.27E-06	1.00	**-2.40E-03**	1.64E-03	0.14	1.00
cg27374726	10	*Unannotated*	-0.007	1.49E-03	6.52E-06	1.00	-1.72E-03	8.27E-04	0.04	0.95
cg01009875	**1**	***TMCO1***	**-0.002**	4.31E-04	7.17E-06	1.00	**5.83E-07**	1.31E-04	1.00	1.00
cg17566334	**6**	***PACRG***	**0.007**	1.55E-03	7.52E-06	1.00	**-3.25E-04**	1.34E-03	0.81	1.00
cg07184465	5	*SPZ1*	-0.005	1.21E-03	8.27E-06	1.00	-1.67E-03	9.88E-04	0.09	1.00
cg11851382	1	*PPAP2B*	-0.006	1.38E-03	8.81E-06	1.00	-2.31E-03	9.03E-04	0.01	0.27
cg08273233	**6**	***HTR1E***	**-0.006**	1.45E-03	8.85E-06	1.00	**6.70E-04**	1.15E-03	0.56	1.00
cg20154947	**8**	***PLEC1***	**-0.002**	4.02E-04	9.02E-06	1.00	**6.80E-05**	9.62E-05	0.48	1.00
cg13927560	**4**	***TMEM33***	**-0.002**	4.64E-04	9.05E-06	1.00	**6.18E-05**	1.85E-04	0.74	1.00
cg01317029	3	*FAM131A*	0.006	1.26E-03	9.48E-06	1.00	1.13E-04	8.61E-04	0.90	1.00
cg17155612	**19**	***LOC148189***	**-0.002**	5.33E-04	9.55E-06	1.00	**1.13E-04**	1.21E-04	0.35	1.00

†Observed exposure versus outcome association. ‡Observed instrumental-variable versus outcome association (observed IV-outcome).

* Bonferroni-adjusted P-value.

In conclusion for this section, the polygenic score was not a good predictor of variation in methylation at top DMPs identified in the meta-EWAS of T2D, or in two other observational datasets, indicating that the predicted IV-outcome association could be underpowered, with estimates biased towards the observational estimates. Results of the single sample MR using the PRS are presented in section 7.5.

The following two sections describe sensitivity analyses conducted to determine if including the polygenic score as a covariate in the regression model increased the strength of the associations previously identified in the EWAS of T2D(section 7.4.5). In addition, results of DMR analysis using summary data from the PRS EWAS are described in section 7.4.6.

7.4.5 Sensitivity analysis using the polygenic score as a covariate in the EWAS of T2D

Using a robust linear regression (RLR) to estimate the association between T2D and top-ranking methylation sites, a strong correlation was identified between estimates obtained in a basic regression model adjusted for age, sex, 8 SVs, 6 Houseman cells, BMI and smoking, and estimates of a second model additionally adjusted for the polygenic score (r range 0.99 to 1.00, p range $3.56x10^{-16}$ to $2.0x10^{-16}$). Similarly, p-values obtained between regression models were strongly correlated (rho range 0.98 to 1.0, $p<2.0x10^{-16}$). Thus, results suggested that adding the polygenic risk score to the basic model was not contributing substantially to explain further difference in methylation at the top-ranking CpG sites relative to the difference already captured by the case control analysis. In addition, the use of a RLR improved the strength of top-ranking associations previously identified with borderline significance in the meta-EWAS while using an ordinary least square (OLS) regression. Results of the RLR with and without the polygenic score, are presented in the appendix Table S8-38.

7.4.6 PRS-associated DMRs

One DMR located in chromosome 2 was identified in strong association with the polygenic score after Sidak correction for multiple testing (Chr2:233,390,771bp-233,390,860bp, $p=2.076x10^{-5}$). This DMR mapped within the *CHRND* gene, spanned a region of 89bp, was composed of three DMPs: cg05875017, cg20371266 and cg22276371. The mean difference in methylation at the DMR in *CHRND* per standard deviation of the score was 0.008. DMPs identified within the DMR in *CHRND*, were not identified as top-ranked signals (at $p>1.0x10^{-5}$) in the PRS EWAS. Figure 7-4 shows all the DMRs initially detected by *comb-p* in association with the PRS and provides an expanded view of the strongest DMR detected in *CHRND*.

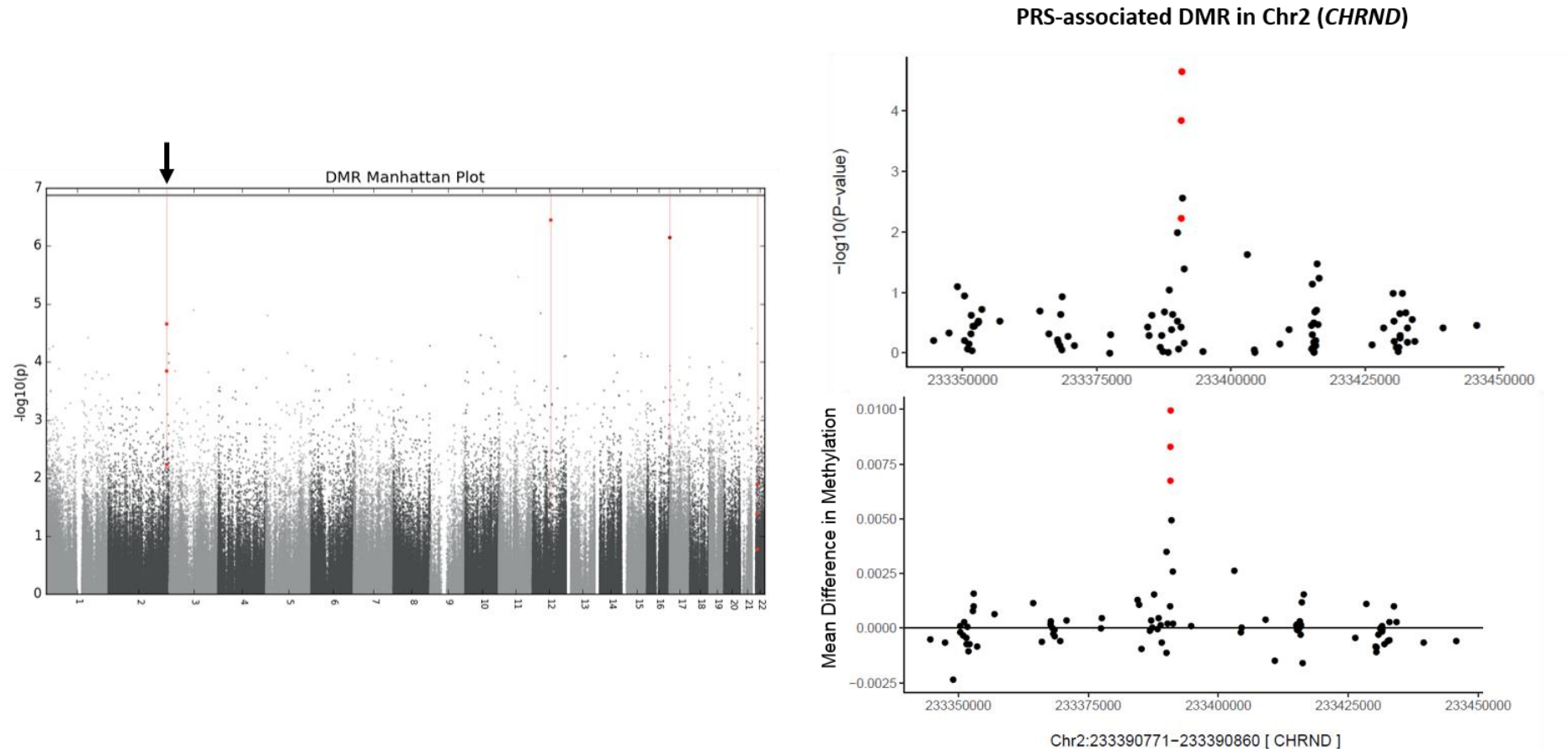


Figure 7-4 Manhattan plot (left-hand-side) showing DMRs identified by comb-p in association with the polygenic risk score for T2D (PRS1). Red lines in the plot represent the genomic location of all the DMRs initially identified, while the black arrow indicates the region in Chr2 that surpassed Sidak correction for multiple testing ($p < 0.05$). Plots on the right-hand-side show and expanded view of the Sidak significant DMR in Chr2 (CHRND gene). The top plot shows the genomic position of DMPs within the DMR against their p-value based on estimates of the PRS EWAS; the bottom plot illustrates the genomic position of the DMPs against their mean difference in methylation per SD increase in the score. Red dots represent index DMPs within the DMR in Chr2.

7.5 Single sample MR was underpowered to detect causality between T2D and differential methylation based on observational findings

Overall, the 2SLS-IV regression did not identify strong evidence of causality between T2D and differential methylation at top-ranking DMPs identified in the observational analysis. One reason for the lack of causality was weakness of the instrument, as an F-statistic < 10 was generally observed in the 2SLS-IV analysis. Despite no evidence of a causal association between T2D and differential DNA methylation, there was consistency in the direction of effect between the observational and the causal estimates, with positive but weak correlation among them. Generally, the magnitude of the effect was higher in the causal compared to the observed estimate. In addition, results of the endogeneity test were underpowered to detect significant difference between effect estimates of the observational and the causal analysis, but this finding was not indicative of a causal association between the exposure (T2D) and the outcome (DNAm). The following sections describe main results of the predicted T2D~methylation association using top DMPs identified in the meta-EWAS of T2D, and in two other observational analyses.

7.5.1 2SLS-IV analysis did not support causality in the association between T2D and methylation for top-ranked DMPs detected in the meta-EWAS of T2D

After correction for multiple testing ($p<2.0x10^{-3}$, $\alpha=0.05/25$ DMPs), only borderline evidence of causality was detected between T2D and difference in methylation at the DMP cg15560632 in *LRCH4* using an uncorrected 2SLS-IV (p=0.04), but not in the corrected IV-regression (*IV-reg* function, p=0.12) (Table 7-8). Overall, large p-values were identified in this MR analysis (range unadjusted-p 0.12 to 0.99). The direction of effect was consistent between observed and predicted estimates for 17/25 associations. The magnitude of the absolute effect was on average 0.05 higher (effect estimate decreased 0.01 and increased 0.17) in the predicted compared the observed estimate, and there was a positive but weak correlation between them (r=0.31 p=0.13). Mean variation in methylation explained by the IV (PRS) was 13% (R^2 range $4.8x10^{-4}$ to 50%). The weak instrument test revealed that the PRS was not a good instrument to predict difference in methylation (Weak instrument p range 0.01 to 0.03), and this result was consistent with a mean value of 0.72 for the Wald test, which is an estimate equivalent to the F-statistic to determine the strength of the instrument. In addition, the endogeneity test had limited power to detect strong difference between observed and causal estimates (Wu Hausman test p range 0.08 to 0.97). Figure 7-5 shows the distribution of effect estimates for the exposure-outcome association between the observational and the causal analysis, highlighting the larger standard errors of the predicted estimates.

Table 7-8 Comparison of observed versus predicted estimates for 25 top-ranked DMPs identified in the meta-EWAS of T2D (5 cohorts, n=5,147). Predicted estimates were calculated using a 2SLS-IV regression in a subsample of 862 middle-age adults in ALSPAC. P values reported are unadjusted, and associations were regarded significant at p < 2.0x10^{-3}. Highlighted in bold is the DMP identified with suggestive evidence of causality (unadjusted p <0.05).

CpG	Gene	Meta-EWAS T2D (n=5,147)		Uncorrected 2SLS-IV† (n=862)		Corrected 2SLS-IV‡ (n=862)					
		Estimate (95%CI)	P	Estimate (95%CI)	P	Estimate (95%CI)	P	$P_{Hausman}$‡‡	P_{IV}	Wald-test	R^2
cg01009875	*TMCO1*	-0.002(-0.003,-0.001)	7.17E-06	-0.022(-0.049,0.006)	0.13	-0.017(-0.047,0.014)	0.29	0.29	0.02	1.12	-0.19
cg11851382	*PPAP2B*	-0.006(-0.009,-0.003)	8.81E-06	-0.137(-0.328,0.054)	0.16	-0.147(-0.371,0.076)	0.20	0.20	0.02	1.67	-0.28
cg19693031	*TXNIP*	-0.013(-0.017,-0.008)	4.26E-09	-0.174(-0.401,0.053)	0.13	-0.175(-0.429,0.079)	0.18	0.18	0.01	1.82	-0.27
cg22628512	*Unannotated*	0.005(0.003,0.007)	5.66E-06	-0.003(-0.109,0.104)	0.96	0.002(-0.108,0.112)	0.97	0.88	0.02	1.4E-03	2.2E-03
cg25741837	*SMYD5*	0.008(0.005,0.011)	4.44E-06	0.173(-0.078,0.424)	0.18	0.181(-0.108,0.471)	0.22	0.23	0.02	1.51	-0.24
cg00082384*	*NISCH*	0.008(0.005,0.012)	2.86E-06	-0.043(-0.218,0.132)	0.63	-0.066(-0.258,0.126)	0.50	0.34	0.02	0.45	-0.15
cg01317029	*FAM131A*	0.006(0.003,0.008)	9.48E-06	0.012(-0.131,0.155)	0.87	0.011(-0.137,0.159)	0.88	0.96	0.02	0.02	0.01
cg13927560	*TMEM33*	-0.002(-0.003,-0.001)	9.05E-06	-0.004(-0.047,0.04)	0.86	-2.82E-04(-0.045,0.044)	0.99	0.97	0.02	1.5E-04	1.7E-04
cg07184465	*SPZ1*	-0.005(-0.008,-0.003)	8.27E-06	-0.114(-0.328,0.101)	0.30	-0.155(-0.402,0.093)	0.22	0.28	0.02	1.50	-0.19
cg14186584	*Unannotated*	-0.002(-0.002,-0.001)	4.01E-06	-0.012(-0.039,0.014)	0.36	-0.011(-0.039,0.018)	0.46	0.55	0.02	0.54	-0.06
cg07400328	*MUTED*	-0.003(-0.004,-0.001)	5.03E-06	-0.014(-0.075,0.047)	0.66	-0.013(-0.076,0.051)	0.70	0.76	0.02	0.15	-0.01
cg08273233*	*HTR1E*	-0.006(-0.009,-0.004)	8.85E-06	0.039(-0.203,0.282)	0.75	1.3E-05(-0.251,0.251)	1.00	0.89	0.02	9.8E-09	-4.8E-06
cg13826139	*Unannotated*	-0.006(-0.008,-0.004)	1.27E-07	-0.079(-0.228,0.071)	0.30	-0.073(-0.237,0.091)	0.38	0.43	0.02	0.77	-0.11
cg17566334*****	*PACRG*	0.007(0.004,0.01)	7.52E-06	-0.101(-0.496,0.294)	0.62	-0.121(-0.546,0.303)	0.58	0.49	0.02	0.31	-0.08
cg15560632	***LRCH4***	-0.001(-0.001,-0.001)	4.58E-06	**-0.008(-0.015,-0.001)**	**0.04**	-0.008(-0.017,0.002)	0.12	0.08	0.02	2.46	-0.50
cg20154947*	*PLEC1*	-0.002(-0.003,-0.001)	9.02E-06	-0.006(-0.03,0.018)	0.62	0.002(-0.023,0.027)	0.85	0.71	0.02	0.04	-0.02
cg27237541	*MYO3A*	-0.009(-0.012,-0.005)	1.99E-06	-0.06(-0.291,0.17)	0.61	-0.079(-0.334,0.175)	0.54	0.69	0.02	0.37	-0.02
cg27374726	*Unannotated*	-0.007(-0.01,-0.004)	6.52E-06	-0.117(-0.282,0.048)	0.16	-0.134(-0.333,0.064)	0.19	0.13	0.02	1.76	-0.38
cg00574958	*CPT1A*	-0.005(-0.007,-0.003)	1.11E-06	-0.030(-0.1,0.04)	0.40	-0.017(-0.097,0.063)	0.68	0.79	0.03	0.17	-0.01
cg19611616	*STK38L*	-0.003(-0.005,-0.002)	2.56E-06	-0.038(-0.111,0.035)	0.31	-0.047(-0.127,0.034)	0.26	0.27	0.02	1.29	-0.21
cg06468695	*CCDC42*	0.005(0.003,0.007)	6.19E-06	0.035(-0.157,0.228)	0.72	0.049(-0.153,0.251)	0.64	0.65	0.02	0.22	-0.04
cg17155612*	*LOC148189*	-0.002(-0.003,-0.001)	9.55E-06	0.003(-0.021,0.026)	0.82	0.005(-0.02,0.029)	0.71	0.53	0.02	0.13	-0.06
cg06039489	*C20orf26*	0.014(0.008,0.019)	6.27E-06	-0.091(-0.344,0.163)	0.48	-0.125(-0.406,0.156)	0.38	0.35	0.02	0.76	-0.15
cg14275576*	*Unannotated*	-0.002(-0.003,-0.001)	1.24E-06	0.012(-0.048,0.071)	0.70	0.025(-0.04,0.091)	0.45	0.37	0.02	0.56	-0.14
cg06500161*	*ABCG1*	0.007(0.004,0.010)	3.30E-06	-0.041(-0.243,0.161)	0.69	-0.07(-0.293,0.153)	0.54	0.34	0.02	0.38	-0.14

† Uncorrected 2SLS-IV refers to the manual estimation of the exposure-outcome effect using fitted values obtained from an OLS regression in the first stage of the 2SLS analysis. ‡ Estimates obtained using the *ivreg* function in the R package AER, where the two-stages of the IV analysis are automated, and a corrected SE for the predicted estimate is reported in addition to diagnostics tests. ‡‡ $P_{Hausman}$ for the Wu-Hausman test or endogeneity test. P_{IV} is the weak instrument test. Strong evidence of a weak instrument if $P_{IV} < 0.05$.*DMPS with opposite direction of effect between the MR and the observational analysis.

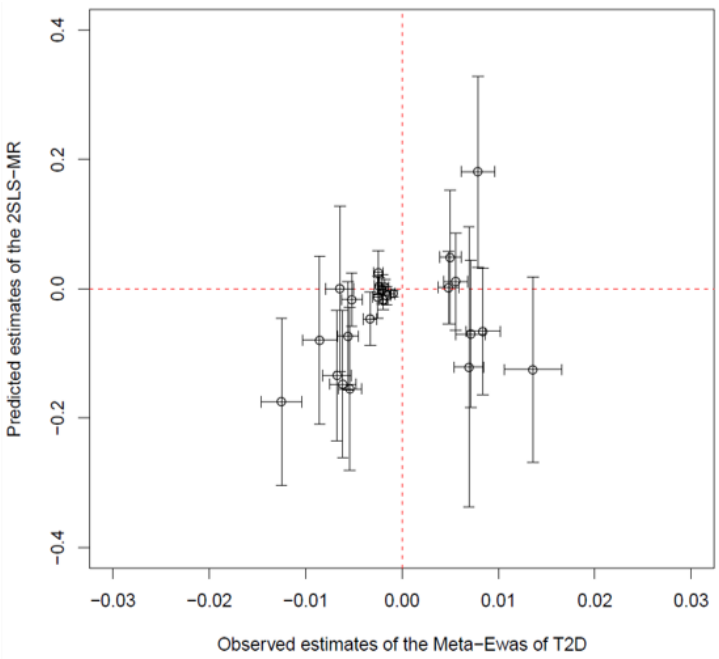


Figure 7-5 Distribution of observed (x-axis) versus predicted estimates (y-axis) of the exposure-outcome association between T2D and DNA methylation for 25 top-ranked DMPs detected observationally in the meta-EWAS of T2D. Causal estimates were obtained using a 2SLS-IV regression. Points within the top-right and bottom-left quadrants indicate CpG sites where similar direction of effect was identified between the observed and the causal analysis. Error bars represent the standard error of the observed (horizontal bars) and the causal (vertical bars) estimates.

Results of the 2SLS-IV regression for top DMPs identified in two additional observational analyses

None of the 58 top-ranked DMPs detected in the sensitivity meta-EWAS of T2D, were identified in strong association with T2D in the causal analysis after Bonferroni correction for multiple testing ($p< 8.6x10^{-4}$, $\alpha=0.05/58$) (Table 7-9). Borderline evidence of causality was identified in the uncorrected 2SLS-IV regression for the DMP cg15560632 in *LRCH4* (estimate=-0.01, unadjusted-p=0.04) and cg20812370 in *PBX1* (estimate=-0.28, unadjusted-p=0.01). In the corrected IV-regression, borderline association was also detected at the DMP in *PBX1* (unadjusted-p=0.07) (Table 7-9). For top DMPs identified in the EWAS of T2D in ALSPAC, nominal evidence of causality was detected at the DMPs cg15986668 in *NFYC* (estimate=-0.38, unadjusted-p=0.09) and cg04656330 in *PNKD* (estimate=-0.01, unadjusted-p=0.09) (Table 7-10). For the DMP in *PNKD*, nominal association was also previously identified in the observed IV-outcome regression (see section 7.4.4). Taking together results at the DMP in *PNKD*, it is possible that the predicted effect of the IV (PRS) on methylation at this site was mediated by T2D.

In general, for the two MR analyses conducted, there was consistency in the direction of effect between observed and causal estimates. The magnitude of the absolute effect was on average 0.09 higher (effect range decreased 0.01 and increased 0.32) in the causal compared to the observed

estimate, and there was positive correlation between effect estimates across analyses (sensitivity meta-EWAS r=0.40, p=0.0002, and EWAS ALSPAC r=0.50, p>0.05). On average, absolute variation in methylation explained by the PRS was between 17% and 22%, and results of the weak instrument test indicated evidence of weak instrument (p<0.05), which was consistent with values of the Wald test (Wald range 0.91 to 1.27). Furthermore, the endogeneity test was underpowered to detect strong difference between observed and causal estimates, except for associations at the DMPs cg20812370 in *PBX1* (p=0.02), cg00320980 (p=0.03), and cg04656330 in *PNKD* (p=0.05). Significance of the endogeneity test at these DMPs suggested evidence of confounding in the observed versus the causal estimates. Figure 7-6 shows the distribution of observed versus causal estimates for top DMPs detected in two additional observational analyses, highlighting the larger standard errors (error bars) identified in the causal estimates.

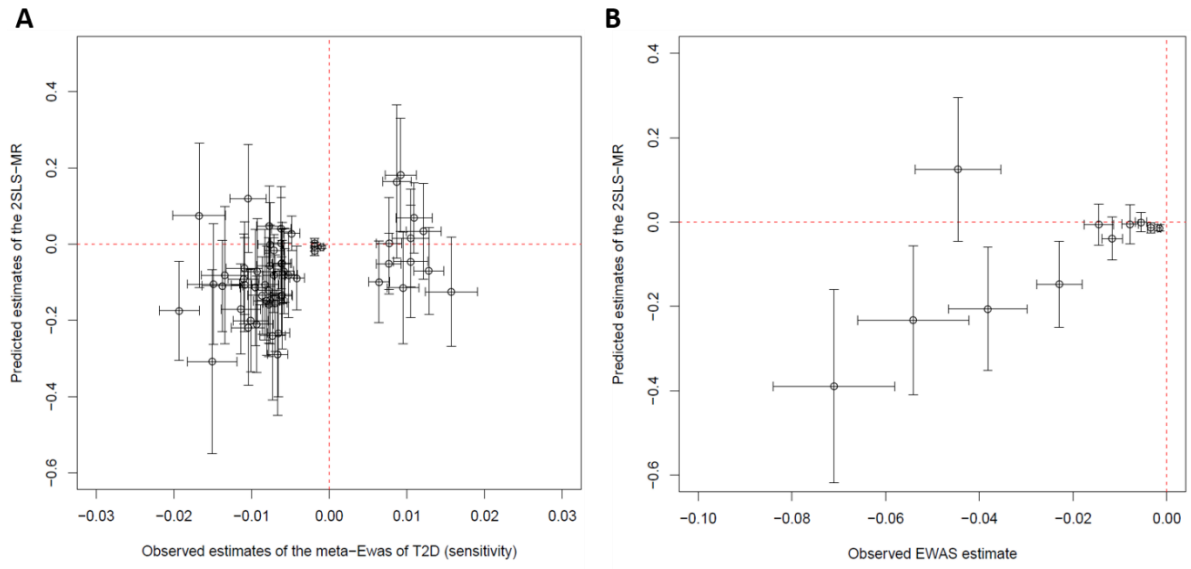


Figure 7-6 Distribution of observed (x-axis) versus predicted estimates (y-axis) of the exposure-outcome association for top-ranked DMPs detected observationally in A) the sensitivity analysis of the meta-EWAS of T2D (n=58 DMPs) and B) the EWAS of T2D in ALSPAC (n=11 DMPs). Standard errors (error bars) were larger for estimates in the causal (vertical bars) compared to the observational analysis (horizontal bars). Points within the top-right and bottom-left quadrants indicate CpG sites where similar direction of effect was identified between the observed and the causal analysis.

Table 7-9 Comparison of observed versus predicted estimates obtained in a 2SLS-IV regression for 58 top-ranked DMPs identified in a sensitivity meta-EWAS of T2D (4 cohorts, n=3,428). MR conducted using a subsample of 862 adults in ALSPAC. Associations were considered significant at p < 8.60x10^{-4}. Highlighted in bold are associations identified with nominal evidence of causality (unadjusted p<0.05).

		Sensitivity Meta-EWAS (n=3,428)		Uncorrected 2SLS-IV (n=862)		Corrected 2SLS-IV (n=862)					
CpG	Gene	Estimate (95%CI)	P	Estimate (95%CI)	P	Estimate (95%CI)	P	$P_{Hausman}$	P_{IV}	Wald-test	R^2
cg06114363	*ZNF683*	-0.01(-0.015,-0.006)	1.37E-06	-0.127(-0.367,0.114)	0.30	-0.219(-0.516,0.078)	0.15	0.12	0.02	2.09	-0.420
cg11851382	*PPAP2B*	-0.008(-0.011,-0.004)	6.42E-06	-0.137(-0.328,0.054)	0.16	-0.147(-0.371,0.076)	0.20	0.20	0.02	1.67	-0.279
cg12593793	*Unannotated*	-0.008(-0.011,-0.004)	2.90E-06	0.028(-0.167,0.222)	0.78	0.047(-0.159,0.253)	0.65	0.59	0.02	0.20	-0.049
cg14476101	*PHGDH*	-0.015(-0.021,-0.008)	9.46E-06	-0.095(-0.388,0.198)	0.53	-0.105(-0.416,0.206)	0.51	0.57	0.02	0.44	-0.052
cg19693031	*TXNIP*	-0.019(-0.024,-0.014)	8.75E-14	-0.174(-0.401,0.053)	0.13	-0.175(-0.429,0.079)	0.18	0.18	0.01	1.82	-0.274
cg20812370	***PBX1***	-0.007(-0.009,-0.004)	7.40E-07	**-0.281(-0.496,-0.067)**	**0.01**	**-0.289(-0.604,0.025)**	**0.07**	0.02	0.02	3.24	-0.996
cg25536676	*DHCR24*	-0.008(-0.011,-0.004)	5.39E-06	-0.08(-0.278,0.119)	0.43	-0.056(-0.265,0.152)	0.60	0.66	0.02	0.28	-0.032
cg00144180	*HDAC4*	0.012(0.008,0.017)	5.64E-08	0.091(-0.146,0.329)	0.45	0.033(-0.212,0.278)	0.79	0.97	0.02	0.07	0.009
cg20316538	*RUFY4*	-0.005(-0.007,-0.003)	6.11E-06	-0.046(-0.254,0.162)	0.66	-0.081(-0.301,0.139)	0.47	0.58	0.02	0.52	-0.046
cg20456243	*SPEG*	-0.007(-0.011,-0.004)	9.99E-06	-0.155(-0.427,0.116)	0.26	-0.24(-0.57,0.09)	0.15	0.13	0.02	2.03	-0.380
cg25741837	*SMYD5*	0.009(0.005,0.013)	4.76E-06	0.173(-0.078,0.424)	0.18	0.181(-0.108,0.471)	0.22	0.23	0.02	1.51	-0.241
cg10584271	*ITIH1*	-0.014(-0.019,-0.009)	1.73E-07	-0.047(-0.271,0.176)	0.68	-0.11(-0.344,0.125)	0.36	0.41	0.01	0.84	-0.111
cg20116935	*SEMA3B*	-0.006(-0.009,-0.003)	8.89E-06	-0.108(-0.276,0.061)	0.21	-0.136(-0.337,0.064)	0.18	0.17	0.02	1.77	-0.315
cg24512093	*ROBO1*	-0.01(-0.013,-0.006)	7.16E-07	-0.107(-0.392,0.178)	0.46	-0.114(-0.413,0.186)	0.46	0.61	0.02	0.56	-0.032
cg07184465	*SPZ1*	-0.007(-0.01,-0.004)	7.18E-06	-0.114(-0.328,0.101)	0.30	-0.155(-0.402,0.093)	0.22	0.28	0.02	1.50	-0.191
cg01963618	*LOC285768*	-0.008(-0.011,-0.004)	1.55E-06	0.023(-0.186,0.233)	0.83	-0.001(-0.217,0.216)	0.99	0.91	0.02	0.00	3.33E-04
cg07068382	*MTCH1*	0.01(0.006,0.015)	9.46E-06	0.024(-0.219,0.266)	0.85	0.016(-0.234,0.265)	0.90	0.99	0.02	0.02	0.002
cg13178597	*RGS17*	-0.01(-0.015,-0.006)	8.57E-06	-0.166(-0.369,0.038)	0.11	-0.201(-0.466,0.064)	0.14	0.10	0.02	2.21	-0.481
cg16192197	*Unannotated*	0.01(0.005,0.014)	3.71E-06	-0.129(-0.387,0.129)	0.33	-0.115(-0.401,0.172)	0.43	0.33	0.02	0.61	-0.159
cg10082515	*Unannotated*	-0.013(-0.019,-0.008)	7.46E-06	-0.028(-0.367,0.311)	0.87	-0.081(-0.436,0.273)	0.65	0.73	0.02	0.20	-0.018
cg15560632	***LRCH4***	-0.001(-0.001,-0.001)	3.83E-06	**-0.008(-0.015,-0.001)**	**0.04**	-0.008(-0.017,0.002)	0.12	0.08	0.02	2.46	-0.500
cg25136644	*ATG9B*	-0.007(-0.01,-0.004)	7.27E-07	-0.062(-0.238,0.114)	0.49	-0.082(-0.274,0.11)	0.40	0.43	0.02	0.70	-0.105
cg07212837	*Unannotated*	0.006(0.004,0.009)	3.28E-06	-0.124(-0.308,0.059)	0.18	-0.099(-0.309,0.11)	0.35	0.27	0.02	0.86	-0.213
cg20154947	*PLEC1*	-0.002(-0.003,-0.001)	4.34E-06	-0.006(-0.03,0.018)	0.62	0.002(-0.023,0.027)	0.85	0.71	0.02	0.04	-0.019
cg00320980	*Unannotated*	-0.009(-0.013,-0.005)	7.97E-06	-0.165(-0.345,0.015)	0.07	-0.21(-0.459,0.038)	0.10	0.03	0.02	2.75	-0.781
cg04567334	*CDH23*	-0.006(-0.008,-0.004)	1.67E-07	-0.037(-0.205,0.131)	0.66	-0.051(-0.228,0.125)	0.57	0.65	0.02	0.32	-0.033
cg08945443	*ZMYND17*	0.010(0.006,0.015)	2.64E-06	0.002(-0.277,0.281)	0.99	-0.046(-0.335,0.243)	0.76	0.78	0.02	0.10	-0.013

Continuation Table 7-9

CpG	Gene	Sensitivity Meta-EWAS (n=3,428)		Uncorrected 2SLS-IV (n=862)		Corrected 2SLS-IV (n=862)					
		Estimate (95%CI)	P	Estimate (95%CI)	P	Estimate (95%CI)	P	$P_{Hausman}$†	P_{IV}‡	Wald-test	R^2
cg19876302	*Unannotated*	-0.008(-0.011,-0.005)	2.22E-06	-0.092(-0.345,0.162)	0.48	-0.149(-0.429,0.131)	0.30	0.36	0.02	1.08	-0.139
cg27374726	*Unannotated*	-0.009(-0.012,-0.005)	2.32E-06	-0.117(-0.282,0.048)	0.16	-0.134(-0.333,0.064)	0.19	0.13	0.02	1.76	-0.382
cg00574958	*CPT1A*	-0.007(-0.01,-0.005)	1.20E-08	-0.03(-0.1,0.04)	0.40	-0.017(-0.097,0.063)	0.68	0.79	0.03	0.17	-0.010
cg11376147	*SLC43A1*	-0.006(-0.008,-0.003)	5.43E-06	-0.09(-0.21,0.03)	0.14	-0.082(-0.221,0.056)	0.25	0.24	0.02	1.35	-0.235
cg15832662	*RTN3*	-0.009(-0.013,-0.005)	8.45E-06	-0.079(-0.339,0.18)	0.55	-0.072(-0.341,0.198)	0.60	0.76	0.02	0.27	-0.008
cg20231084	*Unannotated*	-0.006(-0.009,-0.003)	8.36E-06	0.037(-0.19,0.263)	0.75	0.003(-0.232,0.238)	0.98	0.84	0.02	0.00	-0.002
cg22680424	*HCCA2*	0.008(0.005,0.011)	2.16E-06	-0.03(-0.259,0.2)	0.80	0.001(-0.236,0.239)	0.99	0.89	0.02	0.00	0.001
cg24686009	*RAP1B*	-0.002(-0.003,-0.001)	1.19E-06	-0.018(-0.039,0.003)	0.10	-0.018(-0.043,0.006)	0.14	0.12	0.01	2.16	-0.383
cg24795867	*WNT5B*	-0.006(-0.009,-0.004)	2.47E-06	-0.087(-0.34,0.166)	0.50	-0.133(-0.413,0.146)	0.35	0.38	0.02	0.87	-0.131
cg26270261	*KRT4*	-0.006(-0.009,-0.004)	5.68E-07	0.059(-0.147,0.265)	0.58	0.04(-0.177,0.258)	0.72	0.61	0.02	0.13	-0.042
cg00896068	*Unannotated*	-0.008(-0.011,-0.004)	7.58E-06	-0.105(-0.336,0.127)	0.37	-0.159(-0.425,0.106)	0.24	0.24	0.02	1.38	-0.231
cg11983038	*Unannotated*	-0.017(-0.023,-0.01)	7.23E-07	0.102(-0.251,0.455)	0.57	0.075(-0.296,0.446)	0.69	0.66	0.02	0.16	-0.033
cg00989505	*MIR299*	-0.004(-0.006,-0.002)	9.33E-06	-0.093(-0.233,0.047)	0.19	-0.089(-0.253,0.074)	0.29	0.20	0.02	1.14	-0.278
cg16765088	*Unannotated*	-0.011(-0.014,-0.007)	5.50E-10	-0.09(-0.313,0.134)	0.43	-0.107(-0.348,0.135)	0.39	0.46	0.02	0.75	-0.089
cg00162348	*RNF40*	-0.002(-0.003,-0.001)	6.64E-06	-0.011(-0.032,0.01)	0.29	-0.007(-0.029,0.015)	0.55	0.67	0.02	0.35	-0.026
cg01577083	*Unannotated*	-0.011(-0.016,-0.006)	7.93E-06	-0.13(-0.31,0.051)	0.16	-0.171(-0.403,0.061)	0.15	0.16	0.02	2.08	-0.355
cg16575444	*CX3CL1*	-0.006(-0.008,-0.004)	6.83E-07	-0.039(-0.236,0.158)	0.70	-0.05(-0.259,0.158)	0.64	0.64	0.02	0.22	-0.039
cg08857797	*VPS25*	0.009(0.005,0.012)	2.28E-06	0.116(-0.244,0.476)	0.53	0.163(-0.231,0.558)	0.42	0.40	0.02	0.66	-0.120
cg09185884	*KCTD2*	0.011(0.006,0.015)	2.33E-06	0.056(-0.116,0.228)	0.52	0.069(-0.111,0.249)	0.45	0.62	0.02	0.56	-0.029
cg11024682	*SREBF1*	0.008(0.005,0.011)	1.33E-06	-0.051(-0.193,0.091)	0.48	-0.052(-0.208,0.104)	0.52	0.32	0.02	0.42	-0.152
cg14284506	*Unannotated*	-0.005(-0.007,-0.003)	7.31E-06	0.014(-0.068,0.096)	0.74	0.028(-0.061,0.117)	0.54	0.40	0.02	0.38	-0.115
cg18181703	*SOCS3*	-0.01(-0.015,-0.006)	6.20E-06	0.173(-0.072,0.418)	0.17	0.12(-0.158,0.397)	0.40	0.27	0.02	0.71	-0.202
cg26766064	*MIR657*	-0.007(-0.009,-0.004)	5.17E-06	-0.228(-0.497,0.041)	0.10	-0.234(-0.561,0.094)	0.16	0.13	0.02	1.95	-0.387
cg08570691	*RPL13AP5*	-0.008(-0.012,-0.005)	2.78E-06	-0.05(-0.32,0.22)	0.72	-0.106(-0.391,0.178)	0.46	0.61	0.02	0.54	-0.033
cg11252555	*RPL13AP5*	-0.008(-0.011,-0.004)	7.44E-06	-0.062(-0.315,0.192)	0.63	-0.121(-0.392,0.149)	0.38	0.51	0.02	0.78	-0.061
cg24704287	*Unannotated*	-0.011(-0.015,-0.007)	2.34E-08	-0.087(-0.298,0.124)	0.42	-0.091(-0.323,0.14)	0.44	0.46	0.02	0.60	-0.094
cg06039489	*C20orf26*	0.016(0.009,0.022)	2.71E-06	-0.091(-0.344,0.163)	0.48	-0.125(-0.406,0.156)	0.38	0.35	0.02	0.76	-0.149
cg14003143	*SGK2*	-0.006(-0.008,-0.003)	4.12E-06	-0.05(-0.261,0.16)	0.64	-0.071(-0.29,0.148)	0.53	0.69	0.02	0.40	-0.017
cg06500161	*ABCG1*	0.013(0.009,0.017)	2.34E-11	-0.041(-0.243,0.161)	0.69	-0.07(-0.293,0.153)	0.54	0.34	0.02	0.38	-0.139
cg27037013	*Unannotated*	-0.015(-0.021,-0.009)	2.90E-06	-0.303(-0.708,0.103)	0.14	-0.308(-0.781,0.165)	0.20	0.20	0.02	1.63	-0.270
cg27115863	*Unannotated*	-0.011(-0.015,-0.006)	2.41E-06	-0.076(-0.305,0.154)	0.52	-0.063(-0.3,0.174)	0.60	0.77	0.02	0.27	-0.005

†$P_{Hausman}$ for the Wu-Hausman test or endogeneity test. ‡P_{IV} for the weak instrument test. Strong evidence of a weak instrument if $P_{IV} < 0.05$.

Table 7-10 Comparison of observed versus causal estimates using a 2SLS-IV analysis for 11 top DMPs identified in the EWAS of T2D in ALSPAC. MR conducted in a subsample of adults in ALSPAC (n=862). P values reported are unadjusted, and associations were considered significant at p < 4.5x10^{-3}. Highlighted in bold is the association at the DMP in PNKD identified with nominal evidence of causality (unadjusted p<0.05).

CpG	Gene	EWAS of T2D (n=1,050)		Uncorrected 2SLS-IV† (n=862)		Corrected 2SLS-IV‡ (n=862)					
		Estimate (95%CI)	P	Estimate (95%CI)	P	Estimate (95%CI)	P	$P_{Hausman}$‡‡	P_{IV}*	Wald-test	R^2
cg07251197	*Unannotated*	-0.008(-0.011,-0.004)	8.11E-06	-0.019(-0.108,0.069)	0.67	-0.005 (-0.096,0.087)	0.92	0.98	0.02	0.01	0.001
cg15986668	*NFYC*	-0.071(-0.096,-0.046)	5.48E-08	-0.321(-0.679,0.036)	0.08	-0.389 (-0.839,0.06)	0.09	0.09	0.02	2.89	-0.468
cg04656330	***PNKD***	-0.002(-0.002,-0.001)	7.96E-06	**-0.013(-0.025,-0.001)**	**0.03**	**-0.014 (-0.029,0.002)**	**0.09**	**0.05**	**0.02**	**2.93**	**-0.647**
cg19823491	*OTX1*	-0.006(-0.008,-0.003)	2.99E-06	-0.006(-0.05,0.038)	0.78	-0.0003 (-0.046,0.045)	0.99	0.87	0.02	1.0x10^{-4}	0.001
cg03206717	*SLC25A38*	-0.003(-0.005,-0.002)	2.95E-06	-0.011(-0.03,0.009)	0.29	-0.013 (-0.038,0.011)	0.28	0.34	0.03	1.16	-0.174
cg02307288	*TRPC7*	-0.038(-0.055,-0.022)	5.54E-06	-0.182(-0.424,0.06)	0.14	-0.206 (-0.493,0.081)	0.16	0.17	0.02	1.99	-0.305
cg10870892**	*CTTN*	-0.045(-0.062,-0.027)	1.13E-06	0.121(-0.176,0.419)	0.42	0.125 (-0.21,0.46)	0.46	0.29	0.02	0.54	-0.183
cg14045803	*STARD10*	-0.012(-0.016,-0.007)	1.39E-07	-0.037(-0.109,0.035)	0.31	-0.039 (-0.14,0.061)	0.44	0.51	0.05	0.59	-0.103
cg04016326	*GRIN2B*	-0.054(-0.077,-0.031)	5.71E-06	-0.205(-0.511,0.102)	0.19	-0.233 (-0.579,0.114)	0.19	0.26	0.02	1.73	-0.191
cg00204249	*DNAH17*	-0.015(-0.021,-0.008)	2.76E-06	-0.008(-0.102,0.086)	0.87	-0.005 (-0.101,0.09)	0.91	0.88	0.02	0.01	0.008
cg26652413	*CPAMD8*	-0.023(-0.032,-0.013)	2.51E-06	-0.147(-0.314,0.021)	0.09	-0.148 (-0.347,0.052)	0.15	0.16	0.02	2.09	-0.326

† Uncorrected 2SLS-IV refers to the manual estimation of the exposure-outcome effect using fitted values obtained from an OLS regression in the first stage of the 2SLS analysis. ‡ Estimates obtained using the *ivreg* function in the R package AER, where the two-stages of the IV analysis are automated, and a corrected SE for the predicted estimate is reported in addition to diagnostics tests. ‡‡ $P_{Hausman}$ for the Wu-Hausman test or the endogeneity test. *P_{IV} for the weak instrument test. Strong evidence of a weak instrument if $P_{IV} < 0.05$.**DMP with opposite direction of effect between the MR and the observational analysis.

7.5.2 Power of the 2SLS-IV analysis

On average, there was a 33% power (power range 19% to 49%) to confidently detect (at $p<0.05$) an absolute causal effect of 0.09 of T2D on methylation, this considering a 2.0% variation in T2D explained by the score, a mean variance in methylation of 0.003, and a sample size of 862. Conversely, a sample size of 3,943 will be required to detect a similar causal estimate with 80% power and 0.05 significance. Overall, the single sample MR revealed nominal evidence of causality between T2D and DNA methylation, but weakness of the instrument or a small sample size were likely factors that decreased the strength of the associations detected. To increase power and precision of the causal estimates, a two sample MR (2SMR) was implemented. In a 2SMR, two independent and well-powered samples are used to extract summary data for the genotype-exposure and genotype-outcome associations.

7.6 Two sample MR: variation in middle-age DNA methylation as consequence of T2D

This section outlines results of the 2SMR using summary data from the DIAGRAM consortium to represent the genotype versus T2D association (first sample), and from ALSPAC or the GoDMC consortium to represent the genotype versus DNA methylation association (second sample). Data from the GoDMC consortium considered for this analysis, were SNP-CpG associations that included CpG sites of interest (top-ranking DMPs detected in the observational analysis), and SNPs that were in high LD ($r^2 > 0.6$) with selected T2D SNPs (see section 7.6.5).

7.6.1 Summary data for the genotype-exposure association using DIAGRAM

As mentioned earlier, strong variants associated with T2D were extracted from four different studies included in the DIAGRAM consortium. These variants were imputed in a subsample of adults in ALSPAC and clumped to select only the strongest independent associations (smallest p-value and LD < 0.2) (see section 7.4). After processing, summary statistics were available for 75 SNPs included in the PRS analysis. From these, 65 SNPs were retained for the 2SMR based on their availability of summary data in the genotype-outcome dataset (SNP-CpG analysis) (see section 7.6.3). Top signal in association with T2D was reported at the SNP rs7903146, mapping to the *TCF7L2* gene (OR=1.39, $p=1.20 \times 10^{-139}$), while the largest p-value was reported at the SNP rs16861329 (*ST6GAL1*) (OR=1.03, $p=9.0 \times 10^{-6}$). Mean of the odds ratio was 1.09 (OR range 1.03 to 1.39). Table 7-11 shows summary statistics of the genotype-exposure associations included in the 2SMR, with effect estimates transformed to the log(odds) scale to standardize units of the estimate across the genotype-exposure and genotype-outcome datasets.

Table 7-11 Summary statistics of the genotype-exposure association reported in the DIAGRAM consortium for 65 independent T2D SNPs[20-23]. Estimate: log(odds) of the risk of T2D per increase in the effect allele, EA: effect allele, OA: other allele, EAF: effect allele frequency, N: sample-size. Associations ordered by p-value (from smallest to largest).

SNP	Chr	Mapped gene	Estimate	SE	EA	OA	EAF	P	N
rs7903146	10	*TCF7L2*	0.33	0.02	T	C	0.26	1.20E-139	144,178
rs10811660	9	*Unannotated*	0.24	0.02	G	A	0.83	1.10E-61	219,582
rs35261542	6	*CDKAL1*	0.16	0.01	A	C	0.28	1.50E-50	219,582
rs35510946	3	*IGF2BP2*	0.13	0.01	A	G	0.3	1.10E-39	219,582
rs11187140	10	*Unannotated*	0.11	0.01	G	A	0.63	1.50E-31	219,582
rs13266634	8	*SLC30A8*	0.11	0.01	C	T	0.68	5.00E-28	219,582
rs1513272	7	*JAZF1*	0.1	0.01	C	T	0.52	7.80E-25	219,582
rs9936385	16	*FTO*	0.12	0.02	C	T	0.4	2.60E-23	144,178
rs11712037	3	*PPARG*	0.13	0.02	C	G	0.86	1.70E-20	219,582
rs2972156	2	*Unannotated*	0.09	0.01	G	C	0.62	4.20E-20	219,582
rs11257658	10	*Unannotated*	0.09	0.01	A	G	0.22	1.20E-15	219,582
rs72999033	19	*HAPLN4*	0.15	0.02	T	C	0.07	1.80E-15	219,582
rs7607980	2	*COBLL1*	0.14	0.02	T	C	0.88	8.30E-15	92,794
rs6813195	4	*RPS3AP18; RPS14P6*	0.08	0.01	C	T	0.73	4.10E-14	161,639
rs77981966	2	*THADA*	0.15	0.02	C	T	0.93	4.10E-14	219,582
rs11717195	3	*ADCY5*	0.1	0.02	T	C	0.78	6.50E-14	149,821
rs340874	1	*PROX1*	0.07	0.01	C	T	0.52	5.10E-13	219,582
rs7732130	5	*ZBED3-AS1*	0.08	0.01	G	A	0.28	2.40E-12	219,582
rs17676309	3	*ADAMTS9-AS2; MIR548A2*	0.07	0.01	C	T	0.59	2.80E-12	219,582
rs1387153	11	*MTNR1B*	0.09	0.02	T	C	0.29	1.60E-11	144,178
rs2583941	12	*RPSAP52*	0.1	0.02	A	G	0.09	1.60E-11	219,582
rs10276674	7	*DGKB*	0.08	0.01	C	T	0.18	2.80E-11	219,582
rs3803563	15	*PRC1*	0.08	0.01	A	C	0.18	5.60E-11	219,582
rs12571751	10	*ZMIZ1*	0.08	0.01	A	G	0.51	1.00E-10	149,821
rs878521	7	*YKT6; CAMK2B*	0.07	0.01	A	G	0.24	1.30E-10	219,582
rs1552224	11	*ARAP1*	0.1	0.02	A	C	0.83	1.80E-10	144,178
rs516946	8	*ANK1; MIR486*	0.09	0.02	C	T	0.77	2.50E-10	149,821
rs35720761	2	*THADA*	0.11	0.02	T	C	0.89	3.30E-10	92,794
rs780094	2	*GCKR*	0.06	0.01	C	T	0.61	3.40E-10	219,582
rs243020	2	*Unannotated*	0.06	0.01	G	A	0.46	5.50E-10	219,582
rs35658696	5	*PAM*	0.16	0.03	A	G	0.96	5.70E-10	92,794
rs10842994	12	*KLHL42; PTHLH*	0.1	0.02	C	T	0.8	6.10E-10	149,821
rs5215	11	*KCNJ11*	0.07	0.01	C	T	0.39	8.50E-10	149,821
rs1974620	7	*Unannotated*	0.06	0.01	T	C	0.52	1.00E-09	219,582
rs944801	9	*CDKN2B-AS1*	0.08	0.01	C	G	0.57	2.40E-09	142,671

Continuation Table 7-11.

SNP	Chr	Mapped gene	Estimate	SE	EA	OA	EAF	P	N
rs1496653	3	*UBE2E2; MIR548AC*	0.09	0.02	A	G	0.79	3.60E-09	149,821
rs17106184	1	*FAF1*	0.1	0.02	G	A	0.91	4.10E-09	161,585
rs7177055	15	*Unannotated*	0.08	0.01	A	G	0.72	4.60E-09	149,821
rs7161785	15	*Unannotated*	0.06	0.01	G	C	0.56	4.90E-09	219,582
rs2796441	9	*LOC101927502*	0.07	0.01	G	A	0.63	5.40E-09	147,724
rs41278853	22	*MTMR3*	0.13	0.03	A	G	0.89	5.60E-09	92,794
rs6808574	3	*BCL6; LPP-AS2*	0.07	0.01	C	T	0.6	5.80E-09	140,087
rs7955901	12	*Unannotated*	0.07	0.01	C	T	0.42	6.50E-09	144,178
rs702634	5	*ARL15*	0.06	0.01	A	G	0.71	6.90E-09	154,797
rs4275659	12	*ABCB9*	0.06	0.01	C	T	0.67	9.50E-09	161,459
rs12970134	18	*RPS3AP49; MC4R*	0.08	0.02	A	G	0.26	1.20E-08	138,946
rs1359790	13	*LINC01080; SPRY2*	0.08	0.01	G	A	0.73	1.40E-08	149,821
rs7202877	16	*CTRB1-CTRB2*	0.11	0.02	T	G	0.9	3.50E-08	144,178
rs4812829	20	*HNF4A*	0.07	0.03	A	G	0.16	5.00E-08	77,138
rs7845219	8	*CCNE2; TP53INP1*	0.08	0.02	T	C	0.53	6.00E-08	77,138
rs10510110	10	*MIR3941; ARMS2*	0.05	0.01	C	C	0.43	1.00E-07	77,138
rs7961581	12	*TSPAN8*	0.06	0.01	C	T	0.27	1.80E-07	219,582
rs10190052	2	*FAM150B; TMEM18*	0.07	0.02	C	T	0.87	2.00E-07	77,138
rs9472138	6	*TRNAI25*	0.06	0.01	T	C	0.25	2.00E-07	77,138
rs2028299	15	*AP3S2; C15orf38-AP3S2*	0.04	0.02	C	A	0.29	5.00E-07	77,138
rs2820446	1	*RIMKLBP2; ZC3H11B*	0.05	0.01	C	G	0.72	2.00E-06	77,138
rs319598	5	*PCBD2*	0.05	0.01	C	T	0.53	2.00E-06	77,138
rs4273712	6	*YAP1P3; PRELID1P1*	0.05	0.01	G	A	0.25	3.00E-06	77,138
rs12427353	12	*HNF1A*	0.11	0.03	G	C	0.79	4.00E-06	77,138
rs7041847	9	*GLIS3*	0.05	0.02	A	G	0.51	5.00E-06	77,138
rs6937795	6	*SLC35D3; RPL35AP3*	0.04	0.01	A	C	0.42	7.00E-06	77,138
rs1535500	6	*KCNK16; KCNK17*	0.12	0.03	T	G	0.59	8.00E-06	77,138
rs2284219	7	*CRHR2*	0.05	0.01	G	A	0.66	8.00E-06	77,138
rs10788575	10	*RPL11P3; MED6P1*	0.06	0.01	A	G	0.17	9.00E-06	77,138
rs16861329	3	*ST6GAL1*	0.03	0.04	C	T	0.85	9.00E-06	77,138

7.6.2 Association between the genotype and potential confounders

Association between T2D-SNPs and potential confounders was investigated previously in the subsample of middle-age adults in ALSPAC (see section 7.3). According to this analysis, the SNP rs516946 was the only proxy identified in association with one of the genetic principal components (PC9), which was retained for the 2SMR analysis. However, PCs were not direct confounders of the exposure-outcome association. Thus, including the SNP rs516946 in the 2SMR did not represent a violation of the second MR assumption (i.e. no association of IVs with confounders).

7.6.3 Estimating the association between T2D-SNPs and methylation

As described in the methods section in Chapter 3, an EWAS of T2D-SNPs was performed using 65 T2D-SNPs and 482,015 probes that survived common genetic and methylation QC. Results of this EWAS were used to (1) obtain summary statistics of the observed IV-outcome association for DMPs identified in the observational analysis that were considered in the 2SMR, and to (2) identify strongest genetic influence on middle-age DNA methylation for variants associated with T2D. This EWAS of T2D-SNPs was conducted using a subsample of 1243 adults in ALSPAC. Table 7-12 shows 10

SNPs excluded from the EWAS after failing QC processing for the genetic data (see Chapter 3 "*GoDMC protocol for detecting SNP-CpG associations").*

Table 7-12 T2D-SNPs excluded from the EWAS of T2D-SNPs after failing QC applied to the genetic data before conducting the SNP-CpG analysis.

SNP	Chr	Gene	OR	P	DIAGRAM Study
rs10937721	4	*WFS1*	1.09	$4.30x10^{-18}$	Gaulton *et al.* 2015 [169]
rs4430796	17	*HNF1B*	1.09	$7.80x10^{-18}$	Mahajan *et al*. 2014 [30]
rs2238689	19	*GIPR*	1.08	$8.30x10^{-16}$	Gaulton *et al.* 2015
rs1169288	12	*HNF1A*	1.09	$8.10x10^{-15}$	Gaulton *et al.* 2015
rs9502570	6	*SSR1*	1.06	$1.00x10^{-09}$	Mahajan *et al*. 2014
rs231361	11	*KCNQ1*	1.09	$1.20x10^{-09}$	Morris *et al.* 2012 [101]
rs459193	5	*ANKRD55*	1.08	$6.00x10^{-09}$	Morris *et al.* 2012
rs2283220	11	*KCNQ1*	1.06	$2.40x10^{-07}$	Gaulton *et al.* 2015
rs7795991	7	*ETV1*	1.05	$7.00x10^{-07}$	Mahajan *et al.* 2014
rs2812533	10	*C10orf35*	1.07	$5.00x10^{-06}$	Mahajan *et al.* 2014

7.6.3.1 Main findings of the EWAS of T2D-SNPs

Restricting the analysis to autosomal probes, 110 significant SNP-CpG pairs were identified surpassing a Bonferroni threshold of $p<1.60x10^{-9}$ ($\alpha=0.05/31,330,975$ tests). Associations were further pruned for variants with imbalanced genotype frequencies (genotype frequencies < 0.25 and > 0.75 were excluded). Significant SNP-CpG pairs were composed of 26 unique SNPs and 110 unique DMPs, none of which overlapped with DMPs of interest detected in the observational analysis and considered for the 2SMR. Only borderline association was detected between a T2D SNP and one DMP detected observationally in the sensitivity meta-EWAS of T2D (see section 7.6.3.2).

Strongest DMPs identified in association with genetic variants for T2D could have been good candidates for a 2SMR analysis, with the drawback that they did not represent observational evidence of difference in methylation in response to T2D itself, but to the genetics of T2D. Thus, these associations were not considered for the MR analysis as the main interest was on identifying causality for DMPs detected observationally in association with T2D. Summary of the strongest SNP-CpG pairs identified in the EWAS of T2D-SNPs can be found in the appendix Table S8-39. The following section describes results of the observed IV-outcome association for DMPs identified in the observational analysis.

7.6.3.2 Genetic variants for T2D were not associated with strongest DMPs detected in the meta-EWAS of T2D

No strong association was identified between the genotype for T2D and DMPs detected in the meta-EWAS of T2D after applying Bonferroni correction at $p<3.08x10^{-5}$. Of the 1,625 associations tested

(i.e. 65 SNPs * 25 DMPs), 1,624 were in *trans* and only one was identified in *cis* between rs10842994 (*KLHDC5*) and cg19611616 (*STK38L*) in Chr12. The association with the smallest p-value was detected in *trans* between rs2284219 (*CRHR2*) and cg00574958 (*CPT1A*) (estimate=0.14, $p=8.6x10^{-4}$). For the DMP cg19693031 in *TXNIP*, which was the strongest observational association (adjusted-p=0.002), the strongest association with the genotype was detected in *trans* with the SNP rs10788575 (*PTEN*, estimate=-0.11, unadjusted-p=0.03), and there was some evidence of genotype-frequency imbalance for this association. For the DMP cg13826139, which was the second stronger observational association (adjusted-p=0.048), the strongest association with the genotype was detected in *trans* with the SNP rs7607980 (*COBLL1*, estimate=0.17, $p=4.0x10^{-3}$); there was also some evidence of genotype-frequency imbalance for this SNP-CpG association.

Overall, across the 1,625 SNP-CpG associations interrogated, mean absolute effect estimate was 0.04 (estimate decreased 0.23 and increased 0.21), p-value ranged between $8.6x10^{-4}$ and 0.99, the average variation in methylation explained by the genotype was $8.0x10^{-4}$ (R^2 range $1.0x10^{-9}$ to $8.9x10^{-3}$), mean level of methylation across 25 DMPs of interest was 0.41, and evidence of genotype-frequency imbalance was observed in 600/1,625 SNP-CpG associations. Summary statistics for the top-ten SNP-CpG pairs with the smallest p-value are shown in Table 7-13.

Genotype-methylation association for top DMPs identified in two additional observational analyses

Using a Bonferroni corrected $p<1.33x10^{-5}$, strong association was identified in *trans* between the T2D-SNP rs4275659, and the DMP cg10584271 (*ITIH1*) (estimate=0.22, adjusted-p=0.002, $R^2=0.02$), which was identified in the sensitivity meta-EWAS of T2D (i.e. excluding KORA results). No genotype-frequency imbalance was reported for this association. Relevance of the observed genetic association at the DMP in *ITIH1* will be determined by the strength of the predicted IV~DMP association mediated by T2D, this to disregard any potential pleiotropic effects. Apart from the signal detected at the DMP in *ITIH1*, none of the six strongest associations (adjusted-p<0.05) identified in the sensitivity meta-EWAS of T2D (DMPs mapping to *TXNIP*, *ABCG1*, *CPT1A*, *HDAC4,* and the Intergenic DMPs cg16765088 and cg24704287, see Chapter 6), were found in strong association with genetic variants for T2D at $p<1.33x10^{-5}$ (Table 7-14).

Similarly, there was no evidence of strong association between T2D-SNPs and top DMPs identified in the EWAS of T2D in ALSPAC at Bonferroni corrected $p<6.99x10^{-5}$. The association with the smallest p-value was detected in *trans* between rs7732130 (*ZBED3-AS1*) and the intergenic DMP cg07251197 (estimate=-0.14, $p=1.3x10^{-3}$, $R^2=8.3x10^{-3}$) (Table 7-17). For the DMP in *NFYC,* which was strongly

associated with T2D in the observational analysis (adjusted-p=0.02), only weak association was detected with the genotype, with p-values ranging between $4.0x10^{-3}$ and 0.99.

Table 7-13 Summary statistics of the top-ten strongest associations detected in the EWAS of T2D-SNPs (65 SNPs) in relation to top DMPs identified in the meta-EWAS of T2D (five cohorts: ALSPAC, KORA, LBC1936, Rotterdam-III-1 and Rotterdam-Bios). Estimates are interpreted as the additive effect of the genotype on a unit change in inverse-normal transformed residuals of methylation, EA is the effect allele (minor allele), OA is the major allele, EAF is the effect allele frequency or MAF, P is the unadjusted P from the regression, Class indicates if the SNP-CpG pair was identified in Cis (< 1Mb) or in trans (> 1Mb), GI stands for genotype frequency imbalance: yes if genotype frequency <0.25 or >0.75, no if the opposite, R^2 is the total variation in methylation explained by the SNP on each SNP-CpG association. Associations were considered significant at Bonferroni corrected p < $3.08x10^{-5}$

SNP	DMP	SNP Gene	DMP Gene	EA	OA	EAF	Estimate	SE	P	Bonferroni	R^2	N	Class†	GI
rs2284219	cg00574958	*CRHR2*	*CPT1A*	A	G	0.65	0.14	0.04	8.63E-04	1.00	8.87E-03	1,233	Trans	No
rs1974620	cg06468695	*Unannotated*	*CCDC42*	C	T	0.53	-0.13	0.04	1.27E-03	1.00	8.31E-03	1,243	Trans	No
rs5215	cg15560632	*KCNJ11*	*LRCH4*	C	T	0.63	0.13	0.04	2.09E-03	1.00	7.57E-03	1,243	Trans	No
rs11712037	cg06468695	*PPARG*	*CCDC42*	G	C	0.89	0.19	0.06	2.48E-03	1.00	7.33E-03	1,233	Trans	Yes
rs3803563	cg01009875	*PRC1*	*TMCO1*	A	C	0.83	0.16	0.05	2.65E-03	1.00	7.23E-03	1,236	Trans	Yes
rs7607980	cg13826139	*COBLL1*	*Unannotated*	C	T	0.86	0.17	0.06	3.73E-03	1.00	6.73E-03	1,242	Trans	Yes
rs1496653	cg11851382	*Unannotated*	*PPAP2B*	G	A	0.78	-0.14	0.05	3.97E-03	1.00	6.64E-03	1,243	Trans	Yes
rs1535500	cg00082384	*KCNK16*	*NISCH*	T	G	0.52	-0.11	0.04	4.09E-03	1.00	6.60E-03	1,229	Trans	No
rs9472138	cg13927560	*VEGFA*	*TMEM33*	T	C	0.71	0.12	0.04	5.51E-03	1.00	6.17E-03	1,243	Trans	No
rs7961581	cg00574958	*TSPAN8*	*CPT1A*	C	T	0.74	0.12	0.04	5.53E-03	1.00	6.16E-03	1,235	Trans	No

Table 7-14 Summary statistics of the top-ten strongest associations detected in the EWAS of T2D-SNPs (65 SNPs) in relation to top-ranking DMPs detected in the sensitivity meta-EWAS of T2D (four cohorts, excluding KORA from the analysis). Associations were considered significant at Bonferroni corrected $p<1.33x10^{-5}$.

SNP	DMP	SNP Gene	DMP Gene	EA	OA	EAF	Estimate	SE	P	Bonferroni	R^2	N	Class†	GI‡
rs4275659	**cg10584271**	***Unannotated***	***ITIH1***	**T**	**C**	**0.70**	**0.22**	**0.04**	**5.33E-07**	**2.01E-03**	**2.00E-02**	**1,216**	**Trans**	**No**
rs9936385	cg14284506	*FTO*	*Unannotated*	C	T	0.62	0.16	0.04	1.47E-04	0.55	1.15E-02	1,242	Trans	No
rs12427353	cg08945443	*HNF1A*	*ZMYND17*	C	G	0.80	-0.18	0.05	5.54E-04	1.00	9.53E-03	1,201	Trans	Yes
rs1359790	cg20812370	*SPRY2*	*PBX1*	A	G	0.73	0.16	0.05	5.77E-04	1.00	9.47E-03	1,216	Trans	No
rs780094	cg15832662	*GCKR*	*RTN3*	T	C	0.59	-0.14	0.04	7.54E-04	1.00	9.07E-03	1,243	Trans	No
rs2284219	cg00574958	*CRHR2*	*CPT1A*	A	G	0.65	0.14	0.04	8.63E-04	1.00	8.87E-03	1,233	Trans	No
rs944801	cg16192197	*Unannotated*	*Unannotated*	G	C	0.57	0.13	0.04	8.84E-04	1.00	8.84E-03	1,219	Trans	No
rs13266634	cg04567334	*SLC30A8*	*CDH23*	T	C	0.69	-0.14	0.04	1.28E-03	1.00	8.29E-03	1,243	Trans	No
rs2284219	cg00989505	*CRHR2*	*MIR299*	A	G	0.65	0.13	0.04	1.82E-03	1.00	7.77E-03	1,233	Trans	No
rs5215	cg15560632	*KCNJ11*	*LRCH4*	C	T	0.63	0.13	0.04	2.09E-03	1.00	7.57E-03	1,243	Trans	No

† Class indicates if the SNP-CpG pair was identified in cis (< 1Mb) or in trans (> 1Mb). ‡ GI stands for genotype frequency imbalance, yes if genotype frequency <0.25 and >0.75, no if the opposite

Table 7-15 Summary statistics of the top-ten strongest SNP-CpG associations detected in the EWAS of T2D-SNPs (65 SNPs) in relation to top-ranking DMPs identified in the EWAS of T2D in ALSPAC. Associations were considered significant at Bonferroni corrected $p<6.99x10^{-5}$.

SNP	DMP	SNP Gene	DMP Gene	EA	OA	EAF	Estimate	SE	P	Bonferroni	R^2	N	Class†	GI‡
rs7732130	cg07251197	*ZBED3-AS1*	*Unannotated*	G	A	0.70	-0.14	0.04	1.28E-03	0.91	8.30E-03	1,207	Trans	No
rs35720761	cg15986668	*THADA*	*NFYC*	T	C	0.87	0.17	0.06	3.65E-03	1.00	6.76E-03	1,236	Trans	Yes
rs35261542	cg04656330	*CDKAL1*	*PNKD*	A	C	0.74	-0.13	0.05	4.77E-03	1.00	6.38E-03	1,229	Trans	No
rs7161785	cg03206717	*Unannotated*	*SLC25A38*	C	G	0.57	-0.11	0.04	4.97E-03	1.00	6.32E-03	1,239	Trans	No
rs1552224	cg04656330	*ARAP1*	*PNKD*	C	A	0.84	-0.15	0.05	5.96E-03	1.00	6.05E-03	1,243	Trans	Yes
rs41278853	cg10870892	*MTMR3*	*CTTN*	G	A	0.90	-0.18	0.07	7.50E-03	1.00	5.72E-03	1,242	Trans	Yes
rs77981966	cg15986668	*THADA*	*NFYC*	T	C	0.93	0.21	0.08	7.50E-03	1.00	5.72E-03	1,206	Trans	Yes
rs7955901	cg04656330	*Unannotated*	*PNKD*	C	T	0.57	-0.10	0.04	1.10E-02	1.00	5.18E-03	1,240	Trans	No
rs7961581	cg10870892	*TSPAN8*	*CTTN*	C	T	0.74	0.11	0.04	1.22E-02	1.00	5.03E-03	1,235	Trans	No
rs2583941	cg07251197	*RPSAP52*	*Unannotated*	A	G	0.90	0.16	0.07	1.26E-02	1.00	4.99E-03	1,240	Trans	Yes

† Class indicates if the SNP-CpG pair was identified in cis (< 1Mb) or in trans (> 1Mb). ‡ GI stands for genotype frequency imbalance, yes if genotype frequency <0.25 and >0.75, no if the opposite

7.6.4 Results of the forward 2SMR

After applying data harmonization on MR-Base to account for the differential annotation of SNP-alleles across the genotype-exposure and genotype-outcome datasets, 62/65 T2D-SNPs remained for further analyses. SNPs were excluded based on the incompatibility of alleles across datasets (i.e. SNP rs10510110 with alleles C/C and C/T) and based on palindromic SNPs with unreliable inference of the reference strand when using allele frequencies (SNPs rs944801 and rs7161785 where MAF was close to 0.5). Table 7-16 shows SNPs that were excluded from the analysis during data harmonization on MR-Base.

Table 7-16 List of SNPs excluded from MR-Base during data harmonization. EA: effect allele, OA: other allele.

SNP	Chr	Position	Gene	EA	OA	OR	SE	P	DIAGRAM
rs944801	9	22051670	*CDKN2B-AS1*	C	G	1.08	0.01	2.40E-09	Morris *et al.* 2012
rs10510110	10	124192430	*PLEKHA1*	C	C	1.05	0.01	1.00E-07	Mahajan *et al.* 2014
rs7161785	15	62395224	*Unannotated*	G	C	1.06	0.01	4.90E-09	Gaulton *et al.* 2015

7.6.4.1 Variation in methylation at the DMP cg00082384 in NISCH was the strongest signal identified in response to the effects of T2D

The 2SMR revealed borderline evidence of causality (at $p<0.05$) between T2D and difference in methylation at four DMPs: cg00082384 (*NISCH*), cg19693031 (*TXNIP*), cg06039489 (*C20orf26*) and the DMP cg14275576 (Figure 7-7). Causal estimates for these DMPs did not surpass Bonferroni correction for multiple testing at $p<0.002$ ($\alpha=0.05/25$ DMPs), but they were identified with borderline significance in at least one of the five methods applied, with most of the stronger results obtained using the weighted median method (at least 50% valid instruments). Heterogeneity reported by the Cochran's Q estimate for these DMPs was high (Cochran's Q 55.9 to 79.4) but non-significant (heterogeneity p range 0.06 to 0.66), and the Egger intercept suggested evidence of negative (DMPs cg19693031 and cg14275576) and positive (DMPs cg06039489 and cg00082384) horizontal pleiotropy, but this was not strong ($p>0.002$). Overall, the genetic instruments explained 5% of the mean variation in methylation, and 2% of the mean variation in T2D, reason why the Steiger test suggested that true direction of causality in the 2SMR was from methylation to T2D, and not the opposite.

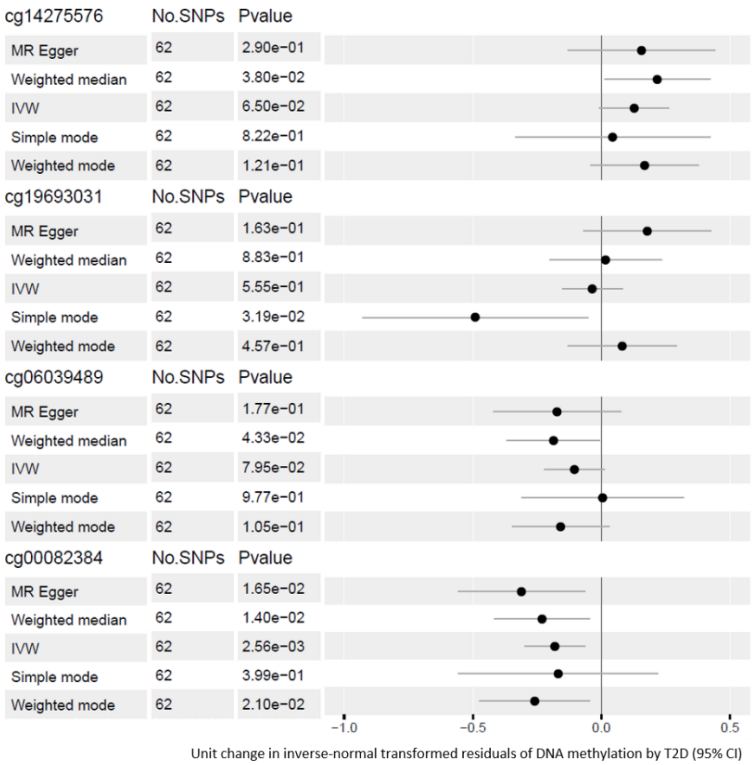


Figure 7-7 One-to-many forest plot illustrating results of the 2SMR for the causal effect of T2D on methylation at four of the 25 top DMPs identified observationally in the meta-EWAS of T2D. DMPs illustrated were found with suggestive evidence of causality in at least one of the five MR methods used. For each DMP, mean of the combined causal effect is depicted by the black point, and lines across the mean are the 95% CI. P-value is the unadjusted P, and results were considered significant at Bonferroni corrected p< 0.002 (α=0.05/25 DMPs tested).

Of the four associations identified with nominal causality, the one reported with the smallest p-value across different methods (IVW-regression, MR-Egger, weighted median and weighted mode), was between T2D and the DMP in *NISCH* (p ranged between 0.003 and 0.02) (Figure 7-7). For this DMP, opposite direction of effect was identified between the observed and the causal estimate, and the Steiger test suggested that true direction of causality was from methylation to T2D (T2D R^2=0.02 versus DMP R^2=0.05), however this evidence was not robust (p>0.002).

Comparing results between methods for the DMP in *NISCH*, the random-effect IVW estimate suggested that T2D was associated with a -0.18 (95% CI= -0.29, -0.06) decrease in inverse-normal transformed residuals of methylation, without evidence for heterogeneity amongst SNPs (Cochran's Q=56.88, p=0.63). Despite no heterogeneity, it was identified asymmetry in the funnel plot for this DMP (Figure 7-8), highlighting the larger negative effect of SNP rs16861329, compared to the effect of other SNPs, indicating evidence of horizontal pleiotropy. In the MR-Egger regression, a positive intercept was detected (Egger-intercept=0.01, SE=0.01), without strong evidence of directional pleiotropy (p=0.25). Compared to the IVW estimate, the magnitude of the MR-Egger estimate was

stronger (estimate=-0.31, 95% CI= -0.56, -0.06), and there was consistency in the direction of effect across the two methods. Magnitude of the MR-Egger estimate was similar to that of the weighted median (estimate=-0.23, 95% CI=-0.42, -0.05) and weighted mode analyses (estimate=-0.26, 95% CI=-0.47, -0.05). Conversely, the IVW estimate was comparable in magnitude to that of the simple mode analysis (estimate=-0.17, 95% CI=-0.56, 0.22), even though results of the simple mode were not robust based on the wide confidence intervals. A leave-one-out sensitivity analysis showed that by removing at least one of three SNPs at a time from the analysis (SNPs rs7903146, rs1535500 and rs7607980), there was a more positive total causal effect of T2D on methylation at the DMP in *NISCH*, compared to the change in the effect observed when removing any other variant (Figure 7-8).

For the remaining DMPs mapping to *TXNIP* (cg19693031), *C20orf26* (cg06039489) and the unannotated DMP cg14275576, a causal effect was detected in at least one of the methods at $p<0.05$. The strongest estimate for the DMP in *TXNIP* was detected using the simple mode estimate, and direction of effect was consistent between the causal and the observed estimates. According to results of the 2SMR, T2D was associated with an average -0.49 (95% CI=-0.93, -0.05) decrease in inverse-normal transformed residuals of methylation at the DMP in *TXNIP*. No evidence for heterogeneity (Cochran's Q=59.24, p=0.54) or directional horizontal pleiotropy (directionality-p=0.06) was detected for this DMP (Table 7-17).

For the DMP in *C20orf26,* the strongest estimate was detected using the weighted median (estimate=-0.18, 95% CI=-0.37, -0.01), and magnitude and direction of effect for the causal estimate was similar across methods, except for estimates of the simple mode analysis (Table 7-17). Direction of effect was not consistent between the causal and the observed estimate, and there was no evidence for heterogeneity (Cochran's Q=55.92, p=0.66) or directional pleiotropy (directionality-p=0.54) for the association at the DMP in *C20orf26*.

Strongest evidence of causality in the intergenic DMP cg14275576 was detected using the weighted median estimate (estimate=0.22, 95% CI=0.01, 0.42). Magnitude of the causal estimate was similar across methods except for the simple mode estimate, and direction of effect was not consistent between the observed and the causal estimate (Table 7-17). For this DMP, there was some evidence of heterogeneity amongst SNPs (Cochran's Q=79.38, p= 0.06), but there was no evidence of directional pleiotropy (directionality-p=0.82).

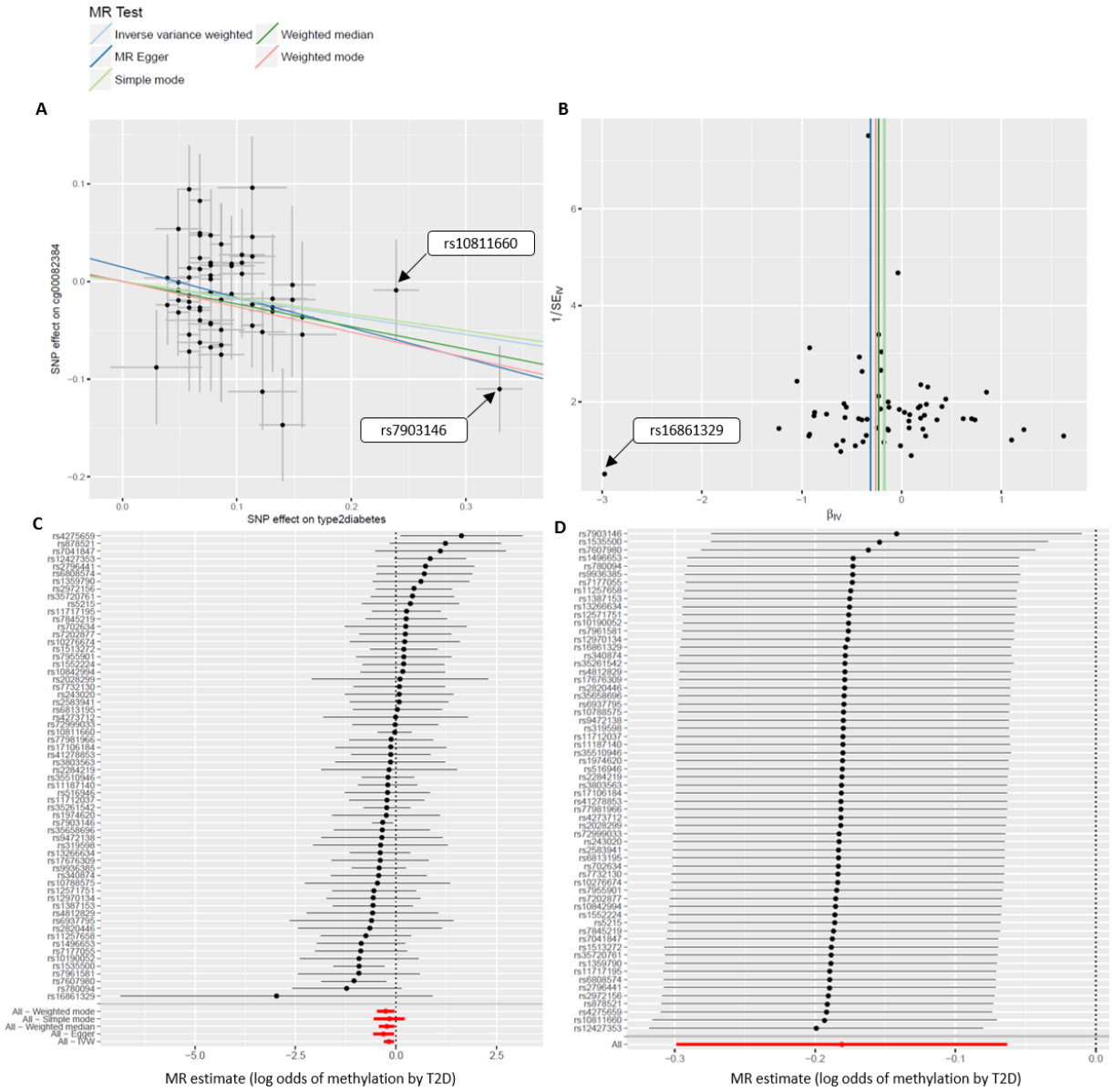


Figure 7-8 Mendelian randomization study for the causal effect of T2D on methylation at the DMP in NISCH (cg00082384). A) scatterplot of the effect of the SNP on the exposure (x-axis) and on the outcome (y-axis), with fitted lines representing results of five different methods; the slope of the fitted line is the combined casual effect across SNPs (n=62). Two SNPs are pointed-out for their heterogenous effect (large effect on the exposure and small effect on the outcome). B) funnel plot showing the causal effect of each SNP considered individually as an instrument (x-axis), against the inverse of the SE of the causal estimate (y-axis). Vertical lines in the plot illustrate the meta-analysed causal effect across SNPs, according to different methods. Asymmetry in the funnel plot caused by the extreme negative effect of SNP rs16861329 on methylation, indicated presence of horizontal pleiotropy. C) forest plot showing the causal effect of each SNP individually as an instrument depicted by the black point, and the horizontal line across is the 95% CI of this estimate. Bottom red lines are the combined causal estimates across SNPs for the different methods applied. Of interest in this plot, is the heterogeneous effect identified at the SNP rs16861329. D) leave-one-out sensitivity analysis. Each black point in the plot is the combined IVW estimate after removing one SNP at a time from the analysis (SNPs on the left-hand side), and the bottom red line is the combined effect across all SNPs using the IVW method. Sequentially removing top-three SNPs in the plot, suggested a less negative effect of T2D on methylation at the DMP in NISCH.

Table 7-17 Results of the 2SMR for the effect of T2D on variation in methylation at three DMPs detected with p<0.05 in at least one MR method. DMPs included as outcomes were detected observationally in the meta-EWAS of T2D. IVW was regarded as main evidence of the 2SMR, while MR-Egger was a sensitivity analysis to account for the effect of horizontal pleiotropy. Associations were regarded significant at p<0.002 after Bonferroni correcton.

CpG	Chr	Gene	Meta-EWAS of T2D			IVW		MR-Egger	
			Beta (95%CI)	P	N	Beta (95%CI)	P	Beta (95%CI)	P
cg19693031	1	*TXNIP*	-0.013(-0.017, -0.008)	4.25E-09	5,147	-0.035(-0.153,0.082)	0.555	0.178(-0.069,0.426)	0.163
cg00082384	3	*NISCH*	0.008(0.005,0.012)	2.86E-06	5,147	**-0.181(-0.298, -0.063)**	**0.003**	**-0.311(-0.558, -0.064)**	**0.017**
cg06039489	20	*C20orf26*	0.014(0.008,0.019)	6.27E-06	5,147	-0.105(-0.223,0.012)	0.080	-0.172(-0.42,0.075)	0.177
cg14275576	20	*Unannotated*	-0.002(-0.003, -0.001)	1.24E-06	5,147	0.127(-0.008,0.263)	0.065	0.156(-0.131,0.443)	0.290

Continuation Table 7-17 Additional MR methods.

CpG	Chr	Gene	Simple mode		Weighted median		Weighted Mode	
			Beta (95%CI)	P	Beta (95%CI)	P	Beta (95%CI)	P
cg19693031	1	*TXNIP*	-0.49(-0.928, -0.053)	**0.032**	0.016(-0.202,0.234)	0.883	0.081(-0.131,0.293)	0.457
cg00082384	3	*NISCH*	-0.168(-0.556,0.22)	0.399	-0.231(-0.415, -0.047)	**0.014**	-0.259(-0.474, -0.045)	**0.021**
cg06039489	20	*C20orf26*	0.005(-0.310,0.319)	0.977	-0.186(-0.367, -0.006)	**0.043**	-0.159(-0.348,0.030)	0.105
cg14275576	20	*Unannotated*	0.044(-0.335,0.423)	0.822	0.217(0.012,0.423)	**0.038**	0.168(-0.041,0.378)	0.121

Continuation Table 7-17 Diagnostic tests to assess heterogeneity (Cochran's Q in IVW regression) and pleiotropy (directionality-p in MR-Egger regression).

CpG	Chr	Gene	Cochran's Q	P_Q	Directionality P
cg19693031	1	*TXNIP*	59.238	0.540	0.059
cg00082384	3	*NISCH*	56.879	0.626	0.246
cg06039489	20	*C20orf26*	55.920	0.660	0.547
cg14275576	20	*Unannotated*	79.381	0.057	0.824

7.6.4.2 Forward 2SMR for DMPs identified in two additional observational analyses

Considering top signals identified in the sensitivity analysis of the meta-EWAS of T2D, which were not reported as top DMPs in the complete meta-analysis (n=48 DMPs), the 2SMR revealed borderline causality at p<0.05 between T2D and five DMPs: cg01577083, cg20456243 (*SPEG*), cg20812370 (*PBX1*), cg24686009 (*RAP1B*) and cg26766064 (*MIR657*), none of these associations surpassing Bonferroni significance at p<0.001 (α=0.05/48 DMPs) (appendix Figure S8-24, Table 7-18). Overall, there was consistency in the direction of effect between the observed and the causal estimate, the magnitude of the causal effect was similar across different methods, and most of the strongest estimates were obtained using the weighted median and weighted mode rather than the IVW and MR-Egger regressions (Table 7-18). Of the five DMPs identified with borderline causality, smallest p-value of association was detected at the DMP cg20812370 in *PBX1* (p ranged $2.0x10^{-3}$ to 0.31), whilst weakest association was detected at the DMP cg24686009 in *RAP1B* (p range 0.05 to 0.58) (appendix Figure S8-24). Looking at the diagnostic tests, there was no evidence for heterogeneity or directional pleiotropy in results of this MR, but the Steiger test strongly suggested that true direction of causality for the association at the DMP cg01577083, was from methylation to T2D, and not the opposite (Steiger $p=5.5x10^{-5}$).

Of interest in this analysis was the replication in the 2SMR of the association at the DMP in *PBX1*, a signal which was previously identified with borderline significance in the single sample MR (estimate=-0.29, unadjusted-p=0.07) (see section 7.5.1). Across analyses, causal estimates were consistent in magnitude and direction of effect, suggesting that T2D was causally associated with a decrease in methylation at *PBX1.* This result was also in agreement with the observational evidence (Table 7-18). In the 2SMR, no evidence for heterogeneity (Cochran's Q=52.92, p=0.76) or horizontal pleiotropy (Egger-intercept=-0.002, p=0.87) was detected at the DMP in *PBX1*, even though there was some asymmetry in the funnel plot caused by the outlier SNPs rs319598 and rs1359790 (Figure 7-9B). A sensitivity analysis excluding one SNP at a time from the MR analysis, revealed that none of the SNPs was completely driving the total effect of T2D on methylation at the DMP in *PBX1* (Figure 7-9D). In contrast, the 2SMR did not support a causal effect of T2D on methylation at the DMP cg10584271 in *ITIH1* (at p<0.05), a DMP that was previously identified in borderline association with the genotype in the observed IV-outcome analysis (see section 7.6.3.2). Thus, the genetic association previously detected at *ITIH1* was probably driven by horizontal pleiotropy, and this concept was further confirmed by the identification of asymmetry in the funnel plot for the DMP in *ITIH1* (appendix Figure S8-25)

Table 7-18 Results of the 2SMR for the effect of T2D on variation in methylation at five DMPs detected with p<0.05 in at least one MR method. DMPs included as outcomes were detected in the sensitivity analysis of meta-EWAS of T2D (excluding KORA). IVW was regarded as main evidence of the 2SMR, while MR-Egger was a sensitivity analysis to account for the effect of horizontal pleiotropy. Associated at Bonferroni corrected p<1.0x10^{-3}.

			Sensitivity meta-EWAS of T2D			IVW		MR-Egger	
CpG	Chr	Gene	Beta (95%CI)	P	N	Beta (95%CI)	P	Beta (95%CI)	P
cg01577083	16	*Unannotated*	-0.011(-0.016,-0.006)	7.93E-06	3,428	-0.151(-0.281,-0.021)	**0.023**	-0.261(-0.535,0.013)	0.067
cg20456243	2	*SPEG*	-0.007(-0.011,-0.004)	9.99E-06	3,428	-0.066(-0.193,0.06)	0.304	-0.372(-0.625,-0.119)	**0.005**
cg20812370	1	*PBX1*	-0.007(-0.009,-0.004)	7.40E-07	3,428	-0.184(-0.301,-0.066)	**0.002**	-0.166(-0.413,0.082)	0.194
cg24686009	12	*RAP1B*	-0.002(-0.003,-0.001)	1.19E-06	3,428	0.004(-0.113,0.122)	0.942	-0.253(-0.500,-0.005)	0.050
cg26766064	17	*MIR657*	-0.007(-0.009,-0.004)	5.17E-06	3,428	-0.100(-0.217,0.018)	0.096	-0.187(-0.434,0.06)	0.143

Continuation Table 7-18 Additional MR methods

			Simple mode		Weighted median		Weighted Mode	
CpG	Chr	Gene	Beta (95%CI)	P	Beta (95%CI)	P	Beta (95%CI)	P
cg01577083	16	*Unannotated*	-0.218(-0.617,0.182)	0.290	-0.227(-0.412,-0.043)	**0.016**	-0.236(-0.449,-0.023)	**0.034**
cg20456243	2	*SPEG*	0.245(-0.209,0.699)	0.294	-0.255(-0.443,-0.067)	**0.008**	-0.284(-0.493,-0.074)	**0.010**
cg20812370	1	*PBX1*	-0.168(-0.49,0.153)	0.309	-0.218(-0.4,-0.037)	**0.018**	-0.205(-0.386,-0.023)	**0.031**
cg24686009	12	*RAP1B*	0.103(-0.261,0.466)	0.581	-0.117(-0.303,0.068)	0.215	-0.141(-0.348,0.066)	0.187
cg26766064	17	*MIR657*	-0.137(-0.569,0.296)	0.538	-0.230(-0.422,-0.037)	**0.019**	-0.229(-0.464,0.005)	0.060

Continuation Table 7-18 Diagnostic tests to assess heterogeneity (Cochran's Q in IVW regression) and pleiotropy (directionality-p in MR-Egger regression).

CpG	Chr	Gene	Cochran's Q	P_Q	Directionality P
cg01577083	16	*Unannotated*	74.826	0.110	0.375
cg20456243	2	*SPEG*	70.392	0.192	0.009
cg20812370	1	*PBX1*	52.923	0.760	0.872
cg24686009	12	*RAP1B*	44.888	0.939	0.024
cg26766064	17	*MIR657*	57.952	0.587	0.434

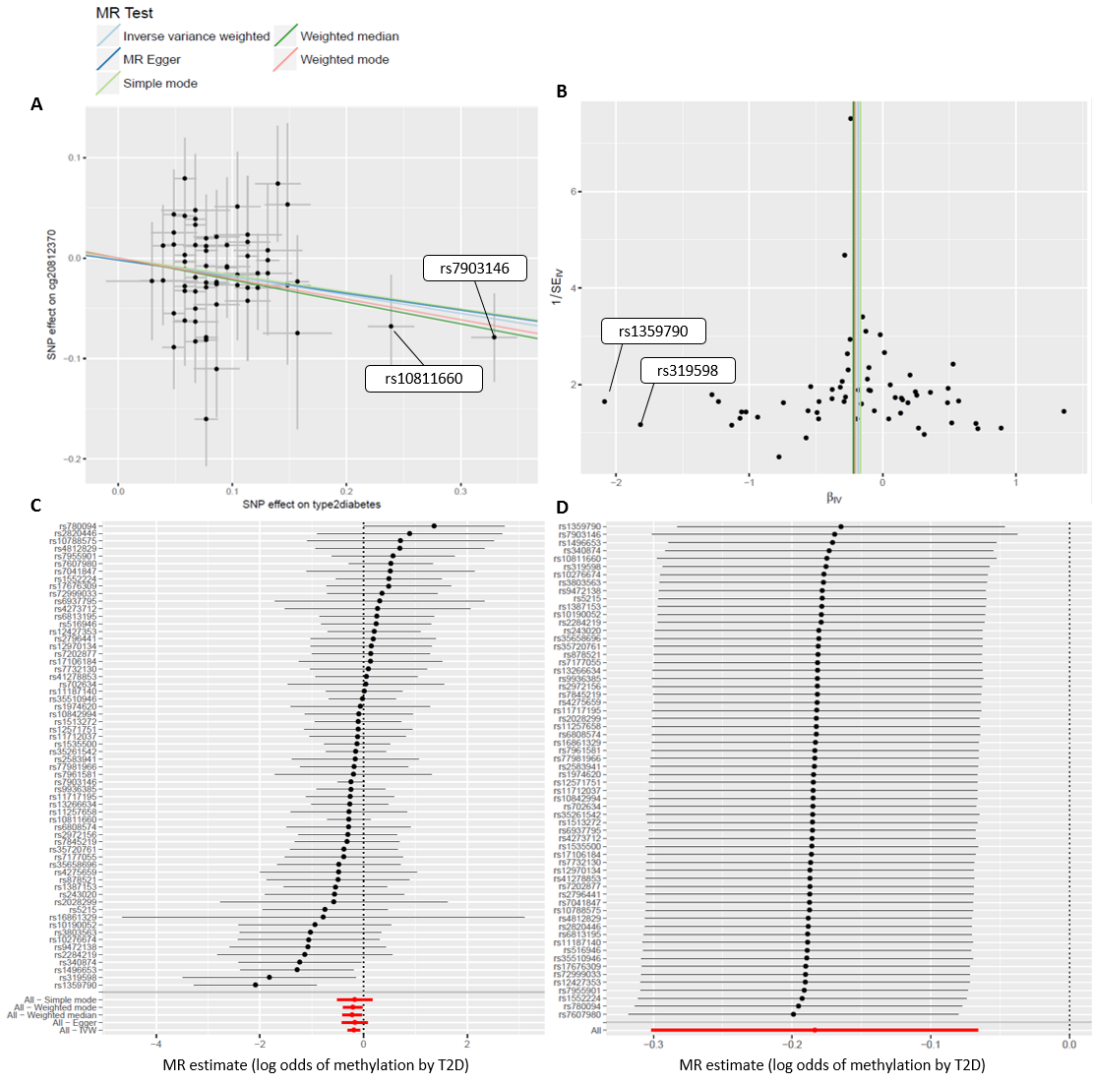


Figure 7-9 Mendelian randomization study for the effect of T2D on methylation at the DMP in PBX1 (cg20812370), detected observationally in the sensitivity meta-EWAS of T2D (excluding KORA). A) scatter plot depicting the effect of the SNPs as individual instruments against the exposure (x-axis) and the outcome (y-axis), and the combined effect of them as a single instrument across different methods, represented in the plot by the fitted regression lines. A negative slope common to all methods, suggested that T2D was negatively associated with inverse-normal transformed residuals of methylation at the DMP in PBX1. Highlighted in the plot were the SNPs rs10811660 and rs7903146 where suggestive heterogeneous effect was identified. B) Funnel plot showing the causal effect of each SNP individually (x-axis) against the inverse of the SE for the combined causal effect (y-axis). Some asymmetry in the funnel plot was identified due to the extreme negative effect of SNPs rs319598 and rs1359790, suggesting evidence of horizontal pleiotropy. D) forest plot illustrating the mean causal effect of each SNP on methylation depicted by the black point, and the horizontal line representing the 95% CI of this effect. At the bottom of the plot, is the combined (meta-analysed) causal effect across SNPs using one of five different methods. Effects crossing the vertical dashed line at "0", or the line of null associations, indicate no evidence of causality for the single SNP, or the combined causal effect. D) leave-one-out sensitivity analysis showing the total effect of T2D on methylation after sequentially excluding one SNP at a time from the analysis, this to identify how robust was the estimate to the effect of individual SNPs, and if the total causal effect was driven by a single instrument. For this analysis, none of the SNPs was driving alone the total causal estimate, and the combined effect using the IVW estimate suggested strong negative effect of T2D on inverse-normal transformed residuals of methylation at the DMP in PBX1.

For signals identified observationally in the EWAS of T2D in ALSPAC (n=11 DMPs), the 2SMR revealed suggestive evidence of causality at p<0.05 for the DMPs cg10870892 (*CTTN*) and cg07251197 (intergenic) (Table 7-19), but none of these associations surpassed Bonferroni correction at p<0.01 (α=0.05/11 DMPs). Stronger causal estimates were observed using the IVW and the weighted median regressions. The magnitude of the absolute effect estimate differed between MR methods, and opposite direction of effect was detected between the causal and the observational estimate for top DMPs detected in the EWAS of T2D in ALSPAC (Table 7-19). Diagnostic tests suggested some evidence of heterogeneity at the DMP cg07251197 (Cochran's Q=79.77, p=0.05), a finding that was visually corroborated by the presence of outlier SNPs in the scatterplot and in the leave-one-out analysis for this site (Figure 7-10). Overall, there was no evidence of directional horizontal pleiotropy based on results of the Egger intercept, even though some asymmetry was detected in the funnel plot for the DMP cg07251197 (Figure 7-10). The Steiger test indicated assessing the opposite direction to the true causal effect for the association at the DMP cg07251197 (Steiger p=1.91×10^{-4}).

7.6.4.3 Summary of findings in the forward 2SMR

Results of the 2SMR did not support causality in the association between T2D and difference in middle-age methylation based on DMPs detected in three observational analyses (Bonferroni corrected p-value 0.001 to 0.01). The Steiger test suggested that for 18 out of 84 DMPs in total analysed in the 2SMR, true direction of effect was from methylation to T2D, rather than the opposite. Despite this, it is important to consider that the Steiger test is susceptible to measurement error in the calculation of the SNP-exposure and SNP-outcome variation. Thus, to accurately establish direction of causality in the context of MR studies, it was necessary to perform a bidirectional 2SMR.

Overall, no evidence of heterogeneity was identified for the effect of 62 SNPs included as instruments in the 2SMR (heterogeneity p-value ranged 0.26 to 0.52). Looking at evidence of horizontal or unbalanced pleiotropy, the MR-Egger intercept suggested evidence of directional pleiotropy (Egger-intercept range -0.004 to 0.01), but this was not strong (directionality p range 0.68 to 0.99). An average I_{GX}^2 of 86.3% across analyses indicated no evidence of weak instrument bias in results of the 2SMR, a finding that was in line with results of the F-statistic (F-statistic range 19.72 to 274.8).

Table 7-19 Results of the 2SMR for the effect of T2D on variation in methylation at two DMPs detected with p<0.05 in at least one MR method. DMPs included as outcomes were identified in the EWAS of T2D in ALSPAC. IVW was regarded as the reference method, while the MR-Egger was a sensitivity analysis to account for the effect of horizontal pleiotropy. Associations were considered significant at p<0.01.

CpG	Chr	Gene	EWAS of T2D ALSPAC			IVW		MR-Egger	
			Beta (95%CI)	P	N	Beta (95%CI)	P	Beta (95%CI)	P
cg07251197	1	*Unannotated*	-0.008(-0.011, -0.004)	8.11E-06	1049	0.055(-0.08,0.189)	0.424	0.130(-0.155,0.414)	0.374
cg10870892	11	*CTTN*	-0.045(-0.062, -0.027)	1.13E-06	1047	0.127(0.009,0.244)	**0.035**	-0.015(-0.262,0.233)	0.908

Continuation Table 7-19 Additional MR methods.

CpG	Chr	Gene	Simple mode		Weighted median		Weighted Mode	
			Beta (95%CI)	P	Beta (95%CI)	P	Beta (95%CI)	P
cg07251197	1	*Unannotated*	0.027(-0.372,0.426)	0.894	0.190(0.008,0.371)	**0.041**	0.145(-0.069,0.359)	0.190
cg10870892	11	*CTTN*	0.021(-0.334,0.376)	0.907	0.010(-0.182,0.202)	0.921	0.012(-0.191,0.215)	0.908

Continuation Table 7-19 Diagnostic tests to assess heterogeneity (Cochran's Q in IVW regression) and pleiotropy (directionality-p in MR-Egger regression).

CpG	Chr	Gene	Cochran's Q	P_Q	Directionality P
cg07251197	1	*Unannotated*	**79.768**	**0.054**	0.559
cg10870892	11	*CTTN*	55.994	0.657	0.208

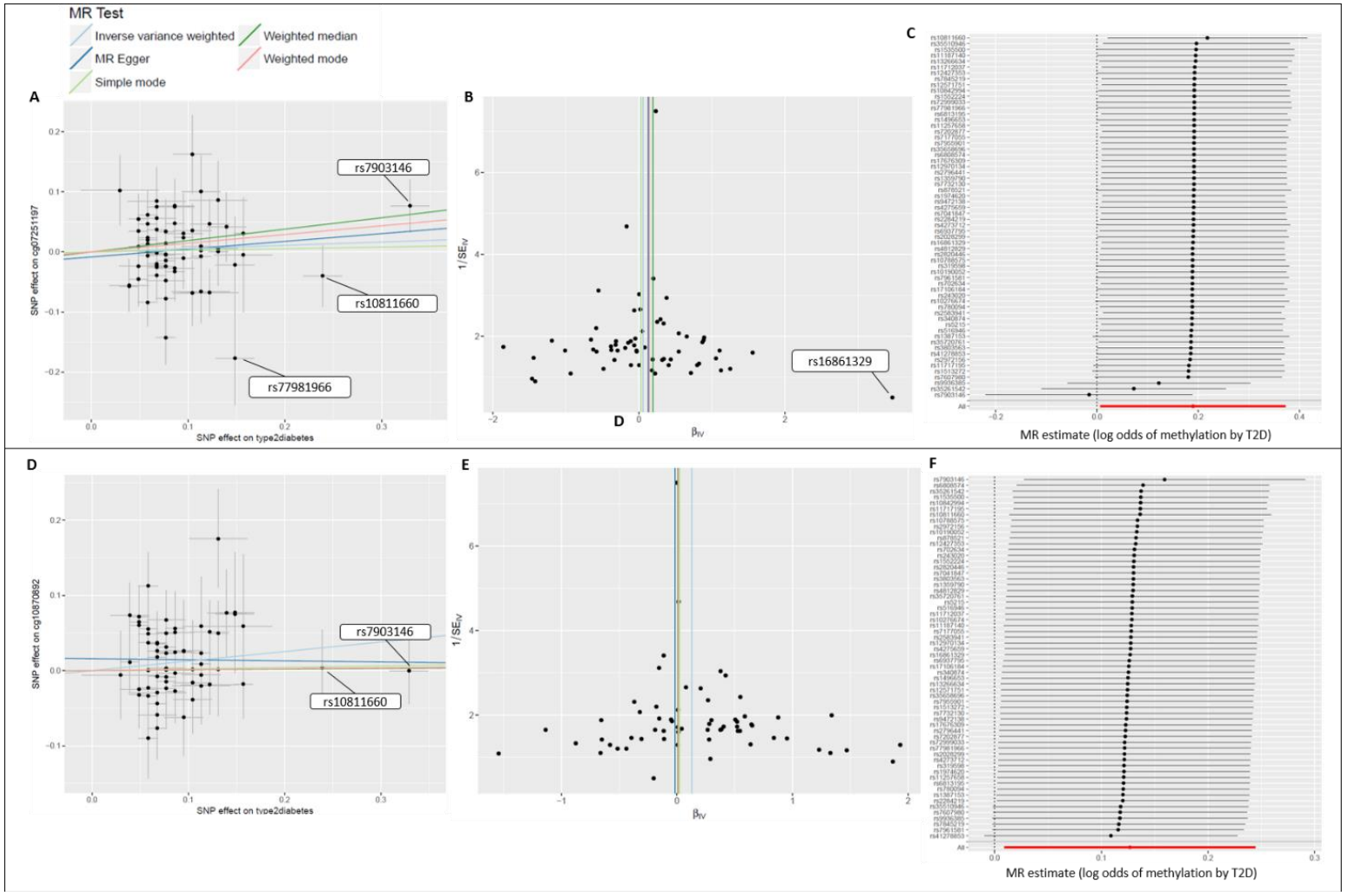


Figure 7-10 Mendelian randomization study for the effect of T2D on methylation at the DMPs cg07251197 and cg10870892 (CTTN). These DMPs were identified observationally in the EWAS of T2D in ALSPAC and were subsequently detected in suggestive association with T2D in the 2SMR according to estimates of the IVW and weighted median. A-C) scatterplot, funnel plot and leave-one-out analysis for the DMP cg07251197 (unannotated). D-F) similar plots for the DMP in CTTN.

7.6.5 Addressing SNP outliers in the forward 2SMR

For some of the strongest associations in the MR analysis, three SNPs were consistently identified as outliers based on a visual inspection of the scatter, forest and funnel plots, suggesting evidence for horizontal pleiotropy. Outlier SNPs were defined as variants with opposite effect on methylation relative to the effect observed for other variants, but that did not necessarily cause a complete change in the direction of the combined causal effect. Small influence of outliers on the total causal effect was demonstrated by the non-significant value of the Egger intercept, and the consistent results obtained in the leave-one-out analysis. In addition, the effect of outlier SNPs varied depending on the DMP where it was identified. The most common outlier SNP was the SNP rs16861329, identified in associations for the DMPs in *NISCH* (cg00082384), *TXNIP* (cg19693031), *C20orf26* (cg06039489), *MIR657* (cg26766064) and the intergenic DMP cg07251197. Two other outlier SNPs at rs1359790 and rs319598 were uniquely identified in results of the DMP cg20812370 in *PBX1*. No detectable sources of effect allele coding errors were seen when comparing outlier SNPs with similar risk variants reported in the GWAS catalog, or in the original studies. These SNPs were also disregarded as palindromic SNPs, which are commonly susceptible to allele coding errors in MR analyses.

To determine potential sources of horizontal pleiotropy, a PheWAS[6] SNP lookup was performed. This lookup revealed that only the outlier SNPs rs1359790 and rs319598 were associated with some other traits at Bonferroni corrected $p< 6.31x10^{-6}$ (α=0.05/7,931 'trait lookups')[6]. However, most of these associations were related with T2D outcomes, like T2D diagnosis, self-report of T2D and medication for T2D, except for three associations detected between the SNP rs1359790 and height (estimate=0.02, SE=0.003, $p=1.60x10^{-7}$) and sitting height (estimate=0.01, SE=0.002, $p=6.45x10^{-9}$) according to a GWAS conducted by Wood *et al*.[24], and between SNP rs1359790 and standing height (estimate=0.02, $p=1.40x10^{-8}$) in a GWAS reported by the UK Biobank (unpublished data, http://www.nealelab.is/uk-biobank/). To determine if height, either sitting or standing, was a mediator of the association between SNP rs1359790 and methylation at the DMP in *PBX1*, an MR analysis was attempted to measure the effect of height on methylation at *PBX1*, but no meQTL was available for this site either in the meQTL database (http://www.mqtldb.org/)[11], or in data from the GoDMC consortium (requested data). Thus, it was not possible to determine if the heterogeneous effect of the SNP rs1359790 on methylation at the DMP in *PBX1* was influenced by height. Because no other outcomes were strongly associated with the outlier SNPs rs1359790 and rs319598 according to the PheWAS lookup, it was possible that the heterogenous effect of these SNPs on

methylation was a true effect mediated by T2D, rather than by the pleiotropic effect of other outcomes.

7.6.6 Forward 2SMR using GoDMC data

LD correlation calculated between T2D-SNPs and GoDMC SNPs strongly associated with DMPs of interest, ranged between 0.0 and 0.2. Because none of the above pairwise comparisons surpassed the cut-off LD r^2 of 0.6 to indicate high correlation among SNPs, none of the SNPs reported in the GoDMC consortium was identified as a good genetic proxy for a T2D SNP. Thus, the forward 2SMR was not conducted using GoDMC data, and main results were reported based on summary statistics of the IV-outcome association obtained in ALSPAC.

7.6.7 Power in the forward 2SMR

The power calculator suggested that a 100% power was obtained to confidently detect (at $p<0.05$) an absolute causal effect of 0.16 (i.e. mean absolute effect across MR methods for four DMPs detected with suggestive evidence of causality) in the association between T2D and inverse-normal transformed residuals of methylation. On the other hand, the minimum sample required to detect a similar causal effect with 95% power and $p<0.05$, were 809 participants.

7.7 Reverse 2SMR

To unveil true direction of causality in the association between T2D and methylation, the opposite direction of the association was investigated, where methylation was regarded as the exposure (multiple DMPs) and T2D as the outcome. As in the forward MR, strong instruments were required to proxy levels of methylation for DMPs detected in the observational analysis. The p-threshold used to include meQTL in the analysis was a $p<1.0x10^{-5}$, which is less stringent than the p-threshold used by the GoDMC consortium to report *cis* ($p<1.0x10^{-8}$) and *trans* ($p<1.0x10^{-14}$) associations. A less stringent p-threshold for meQTL selection was based on the less comprehensive size of the epigenome interrogated in this analysis (i.e. only top DMPs detected observationally).

7.7.1 Selecting instruments for methylation

Initially, 75 unique meQTL were available as instruments for 53 of the 84 top DMPs identified across three observational analyses (i.e. meta-EWAS, sensitivity meta-EWAS, and EWAS of T2D in ALSPAC). None of these meQTL overlapped with a T2D-SNP. Of these 75 meQTL, 15 were excluded based on their correlation with a T2D-SNP at LD $r^2 > 0.01$ (MR-Base LD threshold). After pruning, 60 meQTL (54 in *cis* and six in *trans*) remained as instruments for 41/84 top DMPs in the MR analysis. These 60

meQTL were identified as proxies for methylation at four DMPs detected in the EWAS of T2D in ALSPAC, 12 DMPs detected in the meta-EWAS of T2D, and 25 DMPs detected in the sensitivity meta-EWAS of T2D. Regarding the number of meQTL available per DMP, 29/41 DMPs were instrumented by a single meQTL, 14/41 DMPs had two meQTL, and only one DMP (cg25536676 in *DHCR24*) was instrumented by three meQTL. From the available instruments, the strongest association was identified in *cis* between rs6681644 and cg25536676 (CpG gene *DHCR24*, $p=3.5x10^{-202}$), while the weakest association was identified in *cis* between rs34345524 and cg08273233 (CpG gene *HTR1E*, $p=2.0x10^{-5}$). Table 7-20 compares between two datasets of instruments, one excluding meQTL correlated with T2D-SNPs at LD > 0.01, and a second dataset without pruning. Table 7-21 to Table 7-23 show summary statistics of the observed IV-exposure (SNP~DMP) association for top DMPs that were successfully instrumented by a GoDMC meQTL.

Table 7-20 Comparison of two datasets of instruments available for the reverse MR. The complete dataset includes all meQTL initially reported by GoDMC in association with DMPs of interest at $p<1.0x10^{-5}$. Pruned refers to the dataset instruments excluding meQTL identified in correlation with T2D-SNPs at LD>0.01.

Dataset	meQTL	Unique SNPs	Unique DMPs	Cis/trans	Min # meQTL/DMP	Max # meQTL/DMP	Min P-value	Max P-value
Pruned	60	60	44	54/6	1 (29 DMPs)	3 (1 DMP)	3.54E-202	2.02E-05
Complete	75	75	53	69/6	1 (33 DMPs)	3 (1 DMP)	3.54E-202	2.02E-05

Important to be mentioned is that some of the associations analysed in the forward 2SMR and identified with borderline evidence of causality (see section 7.6.4), were not included in the reverse 2SMR for lack of valid instruments for these DMPs based on data reported by the GoDMC consortium. The following sections describe genotype-exposure and genotype-outcome associations remaining in the reverse 2SMR after applying QC on MR-Base.

Table 7-21 Summary statistics of 17 SNP-CpG associations identified by GoDMC for 12 of the top DMPs detected in the meta-EWAS of T2D (total n=25 DMPs). Estimate: unit change in inverse-normal transformed residuals of methylation per allele increase in the genotype, EA: effect allele (minor allele), OA: major allele, EAF: effect allele frequency, Cis: in cis if the SNP was within 1Mb from the position of the CpG, and in trans if the SNP was > 1Mb from the position of the CpG. Associations were regarded significant at $p<1.0x10^{-5}$

CpG	SNP	Estimate	SE	EA	OA	EAF	P	N	CpG gene	Cis
cg00082384	rs11716756	-0.24	0.01	T	C	0.14	3.54E-202	21,355	*NISCH*	T
cg01317029	rs35668024	-1.04	0.03	A	C	0.97	3.54E-202	25,970	*FAM131A*	T
cg06468695	rs140180165	-0.39	0.03	T	C	0.97	3.54E-202	15,366	*CCDC42*	T
cg06468695	rs72848116	0.22	0.02	T	C	0.94	3.54E-202	16,580	*CCDC42*	T
cg06500161	rs220182	0.06	0.01	T	C	0.55	3.54E-202	24,474	*ABCG1*	T
cg11851382	rs7535757	-0.08	0.01	A	G	0.50	3.54E-202	26,658	*PPAP2B*	T
cg19693031	rs6657798	-0.46	0.01	C	G	0.80	3.54E-202	27,212	*TXNIP*	F
cg25741837	rs62148128	0.71	0.02	A	G	0.06	3.54E-202	20,444	*SMYD5*	T
cg27237541	rs35885100 [a]	-0.09	0.01	D	I	0.56	3.54E-202	16,768	*MYO3A*	T
cg27237541	rs150804707 [a,b]	-0.24	0.01	D	I	0.73	8.93E-121	24,267	*MYO3A*	T
cg25741837	rs6732515	0.54	0.03	A	C	0.98	1.08E-64	22,690	*SMYD5*	T
cg07184465	rs1500138	-0.15	0.01	T	C	0.35	1.78E-58	25,936	*SPZ1*	T
cg01317029	rs28421035	0.22	0.01	T	C	0.09	2.42E-53	25,851	*FAM131A*	F
cg17155612	rs56293553	-0.17	0.01	A	G	0.81	4.67E-51	26,699	*LOC148189*	T
cg06039489	rs6081870	-0.14	0.01	A	G	0.27	7.27E-44	24,388	*C20orf26*	T
cg00082384	rs35911561	0.12	0.02	T	C	0.89	6.63E-16	22,455	*NISCH*	T
cg08273233	rs34345524 [a]	-0.05	0.01	D	I	0.44	2.02E-05	16,750	*HTR1E*	T

[a] Indel SNPs removed from the MR analysis. [b] SNP identified in correlation (LD>0.01) with a nearby SNP and was subsequently removed from further analyses (after applying clumping on MR-Base).

Table 7-22 Summary statistics of 34 SNP-CpG associations identified by GoDMC for25 of the top DMPs detected in the sensitivity meta-EWAS of T2D (total n=58 DMPs). Associations were regarded significant at $p<1.0x10^{-5}$.

CpG	SNP	Estimate	SE	EA	OA	EAF	P	N	CpG gene	Cis
cg00144180	rs11693641	-0.15	0.01	A	C	0.50	3.54E-202	23,360	*HDAC4*	T
cg00896068	rs113786621	-0.34	0.02	T	C	0.08	3.54E-202	25,956	*Unannotated*	T
cg07212837	rs56261297	-0.34	0.01	T	C	0.41	3.54E-202	27,738	*Unannotated*	T
cg10584271	rs115738369	-1.74	0.03	T	C	0.02	3.54E-202	22,070	*ITIH1*	T
cg11024682	rs11652574	1.10	0.03	A	G	0.04	3.54E-202	19,085	*SREBF1*	T
cg12593793	rs11584621	-0.06	0.01	A	T	0.21	3.54E-202	25,084	*Unannotated*	T
cg14476101	rs347903	0.23	0.01	T	C	0.67	3.54E-202	24,554	*PHGDH*	T
cg20231084	rs750129	0.14	0.01	A	G	0.47	3.54E-202	23,944	*Unannotated*	T
cg20456243	rs55760516	-0.12	0.01	A	G	0.67	3.54E-202	27,242	*SPEG*	T
cg24512093	rs9309801	-0.11	0.01	T	C	0.34	3.54E-202	27,235	*ROBO1*	T
cg24512093	rs9831014	0.12	0.01	C	G	0.42	3.54E-202	24,994	*ROBO1*	T
cg25536676	rs6681644	0.26	0.01	C	G	0.42	3.54E-202	27,714	*DHCR24*	T
cg27037013	rs13051329	0.17	0.01	T	C	0.15	3.54E-202	26,837	*Unannotated*	T
cg27037013	rs9976794	-0.07	0.01	A	T	0.53	3.54E-202	26,466	*Unannotated*	T
cg27115863	rs6000773	-0.12	0.01	C	G	0.75	3.54E-202	24,389	*Unannotated*	T
cg16192197	rs9487736	0.36	0.01	A	G	0.14	6.28E-188	27,726	*Unannotated*	T
cg14476101	rs608358	-0.28	0.01	A	C	0.28	5.82E-180	25,566	*PHGDH*	T
cg10082515	rs1525502	-0.25	0.01	T	C	0.41	1.05E-172	24,607	*Unannotated*	T
cg16765088	rs7496161	-0.32	0.01	A	G	0.11	4.70E-122	25,984	*Unannotated*	T
cg01577083	rs1107095	-0.20	0.01	T	C	0.51	7.48E-109	23,764	*Unannotated*	T
cg27115863	rs7602568	0.20	0.01	T	C	0.22	1.01E-85	27,625	*Unannotated*	F
cg18181703	rs4383852	-0.13	0.01	A	G	0.52	2.34E-56	27,746	*SOCS3*	F
cg01963618	rs6596785	0.22	0.01	A	G	0.89	2.23E-50	25,226	*LOC285768*	T
cg08857797	rs1047891	-0.14	0.01	A	C	0.32	8.79E-47	24,138	*VPS25*	F
cg00896068	rs9525281	0.14	0.01	C	G	0.76	3.20E-30	18,956	*Unannotated*	T
cg25536676	rs174551	0.11	0.01	T	C	0.66	5.11E-30	24,653	*DHCR24*	F
cg11252555	rs10421294	-0.14	0.01	A	G	0.10	2.79E-20	25,481	*RPL13AP5*	T
cg13178597	rs540908	-0.10	0.01	A	G	0.82	6.42E-18	26,442	*RGS17*	T
cg11376147	rs2848634	0.09	0.01	A	G	0.75	7.93E-18	27,749	*SLC43A1*	T
cg09185884	rs71380866	0.21	0.03	C	G	0.97	8.70E-15	25,599	*KCTD2*	T
cg11983038	rs74623153	-0.37	0.05	A	G	0.98	5.84E-12	11,047	*Unannotated*	T
cg10584271	rs62250760	0.05	0.01	T	C	0.35	9.14E-08	25,857	*ITIH1*	T
cg25536676	rs79365581	0.35	0.07	T	C	0.98	7.84E-07	5,834	*DHCR24*	T
cg00144180	rs1872614	-0.05	0.01	A	T	0.58	2.22E-06	16,512	*HDAC4*	T

Table 7-23 Summary statistics of five SNP-CpG associations identified in GoDMC for four of the top DMPs detected in the EWAS of T2D in ALSPAC (total n=11 DMPs). Associations were regarded significant at $P<1.0x10^{-5}$.

CpG	SNP	Estimate	SE	EA	OA	EAF	P	N	CpG gene	Cis
cg00204249	rs72903323	-0.15	0.01	A	G	0.17	8.81E-38	26,469	*DNAH17*	T
cg14045803	rs4459332	0.19	0.01	T	C	0.53	1.55E-102	26,453	*STARD10*	T
cg15986668	rs3767953	-0.25	0.02	C	G	0.90	1.32E-46	18,890	*NFYC*	T
cg15986668	rs115582802	-0.19	0.02	T	C	0.10	3.54E-202	21,291	*NFYC*	T
cg26652413	rs773865	0.43	0.01	A	G	0.24	3.54E-202	23,231	*CPAMD8*	T

7.7.2 QC pruning of the genotype-exposure dataset on MR-BASE

Of the 17 SNPs initially available as instruments for 12 top DMPs detected in the meta-EWAS of T2D, three SNPs were removed after QC on MR-Base: two SNPs were excluded based on allele coding errors (i.e. indel SNPs in rs35885100 and rs34345524), and one more was exclude after clumping (SNP rs150804707 with LD > 0.01), leaving in total 14 SNP-CpG associations (14 SNPs and 10 DMPs) for further analyses. Of these, 12/14 associations were *cis* meQTL, and the remaining two were *trans* meQTL. None of the meQTL reported for DMPs identified in the sensitivity meta-EWAS of T2D and in the EWAS of T2D in ALSPAC, were excluded after QC on MR-Base. In total, 34 SNP-CpG associations in the sensitivity meta-EWAS (34 SNPs and 25 DMPs) and five SNP-CpG associations in the EWAS of T2D in ALSPAC (five SNPs and 4 DMPs), remained to conduct the reverse 2SMR. Most of these associations were identified in *cis*, except for four meQTL in *trans* associated with DMPs detected in the sensitivity meta-EWAS.

7.7.3 Selection of studies to extract genotype summary data for T2D

The study conducted by Mahajan *et al*.[20] was selected to extract summary data for the association between the genotype and T2D. This study is part of the DIAGRAM consortium, was composed of 110,452 samples of mixed race (26,488 cases and 83,964 controls) without evidence of comorbidities, and association statistics were available for approximately 2.9 million SNPs. Some other studies in DIAGRAM were available, but most of them included a smaller number of variants (< 2.4 million SNPs) than those considered by Mahajan *et al*.[20], even though they accounted with a larger proportion of T2D cases (i.e. 34,840 cases in Morris *et al*.[22] and 27,206 cases in Gaulton *et al*.[21] versus 26,488 cases in Mahajan *et al*.[20]). To avoid the overlap of samples between DIAGRAM studies, only the study conducted by Mahajan *et al*.[20] was taken forward for the MR analysis.

A second GWAS meta-analysis of T2D was conducted by Wood *et al.*[235] using samples of the UK-Biobank. This study was categorized as of high priority by MR-Base based on the number of samples included (120,286 participants of British ancestry), the number of SNPs with available association estimates (approximately 8.4 million SNPs), the year of publication (2016), and the accessibility to

summary data, despite the small proportion of cases included (4,040 T2D cases and 116,246 controls).

7.7.4 Genotype versus T2D associations retained for the MR analysis after data harmonization

Table 7-24 summarizes the number of genotype-T2D associations successfully extracted and harmonized on MR-Base, based on SNPs identified as instruments in the genotype-exposure dataset. Some of the genotype-outcome associations were duplicated considering the two studies selected to extract summary data for the outcome, while for other associations the effect of the SNP on T2D was proxied by another SNP in high correlation (LD > 0.8) with the target SNP. In general, after data harmonization, each DMP was instrumented by at least one SNP, and the maximum number of SNPs per DMP was three. No strong association was identified between meQTL SNPs and T2D (p range 0.02 to 1.00). appendix Table S8-40 shows the harmonized genotype-exposure and genotype-outcome datasets used in the reverse 2SMR analysis.

Table 7-24 Summary of genotype-outcome (T2D) data extracted on MR-Base based on SNPs included as instruments in three datasets of exposures (top DMPs). Summary data was extracted from two GWAS meta-analyses on T2D reported by Mahajan et al. 2014 and Wood et al. 2016.

Dataset	SNPs Exposure	Exposures (DMPs)	Extracted SNPs†	Remaining DMPs	Proxied SNPs‡	Duplicated G-O associations‡‡	P range††
Meta-EWAS	14	10	13	10	3	9	0.07-1.00
Sensitivity meta-EWAS	34	25	32	24	11	18	0.03-0.95
EWAS ALSPAC	5	4	5	4	4	4	0.02-0.29

†Instruments successfully identified in the outcome dataset. ‡ Number of target SNPs with summary data extracted from another SNP in high correlation identified in the genotype-outcome dataset. ‡‡Duplicated genotype-outcome associations obtained by using two studies to extract the outcome data. ††Minimum and maximum p-value reported for the genotype-T2D association.

7.7.5 Results of the reverse 2SMR

By conducting a reverse 2SMR, true direction of effect was determined for signals that were successfully analysed in both directions of the T2D~methylation association. Because only a small number of instruments were available per DMP in the reverse MR (i.e. maximum three SNPs per DMP), the number of sensitivity analyses that could be applied to detect horizontal pleiotropy and heterogeneity, was limited. As before, an approximation to true direction of causality was provided using the Steiger test, while strength of the instrument was measured using the F-statistic.

7.7.5.1 Increased methylation at the DMP cg11851382 in PPAP2B was strongly associated with a protective effect on T2D

The strongest causal association was identified between methylation at the DMP cg11851382 in *PPA2PB* and T2D (see Figure 7-11,Table 7-25). Results showed that per increase in inverse-normal transformed residuals of methylation at *PPAP2B,* was associated with an average 71% reduced risk of T2D (95% CI= 33%, 88% reduced risk, p=0.004), and this result was significant at Bonferroni corrected p<0.01 (α=0.05/10 DMPs). Furthermore, nominal evidence of causality was identified at the DMPs cg07184465 (*SPZ1,* p=0.07) and cg17155612 (*LOC148189,* p=0.07), and in both cases increased methylation was suggestively associated with higher risk of T2D (see Table 7-25). None of these top three associations overlapped with top signals detected in the forward MR, indicating that difference in methylation at these DMPs could reliably represent a cause of T2D.

Consistency in the direction of effect between the observed and the causal estimate was identified for the DMP in *PPAP2B.* In the observational analysis it was demonstrated that cases of T2D were on average hypomethylated at *PPAP2B* compared to controls (estimate=-0.01, 95% CI= -0.009, -0.003), and in the causal analysis it was shown that hypermethylation of *PPAP2B* was protective against the risk of T2D. Similarly, consistent direction of effect between estimates was detected for the DMPs in *FAM131A*, *CCDC42* and *SMYD5* (Table 7-25), but not for the DMPs in *SPZ1, LOC148189*, *ABCG1* and *C20orf26*. For the DMP in *TXNIP,* it was not possible to compare direction of effect between the observational and the causal estimate as a null effect was detected for this DMP in the reverse MR (OR=1.0, 95%CI=0.79,1.27).

Overall, the Steiger test suggested that direction of causality was correctly assessed for the ten exposure-outcome associations included in the reverse MR (average Steiger p=2.18×10^{-13}). True direction of effect reported by the Steiger test was the result of including instruments that explained a higher proportion of variance in the exposure (methylation R^2 range 0.01 to 0.06), compared to the proportion of variance explained in the outcome (T2D R^2 range 1.4×10^{-8} to 1.4×10^{-4}). Results of the F-statistic (F-statistic range 65.34 to 290.30) suggested less probability of obtaining biased results due to weak instruments in the reverse MR.

In terms of other sensitivity analyses, it was not possible to investigate the effect of directional pleiotropy using the MR-Egger intercept because most of the DMPs analysed in the reverse MR were instrumented by a single SNP. This was also true when trying to apply other methods that allow to relax the assumption of having valid instruments when conducting MR analyses (i.e. weighted

median and weighted/simple mode methods). Heterogeneity of effect using the Cochran's Q test was reported for 3/10 predicted exposure-outcome associations, and there was no evidence of heterogeneity for the effect of SNPs used as instruments for the DMPs cg00082384 (*NISCH*), cg01317029 (*FAM131A*) and cg25741837 (*SMYD5*) (see Table 7-25). In general, inspectional plots were not available for most of the associations, except for those at the DMPs in *NISCH*, *FAM131A* and *SMYD5,* where basic scatter and forest plots were used to illustrate the effect of two different instruments on T2D (appendix Figure S8-26). Similarly, the leave-one-out analysis was not supported for associations inspected in the reverse MR.

When comparing estimates of the reverse and the forward MR for DMPs analysed in both directions of the association, it was demonstrated that the strength of the effect was higher in the forward compared to the reverse MR for the DMPs in *NISCH*, *C20orf26* and *TXNIP*. This finding confirmed results of the forward 2SMR and established T2D as the true hypothesized exposure in these associations

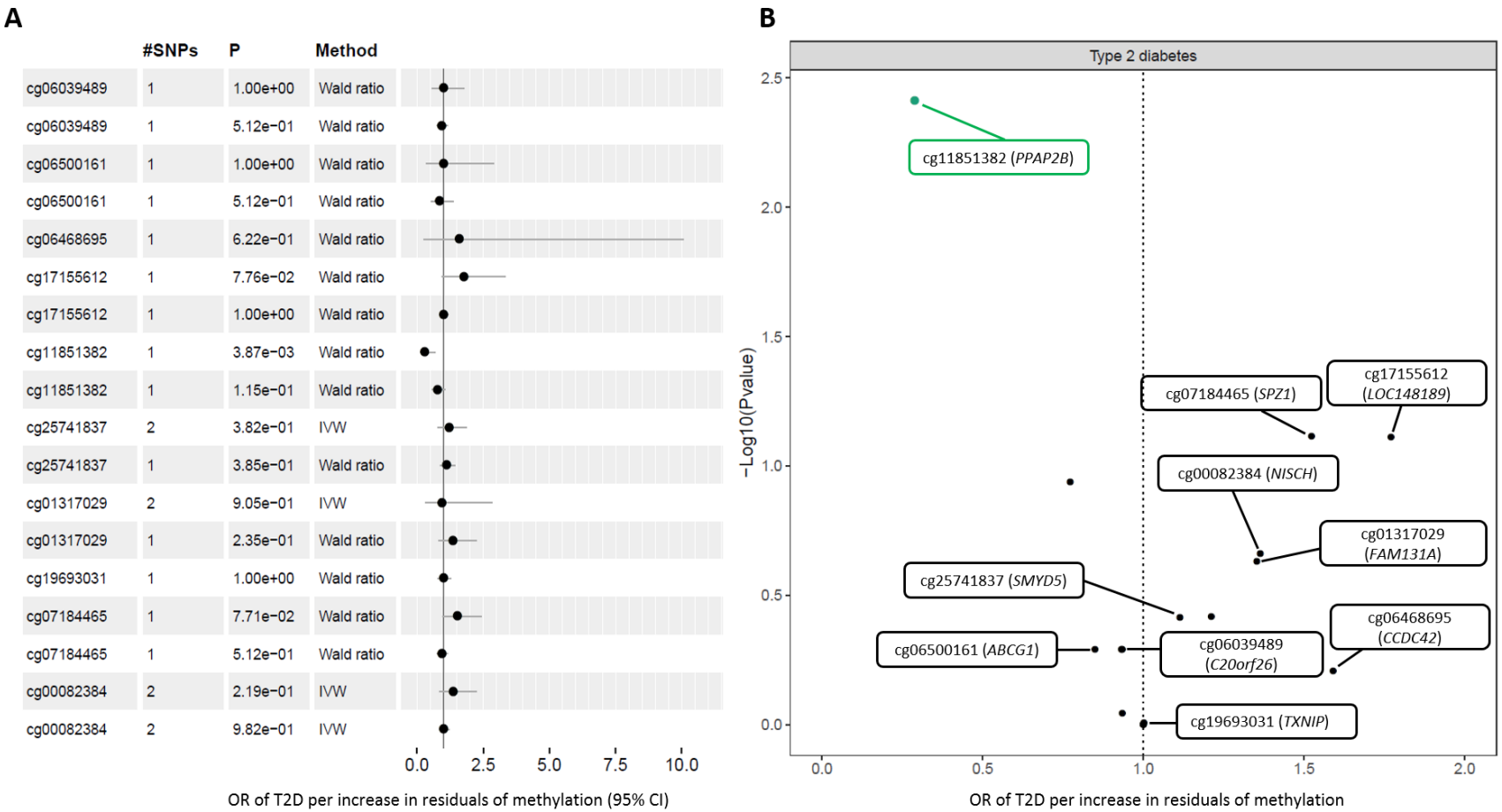


Figure 7-11 Forest plot (A) and volcano plot (B) summarizing results of the reverse 2SMR for the effect of methylation (exposure) on T2D (outcome), using DMPs detected in the meta-EWAS of T2D. Causal estimates were generated using the Wald estimate and the IVW regression. For some of the DMPs, results of the MR were duplicated based on the availability of genotype-outcome data for the same instrument(s) in two independent studies: Mahajan et al. 2014 and Wood et al. 2016. A) black dots represent the total causal estimate, and the 95% CI. Results are shown in odds ratios. B) Volcano plot illustrates the distribution of effect estimates (x-axis) against the -Log10(P-value) (y-axis). Points sitting over the dashed line indicate null associations. Results were considered borderline significant at p <0.05 [-log10(P-value) > 1.30], and Bonferroni significant at p< 0.01 [-Log10(P-value)> 2.0]. Highlighted in green is the strongest association detected at the DMP cg11851382 (PPAP2B).

Table 7-25 Results of the reverse 2SMR for the effect of methylation on the risk of T2D. DMPs included were identified observationally in the meta-EWAS of T2D, and were successfully instrumented by an meQTL reported in GoDMC at p< 1.0x10^{-5}. Summary data for the genotype-T2D association was extracted from two GWAS on T2D reported by Mahajan et al. 2014 and Wood et al. 2016. Associations in the reverse MR were considered significant at p<0.01 after Bonferroni correction (α=0.05/10 DMPs). In bold is the association detected with adjusted-p<0.01.

					Reverse MR (Wald Ratio)		Forward MR (IVW)		Meta-EWAS of T2D		
Exposure	Chr	Gene	Outcome	Study	OR (95%CI)	P	Estimate (95%CI)	P	Estimate(95%CI)	P	N
cg01317029*	3	*FAM131A*	Type 2 diabetes	Mahajan *et al.* 2014	1.35(0.82,2.23)	0.235	0.01(-0.11,0.12)	0.918	0.006(0.003,0.008)	9.48E-06	5,147
cg06039489	20	*C20orf26*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.56,1.79)	1.000	-0.11(-0.22,0.01)	0.080	0.014(0.008,0.019)	6.27E-06	5,147
cg06039489	20	*C20orf26*	Type 2 diabetes	Mahajan *et al.* 2014	0.93(0.76,1.15)	0.512	-0.11(-0.22,0.01)	0.080	0.014(0.008,0.019)	6.27E-06	5,147
cg06468695	17	*CCDC42*	Type 2 diabetes	Wood *et al.* 2016	1.59(0.25,10.08)	0.622	0.06(-0.09,0.21)	0.435	0.005(0.003,0.007)	6.19E-06	5,147
cg06500161	21	*ABCG1*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.35,2.90)	1.000	-0.07(-0.18,0.05)	0.263	0.007(0.004,0.01)	3.30E-06	5,147
cg06500161	21	*ABCG1*	Type 2 diabetes	Mahajan *et al.* 2014	0.85(0.52,1.38)	0.512	-0.07(-0.18,0.05)	0.263	0.007(0.004,0.01)	3.30E-06	5,147
cg07184465	**5**	***SPZ1***	Type 2 diabetes	Wood *et al.* 2016	**1.52(0.96,2.43)**	**0.077**	-0.05(-0.17,0.07)	0.403	-0.005(-0.008,-0.003)	8.27E-06	5,147
cg07184465	5	*SPZ1*	Type 2 diabetes	Mahajan *et al.* 2014	0.93(0.76,1.14)	0.512	-0.05(-0.17,0.07)	0.403	-0.005(-0.008,-0.003)	8.27E-06	5,147
cg11851382	**1**	***PPAP2B***	Type 2 diabetes	Wood *et al.* 2016	**0.29(0.12,0.67)**	**0.004**	0.03(-0.09,0.15)	0.603	-0.006(-0.009,-0.003)	8.81E-06	5,147
cg11851382	1	*PPAP2B*	Type 2 diabetes	Mahajan *et al.* 2014	0.77(0.56,1.07)	0.115	0.03(-0.09,0.15)	0.603	-0.006(-0.009,-0.003)	8.81E-06	5,147
cg17155612	**19**	***LOC148189***	Type 2 diabetes	Wood *et al.* 2016	**1.77(0.94,3.34)**	**0.078**	0.06(-0.06,0.17)	0.358	-0.002(-0.003,-0.001)	9.55E-06	5,147
cg17155612	19	*LOC148189*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.84,1.20)	1.000	0.06(-0.06,0.17)	0.358	-0.002(-0.003,-0.001)	9.55E-06	5,147
cg19693031	1	*TXNIP*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.79,1.27)	1.000	-0.04(-0.15,0.08)	0.555	-0.013(-0.017,-0.008)	4.26E-09	5,147
cg25741837*	2	*SMYD5*	Type 2 diabetes	Mahajan *et al.* 2014	1.11(0.87,1.42)	0.385	0.04(-0.07,0.16)	0.474	0.008(0.005,0.011)	4.44E-06	5,147

*DMPs instrumented by two SNPs.

Continuation Table 7-25 Results of the reverse MR for DMPs proxied by more than one instrument, where results of the reverse MR were obtained using the IVW regression.

					Reverse MR (IVW)				Forward MR (IVW)		Meta-EWAS of T2D	
Exposure	Chr	Gene	Outcome	Study	OR (95%CI)	P	Q†	P_Q	Estimate (95%CI)	P	Estimate(95%CI)	P
cg00082384	3	*NISCH*	Type 2 diabetes	Wood *et al.*2016	1.36(0.83,2.24)	0.219	1.17	0.279	-0.18(-0.3,-0.06)	0.003	0.008(0.005,0.012)	2.86E-06
cg00082384	3	*NISCH*	Type 2 diabetes	Mahajan *et al.*2014	1.00(0.84,1.20)	0.982	0.02	0.889	-0.18(-0.3,-0.06)	0.003	0.008(0.005,0.012)	2.86E-06
cg01317029	3	*FAM131A*	Type 2 diabetes	Wood *et al.*2016	0.93(0.31,2.83)	0.905	2.99	0.084	0.01(-0.11,0.12)	0.918	0.006(0.003,0.008)	9.48E-06
cg25741837	2	*SMYD5*	Type 2 diabetes	Wood *et al.*2016	1.21(0.79,1.86)	0.382	0.41	0.521	0.04(-0.07,0.16)	0.474	0.008(0.005,0.011)	4.44E-06

†Cochran's Q estimate.

7.7.5.2 Reverse 2SMR for DMPs identified in two additional observational analyses

Considering DMPs identified observationally in the sensitivity meta-EWAS of T2D, results of the reverse MR suggested five DMPs with borderline evidence of causality (at $p<0.05$): cg25536676 (*DHCR24*), cg20456243 (*SPEG*) and the intergenic DMPs cg10082515, cg16765088, and cg07212837, but none of them surpassed Bonferroni correction at $p<0.002$ ($\alpha=0.05/24$ DMPs) (Figure 7-12). Of these markers, the strongest association was detected at the DMP cg10082515 ($p=0.005$), and for most of these DMPs hypermethylation was associated with higher risk of T2D, except for the association at the DMP cg10082515, where hypermethylation was protective against T2D risk (see Table 7-26).

Consistency in the direction of effect between the causal and the observed estimate was observed for the DMPs cg07212837 and cg10082515, while opposite direction of effect was identified at the DMP cg16765088, and at the DMPs in *SPEG* and *DHCR24.* None of the top associations (smallest p-value) identified in the reverse MR overlapped with top signals of the forward MR, except for the association at the DMP in *SPEG*, which causal estimate was later demonstrated to be stronger in the forward compared to the reverse MR. Overall, results of the Steiger test suggested correct assessment of the direction of the causal effect for associations in the reverse MR, this considering that the instruments explained a higher proportion of variance in the exposure (methylation R^2 range 0.001 to 0.05), relative to the variance explained in the outcome (T2D R^2 range $3.4x10^{-8}$ to $5.7x10^{-5}$). In addition, values of the F-statistic (F-statistic range 22.41 to 496.40) suggested less probability of obtaining biased results due to weak instruments.

The Wald and the IVW regression were the main methods applied to obtain results of the reverse 2SMR. Other sensitivity methods were reported for the association at the DMP in *DHCR24,* where three SNPs were available as proxies. Estimates obtained for this DMP were consistent across methods (see appendix Table S8-41). For 7/24 associations where the heterogeneity of effect was measured, the Cochran's Q estimate suggested no evidence of heterogeneity (Q range 0.13 to 4.74, p range 0.03 to 1.00), while there was some evidence of negative horizontal pleiotropy for the DMP in *DHCR24*, but this was not strong (Egger intercept=-0.08, p=0.48). Simple scatter and forest plots were used to show the predicted effect of methylation on T2D for DMPs that were instrumented by at least two meQTL (appendix Figure S8-27). Due to the small number of instruments available per DMP, the leave-one-out analysis was not applied.

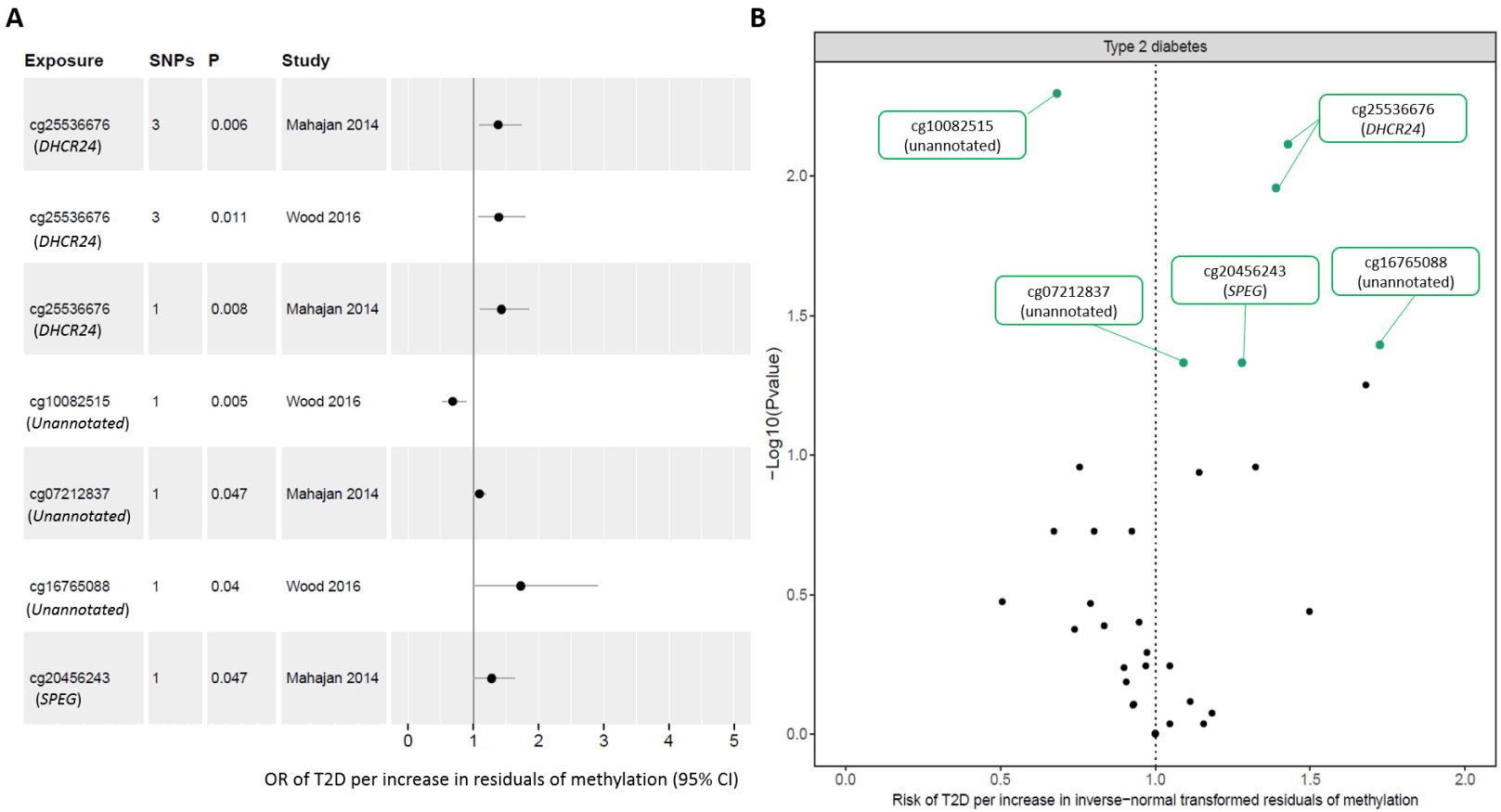


Figure 7-12 Forest plot (A) and Volcano plot (B) for the effect of methylation on T2D risk for DMPs detected in the sensitivity meta-EWAS of T2D (n=24 DMPs). DMPs included in the reverse 2SMR were successfully instrumented by an meQTL reported by the GoDMC consortium. For some DMPs, results of the MR were duplicated due to the availability of genotype-outcome data for the same instrument(s) in two independent studies: Mahajan et al. 2014 and Wood et al. 2016. A) forest-plot showing results of the top five DMPs identified with the strongest evidence of causality (smallest p-value). B) Volcano plot showing 24 associations included in the reverse MR (also duplicated estimates), where the effect estimate (x-axis) was plotted against the -Log10(p-value) (y-axis). Results were considered borderline significant at p<0.05 [-log10(p-value) > 1.30], and Bonferroni significant at p < 0.002 [-Log10(p-value)> 2.69]. Highlighted in green are associations detected with borderline evidence of causality at p<0.05.

Table 7-26 Results of the reverse 2SMR for the effect of methylation on T2D. DMPs included in the analysis were detected observationally in the sensitivity meta-EWAS of T2D, and were successfully instrumented by an meQTL reported in the GoDMC consortium at $p < 1.0x10^{-5}$. Summary data for the genotype-T2D association was extracted from two GWAS meta-analyses on T2D reported by Mahajan et al. 2014 and Wood et al. 2016. Associations in the reverse MR were considered significant at p<0.002 after Bonferroni correction (α=0.05/24 top DMPs instrumented). Highlighted in bold are associations identified with borderline evidence of causality (unadjusted p<0.05).

					Reverse MR (Wald Ratio)		Forward MR (IVW)		Sensitivity Meta-EWAS of T2D		
Exposure	Chr	Gene	Outcome	Study	OR (95%CI)	P	Estimate (95%CI)	P	Estimate(95%CI)	P	N
cg00144180*	2	*HDAC4*	Type 2 diabetes	Mahajan *et al.* 2014	1.14(0.97,1.34)	0.115	0.07(-0.05,0.19)	0.256	0.012(0.008,0.017)	5.64E-08	3,428
cg01577083	16	*Unannotated*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.72,1.38)	1.000	-0.15(-0.28,-0.02)	0.023	-0.011(-0.016,-0.006)	7.93E-06	3,428
cg01963618	6	*LOC285768*	Type 2 diabetes	Wood *et al.* 2016	1.05(0.44,2.48)	0.917	0.00(-0.12,0.12)	0.970	-0.008(-0.011,-0.004)	1.55E-06	3,428
cg01963618	6	*LOC285768*	Type 2 diabetes	Mahajan *et al.* 2014	1.05(0.89,1.22)	0.572	0.00(-0.12,0.12)	0.970	-0.008(-0.011,-0.004)	1.55E-06	3,428
cg07212837	8	*Unannotated*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.83,1.21)	1.000	-0.07(-0.2,0.06)	0.304	0.006(0.004,0.009)	3.28E-06	3,428
cg07212837	8	***Unannotated***	Type 2 diabetes	Mahajan *et al.* 2014	**1.09(1.00,1.19)**	**0.047**	-0.07(-0.2,0.06)	0.304	0.006(0.004,0.009)	3.28E-06	3,428
cg08857797	17	*VPS25*	Type 2 diabetes	Wood *et al.* 2016	1.68(0.99,2.86)	0.056	-0.06(-0.19,0.07)	0.377	0.009(0.005,0.012)	2.28E-06	3,428
cg08857797	17	*VPS25*	Type 2 diabetes	Mahajan *et al.* 2014	1.32(0.94,1.87)	0.110	-0.06(-0.19,0.07)	0.377	0.009(0.005,0.012)	2.28E-06	3,428
cg09185884	17	*KCTD2*	Type 2 diabetes	Wood *et al.* 2016	1.16(0.07,19.01)	0.919	0.00(-0.12,0.12)	0.989	0.011(0.006,0.015)	2.33E-06	3,428
cg10082515	7	***Unannotated***	Type 2 diabetes	Wood *et al.* 2016	**0.68(0.52,0.89)**	**0.005**	-0.05(-0.16,0.07)	0.443	-0.013(-0.019,-0.008)	7.46E-06	3,428
cg10082515	7	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	0.92(0.82,1.04)	0.187	-0.05(-0.16,0.07)	0.443	-0.013(-0.019,-0.008)	7.46E-06	3,428
cg10584271*	3	*ITIH1*	Type 2 diabetes	Mahajan *et al.* 2014	0.67(0.37,1.21)	0.187	-0.07(-0.19,0.06)	0.279	-0.014(-0.019,-0.009)	1.73E-07	3,428
cg11024682	17	*SREBF1*	Type 2 diabetes	Wood *et al.* 2016	0.83(0.54,1.28)	0.408	0.05(-0.07,0.18)	0.414	0.008(0.005,0.011)	1.33E-06	3,428
cg11252555	19	*RPL13AP5*	Type 2 diabetes	Wood *et al.* 2016	0.51(0.13,2.02)	0.336	-0.05(-0.17,0.06)	0.370	-0.008(-0.011,-0.004)	7.44E-06	3,428
cg11252555	19	*RPL13AP5*	Type 2 diabetes	Mahajan *et al.* 2014	0.76(0.54,1.07)	0.110	-0.05(-0.17,0.06)	0.370	-0.008(-0.011,-0.004)	7.44E-06	3,428
cg11376147	11	*SLC43A1*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.40,2.50)	1.000	-0.04(-0.16,0.07)	0.454	-0.006(-0.008,-0.003)	5.43E-06	3,428
cg11376147	11	*SLC43A1*	Type 2 diabetes	Mahajan *et al.* 2014	0.80(0.58,1.11)	0.187	-0.04(-0.16,0.07)	0.454	-0.006(-0.008,-0.003)	5.43E-06	3,428
cg12593793	1	*Unannotated*	Type 2 diabetes	Wood *et al.* 2016	1.18(0.22,6.26)	0.844	0.04(-0.08,0.16)	0.519	-0.008(-0.011,-0.004)	2.90E-06	3,428
cg13178597	6	*RGS17*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.34,2.94)	1.000	-0.03(-0.15,0.08)	0.560	-0.01(-0.015,-0.006)	8.57E-06	3,428
cg16192197	6	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	0.97(0.90,1.06)	0.512	-0.02(-0.14,0.11)	0.784	0.01(0.005,0.014)	3.71E-06	3,428
cg16765088	15	***Unannotated***	Type 2 diabetes	Wood *et al.* 2016	**1.72(1.02,2.90)**	**0.040**	-0.08(-0.2,0.04)	0.188	-0.011(-0.014,-0.007)	5.50E-10	3,428
cg16765088	15	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	0.97(0.87,1.08)	0.572	-0.08(-0.2,0.04)	0.188	-0.011(-0.014,-0.007)	5.50E-10	3,428

* DMPs with results of the MR also reported using the IVW estimate

Continuation Table 7-26.

					Reverse MR (Wald Ratio)		Forward MR (IVW)		Sensitivity Meta-EWAS of T2D		
Exposure	Chr	Gene	Outcome	Study	OR (95%CI)	P	Estimate (95%CI)	P	Estimate(95%CI)	P	N
cg18181703	17	*SOCS3*	Type 2 diabetes	Wood *et al.* 2016	0.79(0.49,1.28)	0.341	0.08(-0.04,0.19)	0.217	-0.01(-0.015,-0.006)	6.20E-06	3,428
cg18181703	17	*SOCS3*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.83,1.21)	1.000	0.08(-0.04,0.19)	0.217	-0.01(-0.015,-0.006)	6.20E-06	3,428
cg20231084	11	*Unannotated*	Type 2 diabetes	Wood *et al.* 2016	0.93(0.56,1.54)	0.780	0.01(-0.11,0.13)	0.871	-0.006(-0.009,-0.003)	8.36E-06	3,428
cg20231084	11	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.84,1.19)	1.000	0.01(-0.11,0.13)	0.871	-0.006(-0.009,-0.003)	8.36E-06	3,428
cg20456243	2	*SPEG*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.55,1.83)	1.000	-0.07(-0.19,0.06)	0.304	-0.007(-0.011,-0.004)	9.99E-06	3,428
cg20456243	2	***SPEG***	Type 2 diabetes	Mahajan *et al.* 2014	**1.28(1.00,1.63)**	**0.047**	-0.07(-0.19,0.06)	0.304	-0.007(-0.011,-0.004)	9.99E-06	3,428
cg24512093	3	*ROBO1*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.80,1.25)	1.000	-0.04(-0.15,0.08)	0.546	-0.01(-0.013,-0.006)	7.16E-07	3,428
cg25536676*	1	***DHCR24***	Type 2 diabetes	Mahajan *et al.* 2014	**1.43(1.10,1.86)**	**0.008**	0.04(-0.08,0.16)	0.490	-0.008(-0.011,-0.004)	5.39E-06	3,428
cg27037013	21	*Unannotated*	Type 2 diabetes	Wood *et al.* 2016	0.74(0.35,1.55)	0.423	-0.04(-0.16,0.07)	0.465	-0.015(-0.021,-0.009)	2.90E-06	3,428
cg27037013	21	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.81,1.23)	1.000	-0.04(-0.16,0.07)	0.465	-0.015(-0.021,-0.009)	2.90E-06	3,428
cg27115863*	22	*Unannotated*	Type 2 diabetes	Mahajan *et al.* 2014	1.00(0.86,1.16)	1.000	-0.08(-0.2,0.04)	0.207	-0.011(-0.015,-0.006)	2.41E-06	3,428

* DMPs with results of the MR also reported using the IVW estimate

Continuation Table 7-26. Results of the reverse 2SMR using the IVW estimate for DMPs instrumented by at least two SNPs. Heterogeneity of effect was reported using the Cochran's Q estimate (Q).

				Reverse MR (IVW)				Forward MR (IVW)		Sensitivity Meta-EWAS of T2D		
Exposure	Chr	Gene	Study	OR (95%CI)	P	Q	P_Q	Estimate (95%CI)	P	Estimate(95%CI)	P	N
cg00144180*	2	*HDAC4*	Wood *et al.* 2016	1.5(0.63,3.58)	0.364	4.735	0.030	0.07(-0.05,0.19)	0.256	0.012(0.008,0.017)	5.64E-08	3,428
cg00896068†	13	*Unannotated*	Wood *et al.* 2016	1.11(0.55,2.25)	0.767	2.244	0.134	-0.08(-0.2,0.05)	0.233	-0.008(-0.011,-0.004)	7.58E-06	3,428
cg10584271*	3	*ITIH1*	Wood *et al.* 2016	0.93(0.54,1.59)	0.787	0.646	0.421	-0.07(-0.19,0.06)	0.279	-0.014(-0.019,-0.009)	1.73E-07	3,428
cg14476101†	1	*PHGDH*	Mahajan *et al.* 2014	0.95(0.84,1.07)	0.398	0.000	1.000	-0.01(-0.14,0.12)	0.901	-0.015(-0.021,-0.008)	9.46E-06	3,428
cg24512093†	3	*ROBO1*	Wood *et al.* 2016	0.91(0.59,1.39)	0.650	0.147	0.702	-0.04(-0.15,0.08)	0.546	-0.01(-0.013,-0.006)	7.16E-07	3,428
cg25536676*‡	1	***DHCR24***	Wood *et al.* 2016	**1.39(1.08,1.79)**	**0.011**	2.411	0.299	0.04(-0.08,0.16)	0.490	-0.008(-0.011,-0.004)	5.39E-06	3,428
cg27115863*	22	*Unannotated*	Wood *et al.* 2016	0.9(0.62,1.31)	0.580	0.132	0.716	-0.08(-0.2,0.04)	0.207	-0.011(-0.015,-0.006)	2.41E-06	3,428

* DMPs with results of the MR also reported using the Wald estimate. † DMPs with results obtained only using the IVW estimate. ‡DMP with estimates reported using additional MR methods.

Furthermore, the reverse MR for observational evidence obtained in the EWAS of T2D in ALSPAC, revealed borderline evidence of causality (at $p<0.05$) for the DMP in *STARD10* (Figure 7-13), but none of the signals analysed surpassing Bonferroni significance at $p<0.01$. For the association at *STARD10*, results suggested that hypermethylation was associated with increased risk of T2D (Table 7-27). Because *STARD10* was not identified among top associations in the forward MR, this suggested that variation in methylation at this DMP was likely to occur before the onset of T2D. For the association in *STARD10* there was no consistency in the direction of effect between the observed and the causal estimate (Table 7-27).

Even though the DMP in *NFYC* was identified as the strongest association in the observational analysis (see chapter 4), this marker was not captured among top signals in the bidirectional MR. Thus, the association previously identified at *NFYC* was probably confounded by other factors related with T2D, like levels of C-reactive protein, as an association between *NFYC* and C-reactive protein was previously identified in ALSPAC (see Chapter 4). Some inspectional plots were generated for the DMP in *NFYC* and can be found in the appendix Figure S8-28.

The Steiger test confirmed true direction of causality for associations inspected in the reverse MR as variance in methylation explained by the instruments was higher (methylation R^2 range 0.01 to 0.04) than the variance explained in T2D (T2D R^2 range 8.65×10^{-8} to 4.46×10^{-5}). In addition, a mean F-statistic of 279.5 suggested less probability of obtaining biased results in the reverse MR.

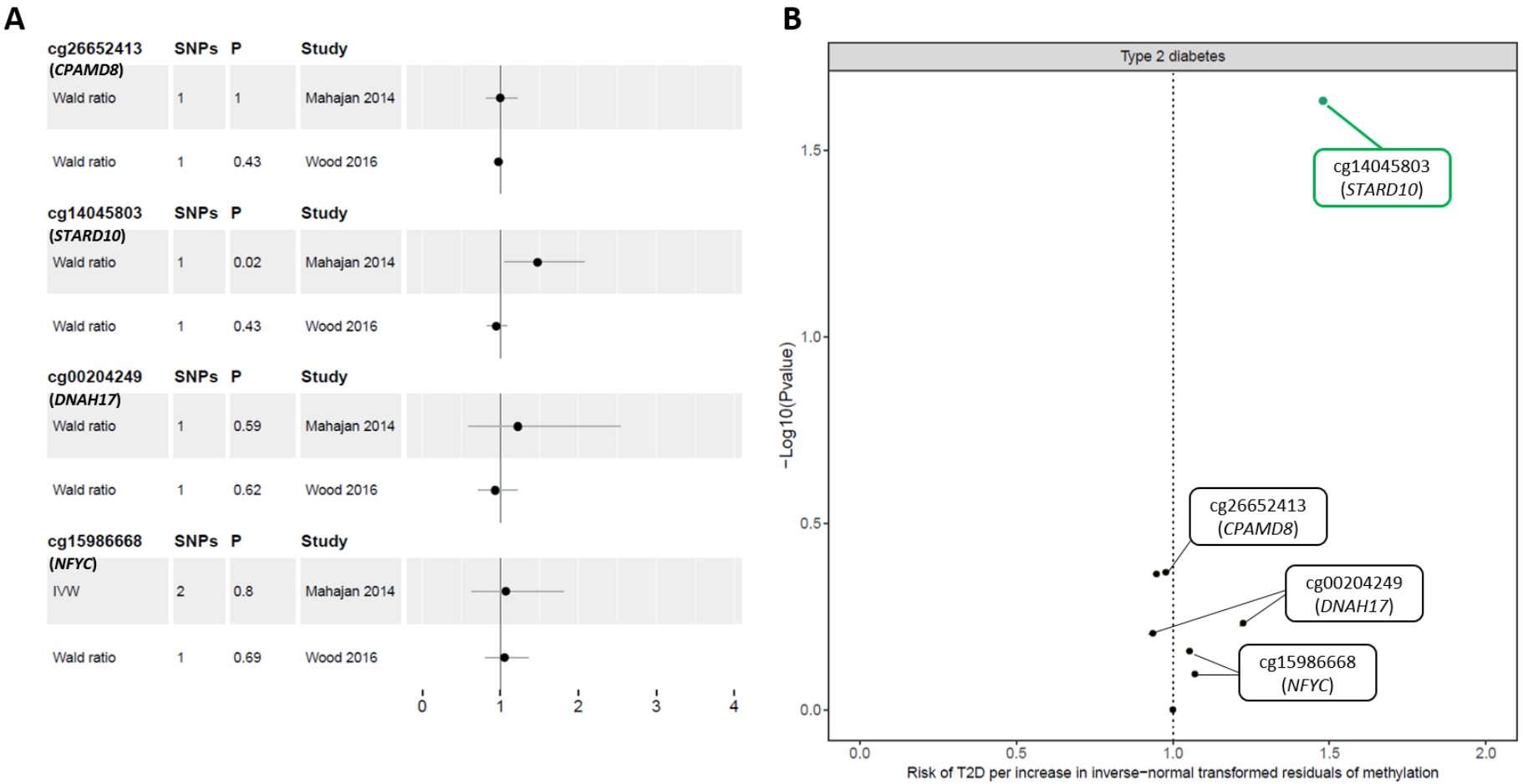


Figure 7-13 Forest plot (A) and volcano plot (B) illustrating results of the reverse 2SMR for the effect of methylation on T2D at four DMPs detected observationally in the EWAS of T2D in ALSPAC. Causal estimates were obtained using the Wald estimate, and the IVW estimate only for the DMP in NFYC. For some DMPs, results of the MR were duplicated due to the availability of genotype-outcome data for the same instrument(s) in two independent studies: Mahajan et al. 2014 and Wood et al. 2016. Highlighted in green is the association at the DMP in STARD10 detected with borderline evidence of causality at p<0.05 (Bonferroni significant at p<0.01).

Table 7-27 Results of the reverse 2SMR for the effect of methylation on T2D at four DMPs detected observationally in the EWAS of T2D in ALSPAC. DMPs included in the analysis were successfully instrumented by an meQTL reported in the GoDMC consortium at $p<1.0x10^{-5}$. Summary data on the outcome was extracted from two GWAS on T2D reported by Mahajan et al. 2014 and Wood et al 2016. Associations were considered significant at $p<0.01$ ($\alpha=0.05/4$ DMPs).

					Reverse MR (Wald Ratio)		Forward MR (IVW)		Observational EWAS of T2D ALSPAC		
Exposure	Chr	Gene	Outcome	Study	OR (95%CI)	P	Estimate (95%CI)	P	Estimate(95%CI)	P	N
cg00204249	17	*DNAH17*	Type 2 diabetes	Wood *et al.* 2016	1.22(0.59,2.54)	0.590	0.07(-0.05,0.18)	0.262	-0.015(-0.021,-0.008)	2.76E-06	1,042
cg00204249	17	*DNAH17*	Type 2 diabetes	Mahajan *et al.* 2014	0.94(0.72,1.22)	0.620	0.07(-0.05,0.18)	0.262	-0.015(-0.021,-0.008)	2.76E-06	1,042
cg14045803	**11**	***STARD10***	Type 2 diabetes	Wood *et al.* 2016	**1.48(1.05,2.07)**	**0.020**	-0.06(-0.18,0.06)	0.297	-0.012(-0.016,-0.007)	1.39E-07	1,033
cg14045803	11	*STARD10*	Type 2 diabetes	Mahajan *et al.* 2014	0.95(0.83,1.08)	0.430	-0.06(-0.18,0.06)	0.297	-0.012(-0.016,-0.007)	1.39E-07	1,033
cg15986668	1	*NFYC*	Type 2 diabetes	Wood *et al.* 2016	1.07(0.63,1.81)†	0.800	0.03(-0.1,0.16)	0.625	-0.071(-0.096,-0.046)	5.48E-08	1,050
cg15986668	1	*NFYC*	Type 2 diabetes	Mahajan *et al.* 2014	1.05(0.81,1.36)	0.690	0.03(-0.1,0.16)	0.625	-0.071(-0.096,-0.046)	5.48E-08	1,050
cg26652413	19	*CPAMD8*	Type 2 diabetes	Wood *et al.* 2016	1.00(0.82,1.22)	1.000	-0.08(-0.2,0.04)	0.190	-0.023(-0.032,-0.013)	2.51E-06	1,050
cg26652413	19	*CPAMD8*	Type 2 diabetes	Mahajan *et al.* 2014	0.98(0.92,1.03)	0.430	-0.08(-0.2,0.04)	0.190	-0.023(-0.032,-0.013)	2.51E-06	1,050

†Estimate obtained using the IVW regression.

7.7.6 Power in the reverse 2SMR

A 57% power was obtained to confidently identify (at p<0.05) a causal effect of 1.18 between methylation and T2D, considering a sample size of 24,227 for the GoDMC dataset, a proportion of cases of 0.32 in the DIAGRAM study, and a mean variance in methylation of 3.0% explained by the instruments. Conversely, a minimum sample of 41,246 will be required to confidently detect (at p<0.05) a similar causal effect with 80% power.

7.8 Comparison of estimates between the causal and the observational analysis for strongest associations detected in the Bidirectional MR

Results of the bidirectional MR for top DMPs detected observationally in the meta-EWAS of T2D, revealed no overlap in the signals identified with significance (adjusted p<0.05) or borderline significance (p<0.05) between the forward (n=4 DMPs) and the reverse MR (n=3 DMPs).These findings suggested that different biological mechanisms could be involved in the association between T2D and DMPs detected downstream T2D onset [cg00082384 (*NISCH*), cg19693031 (*TXNIP*), cg06039489 (*C20orf26*), and cg14275576], compared to the mechanisms responsible for the association between T2D and DMPs identified upstream disease occurrence [cg11851382 (*PPAP2B*), cg07184465 (*SPZ1*) and cg17155612 (*LOC148189*)]. The strongest signal detected in the forward MR was at the DMP in *NISCH*, while in the reverse MR strongest signal surpassing Bonferroni significance was detected at the DMP in *PPAP2B*. To emphasize at this point, is that most of the strongest associations identified in the causal analysis did not surpass epigenome-wide significance in the observational analysis, and the direction of effect between the observed and the causal estimate was generally the opposite, indicating that some level of residual confounding or reverse causation was biasing observational estimates at these DMPs towards the null. The association at the DMP in *TXNIP* was the only one where the observed estimate was identified with epigenome-wide significance ($p < 1.07x10^{-7}$), and where results of the causal analysis supported observational evidence indicating that methylation at this DMP was secondary to T2D.

Of interest was that no association was identified in the bidirectional MR for the DMP cg06500161 (*ABCG1*), despite being this a well-known marker in association with T2D based on observational evidence. Because this signal has been previously identified causally associated with BMI (methylation secondary to BMI)[15], it is possible that the observed association between T2D and *ABCG1* was mediated by BMI, reason why a direct causal effect between T2D and *ABCG1* was not detected in this analysis.

Considering results across analyses, the single sample MR provided less power to detect strong causal associations compared to the 2SMR, even though there was consistency in the direction of effect, but not in the effect size, between estimates of these two analyses. Precision of the estimates varied across analyses and depending on the DMP under study, but in general, estimates of the reverse MR were less precise (i.e. larger confidence intervals) than those of the single sample MR, the forward MR, and the observational analysis, respectively. In terms of magnitude of the absolute effect, estimates of the reverse MR were larger than those of the forward MR, the single sample MR, and the observational analysis, respectively (Figure 7-14).Units of measurement in the reverse MR were interpreted in the odds-ratio scale, rather than in the log(odds) scale. Consistency in the direction of effect between the observed and the causal estimate was observed for associations at the DMPs in *TXNIP* and *PPAP2B*, this based on results of the forward and the reverse MR. For DMPs assessed in both directions of the MR, the direction of the predicted effect was consistent between the forward and the reverse MR for associations at the DMPs in *C20orf26* and *SPZ1*. A forest plot in Figure 7-14 compares estimates between the observed and the causal analysis for top DMPs identified in the bidirectional MR.

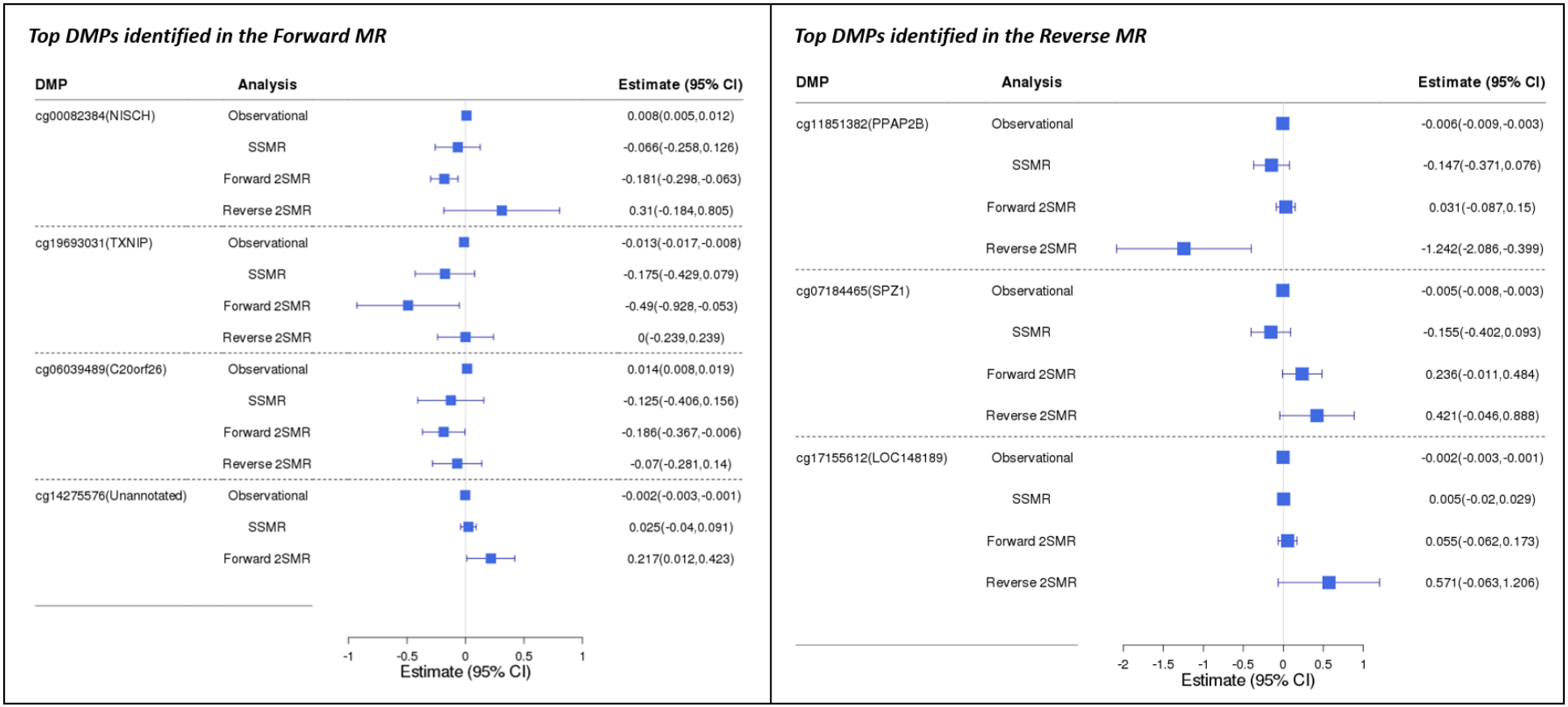


Figure 7-14 Forest-plot comparing estimates of the observational and the causal analysis for DMPs identified with significance (adjusted p<0.05) or borderline significance (p<0.05) in the bidirectional MR. Results are plotted separately for top DMPs identified in the forward MR (left-hand-side) and in the reverse MR (right-hand-side). Among analyses, observational refers to estimates derived from the meta-EWAS of T2D conducted across 5 studies; SSMR is the single sample MR using 2SLS-IV analysis, with individual level data obtained from middle-age adults in ALSPAC; Forward 2SMR is the first direction of the two sample MR where T2D was regarded as the exposure against variation in methylation. Samples included were DIAGRAM (IV-exposure) and ALSPAC (IV-outcome); reverse 2SMR corresponded to the opposite direction of the bidirectional MR using variation in methylation as the exposure and T2D as the outcome. Samples included were GoDMC (IV-exposure) and DIAGRAM (IV-outcome). Results for both directions of the association are presented in the log(odds) scale.

Comparison of estimates across analyses for top DMPs identified in additional observational datasets

Comparing results between the causal and the observational analysis for top DMPs identified in the sensitivity meta-EWAS of T2D (i.e. excluding KORA samples), a total of nine DMPs were detected in the bidirectional MR, one of them with a signal identified in both directions of the association (DMP cg20456243 in *SPEG*). Strongest association observed in the forward MR was at the DMP cg20812370 in *PBX1*, while stronger associations observed in the reverse MR were at the DMPs cg10082515 and cg25536676 (*DHCR24*). Different from associations identified in a previous dataset, signals detected here were more consistent in their direction of effect between the observed and the causal estimate in the forward MR. For associations analysed bidirectionally, there was consistency in the direction of effect between causal estimates for the DMPs cg072212837 and cg16765088. As mentioned before, precision of the estimates varied according to the DMP analysed, but in general, precision was higher in the observed versus the causal estimates (forward MR > reverse MR > single sample MR, respectively). Magnitude of the effect was always higher in the causal compared to observed estimate. In terms of power, the single sample MR was always less powered than the forward 2SMR to detect strong associations, even though there was similarity in the direction of effect and magnitude of effect between estimates of these two analyses. Figure 7-15 compares results across analyses for top DMPs identified in the bidirectional MR.

As in previous analyses, no overlap was detected between results of the forward and the reverse MR for top DMPs identified observationally in the EWAS of T2D in ALSPAC. Strongest association in the forward MR was detected at the DMP cg10870892 in *CTTN*, while in the reverse MR strongest association was detected at the DMP in cg14045803 (*STARD10*). Considering the small number of signals analysed bidirectionally in this dataset, it was more difficult to compare direction of effect and magnitude of effect between causal estimates. For the DMP in *STARD10*, where the MR analysis was conducted bidirectionally, there was consistency in the direction of effect between the observed and the causal estimate in the forward MR, but not in the reverse MR. Comparison of observed and causal estimates for top DMPs of the EWAS of T2D in ALSPAC identified in the bidirectional MR, can be found in Figure 7-16.

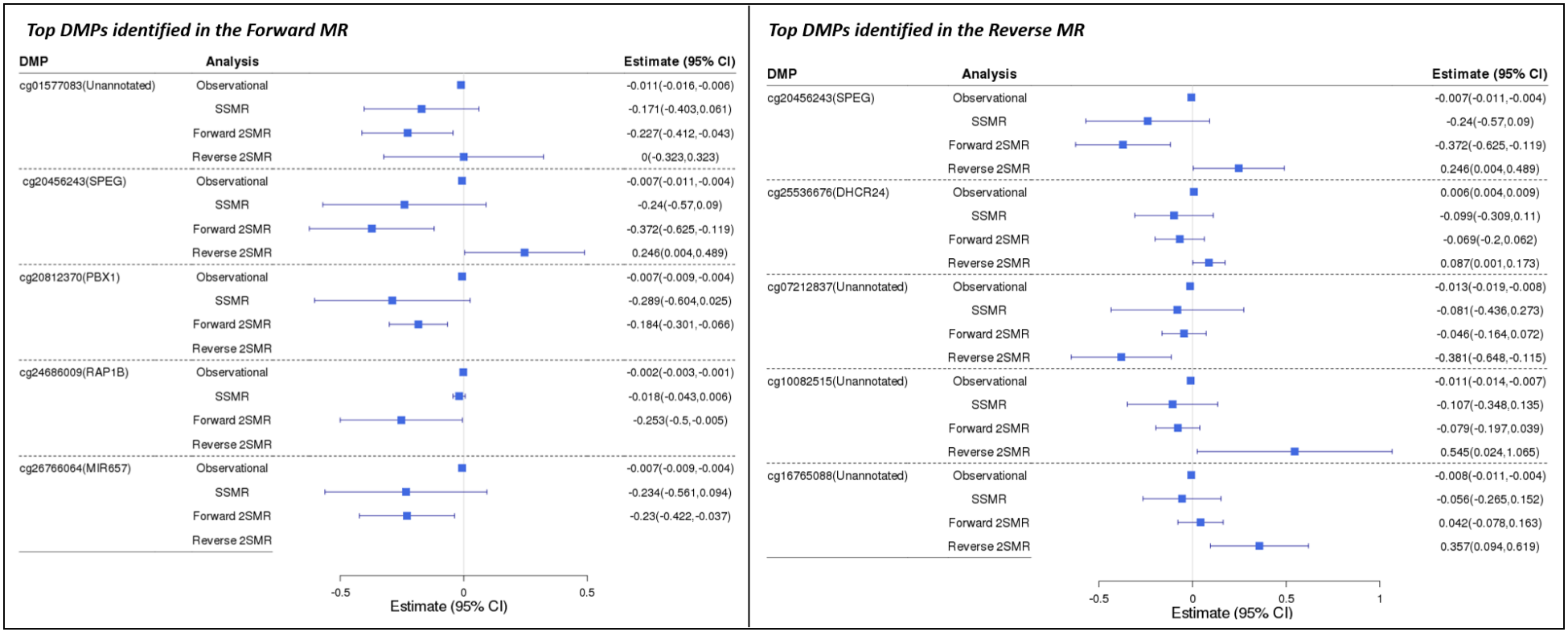


Figure 7-15 Forest-plot comparing estimates of the observed and causal analysis for DMPs identified with borderline significance in the bidirectional MR. DMPs analysed were detected observationally in a sensitivity analysis of the meta-EWAS of T2D (excluding KORA samples). To the left, top DMPs detected in the forward MR, and to the right, top DMPs detected in the reverse MR. The strongest evidence of causality in the forward MR was detected at the DMP in PBX1 (cg20812370), while in the reverse MR stronger association was detected at the DMP in DHCR24 (cg25536676) based on the narrower confidence intervals.

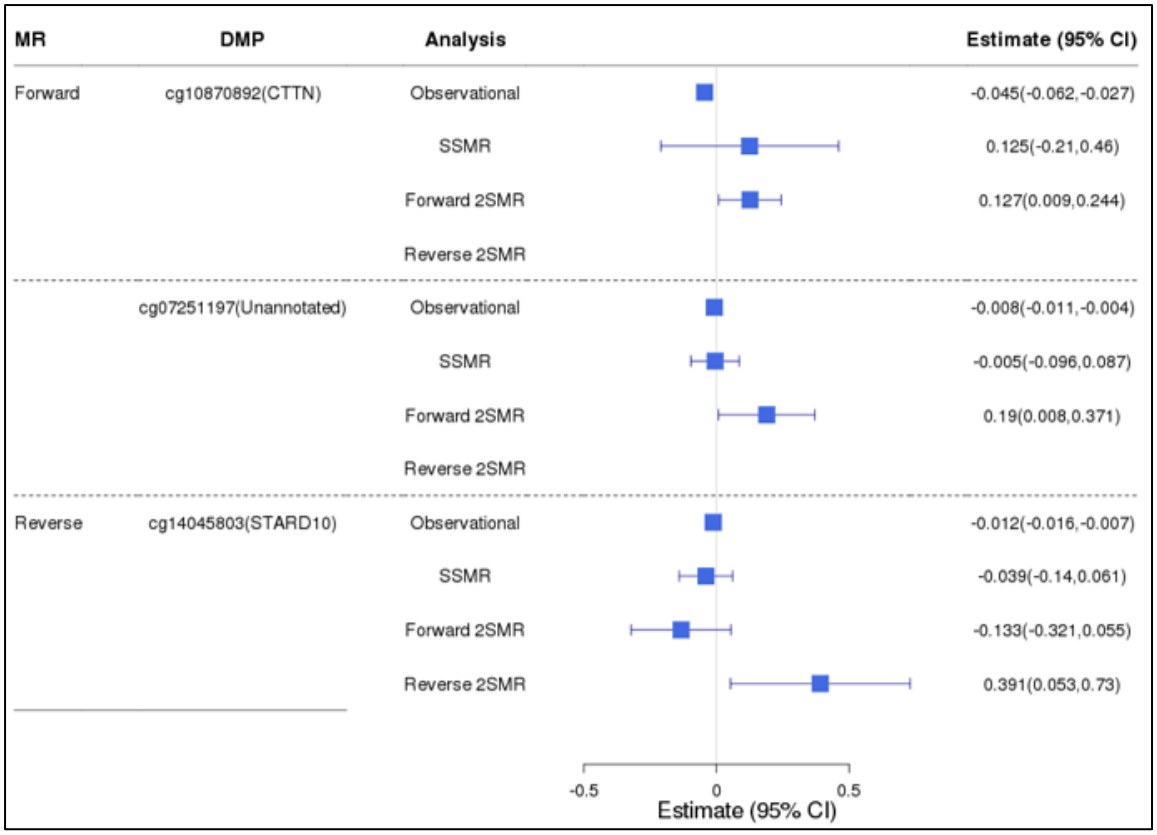


Figure 7-16 Forest-plot comparing observed and causal estimates for top DMPs identified in the bidirectional MR. DMPs analysed were identified observationally in the EWAS of T2D in ALSPAC. Stronger signal in the forward MR was detected at the DMP in CTTN, while in the reverse MR strongest signal was detected at the DMP in STARD10.

7.9 Functional interpretation of results in the causal analysis

To increase the chances of identifying pathways enriched in genes annotated to DMPs detected in the bidirectional MR, top results obtained in the MR analysis were combined across the three different observational datasets, distinguishing between DMPs detected in the forward and the reverse MR (Figure 7-17). Pathway analysis was conducted using GO and KEGG databases and including as background probes all DMPs initially considered for the bidirectional MR analysis across datasets (n=84 DMPs). To identify possible influence of methylation on gene expression, meQTL reported by the GoDMC consortium for DMPs of interest, were examined for their overlap with trans-tissue eQTL reported in the GTEx Portal [163]. Identifying an overlap between an meQTL and an eQTL was suggestive of a causal effect of methylation on gene expression of the nearby gene, especially when both signals were reported in the same tissue.

For DMPs detected in the bidirectional MR, observational evidence of methylation influencing gene expression was looked up in the Bios QTL browser (https://genenetwork.nl/biosqtlbrowser/)[110]. Finally, to identify other outcomes related with T2D that could be influenced by methylation at the

DMPs of interest, meQTL reported for these DMPs were included in a PheWAS SNP lookup using MR-Base (http://phewas.mrbase.org/). Results of this phenotype search were considered significant after multiple testing correction for the number of traits reported for each SNP.

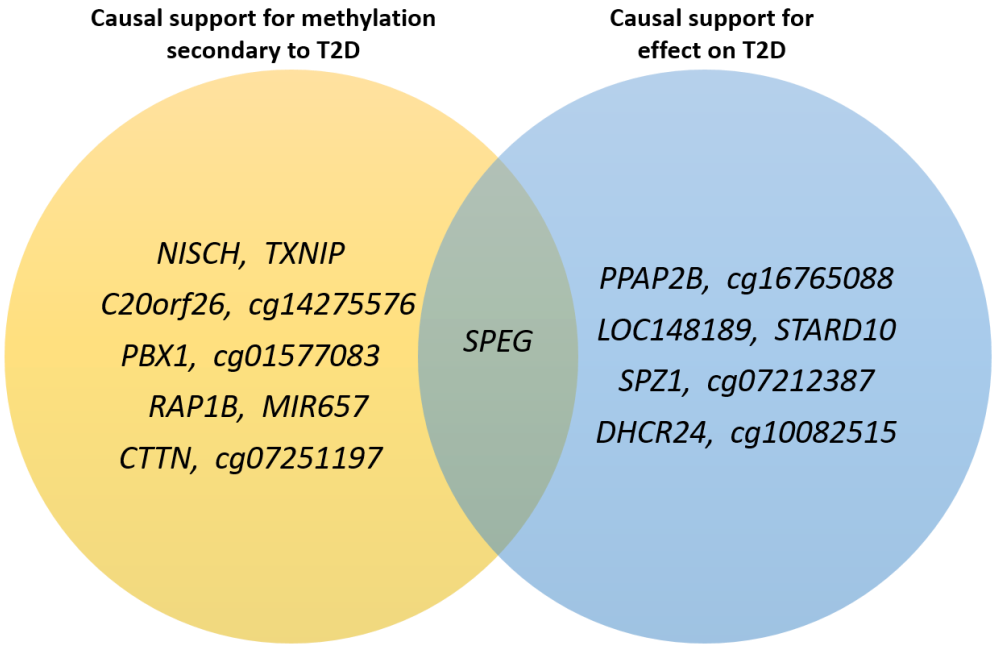


Figure 7-17 Annotated genes and DMPs for associations identified in the bidirectional MR, based on evidence from the T2D epigenome-wide association study. Signals obtained in the causal analysis were combined across the three observational datasets.

Pathway Analysis

Considering the small number of DMPs included in this pathway analysis (11 DMPs annotated to 8 unique genes in the forward MR, and 9 DMPs annotated to 6 unique genes in the reverse MR), there was small power to detect pathways enriched in genes for the DMPs of interest at FDR < 0 .05, and none of the pathways reported by GO or KEGG databases were identified with strong enrichment. For DMPs detected in the forward MR, some of the top pathways identified with suggestive enrichment were related to *pancreatic secretion*, *NOD-like receptor signalling pathway*, *Neurotrophin signalling pathway, MAPK signalling pathway and Ras signalling pathway*, which apart from pancreatic secretion, are biological processes involved in cellular response to inflammatory conditions, neuronal survival, development and function[236], and in signal transduction to control cell growth and differentiation[237, 238], respectively. The NOD-like receptor signalling pathway is related to T2D because hyperglycaemia and the increase in circulatory lipids in insulin sensitive tissues (muscle, pancreas, liver and adipose tissue), promotes the accumulation of pro-inflammatory cytokines and

chemokines that recruit immune cells and cause tissue inflammation[239] (R&D systems, https://www.rndsystems.com/pathways/nod-like-receptor-signaling-pathways).

Regarding DMPs identified in the reverse MR, some of the pathways reported were related to the metabolism of lipids, ABC transporters, and the *AGE-RAGE signalling pathway,* which takes part in the development of T2D complications derived from the abnormal glycation and oxidation of proteins, lipids and nucleic acids, in response to aging and to pathological conditions like hyperglycaemia[240-242]. Abnormal non-enzymatic oxidized metabolites, known as advanced glycation end products (AGEs), attach to immunoglobulin receptors (RAGE) of endothelial cells, smooth muscle cells and monocytes, inducing multiple signalling pathways that promote inflammation (i.e. RAS, MAPK, NF-κB), and the accumulation of ROS elements, which in turn affect the molecular structure of proteins and lipids, altering their physiological function in the cell[242]. The accumulation of AGE products has been shown to be higher in T2D patients compared to controls[242], and among diabetics, these compounds are more concentrated in tissues of patients with secondary cardiovascular complications, compared to patients without diagnosis of adverse secondary outcomes[242]. Thus, it is believed that the accumulation of AGEs is responsible for some of the T2D cardiovascular complications of CHD, heart failure, retinopathy, nephropathy and neuropathy[240-242]. In terms of the role of ABC transporters in T2D, these molecules function in the regulation of cholesterol homeostasis and insulin secretion in β cells[224], and one member of this family (*ABCG1*) has been widely reported in relation to methylation variation associated with T2D[62, 63] and BMI[89, 90, 213]. appendix Table S8-42 describes top 20 pathways reported in KEGG for DMPs detected in the bidirectional MR.

Trans-tissue gene expression

A multi-gene search was performed in the GTEx Portal to determine trans-tissue levels of expression for genes related to DMPs identified in the bidirectional MR. For genes annotated to DMPs detected in the forward MR, *TXNIP* was the only gene which signal was identified across seven tissues selected for their relevance in the pathophysiology of T2D: whole blood, adipose tissue (subcutaneous fat and omentum), liver, pancreas, skeletal muscle and thyroid (Figure 7-18). With respect to genes related to DMPs detected in the reverse MR, strong signals of expression (Transcripts Per Million > 31) were reported for *PPAP2B* in thyroid and adipose tissue, for *SPEG* in skeletal muscle, for *STARD10* in liver and pancreas, and for *DHCR24* in liver and skeletal muscle (Figure 7-18). In general, for the genes investigated, gene expression was less common in blood compared to other tissues examined, with some detectable levels of expression (TPM < 31) reported in *STARD10*, *SPEG* and the *CTTN* gene.

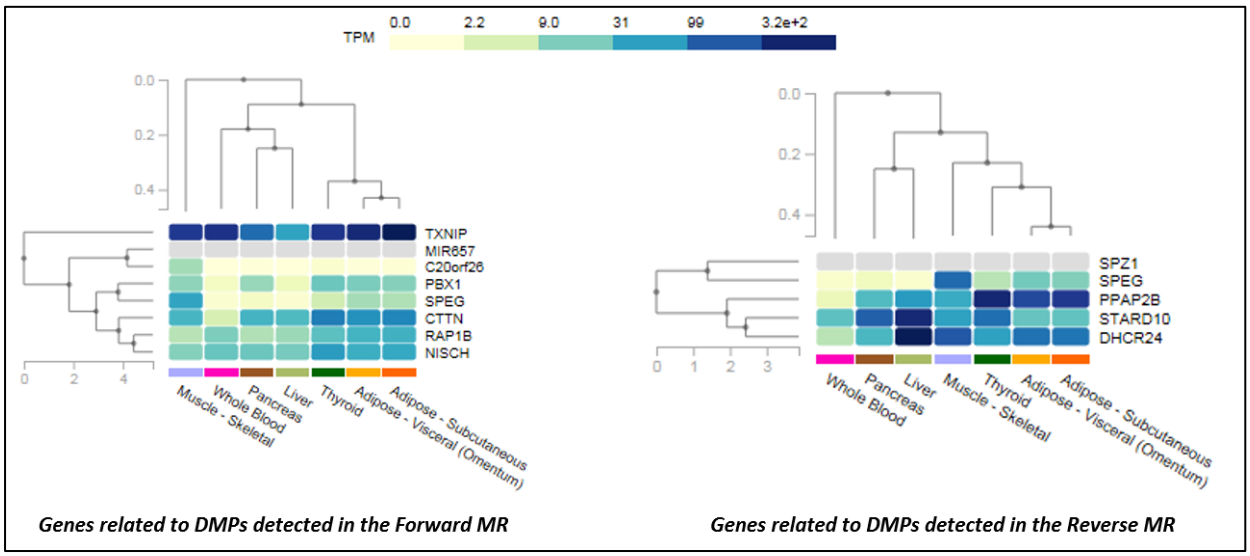


Figure 7-18 Heatmap illustrating levels of gene expression across different tissues for genes annotated to DMPs identified in the bidirectional MR. Heatmap on the left for genes annotated to DMPs identified in the forward MR, and on the right, heatmap for genes annotated to DMPs identified in the reverse MR. Tisuess enquired are labeled at the bottom of the figure, while genes are listed to the right-hand-side of the heatmap. The legend at the top indicates the level of expression of the gene in Transcripts per Million, darker shades indicate higher levels of expression. Figure obtained from the GTEx Portal (https://gtexportal.org/home/multiGeneQueryPage).

eQTL search across tissues

MeQTL reported in GoDMC for top DMPs identified in the bidirectional MR, were compared to trans-tissue eQTL reported in the GTEx Portal for seven tissues of interest (see above), but none of the 16 meQTL of interest was identified as a strong eQTL at q-value < 0.05. When lowering the level of stringency to include eQTL identified with nominal significance, the meQTL rs11716756, identified in association with methylation at the DMP cg00082384 (*NISCH*) using blood samples, was also reported as an eQTL nominally associated (q=0.07) with variation in gene expression at the *IQCF3* gene in skeletal muscle samples. None of the eQTL reported with nominal significance in blood, was related to one of the meQTL of interest. For the signal detected in common at the SNP rs11716756, the effect allele for the meQTL and the eQTL was the same, but direction of effect was the opposite between the two traits, where per increase in the T allele was associated with an average decrease in methylation at the DMP in *NISCH* (estimate=-0.24, p= 3.54×10^{-202}), but it was nominally associated with an average increase in gene expression at the *IQCF3* gene (Log_2 allelic Fold change=1.38, q=0.07). Because the meQTL and the eQTL identified in common at the SNP rs11716756 was detected in different tissues, and because genetic variation in methylation and gene expression is a tissue-specific phenotype, it was not reliable to use the eQTL detected in skeletal muscle as a proxy for an eQTL in blood, the tissue where the meQTL was identified. Thus, the signal detected in

common at the SNP rs11716756 was not considered a good candidate for an MR analysis between methylation at the DMP cg00082384 in *NISCH*, and gene expression of the same gene. No other meQTL was identified in overlap with an eQTL when the list of meQTL was extended to include those correlated with T2D-SNPs at LD > 0.01.

cis-eQTM in blood

A lookup for cis eQTM in the Bios QTL browser[110], revealed that the DMPs cg19693031, cg20456243 and cg17155612 were strongly associated (at FDR<0.05) with variation in gene expression of transcripts in *TXNIP* (ENSG00000117289), *SPEG* (ENSG00000072195) and *LINC0062* (ENSG00000261824), respectively (Table 7-28). Because variation in methylation at the DMPs in *TXNIP* and *SPEG* was demonstrated to occur secondary to T2D, methylation could be a mediator of a putative association between T2D and gene expression of *TXNIP* and *SPEG*. Evidence of an association between expression of *TXNIP* and T2D has been reported in muscle[63, 243], in islets from human cells cultivated *in vitro* and treated with differential glucose concentrations[244], and in islets from animal models with diet-induced T2D[245], but less is known about the expression of *TXNIP* in blood samples of T2D patients. Based on a previous study, common genetic variants detected at the *TXNIP* were not associated with T2D, and normoglycemic participants did not show upregulation of *TXNIP* even if they had genetic predisposition to T2D[243]. Thus, Parihk *et al.* suggested that expression of *TXNIP* was mainly influenced by glucose dysregulation rather than by T2D-associated genetic factors[243]. No evidence has been reported relating the expression of *SPEG* with T2D. For the DMP in *LINC00662* (or *LOC148189*), variation in methylation was demonstrated to occur upstream T2D onset, suggesting that the observed association between methylation and gene expression at *LINC00662* could be independent of the methylation~T2D association, or that T2D could be mediator of the association between methylation and gene expression of *LINC00662*.

Table 7-28 Association between methylation and gene expression for three DMPs identified in the bidirectional MR analysis. Association estimates were reported in the Bios QTL browser based on the study conducted by Bonder et al. 2016.

DMP	Chr	Gene	Transcript	Effect	Estimate	SE	P-value	FDR
cg19693031	1	*TXNIP*	ENSG00000117289	-	-0.12	0.04	7.14E-08	9.28E-06
cg20456243	2	*SPEG*	ENSG00000072195	+	0.09	0.04	2.42E-05	6.02E-03
cg17155612	19	*LINC00662*	ENSG00000261824	-	-0.09	0.04	1.97E-04	3.51E-02

PheWAS lookup for the association between meQTL and other T2D-related outcomes

Results of the PheWAS SNP lookup indicated that six out of 16 meQTL identified as proxies for 13 DMPs detected in the bidirectional MR, were also associated with traits related with the concentration of different lipids in blood, the molecular structure of lipids, inflammatory bowel disease and Crohn's disease, medication for cholesterol, blood pressure or diabetes, and body fat distribution, at Bonferroni corrected $p<2.5x10^{-4}$ (α=0.05/lookup traits per meQTL). None of the traits reported in this lookup was directly related with glycaemic traits or T2D complications. Summary statistics of the association between meQTL and additional traits identified in the PheWAS lookup, are presented in the appendix Table S8-43.

7.10 Chapter summary

The aim of this chapter was to identify causality for signals detected with significance and borderline significance ($p<1.0x10^{-5}$) in a meta-EWAS of T2D, and in two additional datasets (sensitivity meta-EWAS and EWAS of T2D in ALSPAC). Since the causal analysis was performed in two directions, it was possible to distinguish between signals associated with the effects of T2D, from those that occurred prior disease onset.

Polygenic risk score analysis

A weighted polygenic score was generated to strengthen the amount of variation in T2D explained by 56 independent SNPs identified in strong association with the disease across four different DIAGRAM studies. This genetic score was able to capture 2.0% of the total variation in T2D in the subsample of middle-age adults in ALSPAC and was independent of direct confounders of the exposure-outcome association. When evaluating the association between this score and genome-wide variation in methylation (EWAS with the PRS) in the subsample of ALSPAC, no strong association was identified, suggesting that the score had less power than the case control study (EWAS of T2D) to capture variation in methylation across disease groups. Based on top DMPs detected in the observational analysis, smallest P-value for the observed IV-outcome association was detected at the DMP cg11851382 in *PPAP2B* identified in the meta-EWAS and the sensitivity meta-EWAS of T2D, and at the DMP cg04656330 in *PNKD* identified in the EWAS in ALSPAC. In general, magnitude of the effect obtained in the IV-outcome regression was smaller than the effect detected in the observational analysis, but there was some consistency in the direction of effect between estimates across analyses. Additional sensitivity analyses using the polygenic score revealed that there was no benefit in including the score in a regression model to increase the amount of variation

in methylation already explained by the case control study, and a DMR analysis just improved results obtained in an EWAS with the PRS, capturing a unique significant PRS-associated DMR within the region of the *CHRND* gene.

Single sample MR analysis

The single sample MR (SSMR), using a 2SLS-IV analysis, did not support causality in the association between T2D and differential methylation for DMPs identified in the observational analysis. Top associations identified with nominal causality across datasets were detected at the DMPs in cg15560632 in *LRCH4*, cg20812370 in *PBX1,* cg04656330 in *PNKD* and the DMP cg15986668 in *NFYC*, where results suggested that T2D was causally associated with hypomethylation at these sites. Overall, direction of effect of the predicted IV-outcome estimate was consistent with that of the observed estimate across analyses, but magnitude of the effect was always larger in the causal compared to observed estimate. When looking at the strength of the instrument (PRS), results of the Weak instrument test suggested that the PRS was a weak predictor of variation in methylation in response to T2D, and this result was corroborated by obtaining a Wald test value (equivalent to the F-statistic) lower than the threshold used to determine good MR instruments (F-statistic > 10). Finally, it was shown that results of the SSMR were obtained with an average power of 33% at $p<0.05$, and it was estimated that to confidently detect a similar effect with 80% power (at $p<0.05$), the minimum sample required were 3,943 participants.

Overall weakness of instrument identified in the SSMR analysis was also attributed to the small sample size used. Ideally, to precisely detect causal estimates, the sample size required should be on average 50 times larger in the causal compared to the observational analysis[232]. One way to calculate the required sample size is to divide the sample used in the observational analysis, over the R^2 obtained from regressing the IV on the exposure[232]. Based on this rule, a sample of 52,500 (i.e. N=1,050/0.02) was needed to precisely detect causal estimates for DMPs detected in the EWAS of T2D in ALSPAC, while a sample of 257,350 (i.e. N=5,147/0.02) was required to precisely identify causal estimates for DMPs detected in the meta-EWAS of T2D. Considering the above, one of the limitations of this study was the use of a single underpowered sample to identify causality for associations detected observationally in two samples with different detection strengths (i.e. subsample of ALSPAC versus meta-analysis sample). For this reason, causal estimates obtained for signals detected in the meta-EWAS were less precise, than those obtained for signals detected in the EWAS in ALSPAC.

Different approaches can be followed to strengthen causal estimates of the SSMR for associations detected in the meta-EWAS of T2D, one of them would be to replicate the 2SLS-IV analysis in the collaborating studies, and to summarize causal evidence via meta-analysis. One of the limitations of this approach is that it requires that all the studies have genotype and phenotype data to calculate the PRS for T2D, and that the same number of samples account with methylation data to calculate the predictive exposure-outcome association via 2SLS-IV analysis. Selecting on samples with phenotype, genotype and methylation data available, considerably reduces the final size of the sample used in the MR analysis, thus affecting the strength of the predicted estimates. To counteract sample-size limitations of a SSMR, even when it is implemented in the context of a meta-analysis, a 2SMR stands out as a stronger method for causal analysis, as it does not rely on the use of a unique sample to calculate summary statistics for the genotype-exposure and genotype-outcome associations, thus allowing to extract summary data from two independent, even though comparable, well-powered samples, generally coming from big consortiums of GWAS meta-analyses. Furthermore, the existence of platforms like MR-Base that automate steps of data handling and data analysis, largely facilitates the use of 2SMR methods.

Forward two sample MR analysis

When conducting the forward 2SMR, strong estimates were included for the IV-exposure (T2D) association based on summary data extracted from DIAGRAM. However, estimates of the observed IV-outcome (methylation) association, which were derived from an EWAS of T2D-SNPs conducted in ALSPAC, were less powered, and only one signal was detected with borderline significance between the SNP rs4275659 and the DMP cg10584271 in *ITIH1.* The association at *ITIH1* was later shown to be influenced by the effect of horizontal pleiotropy since results of the MR did not provide evidence of causality between T2D and methylation at this DMP. Thus, results of the forward 2SMR were obtained using two samples with different detection strengths. As a sensitivity analysis, estimates of the observed IV-outcome association were intended to be substituted with GoDMC data, but none of the associations reported in this consortium for the DMPs of interest, included SNPs that could be used as proxies for T2D-SNPs.

In general, no evidence for heterogeneity or horizontal pleiotropy was detected for top signals identified in the 2SMR, even though some of the SNPs included in the analysis were identified with heterogeneous effect in the inspectional plots, without evidence of effect-allele coding errors for these SNPs, or an association between them and other outcomes different from T2D. The F-statistic for this analysis (mean= 51.53) suggested less probability of obtaining biased results due to weak

instruments, and a 100% power was reported to detect an absolute causal effect of 0.15 at p<0.05. Main association in the 2SMR was detected at the DMP cg00082384 in *NISCH*, suggesting that T2D was causally associated with a decrease in methylation. For this DMP, magnitude of the causal estimate was consistent across different methods, but direction of effect was conflicting between the observed and the causal estimate. The DMP in *TXNIP* was another signal captured with nominal causality in the 2SMR, and the strongest causal estimate for this DMP suggested a negative effect of T2D on methylation, a result that was consistent with the observational evidence. Two other signals were detected with nominal significance at the DMP in *C20orf26* and the DMP cg14275576, and the effect reported by the causal estimate was in opposite direction to the effect identified in the observational analysis or these two signals.

When conducting the 2SMR on DMPs detected in a sensitivity analysis of the meta-EWAS of T2D, the strongest causal association was identified at the DMP in *PBX1*, a signal that was successfully replicated from the SSMR analysis, and where estimates across analyses suggested that T2D was nominally associated with hypomethylation at *PBX1*, in agreement with results of the observational analysis. Other signals detected with nominal causality were at the DMPs in *SPEG*, *RAP1B*, *MIR657* and the DMP cg01577083, and the strongest estimates for these DMPs were reported using the weighted median, with consistent direction of effect between the observed and the causal estimates. Likewise, the 2SMR conducted on DMPs identified in the EWAS of T2D in ALSPAC, revealed two signals with nominal evidence of causality (at p<0.05) at the DMP in *CTTN* and the DMP cg07251197, suggesting that T2D was associated with an average increase in methylation at these sites, but direction of effect was contradictory between the observed and the causal estimates for these two DMPs.

Reverse two sample MR analysis

The second direction of the association between methylation and T2D was assessed using GoDMC data to extract instruments for DMPs detected in the observational analysis, while summary data for the genetic association with T2D was extracted from two large GWAS on T2D, which were selected based on the number of cases included, and the number of SNPs with available summary statistics. Because the number of instruments available per DMP in the reverse MR was small (1-3 SNPs per DMP), the type of sensitivity methods that could be applied to detect heterogeneity or horizontal pleiotropy was limited. Overall, results of the Steiger test suggested that true direction of causality was assessed for associations included in the reverse MR, considering that the instruments explained a higher proportion of variance in methylation, compared to the variance explained in

T2D. Because the average F-statistic obtained for this analysis was 213.86, it was less likely that results of the MR were affected by weak instrument bias. On average, a 57% power was estimated to confidently detect (at $p<0.05$) a causal effect of 1.18 for the effect of methylation on T2D.

Stronger signal surpassing Bonferroni correction in the reverse MR was detected at the DMP cg11851382 in *PPAP2B*, suggesting that hypermethylation of this site was associated with an increased risk of T2D. Other signals were identified with nominal evidence of causality at the DMPs in *SPZ1* and *LOC148189.* Because these association were uniquely identified in the reverse MR, it was possible that variation in methylation at these DMPs, especially at the DMP in *PPAP2B,* occurred before the onset of T2D. Similarly, results of the reverse MR for signals that were previously identified in the forward MR, allowed to determine their true direction of effect and to confirm that variation in methylation at the DMPs in *NISCH*, *TXNIP* and *C20orf26,* occurred secondary to the onset of T2D.

For other datasets of observational evidence analysed in the reverse MR, it was demonstrated that variation in methylation was predictive of T2D for the intergenic DMPs cg10082515, cg16765088, cg07212837, and for the DMPs mapping to the *DHCR24*, *SPEG,* and *STARD10* genes. None of these DMPs was previously identified as a top signal in the forward MR, except for the association at the DMP in *SPEG*. However, evidence of the bidirectional MR suggested that variation in methylation at the DMP in *SPEG* was more likely to occur as a consequence of T2D considering the higher strength of the causal estimate in the forward MR (T2D→ DNAm), and its consistency in the direction of effect with the observed estimate. For the remaining DMPs detected exclusively in the reverse MR, results suggested that hypermethylation was associated with higher risk of T2D, except for the DMP cg10082515, where hypermethylation was protective against the risk of T2D.

From the associations identified as top signals in the forward MR that were furtherly analysed in the reverse MR, it was possible to confirm that variation in methylation occurred secondary to T2D at the DMP cg01577083 and the DMP in *SPEG*. Even though in the dataset of DMPs detected in the EWAS in ALSPAC the strongest association was detected at the DMP cg15986668 in *NFYC* ($p<1.07x10^{-7}$), this signal was not captured in either direction of the bidirectional MR, suggesting that it was probably confounded by other factors related with T2D that were unaccounted for in the observational analysis, like levels of C-reactive protein.

Results across the causal and the observational analyses

Comparing associations between the SSMR, the 2SMR and the observational analysis, it was evident that estimates in the SSMR were less powered than those in the 2SMR, but there were similarities between them in terms of direction of effect and magnitude of effect, this relative to estimates in the observational analysis. Thus, it was possible that the main limitation of the SSMR was the size of the sample used, rather than the strength of the instrument itself. In terms of precision of the estimates (i.e. narrower confidence intervals), this was generally higher for the observed estimates compared to estimates of the forward MR, the SSMR, and the reverse MR, respectively. Regarding the magnitude of effect, this tended to be larger in the reverse MR compared to the forward MR, the observed estimate and the SSMR, respectively. Rresults in the reverse MR were interpreted in a different scale (odds ratios) to the scale used to interpret results in the remaining analyses (log-odds).

Functional Interpretation of findings in the bidirectional MR

Pathway analysis

Even though there was not enough power to detect pathways enriched in genes for DMPs detected in the bidirectional MR, some of the pathways identified in suggestive enrichment were relevant for the study of T2D. For associations detected in the forward MR, and thus related to variation in methylation occurring as a consequence of T2D, suggestive enrichment was detected for pathways related to inflammatory processes, signal transduction and pancreatic secretion, while for associations detected in the reverse MR, and thus related to variation in methylation occurring upstream the onset of disease, suggestive enrichment was identified for pathways related to the metabolism of lipids, ABC transporters, and to the abnormal glycation and oxidization of lipids, nucleic acids and proteins via the AGE-RAGE signalling pathway. This latter pathway has been extensively reported in relation to the development of T2D cardiovascular complications via de accumulation of AGEs (abnormally glycated end products). Thus, the functional pathway analysis indicated that genes annotated to DMPs that were causal for T2D, were associated with metabolic pathways of lipids and other molecular compounds that, under abnormal conditions, can lead to the development of T2D, while genes annotated to DMPs that were secondary to the effects of T2D, were associated with cell signalling pathways and inflammatory processes in response to the pathological state of the disease (glucotoxicity, lipotoxicity, oxidative stress, endoplasmic reticulum stress).

Trans-tissue eQTL

Looking at the level of expression of genes of interest across different tissues of relevance for T2D, data from the GTEx Portal suggested that *TXNIP* was the only gene which expression could be identified across different tissues. Other genes were more expressed in thyroid and adipose tissue (*PPAP2B*), skeletal muscle (*SPEG* and *DHCR24*), and liver and pancreas (*STARD10*). In general, blood was the tissue where the lowest level of expression for was observed for the genes of interest.

eQTM

Investigating the observational association between methylation and gene expression, data from the Bios QTL browser indicated that methylation at the DMPs in *TXNIP*, *SPEG* and *LINC0062,* was strongly associated with the expression of transcripts for the same genes. Because methylation at *TXNIP* and *SPEG* was demonstrated to occur downstream the onset of T2D, it was possible that methylation could be mediating a putative association between T2D and gene expression of *TXNIP* and *SPEG*. On the contrary, because methylation at the DMP in *LINC0062* was demonstrated to occur prior the onset of T2D, it is likely that the methylation~gene expression association at *LINC0062* is mediated by T2D, or that this association is independent of T2D. Further analyses are required to determine if there is an association between gene expression of *TXNIP*, *SPEG*, *LINC0062,* and T2D, and to demonstrate the possible mediator effect of methylation or T2D in the methylation~gene expression~T2D association.

Genetic overlap between methylation and gene expression

Comparing meQTL reported in GoDMC with eQTL identified across different tissues and reported in the GTEx Portal, a common signal was identified at the SNP rs11716756 in relation to variation in methylation of the DMP cg00082384 in *NISCH* in blood samples, and in suggestive association with variation in gene expression of the *IQCF3* gene in skeletal muscle samples. The effect allele was similar between the meQTL and the eQTL datasets for the SNP rs11716756, but the direction of effect differed among traits. In addition, the signal detected in common at the SNP rs11716756 was not considered a good candidate for a causal analysis because genetically influenced variation in methylation and gene expression was reported in independent tissues.

PheWAS lookup

To furtherly understand the impact of findings from the bidirectional MR, it was investigated other traits that could be associated with signals identified in the bidirectional MR. Thus, a PheWAS SNP lookup was conducted using meQTL for the DMPs of interest, identifying that at least six of these meQTL were associated with traits related to levels of lipids in blood, molecular structure of lipids, fat distribution, medication for cholesterol, T2D or blood pressure, among others. Based on these results, it was concluded that a second step of the MR analysis will be necessary to understand the association between signals detected in the bidirectional MR, and other outcomes that could be in the causal pathway between methylation and T2D.

Results of the bidirectional MR in the context of previous epigenetic studies

Two DMPs mapping to the nischarin gene (*NISCH*) were previously detected (at p<0.05) in association with T2D in a study using discordant MZ twins and adipose tissue samples, with signals replicated in a cross-sectional case control study[79]. However, none of these DMPs identified in adipose tissue, corresponded to the DMP cg00082384 identified in the causal analysis for T2D using blood samples. Since no other signal has been previously reported at this DMP in association with T2D or adiposity traits related with T2D, the signal detected in this study represents a novel finding. The nischarin gene encodes for a noradrenergic imidazoline-1 receptor protein localized in the inner side of the cellular membrane[246, 247]. In mice, this protein is involved in the reorganization of the cytoskeleton and cellular mobility by binding to alpha-5-beta-1 integrin, whereas in humans this protein binds to an adapter of the insulin receptor substrate 4 (*IRS4*)[246] and helps in the translocation of alpha-5 integrin from the membrane to endosomes (uniprot, https://www.uniprot.org/uniprot/Q9Y2I1). The association of *NISCH* with IRS-4 showed no alteration of the tyrosine phosphorylated conformation of the IRS-4 receptor, and no disruption in the binding of this receptor with the downstream signalling proteins PI3K and Grb2[246]. The function of *NISCH* is tissue-dependent, and in breast cancer tissue, the upregulation of this gene prevented tumour cells proliferation and metastasis[247], while in neuronal and cardiac tissue, expression of *NISCH* affected cell growth and differentiation (GeneCards, https://www.genecards.org). The role of *NISCH* on T2D is unknown, but a GWAS reported a variant in this gene associated with waist-hip ratio[248].

Similarly, a DMP mapping to the phosphatidic acid phosphatase 2b (*PPAP2B*) gene, was detected in association with T2D (at p<0.05) in adipose tissue samples according to the EWAS conducted by Nilsson *et al.*[79]. However, there was no similitude between the signal previously detected in adipose tissue (cg19711007), and the one identified in blood (cg11851382) in this study. Additional evidence

for this gene suggests that expression of *PPAP2B* is upregulated in T2D patients with inadequate glucose control (HbA1c > 8.5%), compared to T2D patients with well-controlled glucose (HbA1c < 7.0%)[249]. *PPAP2B* is a cell membrane glycoprotein involved in the conversion of phosphatidic acid to diacylglycerol, in the *novo* synthesis of glycerolipids, and in signal transduction mediated by phospholipase D (GeneCards, https://www.genecards.org), and based on GWAS studies, some variants in this gene have been associated with T2D, glycated haemoglobin levels, and glycaemic traits[249]. The strong association detected between methylation at *PAP2B* and T2D in the reverse causal analysis, constituted a novel finding.

Evidence of an association between methylation at cg19693031 in the thioredoxin interacting protein (*TXNIP*) and T2D, has been extensively reported and validated in epigenome-wide studies of incident[62] and prevalent T2D[46, 64-67], with an effect detected irrespective of the ethnicity of origin (i.e. European, Indian Asian, Arabs and Mexicans). Despite previous evidence, this study is the first to address causality in the association between *TXNIP* and T2D using blood sample, demonstrating that difference in methylation at the DMP cg19693031 occurs secondary to the effects of T2D. This finding contradicts evidence reported by Chambers *et al.*[62], where they identified as their strongest signal for prediction of future risk of T2D, the DMP in *TXNIP*, based on DNA methylation profiled at baseline in normoglycemic, non-diabetic samples[62]. However, various studies have reported methylation at *TXNIP* in association with prevalent T2D[46, 64-66], and some others have failed to replicate the association between *TXNIP* and incident T2D[63]. Thus, it is possible to suggest that even though the study by Chambers had enough power to detect an association between *TXNIP* and incident T2D, the age at which these samples were recruited (approximately 53y), and the time elapsed until some of them undergone T2D (8y of follow-up), indicated that methylation at baseline could have already been affected by prediabetes, or the different metabolic deficiencies that precede T2D (abnormal metabolism of lipids, obesity, accumulation of proinflammatory cytokines, insulin resistance), and that the signal detected at *TXNIP* was indeed a signature of this.

Differential methylation at the DMP in *TXNIP* in association with T2D was demonstrated to occur not only in blood, but also in skeletal muscle and pancreatic tissue from a similar sample according to the study conducted by Dayeh *et al.*[63]. Expression of *TXNIP* was high across different target tissues for T2D based on data available in the GTEx Portal (see section 7.9), and Dayeh *et al.*[63] demonstrated that *TXNIP* expression was upregulated in muscle of T2D patients compared to controls, in line with results shown by Parikh *et al.*[243]. Functionally, *TXNIP* plays an important role in regulating cellular redox, lipid homeostasis and peripheral glucose uptake[62, 65, 243, 250], especially in pancreatic β cells and

other tissues responsive to insulin. The regulatory effect of this protein on glucose uptake is mainly through a negative feedback loop[62], where the expression of *TXNIP* is upregulated by high levels of glucose through the carbohydrate response-element binding protein, a transcription factor that binds to the promoter of *TXNIP* inducing its expression[250], but *TXNIP* in turn downregulates *GLUT1*, an important transporter of glucose within the cell[62]. Downregulation of *TXNIP* has been shown to be protective against obesity-induced diabetes, preventing β cell apoptosis and β-cell mass loss[62].

A DMP in the pre-B-cell leukaemia transcription factor 1 (*PBX1*) was identified in causal association with T2D, and for another DMP mapping to the region of this gene, a signal was previously identified in association with BMI in an EWAS using blood and adipose tissue samples[213]. The association reported between *PBX1* and BMI surpassed adjustment for T2D status[213]. Despite the observational evidence between *PBX1* and BMI, there has been no validation of this association in the context of a causal analysis, as it was done in this study for the association between the DMP cg20812370 in *PBX1* and T2D, representing this a new finding. *PBX1* is involved in pancreatic development and function[213], and some variants for this gene have been associated with obesity[213], while according to animal studies, *PBX1* might influence fatty acid composition in adipose tissue[213].

Methylation at a DMP in the 24-dehydrocholesterol reductase (*DHCR24*), has been previously identified in association with BMI[89, 90] and waist circumference (WC)[213], this latter association dependent on T2D status. For the signal detected in association with BMI, methylation was also associated with gene expression, and gene expression with BMI[90]. Despite this, no causal association was identified between methylation at *DHCR24* and BMI, or WC, whereas in this study we were able to identify a causal association between a DMP in *DHCR24* (cg25536676) and T2D, constituting this a novel finding. Functionally, this gene is involved in the reduction of sterol intermediates during the metabolism of cholesterol[90], and transgenic mice for this gene showed abnormal subcutaneous and mesenteric storage of lipids, reduced body size, and reduced levels of circulating cholesterol[251].

To my knowledge, there has been no report of an association between methylation at *STARD10* and T2D, BMI or other marker of adiposity. In this study, the DMP cg14045803 in *STARD10* was identified in suggestive causality with T2D in the reverse 2SMR. Since this signal represents a novel epigenetic finding, it is necessary to validate this association in an independent study. Functionally, *STARD10* is member of the steroidogenic acute regulatory protein (StAR)-related lipid transfer protein family (GeneCard, https://www.genecards.org), and GWAS studies have reported an association between a variant in this gene and risk of T2D[197]. In humans, it was shown that the risk allele in *STARD10* was

associated with impaired glucose-stimulated secretion of insulin, but a better conversion of proinsulin to insulin[197], while KO mice for StarD10 in β cells showed impaired insulin secretion, impaired glucose-stimulated Ca^{+} dynamics, but improved proinsulin: insulin ratio, as it was observed in humans. Opposite to this, over-expression of StarD10 was associated with improved glucose tolerance in the adult mice[197]. Thus, the study by Carrat *et al.*[197] suggested that genetic variation in *STARD10* was associated with higher risk of T2D via reduction in the expression of this gene in β cells.

Strengths and limitations of this study

One of the strengths of this study was the use of instruments derived from multiple GWAS meta-analyses to ensure that only the best associations were included as instruments for T2D. Because summary data was manually extracted, it ensured that only complete data was included to conduct the MR analysis. Furthermore, because ascertainment of causality between methylation and T2D was based on findings from an observational analysis (hypothesis driven approach), results of the MR were less prone to suffer from multiple testing bias. Adding on, this study addressed causality from the context of a SSMR analysis in ALSPAC, to the context of a 2SMR analysis, allowing to compare the strength of the associations enquired across analyses, and to identify similarities between estimates. Also, identifying signals replicated between the SSMR and the 2SMR, even when they were identified with borderline significance, allowed to build up evidence of causality for the replicated signal. Furthermore, because the 2SMR was assessed in the context of a bidirectional 2SMR, it was possible to discern true direction of effect for signals that were successfully analysed in both directions of the MR.

Another advantage of this study was the use of data from the GoDMC consortium to report summary statistics for the genotype-methylation association. In comparison to a previous dataset of around 742 middle-age females used to identify meQTL [109], the sample from GoDMC comprised more than 20,000 middle-age participants of European Caucasian origin, which in turn provided better power to identify strong genome-wide meQTL associations.

In terms of generating results, particularly for the bidirectional 2SMR, the use of MR-Base was one of the main advantages of this study, as it allowed to adequately process summary data across datasets to avoid problems of incorrect allele coding, incomplete data, and correlated instruments, as well as facilitating calculation of multiple estimates across different statistical methods when enquiring

different exposure-outcome associations. Because estimates were calculated in an automated process in MR-Base, it means that results can be easily replicated for posterior studies.

Even though summary statistics for the genotype-T2D association were extracted from different GWAS to include only the most significant associations, one limitation of this approach was disregarding possible overlap between samples across studies, especially when considering that relevant studies were extracted from a similar GWAS meta-analysis consortium (DIAGRAM). Another limitation of extracting genotype-T2D data from multiple studies, was the inclusion of estimates identified in samples from different ethnic backgrounds, which were not comparable with the characteristics of samples included in the genotype-methylation dataset (i.e. only European Caucasians in ALSPAC and GoDMC). Besides this, using summary data from multiple studies prevented the adequate control for the overlap of samples across the genotype-exposure and genotype-outcome datasets.

Another limitation of this study was the use of a single underpowered sample (ALSPAC) to conduct a 2SLS-IV analysis for observational signals identified in the meta-analysis, thus affecting the precision of the causal estimates obtained. In addition, the SSMR was only conducted in one direction of the association (T2D to methylation), without considering results for the reverse analysis.

One disadvantage of the method implemented in the forward 2SMR, was the use of ALSPAC as a second sample to extract genotype-outcome data. This study provided less power to identify strong estimates compared estimates obtained in DIAGRAM. Power imbalance between the two samples included in the forward 2SMR was not possible to be counteracted by using summary data from GoDMC, since the associations reported by this consortium, in relation to DMPs of interest, did not include SNPs that could be used as proxies for T2D-SNPs. In terms of limitations of the reverse 2SMR, this was the reduced availability of instruments for a subset of the DMPs of interest, most of them instrumented by a single SNP. As a result, not all the associations evaluated in the forward MR were represented in the reverse MR analysis, this preventing true assessment of direction of causality. Furthermore, because few instruments were available per DMP, almost no sensitivity analysis was applied in the reverse MR to account for potential bias due to heterogeneity or horizontal pleiotropy.

Future analyses

Additional analyses that will leverage evidence obtained in this study, will be the assessment of the SSMR in a larger sample, ideally using samples from all the cohorts included in the meta-analysis, providing these sample have availability of phenotype, genotype and methylation data. SSMR results obtained across studies can be then meta-analysed to report a stronger causal estimate for each DMP. To add onto results of the SSMR, it will be necessary to conduct the analysis in the reverse direction of the association, from methylation to T2D, by using a methylation score according to the number of instruments available per DMP of interest.

To further from results obtained in the 2SMR, it will be necessary to strengthen power of the instruments used to proxy methylation in the forward MR by meta-analysing SNP-CpG associations obtained across different studies. Because results of the reverse MR were obtained using methylation instruments reported by the largest consortium available to date for the study of the genetics of DNA methylation (GoDMC), building up evidence from the reverse MR will require using the latest dataset available from GoDMC, or a similar consortium.

In addition, signals detected in the bidirectional MR need to be validated in independent studies. Replicated signals could be taken forward for a longitudinal study of changes in methylation and T2D progression between the baseline (youth, middle-age) and the follow-up time-points (middle-age, elder). Furthermore, signals detected in this causal analysis can be combined in a weighted methylation score to (1) assess the risk of incident T2D in unaffected samples at higher risk of the disease by combining signals detected in the reverse MR (Meth → T2D), and (2) to estimate prevalence of T2D in the population by combining signals detected in the forward MR (T2D → Meth). The derived methylation score can also be applied to assess risk of other outcomes related to T2D macro- and microvascular complications.

Conclusion

Taking together results, the implementation of a bidirectional MR was important to disentangle true direction of effect in circumstances where observationally it is difficult to do so because the exposure measured (middle-age DNA methylation), is already affected by the effects of the outcome (prevalent T2D) due to reverse causation. Thus, the use of a bidirectional MR allowed to determine that in the prevalent state of T2D coexist methylation signals where variation in methylation is determinant of T2D, with signals where variation in methylation occurs as a consequence of the disease. Signals that were identified as happening *a priori* T2D were able to be detected cross-

sectionally because of the stability that these markers may have after disease onset. It was also demonstrated that some of the signals identified causally associated with T2D mapped to genes previously identified in epigenetic studies of BMI or WC (*PBX1, DHCR24*), in epigenetic studies of T2D in adipose tissue (*NISCH, PPAP2B*), some other genes were closely related with the biology of pancreatic cells (*TXNIP, PBX1*), and others with the metabolism of lipids (*DHCR24, PPAP2B*). Validation of signals identified in the bidirectional MR in additional studies will be necessary before considering them as candidate biomarkers of future risk of disease, or disease prevalence. Furthermore, assessment of the impact that these causal markers may have on the pathophysiology of T2D, will require further investigation of their association with the relevant outcomes of gene expression, glycaemic traits, metabolites, and macro and microvascular diabetes complications, ideally under a two-step 2SMR framework.

Chapter 8 Discussion

Type 2 diabetes is a complex multifactorial disease affecting an increasing number of people worldwide, with most of the population affected being middle-age adults. Multiple studies have investigated the influence of DNA methylation in T2D, but the causal role is still unknown. This was a major focus of the thesis.

Overall, the work presented in this thesis sought to explore the relationship between T2D and variation in DNA methylation. Two possible overarching hypotheses were explored; firstly, the possibility that T2D influences DNA methylation and that this provides an index of disease state; secondly, the possibility that DNA methylation variation arises as a consequence of genetic variation or other environmental or lifestyle exposures and this in turn alters risk of subsequent T2D. This latter hypothesis was also explored in relation to glycaemic traits including fasting glucose, fasting insulin, HbA1c and other traits that may precede overt disease but be indicative of early stage disease. The main findings from each chapter are summarized in Table 8-1.

8.1 Epigenetics of prevalent T2D in European samples: evidence from a meta-analysis

Findings

Initial efforts aimed to identify differences in DNA methylation associated with T2D. Modest sample sizes prompted the adoption of a strategy of meta-analysis to maximise available power to identify informative and robust associations.

In brief, a primary analysis was conducted in a subsample of adults in ALSPAC to identify signals in association with prevalent T2D as the exposure, considering that methylation in this dataset was measured at the time of disease occurrence. Only one signal was identified associated with hypomethylation of the *NFYC* locus in T2D cases versus controls. Further sensitivity analysis revealed that the association at this site was independent of BMI and other known risk factors for T2D (i.e. fasting glucose, HOMA-IR, and c-reactive protein). However, from the functional perspective, there was less evidence relating methylation at *NFYC* with the pathophysiology of T2D based on a comparison of meQTL data for this DMP with GWAS data for T2D, T2D-related traits and T2D complications, and based on the function of this gene.

Table 8-1 Summary of main findings from each chapter

Chapter	Main findings
4- Epigenetic analysis of prevalent T2D in ALSPAC	- Detection of a strong signal in association with prevalent T2D in adults in ALSPAC at the *NFYC* locus, which constitutes a novel finding based on recent EWAS repots in European samples. This association was independent of BMI and other common risk factors for T2D. - A second signal in suggestive association was detected at the *STARD10* gene. This locus has been reported in association with T2D based on GWAS data, but not on methylation data. - A DMR analysis provided evidence of several regions with differential methylation in T2D, most of them detected hypomethylated in T2D cases versus controls.
5-Meta-analysis of EWAS in glycaemic traits	- Strongest novel signal in association with fasting insulin and HOMA scores detected at a CpG in the *DDC* locus surpassing adjustment for BMI. - Confirmation of the association between *ABCG1* and fasting insulin and HOMA scores. - Generation of methylation scores for fasting insulin, HOMA scores and HbA1c, without identifying further variance in the trait relative to that captured by established risk factors.
6-Meta-analysis of EWAS in prevalent T2D among Europeans	- Confirmation of the signal at *TXNIP*, a strong marker of differential methylation in T2D that has been identified across different populations. - Identification of a novel signal at *HDAC4* and at two intergenic CpG sites, and confirmation of two other signals widely reported at *ABCG1* and *CPT1A,* based on a results of a sensitivity meta-analysis. - DMR analysis detected the strongest signal at the *ADCY7* locus in hypermethylation, and captured a second signal at the *SLCA15,* confirming a previous association.
7-Causality in DNA methylation and T2D	- A 2SLS MR analysis in ALSPAC revealed consistency in the direction of effect between the observational and the causal analysis, but it was underpowered to detect a significant causal effect. - The bidirectional MR allowed to distinguish differential methylation in T2D occurring as a consequence of the disease (*NISCH* and *TXNIP*), from that causally associated with an effect on T2D (*PPAP2B* and *DHCR24*).

Also, in ALSPAC, a second signal with borderline significance was detected at the *STARD10* gene, but this proved to be confounded by BMI based on a sensitivity analysis. *STARD10* has been previously reported in association with T2D based on GWAS data[197, 252], and in a recent causal analysis it was demonstrated that methylation at a CpG site in *STARD10* was in the causal pathway between genetic variation in this gene and levels of fasting proinsulin[95]. Knowing the role of *STARD10* as an intracellular transporter of lipids, and the influence of risk variants in this gene in the processing of proinsulin and in differential expression of *STARD10* in pancreatic islets of T2D-donors[197], I hypothesize that variation in methylation of the CpG site in *STARD10* might have further impact on the risk of T2D. Future studies are required to further explore this association, considering that the signal detected at this gene represents a novel association with T2D.

Additional examination of epigenetic data in ALSPAC involved a DMR analysis, which resulted in the identification of several regions in strong association with T2D, most of them hypomethylated in T2D cases versus controls. A functional exploration indicated that pathways enriched in differentially methylated sites within these regions were strongly related to the reabsorption of calcium and the synthesis of the parathyroid hormone. Disruption of these pathways has been associated with insulin resistance and impaired insulin-stimulated glucose transport in T2D patients[212]. Since validation of these regions using an independent method was not possible, results of the DMR analysis in ALSPAC were not taken forward for replication.

In an attempt to improve the reliability of signals detected at the single site level in ALSPAC, where the sample size was reduced limiting the ability to detect stronger signals that could be validated in an independent sample, a replication analysis was conducted in four other European cohorts. The individual EWAS in these studies revealed no overlap of top-ranking signals between studies, and only one signal surpassing significance was detected at the *ABCG1* locus in the LBC1936 study, a site that has been widely reported in association with T2D (incident and prevalent). Difference in the type of associations detected between studies might be related to study-specific differences in methylation in response to a different environment, differences in the way T2D was defined (i.e. HbA1c in LBC1936 and medical diagnosis in most of the other studies), difference in age distribution of participants across studies, particular comorbidities unaccounted for in the analysis, or to a combination of these factors. To overcome these limitations, and to increase the power of the signals detected in the individual studies, a meta-analysis was conducted.

The meta-analysis identified as the strongest association the signal at the *TXNIP* gene, and this association was independent of BMI and common risk factors. Thus, this study was able to confirm the association between T2D and *TXNIP* using a large sample, which to my knowledge represents the largest cross-sectional case-control study of epigenetics in T2D conducted to date in European samples. As it has been previously identified, T2D was associated with hypomethylation at the CpG in *TXNIP*, but only a modest effect was detected in this analysis, where on average T2D cases were 1.2% hypomethylated versus controls. The proportion of the variance in T2D explained by *TXNIP* was 2% according to data in ALSPAC. A second signal at the intergenic CpG site cg13826139 was detected with borderline significance in the meta-analysis. To my knowledge, this site is a novel association for T2D. As for *TXNIP*, T2D was associated with a decrease in methylation at cg13826139, and total variance in T2D explained by methylation at this CpG was 1.3%.

A second analysis excluding one of the cohorts based on methodological differences when conducting the EWAS, resulted in a larger number of associations surpassing epigenome-wide significance in the meta-analysis, identified at the previously noted loci in *TXNIP*, *ABCG1* and *CPT1A*, and in three novel loci at *HDAC4* (cg00144180) and in two intergenic CpG sites (cg16765088 and cg24704287). For the novel sites, it was identified that hypermethylation of *HDAC4* was associated with increased risk of T2D, while for the intergenic CpG sites hypermethylation was protective against T2D. Histone deacetylases (HDACs) are enzymes that remove acetyl groups from histone lysine residues, with positive impact on gene transcription[253]. Based on animal models, the overexpression of *HDAC4* leads to a reduction in β-cell mass, and clinical trials have been undertaken to test the utility of inhibitors of HDACs and activators of HDACs in the treatment of T2D[253]. Because methylation at *HDAC4* is a novel signal for T2D, it is important to further explore this association by validation using pyrosequencing, following-up results by *in vitro* functional analyses when possible. Combining results of the seven top-ranking signals detected across meta-analyses, these sites captured around 11.2% of the variance in T2D, although this was less than the variance explained by several established risk factors (21.55%), an expected result knowing the reduced number of signals detected surpassing statistical significance. Further investigation showed that most of these sites were associated with glycaemic traits and other clinical risk factors relevant to T2D.

Interestingly, when looking at differences in methylation in relation to glucose tolerance status, hypermethylation at most of these top CpG sites (except for *ABCG1*, *HDAC4*) was associated with an increase in glucose tolerance, and for some of them, methylation could also be used to distinguish between the prediabetes and the diabetic state (*ABCG1* and cg16765088). Thus, results in this study

were able to corroborate that *ABCG1* is a candidate marker to detect future risk of T2D and overt T2D, as has been shown in previous studies[62-64]. For the CpG in *CPT1A,* the association was exclusively detected between the prediabetic and the normoglycaemic state, indicating that this marker could be of particular use in distinguishing people with future risk of diabetes. A DMR analysis based on summary data from the meta-analysis allowed the identification of further loci in association with T2D, some of them in overlap with the region of previously detected CpG sites in *CPT1A* and *TXNIP*. The strongest region detected was associated with hypermethylation of the *ADCY7* gene in T2D cases versus controls. In a second DMR analysis using results from the sensitivity meta-analysis (i.e. excluding KORA), one DMR was identified in the region of the *SLC1A5* gene, including the CpG cg21766592, which has been previously reported in association with T2D by Kulkarni *et al.*[64]. The proportion of variance in T2D explained by the most clearly associated DMRs was less than the variance explained by the single CpG sites, and this could have been due to using average methylation across CpG sites in the region to estimate variance in the outcome. Altogether, results presented for the single CpG site analysis and the DMR analysis support the knowledge that methylation is associated with T2D, and that sites identified can be related with T2D via mechanisms closely linked to risk factors for the disease.

The loci identified were then subject to various forms of *in silico* interrogation to elicit information about their likely functional role, their relationship with gene expression data and their association with genetic variation. For instance, findings in blood samples suggested that methylation at *ABCG1*, *TXNIP* and *CPT1A* was inversely associated with gene expression in non-diabetic participants. For *ABCG1*, results in blood were consistent with reports from Kriebel *et al.*[88] and Chambers *et al.*[62] using non-diabetic samples. Methylation at *TXNIP* was also previously identified in association with gene expression in the liver[62]. Even though the enrichment analysis results were not fruitful, as they were most likely underpowered, some of the pathways identified were linked to the metabolism of lipids, insulin resistance, and the tumour necrosis factor (TNF) signalling pathway. These are all informative pathways in the pathophysiology of T2D, knowing that insulin resistance is one of the signatures of T2D, dysregulation in the metabolism of lipids is also a characteristic of T2D patients, and TNF is a cytokine produced in response to inflammation, which is generated in low-grades under hyperglycaemic conditions[253].

The use of GWAS and meQTL data from publicly available datasets, allowed the identification of some overlap between genetic variants for *TXNIP* and *ABCG1*, and genetic variation for specific glycaemic traits. However, none of the meQTL investigated overlapped with reported risk variants

for T2D according to results from a recent GWAS[30], and based on a list of the strongest SNPs in T2D identified across GWAS and selected to construct a polygenic risk score in this study. This lack of overlap between genetic and epigenetic loci for T2D is a result consistent with findings from Kulkarni *et al.*, who did not identify similarity between 51 methylation loci associated with T2D, and 63 genetic loci reported in a recent GWAS[64]; it is also consistent with the study by Hidalgo *et al.*[87], which did not identify an association between CpG sites and GWAS loci for insulin resistance. Despite this, other studies have identified ~20% overlap between GWAS loci and methylation loci for T2D[61, 254], but this was likely due to differences in the detection method used (MeDIP-Seq by Yuan *et al.* versus 450K array by Kulkarni *et al.* and this study), due to a small sample-size (<46 cases in Xu *et al.* and Yuan *et al.*), or to an incorrect adjustment for technical covariates (Xu *et al*. analysis was unadjusted for batch effects, cell composition and differences in probe type) and lack of replicability of results. As mentioned previously by Kulkarni, it is likely that independent genetic and epigenetic loci influence the risk of T2D, but it is also possible that their action is within the same or interacting biological pathways[64].

Strengths

The large sample-size included in this meta-analysis, with a representative number of T2D cases (n=496), was one of the advantages this study. Relative to other cross-sectional studies in prevalent T2D, like the one conducted by Kulkarni *et al*.[64] (n=179), Meeks *et al*.[68] (n=256), Soriano-Tárraga *et al*. (n=151) and Florath *et al*. (n=153), the larger number of cases incorporated in this study allowed the identification of various signals at the epigenome-wide level, which are good candidates for validation. Another strength was the inclusion in the meta-analysis of cohorts with similar distribution of important covariates (age, sex, BMI, FG), in addition to implementing a similar protocol to run the analyses across cohorts, which helped to reduce heterogeneity in the results. Furthermore, the implementation of a DMR analysis allowed me to reinforce findings from the single-site analysis, and to have a broader view of the effect of T2D on DNA methylation, but this analysis is still limited by the coverage of the 450K array. Adding to this, the use of different *in silico* approaches for functional exploration provided a better biological context of the mechanisms linking T2D with methylation.

Weaknesses

The inclusion of only Europeans in the meta-analysis was a likely weakness of this study, which could have lowered the generalizability of results in other populations. However, recent studies have demonstrated that methylation markers in T2D, especially those detected at *ABCG1*, *TXNIP* and

CPT1A, can be identified in populations of different ancestral origin. This means that it is also likely that some of the novel loci reported in this study, such as the signal at *HDAC4*, may also be important markers of T2D in other populations. To determine the generalizability of these results, further replication of novel signals in other populations will be required.

Another limitation was the inability to appraise the association between methylation and T2D by distinguishing between participants with adequate and poorly controlled glucose, and this analysis was not possible mainly because of sample size limitations. However, such analysis will be important to provide indicators of patients with adverse prognosis of the disease who could be at higher risk of developing T2D-associated complications. Additionally, the reduced number of incident cases of T2D (i.e. newly diagnosed T2D) in ALSPAC, prevented the study of this phenotype which is also of importance when looking for predictive markers of future liability of T2D. As the cohort continues to accrue middle-age participants (the study participants are now around age 30), more cases can be captured in future follow-ups, some of them likely to be incident rather than persistent cases of T2D.

Lastly, methylation is a tissue-specific marker, which is often sampled in blood rather than in target tissues relevant for the outcome of interest, in this case human pancreatic islets, skeletal muscle cells, liver or adipose tissue for T2D. In this study we partially addressed this limitation by comparing levels of methylation in blood with other target internal tissues using a publicly available dataset. However, DNA methylation in this dataset was available in a very small 'reference' sample of participants, and the level of correlation detected across tissues is potentially less informative of what would be observed if including in the comparison participants with T2D. Despite this, some studies have indicated that for specific markers such as *TXNIP,* differential methylation in blood can also be mirrored in other tissues including the liver[62], pancreatic islets and skeletal muscle[63], with consistent directionality with blood methylation.

In context

Various EWAS on T2D have been conducted in relation to prevalent T2D using cross-sectional studies across different populations (i.e. Europeans, Hispanic-Americans, Africans, Arabs), but this thesis represents the largest sample of T2D cases included in a meta-analysis of Europeans to date, which increases the power and reliability of the findings presented. Furthermore, most of the signals identified are consistent with the current literature in addition to some new associations that are candidates for validation. As it becomes clearer, markers identified in relation to incident T2D can also be captured in relation to prevalent T2D, indicating the stability of these signals once disease

has occurred. Therefore, biomarkers for future risk of disease can also be obtained from well-powered cross-sectional studies knowing the advantages that these studies have in terms of sample-size and cost versus longitudinal studies, which can give more information about the epigenetic dynamics of complex traits, but are also more expensive and difficult to conduct. As the number of exploratory EWAS in T2D continues to grow, it will be necessary to combine efforts across studies to enlarge the sample size included in current EWAS, and this might be possible by establishing consortium for the study of the Epigenetics in T2D, similar to what has been done to study the genetics of the disease in the GWAS era.

Options for the future

Future avenues for the epigenetic findings in T2D will be the validation of signals using alternative detection methods (i.e. pyrosequencing), and the replication of novel markers in other populations. In addition, the possibility of increasing the sample size analysed by including participants from different ethnicities needs to be explored, and this could improve the generalizability of the results. Furthermore, the possibility of testing differential methylation and expression across tissues should be explored. Lastly, the *in silico* analysis used to identify shared genetics between methylation, gene expression and T2D, another approach that could be pursued is the use of colocalization methods, which allow you to determine if the same variant affecting methylation also affects the trait of interest, indicating that variation in the traits is driven by a common causal effect[95, 255].

8.2 Epigenetics of glycaemic traits: evidence from a meta-analysis

Findings

Various EWAS using glycaemic traits measures across ALSPAC and SABRE allowed the identification of the noted locus at *ABCG1* (cg06500161), and the novel loci at *DDC* (cg18232548) and *UFM1* (cg19750657) in association with fasting insulin and HOMA-IR, two of them also associated with HOMA-B (*ABCG1* and *DDC*). The signal at *ABCG1* has been previously reported in association with various glycaemic traits including fasting glucose[46, 64, 88], 2-h glucose[88], fasting insulin[88, 256] and HOMA scores[88]. Direction of effect reported for *ABCG1* in this study is consistent with previous findings, showing that hypermethylation of *ABCG1* is related with increased levels of the glycaemic traits. Variation in methylation at *ABCG1* could be therefore an early indicator of future liability of T2D in healthy normoglycaemic participants, considering that hypermethylation at this site has also been reported in association with incident and prevalent T2D. Different from work by Hidalgo *et al*.[87], correlation between genetic variation at *ABCG1* and variation in the glycaemic traits was not

identified in this study. However, there was some evidence of overlap between meQTL for the CpG in *DDC* and GWAS variants for the phenotypes of interest, suggesting a mediating effect of methylation at this locus in the genetic effect on the glycaemic traits. Furthermore, shared genetics between methylation at *DDC* and gene expression of the nearby gene *FIGNL1* was also identified. *DDC* is an enzyme that participates in the decarboxylation of multiple compounds to dopamine, tryptamine and serotonin[186], and defects in this gene leads to deficiency in serotonin and catecholamine (i.e. neurotransmitters)[186]. At present, no other studies have reported an association between methylation at *DDC* and glycaemic traits, and GWAS data appoints some variants in this gene in relation to BMI and blood metabolite measurement, among others (GWAS catalog, https://www.ebi.ac.uk/gwas). Further investigation is required to demonstrate the replicability of this signal in other datasets, in addition to that of methylation at *UFM1* in relation to fasting insulin and HOMA-IR.

The use of DNA methylation scores to predict either exposure history or future disease risk is gaining considerable attention[257]. The use of DNA methylation scores constructed from glycaemic traits-associated loci was assessed but this proved to have little success, most likely due to the small amount of variance in the trait explained by the markers identified. Furthermore, variance captured by the score was not independent of common risk factors (i.e. BMI, sex, smoking), and this was an expected result knowing the association between the individual CpG sites in the score, and some of the risk factors included in the polygenic model. There were few instances in which including the score in the predictive model enhanced the performance of the model, and this was the case for the score of fasting insulin and the score of HbA1c.

Strengths

Instead of suing ALSPAC data alone, results were analysed in a second dataset, and summarized via meta-analysis, which provided higher power and reliability of the signals identified. Because access to individual level data from SABRE was provided, analyses were conducted in this dataset similar to what was done in ALSPAC for the multiple glycaemic traits. In addition, the association at specific traits only available in SABRE such as HbA1c and 2-h insulin was investigated. Another strength of this study was the construction of methylation scores to establish the amount of variance in the trait that was captured by the methylation hits, and their independence from common risk factors. The score analysis, to my knowledge, has not been yet implemented in published studies of epigenetic in glycaemic traits, which adds an extra value to this work.

Weaknesses

Despite the use of a meta-analysis to increase power, and despite including a representative sample to conduct the analysis (n=980 to 1384) relative to other studies on glycaemic traits (n=617 to 1440)[88], one major limitation was the use of datasets of different sex to estimate initial associations, knowing that differences in methylation are also driven by sex[42]. Therefore, more distinct associations were expected across datasets. One possible way to have overcome this issue was to use a random effect instead of a fixed-effect model in the meta-analysis to account for larger heterogeneity across datasets, but because the sample size for the analysis was small, and because random-effect models have less power compared to fixed-effect models in detecting associations for discovery purposes[145], this model was not selected to conduct further analyses. However, a random effect model will be of use if larger samples become available for this study. As with the meta-analysis of prevalent T2D, the use of only European samples restricted the generalizability of results obtained.

Options for the future

The opportunity to include larger samples through collaboration across studies needs to be considered to strengthen the ability to detect markers in disease-free participants that could be used to generate predictive scores of future risk of T2D. In addition, differential methylation for the glycaemic traits can be explored in additional target tissues for T2D, even though blood is the most accessible tissue for the measurement of these markers. Finally, it would be interesting to assess changes in gene expression in relation to differential methylation to contribute to the understanding of the mechanisms linking difference in methylation with variation in the glycaemic trait.

8.3 Summary of top signals identified across EWAS

Associations detected with epigenome-wide significance throughout this thesis are summarized in Table 8-2. Some of the CpG sites identified are well-established T2D or glycaemic traits-associated CpG sites (i.e. *ABCG1* with T2D, fasting insulin and HOMA-IR), while others are novel signals for the specific phenotype (i.e. *DDC* and *UFM1* with HOMA scores). Additional associations previously reported between top-ranking CpG sites and other metabolic, anthropometric and lifestyle traits (i.e. *TXNIP* and triglycerides), are also described in Table 8-2.

Table 8-2 Top-ranking CpG sites detected with epigenome-wide significance across the different EWAS conducted in this study.

Chapter	Top hit	Reported association	Previous evidence	Other traits
4	*NFYC* (cg15986668)	Prevalent T2D	Novel marker for T2D	Sustained maternal smoking and new-born methylation[258]
5	*ABCG1* (cg06500161)	Fasting insulin, HOMA-IR, HOMA-B	Kriebel *et al*.[88], Kulkarni *et al*.[64], Hidalgo *et al*.[87], Arpón *et al*.[259]	Fasting glucose[46, 88], BMI[89, 90, 213, 260-263], triglycerides[264-266], HDL [264-267] and the metabolic syndrome[268]
	DDC (cg18232548)	Fasting insulin, HOMA-IR, HOMA-B	Novel signal for the glycaemic traits	None
	UFM1 (cg19750657)	Fasting insulin, HOMA-IR	Novel signal for the glycaemic traits	BMI[89, 90, 260] and adiponectin[269]
6	*TXNIP* (cg19693031)	Prevalent T2D	Chambers *et al*.[62], Kulkarni *et al*.[64], Soriano-Tárraga *et al*.[65], Florath *et al.*[66], Walaszczyk *et al.*[46], Al Muftah *et al.*[67], Meeks *et al*.[68]	HbA1c[68], BMI[89], triglycerides[264, 265] and other fatty acids[270]
	ABCG1 (cg06500161)	Prevalent T2D	Chambers *et al*.[62], Kulkarni *et al*.[64], Walaszczyk *et al.*[46], Dayeh *et al*.[63], Al Muftah *et al.*[67]	In addition: waist-hip, waist-height ratio [267], waist circumference[213, 263, 267], hypertriglyceridemic waist (HTGW)†[271] phenotype, postprandial lipemia[272], and some metabolites[270]
	Intergenic (cg16765088)	Prevalent T2D	Novel signal for T2D	None
	CPT1A (cg00574958)	Prevalent T2D	Meeks *et al*.[68], Kulkarni *et al*.[64]	BMI[89, 90, 213, 260, 263], triglycerides [264, 265 266, 273], waist circumference[263], HTGW[271], postprandial lipemia[272], VLDL[270, 273, 274], LDL[274], adiponectin[269], the metabolic syndrome[275], and CVD risk[276]
	Intergenic (cg24704287)	Prevalent T2D	Novel signal for T2D	sTNFR‡[277], CV risk[278] and smoking*[279]
	HDAC4 (cg00144180)	Prevalent T2D	Novel signal for T2D	BMI[89]

†Hypertriglyceridemic waist (HTGW) phenotype. ‡Soluble Tumour Necrosis Factor Receptor 2 Levels. *Association identified when comparing current versus never smoking.

8.4 Exploring causality of DNA methylation and T2D

Findings

A comprehensive range of Mendelian randomization approaches were applied to the data in an attempt to ascertain the direction of any causal pathways between DNA methylation and T2D. Associations were causally appraised using a single sample MR and a bidirectional two sample MR. In general, results of the single sample MR were underpowered and unable to detect a causal association between T2D as the exposure and DNAm as the outcome, even though there was consistency in the direction of effect between the observational (meta-EWAS of T2D) and the causal

estimate. From the bidirectional MR, it was possible to disentangle the direction of causality for some of the DMPs detected in the meta-analysis. For instance, the strongest evidence of association in the forward MR (T2D→DNAm) was identified at the DMP in *NISCH* (cg00082384), which was detected with borderline significance in the observational analysis. In addition, there was some evidence that methylation at the well-known DMP in *TXNIP* occurred as a consequence of T2D and the undergoing metabolic abnormalities of the disease. For the reverse MR (DNAm→T2D), strongest evidence of causality was identified at the DMP in *PPAP2B* (cg11851382) and *DHCR24* (cg25536676), both captured with borderline significance in the observational analysis. Interestingly, for the DMP in *SPEG* (cg20456243), the association was observed in both directions of the MR analysis. However, comparing the strength of the estimates and the consistency in the direction of effect between the observational and the causal estimate, it was deemed that methylation in *SPEG* occurred as a consequence of T2D. Overall, results of the bidirectional MR revealed that in the prevalent state of the disease, changes in methylation that are determinant of the disease coexist with those that occur as a consequence of it. The reason why methylation changes that are happening *a priori* are still detected afterwards, may be because of the stability in the levels of methylation at these CpG sites once T2D occurs, and support of this is the identification of differential methylation at *TXNIP* and *ABCG1* in studies of incident T2D[62, 63] and prevalent T2D[64-66, 68].

Strengths

Implementation of a polygenic risk score in the single sample MR allowed me to capture more variation in the outcome than the single SNPs, and it increased the precision of the genotype-exposure estimate. Furthermore, a polygenic risk score is particularly useful for modifiable exposures where the genetic component only accounts for a small proportion of the total variation in the trait, as it is the case for T2D (~15%). Another advantage of this study was to incorporate a two sample MR, which improved power to detect causal effects in the association between methylation and T2D. Adding to this, was the implementation of a reverse MR analysis which allowed me to determine the true direction of the causal effect. Different MR methods and tests were applied to account for possible issues in MR analyses such as pleiotropy, presence of invalid instruments, weak instruments or directionality issues.

An additional strength was the identification of genetic instruments for DMPs of interest using data from the GoDMC consortium, which to date represents the largest study of the genetics of DNA methylation, allowing identification of meQTL with enough strength to be used in MR analyses. Furthermore, different *in silico* functional explorations were implemented using results of the causal

analysis to determine the biological implications of these associations in relation to the pathophysiology of T2D.

Weaknesses

A polygenic risk score implemented in the single sample MR analysis increased the estimation of the IV-exposure association. The drawback of this, in the context of MR studies, is that it can also increase the chances of introducing horizontal pleiotropy. Horizontal pleiotropy occurs when the genetic instrument is associated with the outcome through pathways that do not include the exposure of interest, thus violating one of the MR assumptions.

Another limitation was the implementation of a risk score generated in ALSPAC to determine causality for observational associations detected in the meta-analysis, which likely rendered more imprecise estimates in the 2SLS MR analysis. Ideally, the 2SLS MR should be conducted separately in the different cohorts, combining results of the causal estimates across the cohorts via meta-analysis; this approach will likely increase the power to detect causal effects in the 2SLS MR. However, this type of analysis is more difficult to conduct and normally relies on the availability of genotype, methylation, and phenotype data in the same sample, which limits the size of the sample to include in the causal analysis.

A further limitation was the use of two samples with different detection strengths to conduct the two sample MR. The first sample was DIAGRAM and the second sample was ALSPAC, knowing upfront that estimates from ALSPAC will be less powered than those in DIAGRAM, which could have reduced the ability to detect strong causal estimates in the forward 2SMR. For the reverse MR, the only limitation was the reduced number of instruments identified per DMP. As a result, not all the DMPs of interest in the forward MR were analysed in the reverse MR, and for those that were analysed, the number of sensitivity tests that were applied to account for heterogeneity and pleiotropy was limited.

In context

Most of the studies conducted to date in T2D are observational studies, but none of them have appraised causality or the direction of the causal effect for the identified associations. Despite this, there are examples of causal analyses for epigenetic studies in BMI[89, 90, 98], some of them reporting signals that coincide with those observed in epigenetic studies of T2D. In addition, studies by Olsson and colleagues have applied causal inference tests to determine the mediating role of DNA methylation in the genetic association with gene expression and insulin secretion using meQTL

identified in human pancreatic islets of T2D donors[106]. However, the present study is the first one that investigates causality and direction of causality directly at the DMPs identified in the observational analysis, which offers the possibility to prioritize DMPs to be used for the early detection and treatment of T2D, and to select those that can be used as indicators of disease status.

Options for the future

Validation of the associations detected in this causal analysis will require the use of larger datasets to identify stronger genetic associations when DNA methylation is regarded as the outcome. One approach to do this is to combine genotype-methylation estimates across different cohorts and to use the meta-analysed summary data to conduct the MR analysis. In addition, as larger consortiums for the study of the genetics of DNA methylation become available, it is possible to include more instruments for methylation when this is regarded as the exposure in the MR analysis. The use of more instruments for methylation exposures (meQTL) would also allow me to apply further sensitivity analyses in the MR, and to implement methylation scores (akin allele scores) to strengthen causal inference. By consolidating results of the bidirectional MR, stronger methylation markers can be taken forward to estimate their association with T2D-secondary outcomes in the context of a two-step two sample MR. In addition, stronger markers can be followed-up to determine longitudinal changes in methylation and the risk of T2D, or to generate scores of methylation to determine future risk of T2D in disease-free participants, or to detect prevalence of T2D in undiagnosed participants.

8.5 Conclusion

The findings of this thesis support the knowledge that DNA methylation is an important factor in the study of T2D, with variation that can be captured either prior the onset of the disease using glycaemic traits in normoglycaemic participants, or after disease occurrence. Furthermore, due to the stability in the levels of methylation for some of the strongest markers of T2D, difference in methylation in the prevalent state of the disease can mirror difference in methylation prior disease onset, indicating that cross-sectional case controls studies are still an important approach to detect methylation marks for the early detection of T2D. The causal analysis further supported this concept by identifying that in prevalent T2D variation in methylation that is both a determinant of the disease and a consequence of it, coexist. Additional studies are required to validate novel findings of this study, and to replicate validated signals in other populations.

8.6 Plan for publication of research findings

The next steps in developing and disseminating the contents of this thesis will involve preparing various parts of the thesis for publication. Four manuscripts are planned:

1. *A review of current evidence for a role of DNA methylation variation in T2D*. This will encompass much of the literature reviewed in the introductory chapter of this thesis. An invitation to complete a review on this topic has been accepted in the journal of "Current Genetic Medicine Reports".
2. *An EWAS meta-analysis of T2D in 4 cohort studies*. Building on existing, less comprehensive EWAS, this will report novel findings from the large scale EWAS described in chapter 6 of this thesis.
3. *An EWAS of glycaemic traits*. This will report the novel findings reported in chapter 5 of this thesis.
4. Applying causal inference methods to decipher the direction of association between T2D and DNA methylation: A Mendelian randomization study. This will lay out the comprehensive approaches taken in this thesis (Chapter 7) to assert causality and direction of the postulated causal relationships. Where possible it will draw upon additional data sources to enhance power, especially the 'forward MR' analyses where DNA methylation is the outcome measured and data are currently limited in comparison to the large sample sizes available on T2D cases and controls.

References

1. Nolan CJ, Damm P, Prentki M. Type 2 diabetes across generations: from pathophysiology to prevention and management. The Lancet.378(9786):169-81.

2. American Diabetes Association (ADA). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. Diabetes Care. 2018;41(Supplement 1):S13-S27.

3. Hameed I, Masoodi SR, Mir SA, Nabi M, Ghazanfar K, Ganai BA. Type 2 diabetes mellitus: From a metabolic disorder to an inflammatory condition. World Journal of Diabetes. 2015;6(4):598-612.

4. Relton CL, Davey Smith G. Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. PLOS Medicine. 2010;7(10):e1000356.

5. Perry JRB, Voight BF, Yengo L, Amin N, Dupuis J, Ganser M, et al. Stratifying Type 2 Diabetes Cases by BMI Identifies Genetic Risk Variants in LAMA1 and Enrichment for Risk Variants in Lean Compared to Obese Cases. PLOS Genetics. 2012;8(5):e1002741.

6. Polonsky KS, Sturis J, Bell GI. Non-Insulin-Dependent Diabetes Mellitus — A Genetically Programmed Failure of the Beta Cell to Compensate for Insulin Resistance. New England Journal of Medicine. 1996;334(12):777-83.

7. WHO. Global report on diabetes2016 18/12/2018:[1-84 pp.].

8. Rathmann W, Scheidt-Nave C, Roden M, Herder C. Type 2 diabetes: prevalence and relevance of genetic and acquired factors for its prediction. Dtsch Arztebl Int. 2013;110(19):331-7.

9. Wilmot E, Idris I. Early onset type 2 diabetes: risk factors, clinical impact and management. Therapeutic Advances in Chronic Disease.5(6):234-44.

10. American Diabetes A. Diagnosis and classification of diabetes mellitus. Diabetes care. 2010;33 Suppl 1(Suppl 1):S62-S9.

11. Kautzky-Willer A, Harreiter J, Pacini G. Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. Endocr Rev. 2016;37(3):278-316.

12. Pfutzner A, Kunt T, Hohberg C, Mondok A, Pahler S, Konrad T, et al. Fasting intact proinsulin is a highly specific predictor of insulin resistance in type 2 diabetes. Diabetes Care. 2004;27(3):682-7.

13. Pfützner A, Forst T. Elevated Intact Proinsulin Levels Are Indicative of Beta-Cell Dysfunction, Insulin Resistance, and Cardiovascular Risk: Impact of the Antidiabetic Agent Pioglitazone. Journal of Diabetes Science and Technology. 2011;5(3):784-93.

14. International Diabetes Federation Diabetes Atlas. 2017 01/01/19:[1-145 pp.]. Available from: https://www.idf.org/e-library/epidemiology-research/diabetes-atlas.

15. 3. Comprehensive Medical Evaluation and Assessment of Comorbidities: Standards of Medical Care in Diabetes—2018. Diabetes Care. 2018;41(Supplement 1):S28.

16. Wong E, Backholer K, Gearon E, Harding J, Freak-Poli R, Stevenson C, et al. Diabetes and risk of physical disability in adults: a systematic review and meta-analysis. Lancet Diabetes Endocrinol. 2013;1(2):106-14.

17. Hattersley AT, Tooke JE. The fetal insulin hypothesis: an alternative explanation of the association of low bir thweight with diabetes and vascular disease. The Lancet. 1999;353(9166):1789-92.

18. Hattersley AT, Beards F, Ballantyne E, Appleton M, Harvey R, Ellard S. Mutations in the glucokinase gene of the fetus result in reduced birth weight. Nature Genetics. 1998;19:268.

19. Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, et al. Genome-wide associations for birth weight and correlations with adult disease. Nature. 2016;538:248.

20. Crowther CA, Hiller JE, Moss JR, McPhee AJ, Jeffries WS, Robinson JS. Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. N Engl J Med. 2005;352(24):2477-86.

21. Landon MB, Spong CY, Thom E, Carpenter MW, Ramin SM, Casey B, et al. A Multicenter, Randomized Trial of Treatment for Mild Gestational Diabetes. New England Journal of Medicine. 2009;361(14):1339-48.

22. Ma R, Tong P. Epidemiology of Type 2 Diabetes. In: Holt R, Cockram C, Flyvbjerg A, Goldstein B, editors. Textbook of Diabetes. Fifth ed. USA: Wiley-Blackwell; 2017. p. 1104

23. Li G, Zhang P, Wang J, Gregg EW, Yang W, Gong Q, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet. 2008;371(9626):1783-9.

24. Li G, Zhang P, Wang J, An Y, Gong Q, Gregg EW, et al. Cardiovascular mortality, all-cause mortality, and diabetes incidence after lifestyle intervention for people with impaired glucose tolerance in the Da Qing Diabetes Prevention Study: a 23-year follow-up study. Lancet Diabetes Endocrinol. 2014;2(6):474-80.

25. Salpeter SR, Buckley NS, Kahn JA, Salpeter EE. Meta-analysis: metformin treatment in persons at risk for diabetes mellitus. Am J Med. 2008;121(2):149-57.e2.

26. NCD-RisC. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. The Lancet. 2016;387(10027):1513-30.

27. Rathmann W, Strassburger K, Heier M, Holle R, Thorand B, Giani G, et al. Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. Diabet Med. 2009;26(12):1212-9.

28. Raciti GA, Longo M, Parrillo L, Ciccarelli M, Mirra P, Ungaro P, et al. Understanding type 2 diabetes: from genetics to epigenetics. Acta Diabetol. 2015;52(5):821-7.

29. Keating BJ. Advances in risk prediction of type 2 diabetes: integrating genetic scores with Framingham risk models. Diabetes. 2015;64(5):1495-7.

30. Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46(3):234-44.

31. Alam F, Islam MA, Gan S, Mohamed M, Sasongko TH. DNA Methylation: An Epigenetic Insight into Type 2 Diabetes Mellitus2016.

32. Haffner SM. Epidemiology of type 2 diabetes: risk factors. Diabetes Care.21 Suppl 3:C3-6.

33. Swerdlow DI. Mendelian Randomization and Type 2 Diabetes. Cardiovasc Drugs Ther. 2016;30(1):51-7.

34. Clausen TD, Mathiesen ER, Hansen T, Pedersen O, Jensen DM, Lauenborg J, et al. High prevalence of type 2 diabetes and pre-diabetes in adult offspring of women with gestational diabetes mellitus or type 1 diabetes: the role of intrauterine hyperglycemia. Diabetes Care. 2008;31(2):340-6.

35. Chen H, Simar D, Morris MJ. Hypothalamic neuroendocrine circuitry is programmed by maternal obesity: interaction with postnatal nutritional environment. PloS one. 2009;4(7):e6259.

36. Pinney SE, Simmons RA. Epigenetic mechanisms in the development of type 2 diabetes. Trends in Endocrinology & Metabolism. 2010;21(4):223-9.

37. Musso G, Gambino R, Cassader M. Obesity, diabetes, and gut microbiota: the hygiene hypothesis expanded? Diabetes Care. 2010;33(10):2277-84.

38. Agardh E, Allebeck P, Hallqvist J, Moradi T, Sidorchuk A. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. Int J Epidemiol. 2011;40(3):804-18.

39. Luo J, Rossouw J, Tong E, Giovino GA, Lee CC, Chen C, et al. Smoking and diabetes: does the increased risk ever go away? American journal of epidemiology. 2013;178(6):937-45.

40. Pan A, Wang Y, Talaei M, Hu FB, Wu T. Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. Lancet Diabetes Endocrinol. 2015;3(12):958-67.

41. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529-41.

42. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. Nature Reviews Genetics. 2013;14:585.

43. Michels K, B. . Considerations in the Design, Conduct, and Interpretation of Studies in Epigenetic Epidemiology. In: Michels K, B. , editor. Epigenetic Epidemiology. 2012 Edition ed: Springer; 2012. p. 460.

44. Breton CV, Marsit CJ, Faustman E, Nadeau K, Goodrich JM, Dolinoy DC, et al. Small-Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. Environ Health Perspect. 2017;125(4):511-26.

45. Barault L, Rancourt R, C. Laboratory Methods in Epigenetic Epidemiology. In: Michels K, B. , editor. Epigenetic Epidemiology. 2012 Edition ed: Springer; 2012. p. 460.

46. Walaszczyk E, Luijten M, Spijkerman AMW, Bonder MJ, Lutgers HL, Snieder H, et al. DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA1c levels: a systematic review and replication in a case-control sample of the Lifelines study. Diabetologia. 2018;61(2):354-68.

47. Zou L, Yan S, Guan X, Pan Y, Qu X. Hypermethylation of the PRKCZ Gene in Type 2 Diabetes Mellitus. Journal of Diabetes Research. 2013;2013:4.

48. Gu T, Gu HF, Hilding A, Sjoholm LK, Ostenson CG, Ekstrom TJ, et al. Increased DNA methylation levels of the insulin-like growth factor binding protein 1 gene are associated with type 2 diabetes in Swedish men. Clin Epigenetics. 2013;5(1):21.

49. Gu HF, Gu T, Hilding A, Zhu Y, Kärvestedt L, Ostenson C-G, et al. Evaluation of IGFBP-7 DNA methylation changes and serum protein variation in Swedish subjects with and without type 2 diabetes. Clin Epigenetics. 2013;5(1):20-.

50. Canivell S, Ruano EG, Sisó-Almirall A, Kostov B, González-de Paz L, Fernandez-Rebollo E, et al. Differential Methylation of TCF7L2 Promoter in Peripheral Blood DNA in Newly Diagnosed, Drug-Naïve Patients with Type 2 Diabetes. PloS one. 2014;9(6):e99310.

51. Ronn T, Ling C. DNA methylation as a diagnostic and therapeutic target in the battle against Type 2 diabetes. Epigenomics. 2015;7(3):451-60.

52. Gemma C, Sookoian S, Dieuzeide G, Garcia SI, Gianotti TF, Gonzalez CD, et al. Methylation of TFAM gene promoter in peripheral white blood cells is associated with insulin resistance in adolescents. Mol Genet Metab. 2010;100(1):83-7.

53. Davegårdh C, García-Calzón S, Bacos K, Ling C. DNA methylation in the pathogenesis of type 2 diabetes in humans. Molecular Metabolism. 2018.

54. Yang BT, Dayeh TA, Kirkpatrick CL, Taneera J, Kumar R, Groop L, et al. Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA1c levels in human pancreatic islets. Diabetologia. 2011;54(2):360-7.

55. Yang BT, Dayeh TA, Volkov PA, Kirkpatrick CL, Malmgren S, Jing X, et al. Increased DNA methylation and decreased expression of PDX-1 in pancreatic islets from patients with type 2 diabetes. Mol Endocrinol. 2012;26(7):1203-12.

56. Ling C, Del Guerra S, Lupi R, Ronn T, Granhall C, Luthman H, et al. Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion. Diabetologia. 2008;51(4):615-22.

57. Hall E, Dayeh T, Kirkpatrick CL, Wollheim CB, Dekker Nitert M, Ling C. DNA methylation of the glucagon-like peptide 1 receptor (GLP1R) in human pancreatic islets. BMC Med Genet. 2013;14:76.

58. Kulkarni SS, Salehzadeh F, Fritz T, Zierath JR, Krook A, Osler ME. Mitochondrial regulators of fatty acid metabolism reflect metabolic dysfunction in type 2 diabetes mellitus. Metabolism. 2012;61(2):175-85.

59. Gillberg L, Ling C. The potential use of DNA methylation biomarkers to identify risk and progression of type 2 diabetes. Front Endocrinol (Lausanne). 2015;6:43.

60. Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. Human Molecular Genetics. 2012;21(2):371-83.

61. Yuan W, Xia Y, Bell CG, Yet I, Ferreira T, Ward KJ, et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. Nat Commun. 2014;5:5719.

62. Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol. 2015;3(7):526-34.

63. Dayeh T, Tuomi T, Almgren P, Perfilyev A, Jansson P-A, de Mello VD, et al. DNA methylation of loci within ABCG1 and PHOSPHO1 in blood DNA is associated with future type 2 diabetes risk. Epigenetics. 2016;11(7):482-8.

64. Kulkarni H, Kos MZ, Neary J, Dyer TD, Kent JW, Jr., Goring HH, et al. Novel epigenetic determinants of type 2 diabetes in Mexican-American families. Hum Mol Genet. 2015;24(18):5330-44.

65. Soriano-Tarraga C, Jimenez-Conde J, Giralt-Steinhauer E, Mola-Caminal M, Vivanco-Hidalgo RM, Ois A, et al. Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. Hum Mol Genet. 2016;25(3):609-19.

66. Florath I, Butterbach K, Heiss J, Bewerunge-Hudler M, Zhang Y, Schöttker B, et al. Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. Diabetologia. 2016;59(1):130-8.

67. Al Muftah WA, Al-Shafai M, Zaghlool SB, Visconti A, Tsai P-C, Kumar P, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clin Epigenetics. 2016;8:13.

68. Meeks KAC, Henneman P, Venema A, Addo J, Bahendeka S, Burr T, et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. Int J Epidemiol. 2018.

69. Simar D, Versteyhe S, Donkin I, Liu J, Hesson L, Nylander V, et al. DNA methylation is altered in B and NK lymphocytes in obese and type 2 diabetic human. Metabolism. 2014;63(9):1188-97.

70. Ribel-Madsen R, Fraga MF, Jacobsen S, Bork-Jensen J, Lara E, Calvanese V, et al. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. PloS one. 2012;7(12):e51302.

71. Nilsson E, Jansson PA, Perfilyev A, Volkov P, Pedersen M, Svensson MK, et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. Diabetes. 2014;63(9):2962-76.

72. You D, Nilsson E, Tenen DE, Lyubetskaya A, Lo JC, Jiang R, et al. Dnmt3a is an epigenetic mediator of adipose insulin resistance. Elife. 2017;6.

73. Rönn T, Volkov P, Gillberg L, Kokosar M, Perfilyev A, Jacobsen AL, et al. Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. Human Molecular Genetics. 2015;24(13):3792-813.

74. Agha G, Houseman EA, Kelsey KT, Eaton CB, Buka SL, Loucks EB. Adiposity is associated with DNA methylation profile in adipose tissue. Int J Epidemiol. 2015;44(4):1277-87.

75. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet. 2014;383(9933):1990-8.

76. Ronn T, Volkov P, Davegardh C, Dayeh T, Hall E, Olsson AH, et al. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLoS Genet. 2013;9(6):e1003572.

77. Gillberg L, Perfilyev A, Brons C, Thomasen M, Grunnet LG, Volkov P, et al. Adipose tissue transcriptomics and epigenomics in low birthweight men and controls: role of high-fat overfeeding. Diabetologia. 2016;59(4):799-812.

78. Volkmar M, Dedeurwaerder S, Cunha DA, Ndlovu MN, Defrance M, Deplus R, et al. DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. Embo j. 2012;31(6):1405-26.

79. Dayeh T, Volkov P, Salo S, Hall E, Nilsson E, Olsson AH, et al. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS Genet. 2014;10(3):e1004160.

80. Daneshpajooh M, Bacos K, Bysani M, Bagge A, Ottosson Laakso E, Vikman P, et al. HDAC7 is overexpressed in human diabetic islets and impairs insulin secretion in rat islets and clonal beta cells. Diabetologia. 2017;60(1):116-25.

81. Volkov P, Bacos K, Ofori JK, Esguerra JLS, Eliasson L, Ronn T, et al. Whole-Genome Bisulfite Sequencing of Human Pancreatic Islets Reveals Novel Differentially Methylated Regions in Type 2 Diabetes Pathogenesis. Diabetes. 2017;66(4):1074-85.

82. Nitert MD, Dayeh T, Volkov P, Elgzyri T, Hall E, Nilsson E, et al. Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. Diabetes. 2012;61(12):3322-32.

83. Barres R, Osler ME, Yan J, Rune A, Fritz T, Caidahl K, et al. Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density. Cell Metab. 2009;10(3):189-98.

84. Nilsson E, Matte A, Perfilyev A, de Mello VD, Kakela P, Pihlajamaki J, et al. Epigenetic Alterations in Human Liver From Subjects With Type 2 Diabetes in Parallel With Reduced Folate Levels. J Clin Endocrinol Metab. 2015;100(11):E1491-501.

85. Shoelson SE, Lee J, Goldfine AB. Inflammation and insulin resistance. J Clin Invest. 2006;116(7):1793-801.

86. Kirchner H, Sinha I, Gao H, Ruby MA, Schonke M, Lindvall JM, et al. Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. Mol Metab. 2016;5(3):171-83.

87. Hidalgo B, Irvin MR, Sha J, Zhi D, Aslibekyan S, Absher D, et al. Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study. Diabetes. 2014;63(2):801-7.

88. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, et al. Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. PloS one. 2016;11(3):e0152314.

89. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541(7635):81-6.

90. Mendelson MM, Marioni RE, Joehanes R, Liu C, Hedman AK, Aslibekyan S, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. PLoS Med. 2017;14(1):e1002215.

91. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705-19.

92. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Meth. 2013;10(10):949-55.

93. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.

94. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133-63.

95. Richardson TG, Haycock PC, Zheng J, Timpson NJ, Gaunt TR, Davey Smith G, et al. Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. Human Molecular Genetics. 2018:ddy210-ddy.

96. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). International Journal of Epidemiology. 2015;44(4):1181-90.

97. Elliott HR, Shihab HA, Lockett GA, Holloway JW, McRae AF, Smith GD, et al. The Role of DNA Methylation in Type 2 Diabetes Aetiology – Using Genotype as a Causal Anchor. Diabetes. 2017.

98. Richmond RC, Sharp GC, Ward ME, Fraser A, Lyttleton O, McArdle WL, et al. DNA Methylation and BMI: Investigating Identified Methylation Sites at HIF3A in a Causal Framework. Diabetes. 2016;65(5):1231-44.

99. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature genetics. 2008;40(5):638-45.

100. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.

101. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012;44(9):981-90.

102. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.

103. Kwak SH, Park KS. Recent progress in genetic and epigenetic research on type 2 diabetes. Experimental & Molecular Medicine. 2016;48(3):e220.

104. Ling C, Pasquali L. Epigenetics in Type 2 Diabetes. In: Florez JC, editor. The Genetics of Type 2 Diabetes and Related Traits: Biology, Physiology and Translation. Cham: Springer International Publishing; 2016. p. 241-58.

105. Volkov P, Olsson AH, Gillberg L, Jørgensen SW, Brøns C, Eriksson K-F, et al. A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits. PloS one. 2016;11(6):e0157776.

106. Olsson AH, Volkov P, Bacos K, Dayeh T, Hall E, Nilsson EA, et al. Genome-Wide Associations between Genetic and Epigenetic Variation Influence mRNA Expression and Insulin Secretion in Human Pancreatic Islets. PLOS Genetics. 2014;10(11):e1004735.

107. Dayeh TA, Olsson AH, Volkov P, Almgren P, Rönn T, Ling C. Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. Diabetologia. 2013;56(5):1036-46.

108. Ling C, Poulsen P, Simonsson S, Ronn T, Holmkvist J, Almgren P, et al. Genetic and epigenetic factors are associated with expression of respiratory chain component NDUFB6 in human skeletal muscle. J Clin Invest. 2007;117(11):3427-35.

109. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.

110. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics. 2016;49:131.

111. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2013;42(1):111-27.

112. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97-110.

113. Levy JC, Matthews DR, Hermans MP. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. Diabetes Care. 1998;21(12):2191-2.

114. Li CL, Tsai ST, Chou P. Relative role of insulin resistance and beta-cell dysfunction in the progression to type 2 diabetes--The Kinmen Study. Diabetes Res Clin Pract. 2003;59(3):225-32.

115. Bath SC, Steer CD, Golding J, Emmett P, Rayman MP. Effect of inadequate iodine status in UK pregnant women on cognitive outcomes in their children: results from the Avon Longitudinal Study of Parents and Children (ALSPAC). Lancet. 2013;382(9889):331-7.

116. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333-46.

117. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692-702.

118. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018.

119. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

120. Association AD. 5. Prevention or Delay of Type 2 Diabetes. Diabetes Care. 2017;40(Supplement 1):S44.

121. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one. 2013;8(5):e63812.

122. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6(1):4.

123. Holle R, Happich M, Lowel H, Wichmann HE. KORA--a research platform for population based health research. Gesundheitswesen. 2005;67 Suppl 1:S19-25.

124. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. International Journal of Epidemiology. 2012;41(6):1576-84.

125. Hofman A, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, Ikram MA, et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol. 2013;28(11):889-926.

126. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. Eur J Epidemiol. 2017;32(9):807-50.

127. Tillin T, Forouhi NG, McKeigue PM, Chaturvedi N. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. International Journal of Epidemiology. 2012;41(1):33-42.

128. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome Research. 2014;24(11):1725-33.

129. Roche Diagnostics RHMPa. HbA1c Tina-quant Heamogobin A1c III. 2011. p. 1-6.

130. Chaturvedi N, McKeigue PM, Marmot MG. Relationship of glucose intolerance to coronary risk in Afro-Caribbeans compared with Europeans. Diabetologia. 1994;37(8):765-72.

131. Quinn GP, Keough MJ. Multidimensional scaling and cluster analysis. Experimental Design and Data Analysis for Biologists. Cambridge: Cambridge University Press; 2002. p. 473-93.

132. Evert S, editor Distributional semantics in R with the wordspace package. The 25th International Conference on Computational Linguistics: System Demonstrations; 2014; Dublin, Ireland.

133. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016.

134. Venables WNR, B.D. . Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.

135. Sharp GC, Arathimos R, Reese SE, Page CM, Felix J, Küpers LK, et al. Maternal alcohol consumption and offspring DNA methylation: findings from six general population-based birth cohorts. Epigenomics. 2018;10(1):27-42.

136. Tukey JW. Exploratory Data Analysis: Addison-Wesley Publishing Company; 1977.

137. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.

138. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology. 2012;41(1):200-9.

139. Naeem H, Wong NC, Chatterton Z, Hong MKH, Pedersen JS, Corcoran NM, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics. 2014;15(1):1-15.

140. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. 2014.

141. Simko TWV. R package "corrplot": Visualization of a Correlation Matrix Version 0.84 ed2017.

142. Warnes GR, Bolker, B., Bonebakker, l., Gentleman, R., Wolfgang Huber A. L., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M . & B. Venables. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1 ed2015.

143. Van der Most P KL, Snieder H and Nolte I QCEWAS: automated quality control of results of epigenome-wide association studies 2016 [R package version 1.1-0:[Available from: https://CRAN.R-project.org/package=QCEWAS.

144. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.

145. Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379-89.

146. Viechtbauer W. Conducting Meta-Analyses in R with The metafor Package. Journal of Statistical Software. 2010;36:1-48.

147. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin. 2015;8(1):6.

148. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinformatics. 2012;28(22):2986-8.

149. Breeze Charles E, Paul Dirk S, van Dongen J, Butcher Lee M, Ambrose John C, Barrett James E, et al. eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. Cell Reports. 2016;17(8):2137-50.

150. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics. 2016;32(4):587-9.

151. Halachev K, Bast H, Albrecht F, Lengauer T, Bock C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biology. 2012;13(10):R96-R.

152. ENCODE project Common cell Lines [updated 09-03-2012. Available from: https://www.genome.gov/26524238/encode-project-common-cell-types/.

153. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115-21.

154. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846-7.

155. Hayashi T, Urayama O, Kawai K, Hayashi K, Iwanaga S, Ohta M, et al. Laughter regulates gene expression in patients with type 2 diabetes. Psychother Psychosom. 2006;75(1):62-5.

156. Matone A, Derlindati E, Marchetti L, Spigoni V, Dei Cas A, Montanini B, et al. Identification of an early transcriptomic signature of insulin resistance and related diseases in lymphomonocytes of healthy subjects. PloS one. 2017;12(8):e0182559.

157. Hansen KD. IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. R package version 0.2.1 ed.

158. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32(2):286-8.

159. Slieker RC, Bos SD, Goeman JJ, Bovée JVMG, Talens RP, van der Breggen R, et al. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. Epigenetics & Chromatin. 2013;6:26-.

160. Vasiliki Lagou RM, Jouke-Jan J Hottenga, et al. Fasting glucose and Fasting insulin sex-specific and sex-differentiated GWAS meta-analysis public data release May 2018: Wellcome Sanger Institute; 2018 [Available from: ftp://ftp.sanger.ac.uk/pub/magic/MAGIC-sex_dimorphic_fasting_glucose_insulin_README.pdf.

161. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature genetics. 2010;42(2):105-16.

162. Soranzo N, Sanna S, Wheeler E, Gieger C, Radke D, Dupuis J, et al. Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. Diabetes. 2010;59(12):3229-39.

163. GTEx. The Genotype-Tissue Expression (GTEx) project 2017 [updated 2017. Available from: https://www.gtexportal.org.

164. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics. 2014;23(R1):R89-R98.

165. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7:e34408.

166. Song Y, Yeung E, Liu A, Vanderweele TJ, Chen L, Lu C, et al. Pancreatic beta-cell function and type 2 diabetes risk: quantify the causal effect using a Mendelian randomization approach based on meta-analyses. Hum Mol Genet. 2012;21(22):5010-8.

167. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. International Journal of Epidemiology. 2012;41(1):161-76.

168. Caramaschi D, Sharp GC, Nohr EA, Berryman K, Lewis SJ, Davey Smith G, et al. Exploring a causal role of DNA methylation in the relationship between maternal vitamin B12 during pregnancy and child's IQ at age 8, cognitive performance and educational attainment: a two-step Mendelian randomization study. Human molecular genetics. 2017;26(15):3001-13.

169. Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Magi R, Reschen ME, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nat Genet. 2015;47(12):1415-25.

170. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41-7.

171. Weale ME. Quality control for genome-wide association studies. Methods Mol Biol. 2010;628:341-72.

172. Brion M-JA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. International Journal of Epidemiology. 2013;42(5):1497-501.

173. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. PLOS Genetics. 2013;9(3):e1003348.

174. Zeileis CKaA. Applied Econometrics with R: Springer-Verlag; 2008.

175. Stef van Buuren KG-O. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67.

176. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353-8.

177. MR Base for two sample MR 2018 [Available from: https://mrcieu.github.io/TwoSampleMR.

178. Zheng J, Baird D, Borges M-C, Bowden J, Hemani G, Haycock P, et al. Recent Developments in Mendelian Randomization Studies. Current Epidemiology Reports. 2017;4(4):330-45.

179. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015;44(2):512-25.

180. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLOS Genetics. 2017;13(11):e1007081.

181. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol. 2017;46(6):1734-9.

182. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555-7.

183. Timpson NJ, Nordestgaard BG, Harbord RM, Zacho J, Frayling TM, Tybjærg-Hansen A, et al. C-reactive protein levels and body mass index: Elucidating direction of causation through reciprocal Mendelian randomization. International journal of obesity (2005). 2011;35(2):300-8.

184. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic Acids Research. 2018;46(W1):W109-W13.

185. Liaw AM, W. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

186. Science WIo. GeneCards: Human gene database 1996-2018 [Build 6:[Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=NFYC&keywords=NFYC.

187. Dong X, Weng Z. The correlation between histone modifications and gene expression. Epigenomics. 2013;5(2):113-6.

188. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Research. 2012;40(D1):D912-D7.

189. Lu Q, Richardson B. DNaseI Hypersensitivity Analysis of Chromatin Structure. In: Tollefsbol TO, editor. Epigenetics Protocols. Totowa, NJ: Humana Press; 2004. p. 77-86.

190. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC genome browser database: extensions and updates 2013. Nucleic Acids Res. 2012;41.

191. Ramsuran V, Kulkarni H, He W, Mlisana K, Wright EJ, Werner L, et al. Duffy-null-associated low neutrophil counts influence HIV-1 susceptibility in high-risk South African black women. Clin Infect Dis. 2011;52(10):1248-56.

192. Horikawa Y, Wood CG, Yang H, Zhao H, Ye Y, Gu J, et al. Single nucleotide polymorphisms of microRNA machinery genes modify the risk of renal cell carcinoma. Clin Cancer Res. 2008;14(23):7956-62.

193. Olayioye MA, Vehring S, Muller P, Herrmann A, Schiller J, Thiele C, et al. StarD10, a START domain protein overexpressed in breast cancer, functions as a phospholipid transfer protein. J Biol Chem. 2005;280(29):27436-42.

194. Ensembl. Gene Ontology track Release 92 [Available from: http://www.ensembl.org/Homo_sapiens/Gene/Ontologies/biological_process?g=ENSG00000214530;r=11:72754729-72794168;t=ENST00000543304.

195. Imamura M, Takahashi A, Yamauchi T, Hara K, Yasuda K, Grarup N, et al. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. Nat Commun. 2016;7:10531.

196. Wood AR, Jonsson A, Jackson AU, Wang N, van Leewen N, Palmer ND, et al. A Genome-Wide Association Study of IVGTT-Based Measures of First-Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants. Diabetes. 2017;66(8):2296-309.

197. Carrat GR, Hu M, Nguyen-Tu M-S, Chabosseau P, Gaulton KJ, van de Bunt M, et al. Decreased STARD10 Expression Is Associated with Defective Insulin Secretion in Humans and Mice. The American Journal of Human Genetics. 2017;100(2):238-56.

198. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010-22.

199. Hahn S. Structure and mechanism of the RNA Polymerase II transcription machinery. Nature structural & molecular biology. 2004;11(5):394-403.

200. Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends Genet. 2016;32(1):42-56.

201. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 2007;129(4):823-37.

202. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

203. Law SW, Lackner KJ, Hospattankar AV, Anchors JM, Sakaguchi AY, Naylor SL, et al. Human apolipoprotein B-100: cloning, analysis of liver mRNA, and assignment of the gene to chromosome 2. Proceedings of the National Academy of Sciences of the United States of America. 1985;82(24):8340-4.

204. Metin DM, Orkide D. Apolipoproteins: Biochemistry, methods and clinical significance. Biochemical Education. 1989;17(2):63-8.

205. Lee B, Pratumvinit B, Thongtang N. The role of apoB measurement in Type 2 diabetic patients. Clinical Lipidology. 2015;10(2):137-44.

206. Onat A, Can G, Hergenc G, Yazici M, Karabulut A, Albayrak S. Serum apolipoprotein B predicts dyslipidemia, metabolic syndrome and, in women, hypertension and diabetes, independent of markers of central obesity and inflammation. International journal of obesity (2005). 2007;31(7):1119-25.

207. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan Ja, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nature genetics. 2012;44(9):991-1005.

208. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13.

209. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3.

210. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013;41(D1):D991-D5.

211. Kumar R, Thompson JR. The Regulation of Parathyroid Hormone Secretion and Synthesis. Journal of the American Society of Nephrology : JASN. 2011;22(2):216-24.

212. Taylor WH, Khaleeli AA. Coincident diabetes mellitus and primary hyperparathyroidism. Diabetes Metab Res Rev. 2001;17(3):175-80.

213. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Hum Mol Genet. 2015;24(15):4464-79.

214. Hall E, Dekker Nitert M, Volkov P, Malmgren S, Mulder H, Bacos K, et al. The effects of high glucose exposure on global gene expression and DNA methylation in human pancreatic islets. Molecular and Cellular Endocrinology. 2018;472:57-67.

215. Toperoff G, Kark JD, Aran D, Nassar H, Abu Ahmad W, Sinnreich R, et al. Premature aging of leukocyte DNA methylation is associated with type 2 diabetes prevalence. Clin Epigenetics. 2015;7.

216. Knopfholz J, #xe9, Disserol CC, #xe9, Diniz s, Pierin AJ, et al. Validation of the Friedewald Formula in Patients with Metabolic Syndrome. Cholesterol. 2014;2014:5.

217. Prokopenko I, Poon W, Mägi R, Prasad B R, Salehi SA, Almgren P, et al. A central role for GRB10 in regulation of islet function in man. PLoS genetics. 2014;10(4):e1004235-e.

218. Ward A. New Role for Grb10 Signaling in the Pancreas. Diabetes. 2012;61(12):3066.

219. QuickGO. Gene ontology and GO annotations Genome Campus, Hinxton, Cambridgeshire: EMBL-EBI; 2018 [GO version 2018-07-28:[Available from: https://www.ebi.ac.uk/QuickGO/.

220. Cell Signaling Technology 2006 [updated November 2012. Available from: https://www.cellsignal.co.uk/contents/science-cst-pathways-cellular-metabolism/ampk-signaling-interactive-pathway/pathways-ampk.

221. Zhang BB, Zhou G, Li C. AMPK: An Emerging Drug Target for Diabetes and the Metabolic Syndrome. Cell Metabolism. 2009;9(5):407-16.

222. Kim W, Shin Y-K, Kim B-J, Egan JM. Notch signaling in pancreatic endocrine cell and diabetes. Biochemical and biophysical research communications. 2010;392(3):247-51.

223. Tyagi S, Gupta P, Saini AS, Kaushal C, Sharma S. The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. Journal of Advanced Pharmaceutical Technology & Research. 2011;2(4):236-40.

224. Schou J, Tybjaerg-Hansen A, Moller HJ, Nordestgaard BG, Frikke-Schmidt R. ABC transporter genes and risk of type 2 diabetes: a study of 40,000 individuals from the general population. Diabetes Care. 2012;35(12):2600-6.

225. Sharp GC, Salas LA, Monnereau C, Allard C, Yousefi P, Everson TM, et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. Human Molecular Genetics. 2017;26(20):4067-85.

226. Lever J, Krzywinski M, Altman N. Principal component analysis. Nature Methods. 2017;14:641.

227. Richmond RC, Al-Amin A, Smith GD, Relton CL. Approaches for drawing causal inferences from epidemiological birth cohorts: a review. Early Hum Dev. 2014;90(11):769-80.

228. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014;15.

229. Smith GD, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? BMJ. 2005;330(7499):1076-9.

230. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22.

231. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. Epidemiology (Cambridge, Mass). 2014;25(3):427-35.

232. Grover S. Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. S. Burgess and S. G. Thompson (2015). London, UK: Chapman & Hall/CRC Press. 224 pages, ISBN: 9781466573178. Biometrical Journal. 2017;59(5):1086-7.

233. Vaxillaire M, Yengo L, Lobbens S, Rocheleau G, Eury E, Lantieri O, et al. Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. Diabetologia. 2014;57(8):1601-10.

234. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46:1173.

235. Wood AR, Tyrrell J, Beaumont R, Jones SE, Tuke MA, Ruth KS, et al. Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia. 2016;59(6):1214-21.

236. Reichardt LF. Neurotrophin-regulated signalling pathways. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2006;361(1473):1545-64.

237. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Research. 2002;12:9.

238. Simanshu DK, Nissley DV, McCormick F. RAS Proteins and Their Regulators in Human Disease. Cell. 2017;170(1):17-33.

239. Donath MY, Shoelson SE. Type 2 diabetes as an inflammatory disease. Nat Rev Immunol. 2011;11(2):98-107.

240. Calcutt NA, Cooper ME, Kern TS, Schmidt AM. Therapies for hyperglycaemia-induced diabetic complications: from animal models to clinical trials. Nature Reviews Drug Discovery. 2009;8:417.

241. Yamagishi S-i. Role of advanced glycation end products (AGEs) and receptor for AGEs (RAGE) in vascular damage in diabetes. Experimental Gerontology. 2011;46(4):217-24.

242. Hegab Z, Gibbons S, Neyses L, Mamas MA. Role of advanced glycation end products in cardiovascular disease. World journal of cardiology. 2012;4(4):90-102.

243. Parikh H, Carlsson E, Chutkow WA, Johansson LE, Storgaard H, Poulsen P, et al. TXNIP regulates peripheral glucose metabolism in humans. PLoS medicine. 2007;4(5):e158-e.

244. Shalev A, Pise-Masison CA, Radonovich M, Hoffmann SC, Hirshberg B, Brady JN, et al. Oligonucleotide microarray analysis of intact human pancreatic islets: identification of glucose-responsive genes and a highly regulated TGFbeta signaling pathway. Endocrinology. 2002;143(9):3695-8.

245. Shaked M, Ketzinel-Gilad M, Ariav Y, Cerasi E, Kaiser N, Leibowitz G. Insulin counteracts glucotoxic effects by suppressing thioredoxin-interacting protein production in INS-1E beta cells and in Psammomys obesus pancreatic islets. Diabetologia. 2009;52(4):636-44.

246. Sano H, Liu SC, Lane WS, Piletz JE, Lienhard GE. Insulin receptor substrate 4 associates with the protein IRAS. J Biol Chem. 2002;277(22):19439-47.

247. Chen J, Feng WL, Mo WJ, Ding XW, Xie SN. Expression of integrin-binding protein Nischarin in metastatic breast cancer. Mol Med Rep. 2015;12(1):77-82.

248. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010;42(11):949-60.

249. Corbi S, Bastos A, Nepomuceno R, Cirelli T, Santos RAd, Takahashi CS, et al. Expression Profile of Genes Potentially Associated with Adequate Glycemic Control in Patients with Type 2 Diabetes Mellitus %J Journal of Diabetes Research. 2017;2017:9.

250. Leibowitz G, Ktorza, A. & E. Cerasi. The role of TXNIP in the pathophysiology of diabetes and its vascular complications: a concise review. Medicographia. 2014;36:391-7.

251. Wechsler A, Brafman A, Shafir M, Heverin M, Gottlieb H, Damari G, et al. Generation of viable cholesterol-free mice. Science. 2003;302(5653):2087.

252. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature Genetics. 2010;42:579.

253. Sommese L, Zullo A, Mancini FP, Fabbricini R, Soricelli A, Napoli C. Clinical relevance of epigenetics in the onset and management of type 2 diabetes mellitus. Epigenetics. 2017;12(6):401-15.

254. Xu X, Su S, Barnes VA, De Miguel C, Pollock J, Ownby D, et al. A genome-wide methylation study on obesity: differential variability and differential methylation. Epigenetics. 2013;8(5):522-33.

255. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet. 2017;49(4):600-5.

256. Hidalgo Bertha A, Hivert M-F, Wessel J, Guan W, Gondalia Rahul B, Salfati Elias L, et al. Abstract P102: Epigenome-wide Association Study of Measures of Fasting Glucose, Fasting Insulin, and Hba1c in Non-diabetic Individuals of European, African, and Hispanic Ancestry in the Charge Consortium. Circulation. 2017;135(suppl_1):AP102-AP.

257. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biology. 2018;19(1):136.

258. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am J Hum Genet. 2016;98(4):680-96.

259. Arpón A, Milagro FI, Ramos-Lopez O, Mansego ML, Santos JL, Riezu-Boj J-I, et al. Epigenome-wide association study in peripheral white blood cells involving insulin resistance. Sci Rep. 2019;9(1):2445-.

260. Geurts YM, Dugué PA, Joo JE, Makalic E, Jung CH, Guan W, et al. Novel associations between blood DNA methylation and body mass index in middle-aged and older adults. International Journal Of Obesity. 2017;42:887.

261. Sayols-Baixeras S, Subirana I, Fernández-Sanlés A, Sentí M, Lluís-Ganella C, Marrugat J, et al. DNA methylation and obesity traits: An epigenome-wide association study. The REGICOR study. Epigenetics. 2017;12(10):909-16.

262. Wilson LE, Harlid S, Xu Z, Sandler DP, Taylor JA. An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort. International journal of obesity (2005). 2017;41(1):194-9.

263. Dhana K, Braun KVE, Nano J, Voortman T, Demerath EW, Guan W, et al. An Epigenome-Wide Association Study of Obesity-Related Traits. American Journal of Epidemiology. 2018;187(8):1662-9.

264. Dekkers KF, van Iterson M, Slieker RC, Moed MH, Bonder MJ, van Galen M, et al. Blood lipids influence DNA methylation in circulating cells. Genome Biology. 2016;17(1):138.

265. Pfeiffer L, Wahl S, Pilling LC, Reischl E, Sandling JK, Kunze S, et al. DNA methylation of lipid-related genes affects blood lipid levels. Circulation Cardiovascular genetics. 2015;8(2):334-42.

266. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics. 2017;9:15.

267. Campanella G, Gunter MJ, Polidoro S, Krogh V, Palli D, Panico S, et al. Epigenome-wide association study of adiposity and future risk of obesity-related diseases. International journal of obesity (2005). 2018;42(12):2022-35.

268. Akinyemiju T, Do AN, Patki A, Aslibekyan S, Zhi D, Hidalgo B, et al. Epigenome-wide association study of metabolic syndrome in African-American adults. Clin Epigenetics. 2018;10:49-.

269. Aslibekyan S, Do AN, Xu H, Li S, Irvin MR, Zhi D, et al. CPT1A methylation is associated with plasma adiponectin. Nutrition, metabolism, and cardiovascular diseases : NMCD. 2017;27(3):225-33.

270. Petersen AK, Zeilinger S, Kastenmuller G, Romisch-Margl W, Brugger M, Peters A, et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet. 2014;23(2):534-45.

271. Mamtani M, Kulkarni H, Dyer TD, Goring HH, Neary JL, Cole SA, et al. Genome- and epigenome-wide association study of hypertriglyceridemic waist in Mexican American families. Clin Epigenetics. 2016;8:6.

272. Lai CQ, Wojczynski MK, Parnell LD, Hidalgo BA, Irvin MR, Aslibekyan S, et al. Epigenome-wide association study of triglyceride postprandial responses to a high-fat dietary challenge. Journal of lipid research. 2016;57(12):2200-7.

273. Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, et al. Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. Circulation. 2014;130(7):565-72.

274. Frazier-Wood AC, Aslibekyan S, Absher DM, Hopkins PN, Sha J, Tsai MY, et al. Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. Journal of lipid research. 2014;55(7):1324-30.

275. Das M, Sha J, Hidalgo B, Aslibekyan S, Do AN, Zhi D, et al. Association of DNA Methylation at CPT1A Locus with Metabolic Syndrome in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) Study. PloS one. 2016;11(1):e0145789.

276. Fernandez-Sanles A, Sayols-Baixeras S, Curcio S, Subirana I, Marrugat J, Elosua R. DNA Methylation and Age-Independent Cardiovascular Risk, an Epigenome-Wide Approach: The REGICOR Study (REgistre GIroni del COR). Arteriosclerosis, thrombosis, and vascular biology. 2018;38(3):645-52.

277. Mendelson MM, Johannes R, Liu C, Huan T, Yao C, Miao X, et al. Epigenome-Wide Association Study of Soluble Tumor Necrosis Factor Receptor 2 Levels in the Framingham Heart Study. 2018;9(207).

278. Istas G, Declerck K, Pudenz M, Szic KSv, Lendinez-Tortajada V, Leon-Latre M, et al. Identification of differentially methylated BRCA1 and CRISP2 DNA regions as blood surrogate markers for cardiovascular disease. Sci Rep. 2017;7(1):5120.

279. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circulation Cardiovascular genetics. 2016;9(5):436-47.

Appendices

Table S8-1 Association summary statistics for 148 T2D genetic proxies selected from four DIAGRAM studies (Diabetes Genetics Replication and Meta-analysis). EA is the reported effect allele and OA is the other allele, P is the significance threshold. Two p-value thresholds were considered: the genome-wide significance threshold at $p<5.0x10^{-8}$ and the locus-wide significance threshold at $p<1.0x10^{-5}$.

SNP	Chr	Position	EA	OA	N	OR	95%CI	P	DIAGRAM Study†
rs340874	1	214159256	C	T	219582	1.07	(1.05-1.09)	5.10E-13	Gaulton_etal_2015
rs17106184	1	50682573	G	A	161585	1.1	(1.07-1.14)	4.10E-09	Mahajan_etal_2014
rs2075423	1	212221342	G	T	149821	1.07	(1.05-1.10)	8.10E-09	Morris_etal_2012
rs2820446	1	219575476	C	G	77138	1.05	(1.03, 1.07)	2.00E-06	Mahajan_etal_2014
rs2972156	2	227117778	G	C	219582	1.09	(1.07-1.11)	4.20E-20	Gaulton_etal_2015
rs77981966	2	43,777,964	C	T	219582	1.16	(1.11-1.20)	4.10E-14	Gaulton_etal_2015
rs75297654	2	165545615	C	T	219582	1.11	(1.08-1.14)	6.30E-13	Gaulton_etal_2015
rs10203174	2	43543534	C	T	149821	1.14	(1.10-1.19)	9.50E-12	Morris_etal_2012
rs11899863	2	43472323	C	T	144178	1.15	(1.10-1.20)	9.50E-11	Morris_etal_2012
rs35720761	2	43519977	T	C	92794	1.12	(1.07-1.16)	3.30E-10	Fuchsberger_etal_2016
rs780094	2	27741237	C	T	219582	1.06	(1.04-1.08)	3.40E-10	Gaulton_etal_2015
rs7578326	2	226728897	A	G	144178	1.08	(1.06-1.11)	3.80E-10	Morris_etal_2012
rs243020	2	60585028	G	A	219582	1.06	(1.04-1.08)	5.50E-10	Gaulton_etal_2015
rs1260326	2	27730940	T	C	92794	1.07	(1.04-1.10)	1.20E-09	Fuchsberger_etal_2016
rs13389219	2	165237122	C	T	144178	1.07	(1.05-1.10)	1.00E-08	Morris_etal_2012
rs243088	2	60422249	T	A	144178	1.07	(1.04-1.09)	1.80E-08	Morris_etal_2012
rs243019	2	60439310	C	T	147399	1.07	(1.04-1.09)	2.20E-08	Morris_etal_2012
rs3923113	2	165210095	A	C	133303	1.07	(1.05-1.10)	3.30E-08	Morris_etal_2012
rs10190052	2	646674	C	T	77138	1.07	(1.04, 1.10)	2.00E-07	Mahajan_etal_2014
rs2943640	2	226801829	C	A	149821	1.1	(1.07-1.12)	2.70E-14	Morris_etal_2012
rs7607980	2	165551201	T	C	92794	1.15	(1.11-1.19)	8.30E-15	Fuchsberger_etal_2016
rs35510946	3	185518910	A	G	219582	1.14	(1.12-1.16)	1.10E-39	Gaulton_etal_2015
rs4402960	3	186994381	T	G	148167	1.13	(1.10-1.16)	2.40E-23	Morris_etal_2012
rs6769511	3	187012984	C	T	144178	1.13	(1.10-1.16)	2.00E-21	Morris_etal_2012
rs11712037	3	12344730	C	G	219582	1.14	(1.11-1.17)	1.70E-20	Gaulton_etal_2015
rs11717195	3	124565088	T	C	149821	1.11	(1.08-1.14)	6.50E-14	Morris_etal_2012
rs11708067	3	124548468	A	G	149821	1.11	(1.08-1.14)	7.20E-14	Morris_etal_2012
rs1801282	3	12368125	C	G	149821	1.13	(1.09-1.17)	1.10E-12	Morris_etal_2012
rs17676309	3	64730121	C	T	219582	1.07	(1.05-1.09)	2.80E-12	Gaulton_etal_2015
rs6795735	3	64680405	C	T	149821	1.08	(1.06-1.11)	7.40E-11	Morris_etal_2012
rs1470579	3	187011774	C	A	69033	1.12	(1.08-1.16)	7.50E-11	Morris_etal_2012
rs1496653	3	23429794	A	G	149821	1.09	(1.06-1.12)	3.60E-09	Morris_etal_2012
rs6808574	3	189223217	C	T	140087	1.07	(1.04-1.09)	5.80E-09	Mahajan_etal_2014
rs7612463	3	23294959	C	A	77138	1.1	(1.04, 1.16)	7.00E-09	Mahajan_etal_2014
rs16861329	3	186948673	C	T	77138	1.03	(0.96, 1.10)	9.00E-06	Mahajan_etal_2014
rs10937721	4	6306763	C	G	219582	1.09	(1.07-1.11)	4.30E-18	Gaulton_etal_2015
rs4458523	4	6340887	G	T	148314	1.1	(1.07-1.12)	2.00E-15	Morris_etal_2012
rs1801214	4	6353923	T	C	144178	1.1	(1.08-1.13)	3.30E-15	Morris_etal_2012
rs6813195	4	153739925	C	T	161639	1.08	(1.06-1.10)	4.10E-14	Mahajan_etal_2014
rs1801212	4	6302519	A	G	92794	1.09	(1.06-1.12)	9.00E-14	Fuchsberger_etal_2016
rs734312	4	6303354	A	G	92794	1.06	(1.04-1.09)	6.90E-11	Fuchsberger_etal_2016
rs7732130	5	76435004	G	A	219582	1.08	(1.05-1.10)	2.40E-12	Gaulton_etal_2015
rs6878122	5	76463067	G	A	142081	1.1	(1.07-1.13)	5.00E-11	Morris_etal_2012
rs4457053	5	76460705	G	A	142081	1.09	(1.06-1.12)	1.80E-10	Morris_etal_2012
rs35658696	5	102338811	A	G	92794	1.17	(1.11-1.24)	5.70E-10	Fuchsberger_etal_2016
rs459193	5	55842508	G	A	144178	1.08	(1.05-1.11)	6.00E-09	Morris_etal_2012
rs702634	5	53307177	A	G	154797	1.06	(1.04-1.09)	6.90E-09	Mahajan_etal_2014
rs36046591	5	102537285	A	G	92794	1.19	(1.12-1.26)	3.30E-08	Fuchsberger_etal_2016
rs319598	5	134904545	C	T	77138	1.05	(1.03, 1.07)	2.00E-06	Mahajan_etal_2014
rs35261542	6	20675792	A	C	219582	1.17	(1.14-1.19)	1.50E-50	Gaulton_etal_2015
rs7756992	6	20787688	G	A	149821	1.17	(1.14-1.20)	7.00E-35	Morris_etal_2012
rs9368222	6	20794975	A	C	148167	1.17	(1.14-1.20)	7.00E-34	Morris_etal_2012
rs10440833	6	20796100	A	T	63390	1.22	(1.17-1.27)	3.60E-22	Morris_etal_2012

Continuation Table S8-1

SNP	Chr	Position	EA	OA	N	OR	95%CI	P	DIAGRAM Study†
rs9502570	6	7258384	A	G	77138	1.06	(1.04, 1.08)	1.00E-09	Mahajan_etal_2014
rs9505118	6	7235436	A	G	158348	1.06	(1.04-1.08)	1.40E-09	Mahajan_etal_2014
rs9379084	6	7231843	A	G	92794	1.13	(1.09-1.18)	4.00E-09	Fuchsberger_etal_2016
rs3130501	6	31244432	G	A	155815	1.07	(1.04-1.09)	4.20E-09	Mahajan_etal_2014
rs9472138	6	43844025	T	C	77138	1.06	(1.04, 1.08)	2.00E-07	Mahajan_etal_2014
rs4273712	6	126643364	G	A	77138	1.05	(1.03, 1.07)	3.00E-06	Mahajan_etal_2014
rs6937795	6	136970143	A	C	77138	1.04	(1.02, 1.06)	7.00E-06	Mahajan_etal_2014
rs1535500	6	39316274	T	G	77138	1.13	(1.08, 1.19)	8.00E-06	Mahajan_etal_2014
rs1513272	7	28200097	C	T	219582	1.1	(1.08-1.12)	7.80E-25	Gaulton_etal_2015
rs849135	7	28162938	G	A	144178	1.11	(1.08-1.13)	3.10E-17	Morris_etal_2012
rs10276674	7	14922007	C	T	219582	1.08	(1.06-1.11)	2.80E-11	Gaulton_etal_2015
rs17168486	7	14864807	T	C	144178	1.11	(1.07-1.14)	5.90E-11	Morris_etal_2012
rs878521	7	44255643	A	G	219582	1.07	(1.05-1.10)	1.30E-10	Gaulton_etal_2015
rs849134	7	28162747	A	G	63390	1.12	(1.08-1.16)	3.20E-10	Morris_etal_2012
rs1974620	7	15065467	T	C	219582	1.06	(1.04-1.08)	1.00E-09	Gaulton_etal_2015
rs2233580	7	127253550	T/C	T/C	92794	1.79	(1.47-2.19)	9.30E-09	Fuchsberger_etal_2016
rs7795991	7	13861106	G	A	77138	1.05	(1.03, 1.07)	7.00E-07	Mahajan_etal_2014
rs2284219	7	30674820	T	C	77138	1.05	(1.03, 1.08)	8.00E-06	Mahajan_etal_2014
rs13266634	8	118184783	C	T	219582	1.12	(1.09-1.14)	5.00E-28	Gaulton_etal_2015
rs3802177	8	118254206	G	A	142307	1.14	(1.11-1.17)	1.30E-21	Morris_etal_2012
rs516946	8	41638405	C	T	149821	1.09	(1.06-1.12)	2.50E-10	Morris_etal_2012
rs7845219	8	94925274	T	C	77138	1.08	(1.04, 1.12)	6.00E-08	Mahajan_etal_2014
rs1561927	8	128555832	C	T	77138	1.06	(1.04, 1.09)	1.00E-07	Mahajan_etal_2014
rs10811660	9	22134068	G	A	219582	1.27	(1.23-1.30)	1.10E-61	Gaulton_etal_2015
rs10811661	9	22124094	T	C	149821	1.18	(1.15-1.22)	3.70E-27	Morris_etal_2012
rs10757283	9	22134172	T	C	219582	1.12	(1.10-1.14)	3.60E-26	Gaulton_etal_2015
rs10965250	9	22123284	G	A	144178	1.19	(1.15-1.23)	1.80E-25	Morris_etal_2012
chr9:4294707:I	9	4294707	I	R	219582	1.07	(1.05-1.09)	3.10E-11	Gaulton_etal_2015
rs60980157	9	139235415	T	C	92794	1.09	(1.06-1.12)	1.70E-09	Fuchsberger_etal_2016
rs944801	9	22041670	C	G	142671	1.08	(1.05-1.10)	2.40E-09	Morris_etal_2012
rs2796441	9	83498768	G	A	147724	1.07	(1.05-1.10)	5.40E-09	Morris_etal_2012
rs17791513	9	79290675	A	G	77138	1.21	(1.13, 1.31)	3.00E-08	Mahajan_etal_2014
rs7041847	9	4287466	A	G	77138	1.05	(1.01, 1.09)	5.00E-06	Mahajan_etal_2014
rs7903146	10	114748339	T	C	144178	1.39	(1.35-1.42)	1.20E-139	Morris_etal_2012
rs11187140	10	94466910	G	A	219582	1.12	(1.10-1.14)	1.50E-31	Gaulton_etal_2015
rs1111875	10	94452862	C	T	149821	1.11	(1.09-1.14)	2.00E-19	Morris_etal_2012
rs5015480	10	94455539	C	T	69033	1.15	(1.11-1.19)	2.20E-16	Morris_etal_2012
rs11257658	10	12309268	A	G	219582	1.09	(1.07-1.12)	1.20E-15	Gaulton_etal_2015
rs12571751	10	80612637	A	G	149821	1.08	(1.05-1.10)	1.00E-10	Morris_etal_2012
rs11257655	10	12265895	T	C	77138	1.06	(1.01, 1.11)	3.00E-09	Mahajan_etal_2014
rs10510110	10	122432914	C	T	77138	1.05	(1.03, 1.07)	1.00E-07	Mahajan_etal_2014
rs2812533	10	69692529	C	T	77138	1.07	(1.04, 1.09)	5.00E-06	Mahajan_etal_2014
rs10788575	10	88008827	A	G	77138	1.06	(1.03, 1.08)	9.00E-06	Mahajan_etal_2014
rs74046911	11	2858636	C	T	219582	1.29	(1.23-1.35)	9.60E-26	Gaulton_etal_2015
chr11:2692322:D	11	2692322	D	R	219582	1.08	(1.06-1.10)	2.30E-15	Gaulton_etal_2015
chr11:72460930:I	11	72460930	R	I	219582	1.1	(1.07-1.12)	6.70E-14	Gaulton_etal_2015
rs10830963	11	92348358	G	C	143723	1.1	(1.07-1.13)	5.30E-13	Morris_etal_2012
rs163184	11	2803645	G	T	142181	1.09	(1.06-1.11)	1.20E-11	Morris_etal_2012
rs1387153	11	92313476	T	C	144178	1.09	(1.06-1.12)	1.60E-11	Morris_etal_2012
rs1552224	11	72110746	A	C	144178	1.11	(1.07-1.14)	1.80E-10	Morris_etal_2012
rs2237895	11	2857194	C	A	219582	1.07	(1.05-1.10)	5.30E-10	Gaulton_etal_2015
rs5215	11	17365206	C	T	149821	1.07	(1.05-1.10)	8.50E-10	Morris_etal_2012
rs5219	11	17409572	T	C	92794	1.07	(1.05-1.10)	9.00E-10	Fuchsberger_etal_2016

Continuation Table S8-1

SNP	Chr	Position	EA	OA	N	OR	95%CI	P	DIAGRAM Study†
rs231361	11	2648076	A	G	144178	1.09	(1.06-1.12)	1.20E-09	Morris_etal_2012
rs231362	11	2648047	G	A	134972	1.08	(1.05-1.11)	1.70E-09	Morris_etal_2012
rs757110	11	17418477	A	C	92794	1.07	(1.04-1.10)	1.70E-08	Fuchsberger_etal_2016
rs2283220	11	2755548	A	G	219582	1.06	(1.03-1.08)	2.40E-07	Gaulton_etal_2015
rs458069	11	2858800	G	C	219582	1.06	(1.04-1.09)	1.00E-06	Gaulton_etal_2015
rs1169288	12	121,416,650	C	A	219582	1.09	(1.07-1.12)	8.10E-15	Gaulton_etal_2015
rs2583941	12	66204598	A	G	219582	1.11	(1.08-1.15)	1.60E-11	Gaulton_etal_2015
chr12:121440833:D	12	121440833	R	D	219582	1.07	(1.05-1.09)	2.90E-10	Gaulton_etal_2015
rs10842994	12	27856417	C	T	149821	1.1	(1.06-1.13)	6.10E-10	Morris_etal_2012
rs2261181	12	64498585	T	C	147824	1.13	(1.08-1.17)	1.20E-09	Morris_etal_2012
rs2612035	12	64478934	G	A	144178	1.12	(1.08-1.17)	3.00E-09	Morris_etal_2012
rs7955901	12	69719560	C	T	144178	1.07	(1.05-1.10)	6.50E-09	Morris_etal_2012
rs4275659	12	122013881	C	T	161459	1.06	(1.04-1.08)	9.50E-09	Mahajan_etal_2014
rs1727313	12	123156306	C	G	77138	1.06	(1.04, 1.08)	1.00E-08	Mahajan_etal_2014
rs7961581	12	71663102	C	T	219582	1.06	(1.03-1.08)	1.80E-07	Gaulton_etal_2015
rs12427353	12	120989098	G	C	77138	1.12	(1.07, 1.18)	4.00E-06	Mahajan_etal_2014
rs1359790	13	79615157	G	A	149821	1.08	(1.05-1.10)	1.40E-08	Morris_etal_2012
rs10507349	13	26207391	G	C	77138	1.06	(1.04, 1.08)	2.00E-07	Mahajan_etal_2014
rs3803563	15	91531352	A	C	219582	1.08	(1.06-1.11)	5.60E-11	Gaulton_etal_2015
rs7177055	15	75619817	A	G	149821	1.08	(1.05-1.10)	4.60E-09	Morris_etal_2012
rs7161785	15	62395224	G	C	219582	1.06	(1.04-1.08)	4.90E-09	Gaulton_etal_2015
rs12899811	15	89345080	G	A	144178	1.08	(1.05-1.10)	6.30E-09	Morris_etal_2012
rs7178572	15	75534245	G	A	144178	1.07	(1.05-1.10)	2.20E-08	Morris_etal_2012
rs2028299	15	89831025	C	A	77138	1.04	(1.00, 1.09)	5.00E-07	Mahajan_etal_2014
rs7163757	15	62099409	C	T	77138	1.06	(1.02, 1.11)	4.00E-06	Mahajan_etal_2014
rs9927317	16	53820996	G	C	219582	1.14	(1.12-1.16)	7.90E-43	Gaulton_etal_2015
rs9936385	16	52376670	C	T	144178	1.13	(1.10-1.16)	2.60E-23	Morris_etal_2012
rs11642841	16	52402988	A	C	144178	1.12	(1.09-1.14)	1.10E-19	Morris_etal_2012
rs7202877	16	73804746	T	G	144178	1.12	(1.07-1.16)	3.50E-08	Morris_etal_2012
rs4430796	17	36098040	G	A	219582	1.09	(1.07-1.11)	7.80E-18	Gaulton_etal_2015
rs11651052	17	33176494	A	G	80788	1.1	(1.07-1.14)	2.00E-11	Morris_etal_2012
rs11651755	17	33173953	C	T	80788	1.1	(1.07-1.13)	1.80E-10	Morris_etal_2012
rs12970134	18	56035730	A	G	138946	1.08	(1.05-1.11)	1.20E-08	Morris_etal_2012
chr18:57739289:D	18	57739289	D	R	219582	1.06	(1.04-1.08)	1.90E-07	Gaulton_etal_2015
rs2238689	19	46178661	C	T	219582	1.08	(1.06-1.11)	8.30E-16	Gaulton_etal_2015
rs72999033	19	19366632	T	C	219582	1.16	(1.12-1.20)	1.80E-15	Gaulton_etal_2015
rs58542926	19	19379549	T	C	92794	1.14	(1.10-1.19)	3.20E-10	Fuchsberger_etal_2016
rs10401969	19	19268718	C	T	149821	1.13	(1.09-1.18)	7.00E-09	Morris_etal_2012
rs4399645	19	46166073	T	C	219582	1.06	(1.04-1.08)	1.40E-08	Gaulton_etal_2015
rs8108269	19	45655255	G	T	77138	1.06	(1.02, 1.11)	5.00E-06	Mahajan_etal_2014
rs4812829	20	44360627	A	G	77138	1.07	(1.01, 1.12)	5.00E-08	Mahajan_etal_2014
rs41278853	22	30416527	A	G	92794	1.14	(1.09-1.19)	5.60E-09	Fuchsberger_etal_2016

† Morris *et al*. 2012[101] Discovery Sample size: 121,171 European ancestry cases, 56,862 European ancestry controls. Replication Sample size: 22,669 European ancestry cases, 58,119 European ancestry controls, 1,178 South Asian ancestry cases, 2,472 South Asian ancestry controls. Mahajan *et al*. 2014[30] Discovery Sample size: 12,171 European ancestry cases, 56,862 European ancestry controls, 6,952 East Asian ancestry cases, 11,865 East Asian ancestry controls, 5,561 South Asian ancestry cases, 14,458 South Asian ancestry controls, 1,804 Mexican ancestry cases, 779 Mexican ancestry controls. Replication Sample size: 21,491 European ancestry cases, 55,647 European ancestry controls. Gaulton *et al*. 2015[169] Discovery Sample size: 27,206 European ancestry cases, 57,574 European ancestry controls. Fuchsberger *et al*. 2016[170] Discovery Sample size: 11,645 T2D cases and 32,769 controls from European, East Asian, South Asian, African American, Hispanic ancestry.

Figure S8-1 Inspection plots of genotype data in middle-age participants in ALSPAC, used in the SNP-CpG analysis. A) Minor allele frequency (MAF) versus imputation quality scores. B) genome-wide distribution of variants per chromosome.

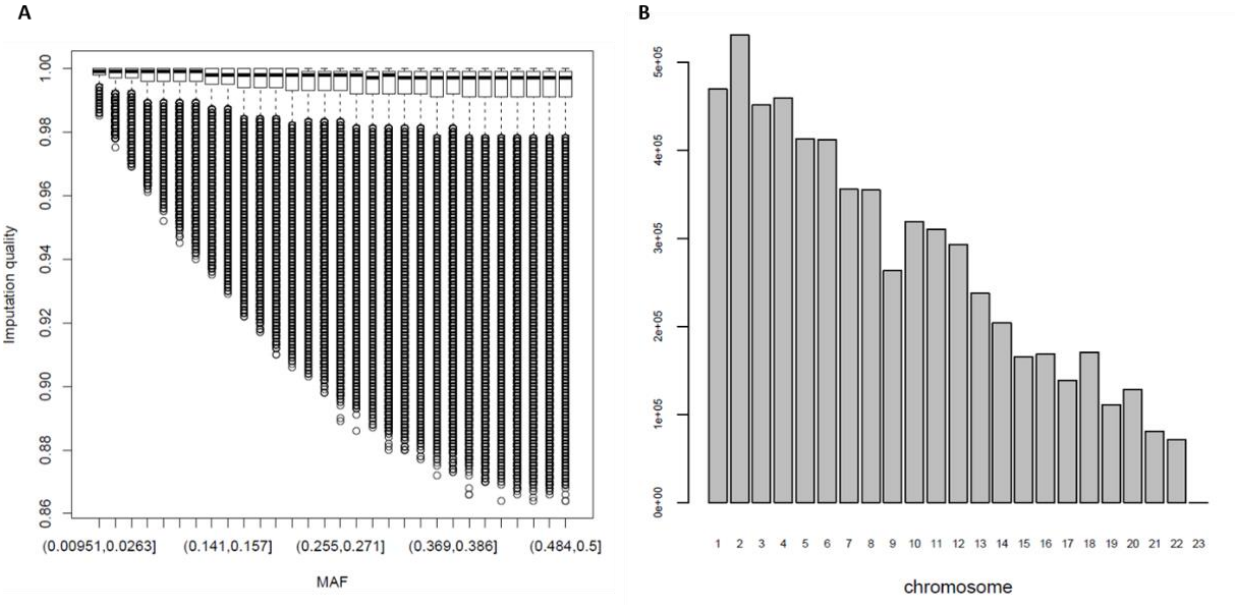


Table S8-2 Summary of QC steps applied to the genetic data used to conduct the EWAS of T2D SNPs. Data corresponds to the subsample of midle-age adults in ALSPAC with available methylation and geneotype data (n=1,252). QC steps as specified in the GoDMC pipeline.

QC implemented	No. SNPs	Sample	No. Excluded	Remaining SNPs	Remaining Samples
Initial numbers	6,102,837	1,252			
Variants with missing genotyping rate > 0.05			800,072	5,302,765	1,252
Samples with missing genotyping rate > 0.05			0	5,302,765	1,252
Hardy-Weinberg Equilibrium (P<1.0e-6)			151	5,302,614	1,252
Minor allele threshold (MAF<0.01)			30,446	5,272,168	1,252
SNPs with another allele coding (A, C, T, G, D, I)			0	5,272,168	1,252
Cryptic relatedness (cut-off 0.125)			4	5,272,168	1,248
Allele mismatch			24	5,272,144	1,248
Outliers after allele freq. check (1000G)			1,437	5,270,707	1,248
SNPs with allele strand flipped	66,341		0	5,270,707	1,248
Final dataset				**5,270,707**	**1,248**

Table S8-3 Comparison of top-ten associations with the smallest p-value obtained across models in the EWAS of T2D using the subsample of ALSPAC/ARIES (n=1,050).

	CpG	Chr	Gene	Basic model			Cell-adjusted model			Smoking-adjusted model			Fully adjusted		
				Beta	SE	P	Beta	SE	P	Beta	SE	P	Beta	SE	P
Top-10 CpGs Basic model	cg15986668	1	*NFYC*	-0.064	1.27E-02	7.00E-07	-0.065	1.28E-02	4.61E-07	-0.065	1.28E-02	5.78E-07	-0.071	1.30E-02	5.48E-08
	cg19823491	2	*OTX1*	-0.006	1.16E-03	1.79E-06	-0.006	1.16E-03	1.62E-06	-0.006	1.16E-03	1.70E-06	-0.006	1.17E-03	2.99E-06
	cg24605023	3	*CADPS*	-0.031	6.13E-03	4.84E-07	-0.028	6.16E-03	4.94E-06	-0.028	6.16E-03	6.06E-06	-0.026	6.24E-03	2.53E-05
	cg03206717	3	*SLC25A38*	-0.003	7.29E-04	4.38E-06	-0.003	7.17E-04	4.76E-06	-0.003	7.18E-04	4.84E-06	-0.003	7.26E-04	2.95E-06
	cg05575921	5	*AHRR*	-0.036	7.82E-03	3.61E-06	-0.036	7.84E-03	3.72E-06	-0.030	7.25E-03	4.86E-05	-0.026	7.31E-03	3.94E-04
	cg10870892	11	*CTTN*	-0.050	9.10E-03	6.24E-08	-0.045	8.96E-03	6.40E-07	-0.045	8.97E-03	7.58E-07	-0.045	9.09E-03	1.13E-06
	cg26353859	12	*SLC16A7*	0.031	6.57E-03	2.99E-06	0.029	6.57E-03	1.42E-05	0.029	6.57E-03	1.49E-05	0.029	6.67E-03	2.07E-05
	cg04016326	12	*GRIN2B*	-0.054	1.16E-02	3.44E-06	-0.055	1.17E-02	2.71E-06	-0.055	1.17E-02	2.78E-06	-0.054	1.19E-02	5.71E-06
	cg17749033	17	Unannotated	-0.019	3.81E-03	1.29E-06	-0.016	3.62E-03	7.16E-06	-0.016	3.62E-03	7.47E-06	-0.016	3.67E-03	1.73E-05
	cg25341923	17	*KRTAP4-7*	-0.016	3.26E-03	1.74E-06	-0.013	3.19E-03	4.27E-05	-0.013	3.20E-03	3.58E-05	-0.013	3.24E-03	8.47E-05
Top-10 CpGs Cell-adjusted model	cg15986668	1	*NFYC*	-0.064	1.27E-02	7.00E-07	-0.065	1.28E-02	4.61E-07	-0.065	1.28E-02	5.78E-07	-0.071	1.30E-02	5.48E-08
	cg19823491	2	*OTX1*	-0.006	1.16E-03	1.79E-06	-0.006	1.16E-03	1.62E-06	-0.006	1.16E-03	1.70E-06	-0.006	1.17E-03	2.99E-06
	cg03206717	3	*SLC25A38*	-0.003	7.29E-04	4.38E-06	-0.003	7.17E-04	4.76E-06	-0.003	7.18E-04	4.84E-06	-0.003	7.26E-04	2.95E-06
	cg05575921	5	*AHRR*	-0.036	7.82E-03	3.61E-06	-0.036	7.84E-03	3.72E-06	-0.030	7.25E-03	4.86E-05	-0.026	7.31E-03	3.94E-04
	cg14290451	6	*RPL10A*	-0.004	9.86E-04	1.61E-05	-0.004	9.45E-04	4.38E-06	-0.004	9.46E-04	5.20E-06	-0.004	9.59E-04	1.35E-05
	cg14045803	11	*STARD10*	-0.010	2.20E-03	1.05E-05	-0.011	2.17E-03	2.70E-07	-0.011	2.17E-03	3.07E-07	-0.012	2.20E-03	1.39E-07
	cg10870892	11	*CTTN*	-0.050	9.10E-03	6.24E-08	-0.045	8.96E-03	6.40E-07	-0.045	8.97E-03	7.58E-07	-0.045	9.09E-03	1.13E-06
	cg04016326	12	*GRIN2B*	-0.054	1.16E-02	3.44E-06	-0.055	1.17E-02	2.71E-06	-0.055	1.17E-02	2.78E-06	-0.054	1.19E-02	5.71E-06
	cg00204249	17	*DNAH17*	-0.013	3.02E-03	1.50E-05	-0.014	3.04E-03	3.29E-06	-0.014	3.04E-03	2.47E-06	-0.015	3.08E-03	2.76E-06
	cg26652413	19	*CPAMD8*	-0.022	4.90E-03	1.14E-05	-0.022	4.78E-03	2.96E-06	-0.022	4.78E-03	3.49E-06	-0.023	4.85E-03	2.51E-06
Top-10 CpGs Smoking- adjusted model	cg15986668	1	*NFYC*	-0.064	1.27E-02	7.00E-07	-0.065	1.28E-02	4.61E-07	-0.065	1.28E-02	5.78E-07	-0.071	1.30E-02	5.48E-08
	cg19823491	2	*OTX1*	-0.006	1.16E-03	1.79E-06	-0.006	1.16E-03	1.62E-06	-0.006	1.16E-03	1.70E-06	-0.006	1.17E-03	2.99E-06
	cg03206717	3	*SLC25A38*	-0.003	7.29E-04	4.38E-06	-0.003	7.17E-04	4.76E-06	-0.003	7.18E-04	4.84E-06	-0.003	7.26E-04	2.95E-06
	cg24605023	3	*CADPS*	-0.031	6.13E-03	4.84E-07	-0.028	6.16E-03	4.94E-06	-0.028	6.16E-03	6.06E-06	-0.026	6.24E-03	2.53E-05
	cg14290451	6	*RPL10A*	-0.004	9.86E-04	1.61E-05	-0.004	9.45E-04	4.38E-06	-0.004	9.46E-04	5.20E-06	-0.004	9.59E-04	1.35E-05
	cg10870892	11	*CTTN*	-0.050	9.10E-03	6.24E-08	-0.045	8.96E-03	6.40E-07	-0.045	8.97E-03	7.58E-07	-0.045	9.09E-03	1.13E-06
	cg14045803	11	*STARD10*	-0.010	2.20E-03	1.05E-05	-0.011	2.17E-03	2.70E-07	-0.011	2.17E-03	3.07E-07	-0.012	2.20E-03	1.39E-07
	cg04016326	12	*GRIN2B*	-0.054	1.16E-02	3.44E-06	-0.055	1.17E-02	2.71E-06	-0.055	1.17E-02	2.78E-06	-0.054	1.19E-02	5.71E-06
	cg00204249	17	*DNAH17*	-0.013	3.02E-03	1.50E-05	-0.014	3.04E-03	3.29E-06	-0.014	3.04E-03	2.47E-06	-0.015	3.08E-03	2.76E-06
	cg26652413	19	*CPAMD8*	-0.022	4.90E-03	1.14E-05	-0.022	4.78E-03	2.96E-06	-0.022	4.78E-03	3.49E-06	-0.023	4.85E-03	2.51E-06

Continuation Table S8-3.

				Basic model			Cell-adjusted model			Adjusted for smoking			Fully adjusted		
	CpG	Chr	Gene	Beta	SE	P	Beta	SE	P	Beta	SE	P	Beta	SE	P
Top-10 CpGs Fully adjusted model	cg15986668	1	*NFYC*	-0.064	1.27E-02	7.00E-07	-0.065	1.28E-02	4.61E-07	-0.065	1.28E-02	5.78E-07	-0.071	1.30E-02	5.48E-08
	cg04656330	2	*PNKD*	-0.002	3.46E-04	9.95E-06	-0.002	3.47E-04	6.98E-06	-0.002	3.47E-04	7.58E-06	-0.002	3.52E-04	7.96E-06
	cg19823491	2	*OTX1*	-0.006	1.16E-03	1.79E-06	-0.006	1.16E-03	1.62E-06	-0.006	1.16E-03	1.70E-06	-0.006	1.17E-03	2.99E-06
	cg03206717	3	*SLC25A38*	-0.003	7.29E-04	4.38E-06	-0.003	7.17E-04	4.76E-06	-0.003	7.18E-04	4.84E-06	-0.003	7.26E-04	2.95E-06
	cg02307288	5	*TRPC7*	-0.036	8.23E-03	1.55E-05	-0.035	8.25E-03	2.20E-05	-0.036	8.26E-03	1.82E-05	-0.038	8.36E-03	5.54E-06
	cg10870892	11	*CTTN*	-0.050	9.10E-03	6.24E-08	-0.045	8.96E-03	6.40E-07	-0.045	8.97E-03	7.58E-07	-0.045	9.09E-03	1.13E-06
	cg14045803	11	*STARD10*	-0.010	2.20E-03	1.05E-05	-0.011	2.17E-03	2.70E-07	-0.011	2.17E-03	3.07E-07	-0.012	2.20E-03	1.39E-07
	cg04016326	12	*GRIN2B*	-0.054	1.16E-02	3.44E-06	-0.055	1.17E-02	2.71E-06	-0.055	1.17E-02	2.78E-06	-0.054	1.19E-02	5.71E-06
	cg00204249	17	*DNAH17*	-0.013	3.02E-03	1.50E-05	-0.014	3.04E-03	3.29E-06	-0.014	3.04E-03	2.47E-06	-0.015	3.08E-03	2.76E-06
	cg26652413	19	*CPAMD8*	-0.022	4.90E-03	1.14E-05	-0.022	4.78E-03	2.96E-06	-0.022	4.78E-03	3.49E-06	-0.023	4.85E-03	2.51E-06

Table S8-4 Correlation between methylation in the strongest marker detected in the EWAS at cg15986668 in NFYC, and additional risk factors for T2D in the subsample of ALSPAC/ARIES (n=1,050). Correlations were calculated using the Pearson correlation; values of fasting glucose, insulin, HOMA-IR, HOMA-B, C-reactive protein and diastolic blood pressure, were transformed using the natural logarithm before the analysis.

	Correlation Coefficient	P-value
Total cholesterol (mmol/l)	0.01	0.71
LDL (mmol/l)	0.02	0.49
HDL (mmol/l)	-0.04	0.15
Triglycerides (mmol/l)	-0.02	0.61
Fasting Glucose (mmol/l)	-0.13	2.98×10^{-5}
Insulin (mmol/l)	-0.07	0.08
HOMA-IR	-0.10	0.01
HOMA-B	0.01	0.83
Waist circumference (cm)	0.06	0.07
Waist-hip ratio	0.01	0.79
Diastolic BP (mmHg)	0.05	0.12
Systolic BP (mmHg)	0.04	0.22
C-reactive Protein (mg/l)	0.07	0.02

Figure S8-2 Directed-acyclic graphs showing potential direction in the association between T2D, methylation at NFYC, and additional risk factors of T2D found in strong correlation with methylation at this DMP. There was an inverse association between methylation and fasting glucose and HOMA-IR, while a positive association was detected between methylation and C-reactive protein, and this latter was independent of T2D.

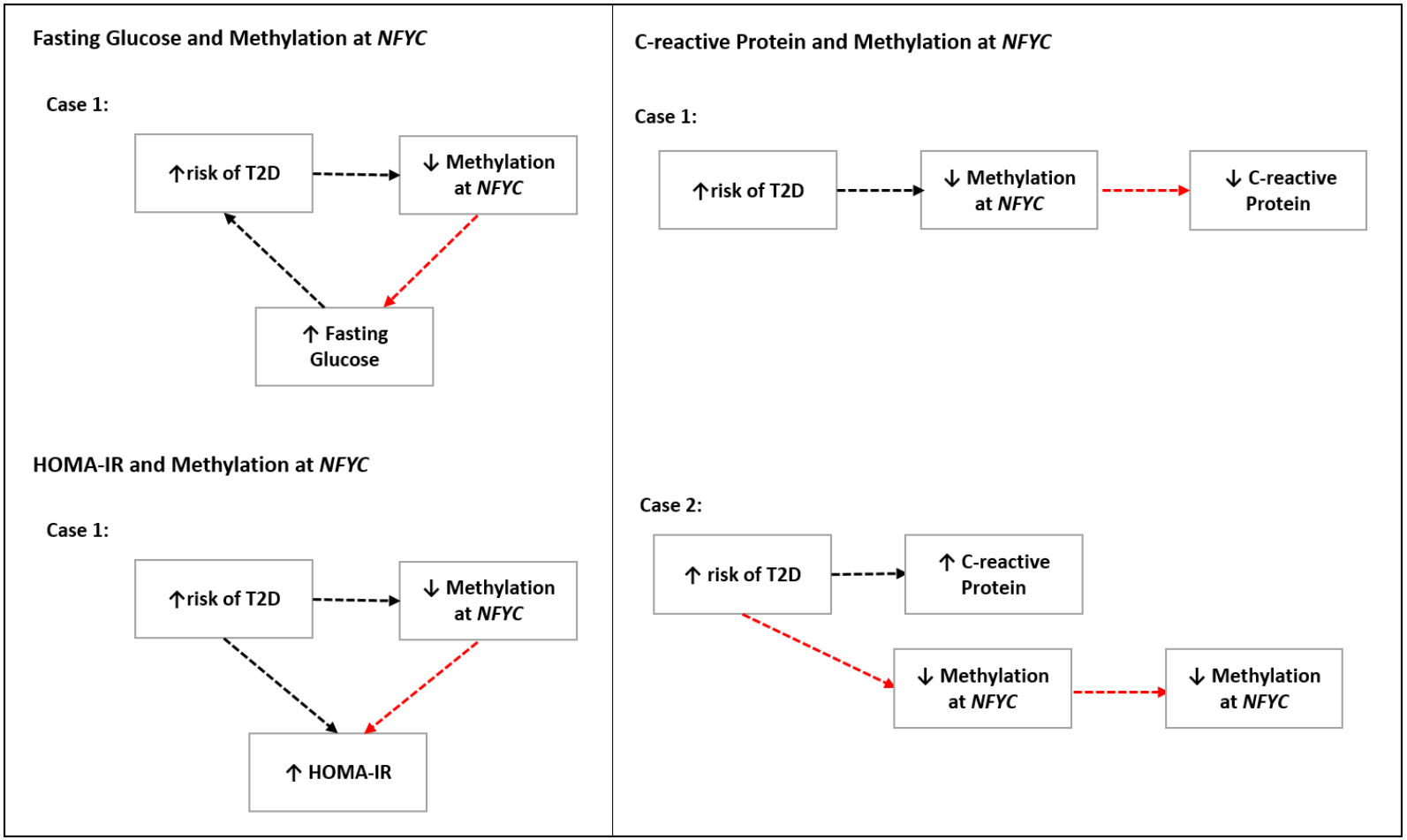


Figure S8-3. Q-Q plot and Manhattan plot of the EWAS in T2D conducted in participants in ALSPAC/ARIES (cases=48, controls=1,002). Results were adjusted age, sex, 7SVs, 6 predicted cell-counts, BMI and smoking (non-smoker, smoker). Q-Q plot (left-hand side) shows the distribution of observed versus expected -log10(P-values), the red line represents the distribution of observed P-values under the null hypothesis of no-associations. Observed P-values consistently deviated from the line of expected P-values representing the line of null associations; On one side, this deviation is a signature of high genomic inflation in the EWAS (Lambda= 1.48), and on the other side, it illustrated the presence of one DMP in strong association with T2D (DMP in NFYC), and another DMP with borderline association with T2D (DMP in STARD10). The Manhattan plot (right-hand side) shows the distribution of P-values across genomic coordinates (i.e. chromosomes) for each CpG site included in the EWAS. Horizontal blue line is the line of borderline association ($P=1.00x10^{-6}$), while the horizontal red line is the Bonferroni corrected P-value at $P < 1.07x10^{-7}$. One DMP found within the region of NFYC was identified with EWAS significance ($P= 5.48x10^{-8}$).

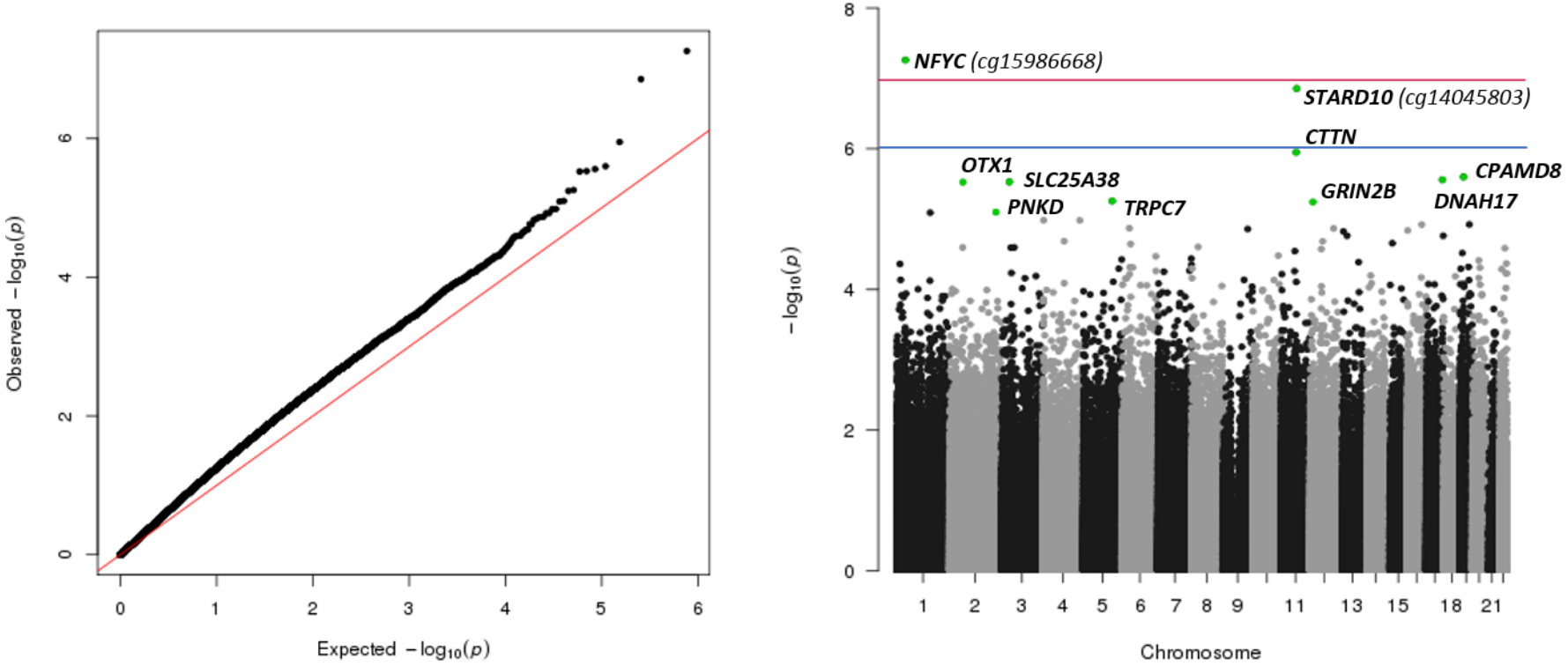


Figure S8-4. Association between methylation and T2D for CpG sites located within candidate genetic loci for T2D. Plots illustrate the genomic position of each CpG site against the effect size (left) and -log10(p-value) (right) according to results of the fully-adjusted EWAS in T2D conducted in the subsample of ALSPAC/ARIES (n=1,050). CpG sites are represented by small bars within the plot. None of the CpG sites within the candidate loci analysed reached EWAS significance at Bonferroni corrected p< 1.07x10^{-7}or at -log10(p-value) ≥7.0. Loci represented from the top left to the bottom right are SLC30A8 (Chr8, n= 7 probes), JAZF1 (Chr7, n= 56 probes), HNF1B (Chr17, n= 26 probes), KCNQ1 (Chr11, n= 288 probes) and THADA (Chr2, n= 48 probes).

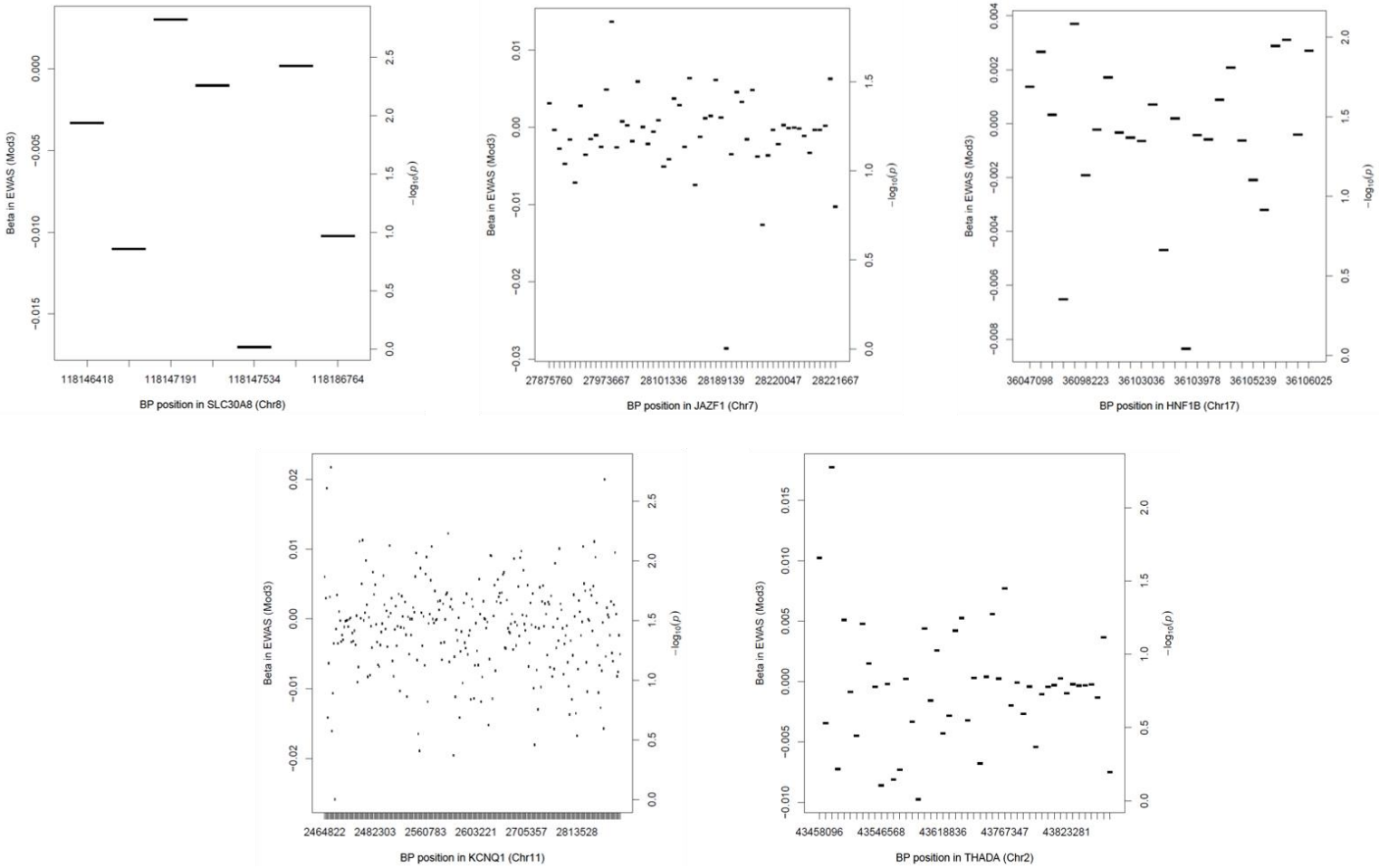
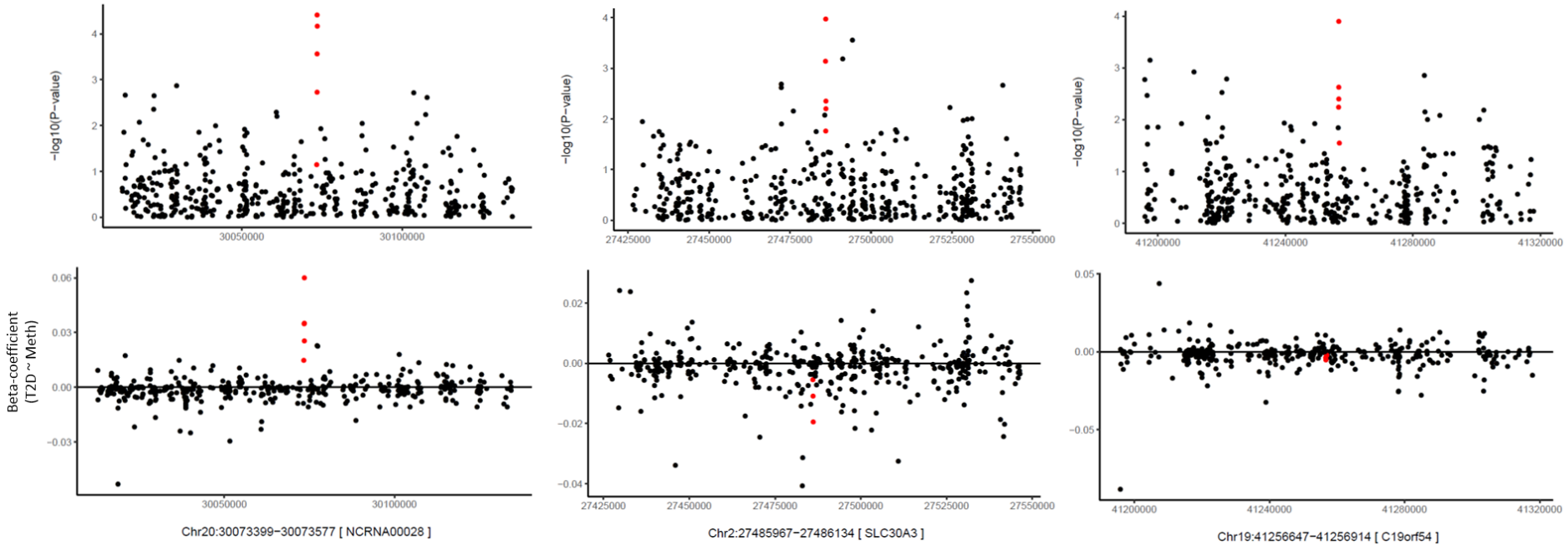
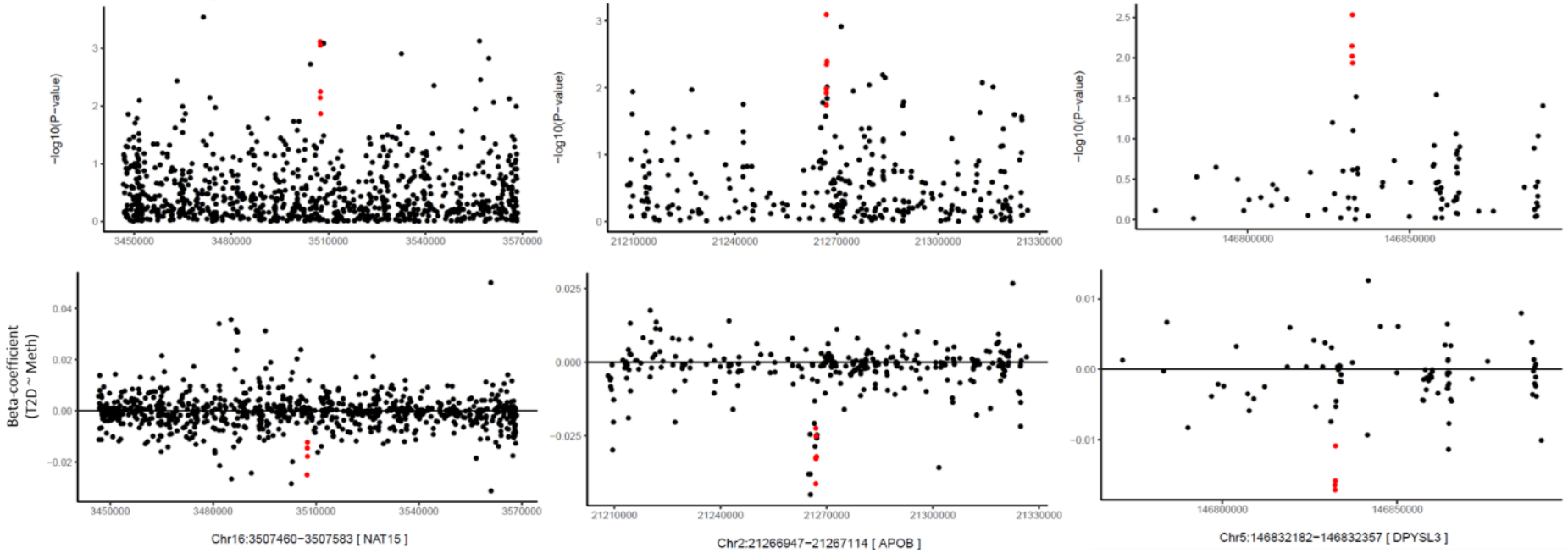


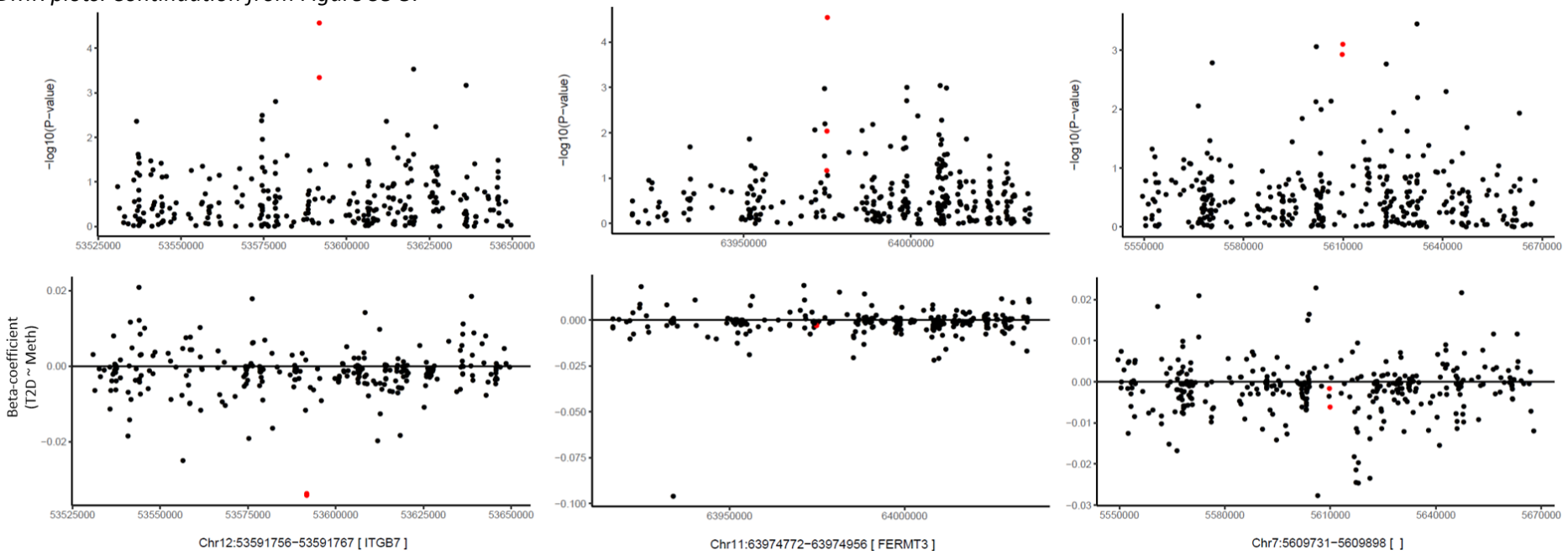
Figure S8-5. DMR plot showing the effect-direction and -log10(P-value) of DMPs mapping within 12 DMRs identified in association with prevalent T2D. DMRs were generated in comb-p using summary data of a fully-adjusted EWAS (covariates: age, sex, 7 SVs, 6 predicted cell-count, smoking and BMI). Black dots represent CpG sites within the DMR without differential methylation (background CpG sites), while red dots are CpG sites found with differential methylation. On each panel, x-axis: chromosome location of the DMR; y-axis (upper plot): -log10(P-value), based on P-values reported in the EWAS, and y-axis (lower plot): effect size and direction of effect of DMPs within the DMR.



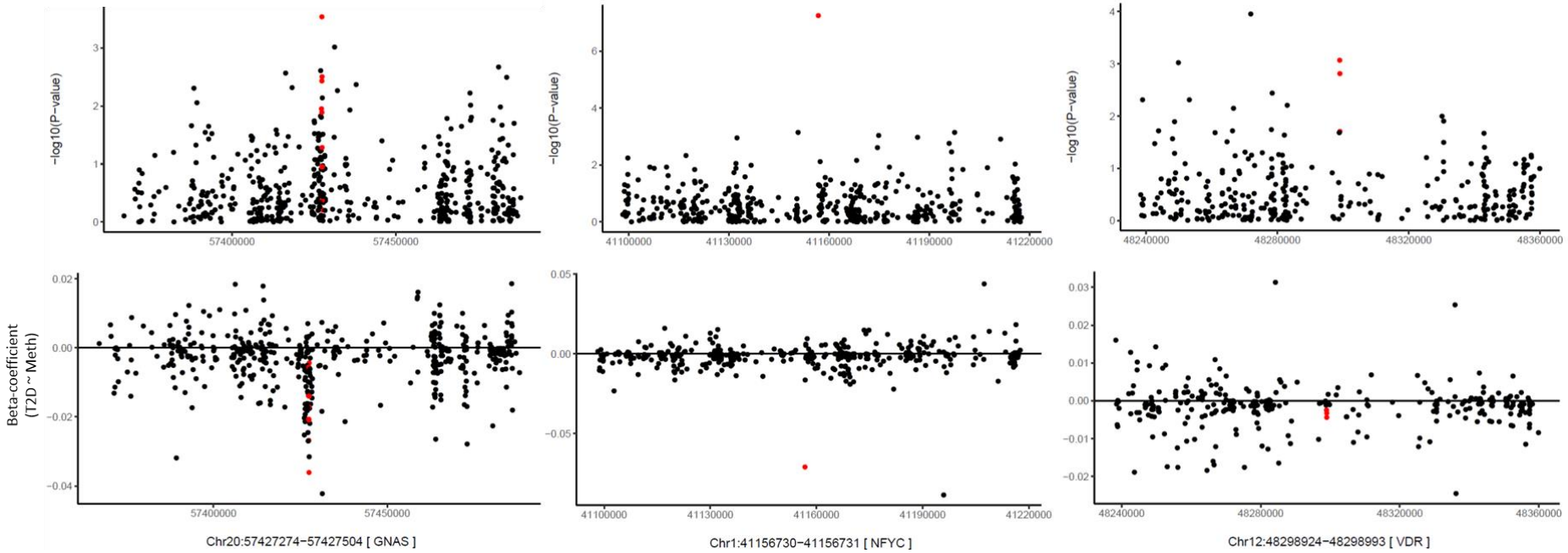
DMR plots: Continuation from Figure S8-5.



DMR plots: Continuation from Figure S8-5.



DMR plots: Continuation from Figure S8-5.



DMR plots: Continuation from Figure S8-5.

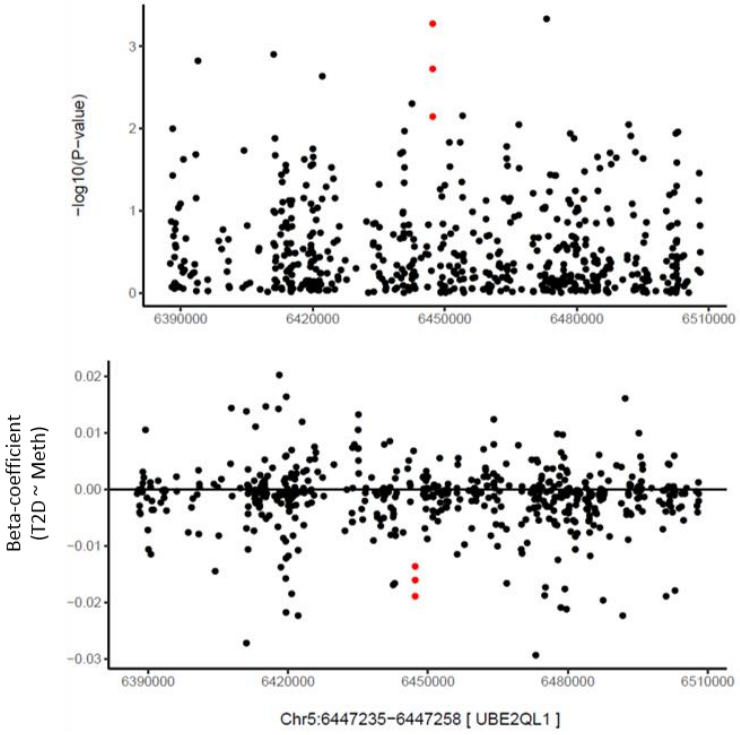


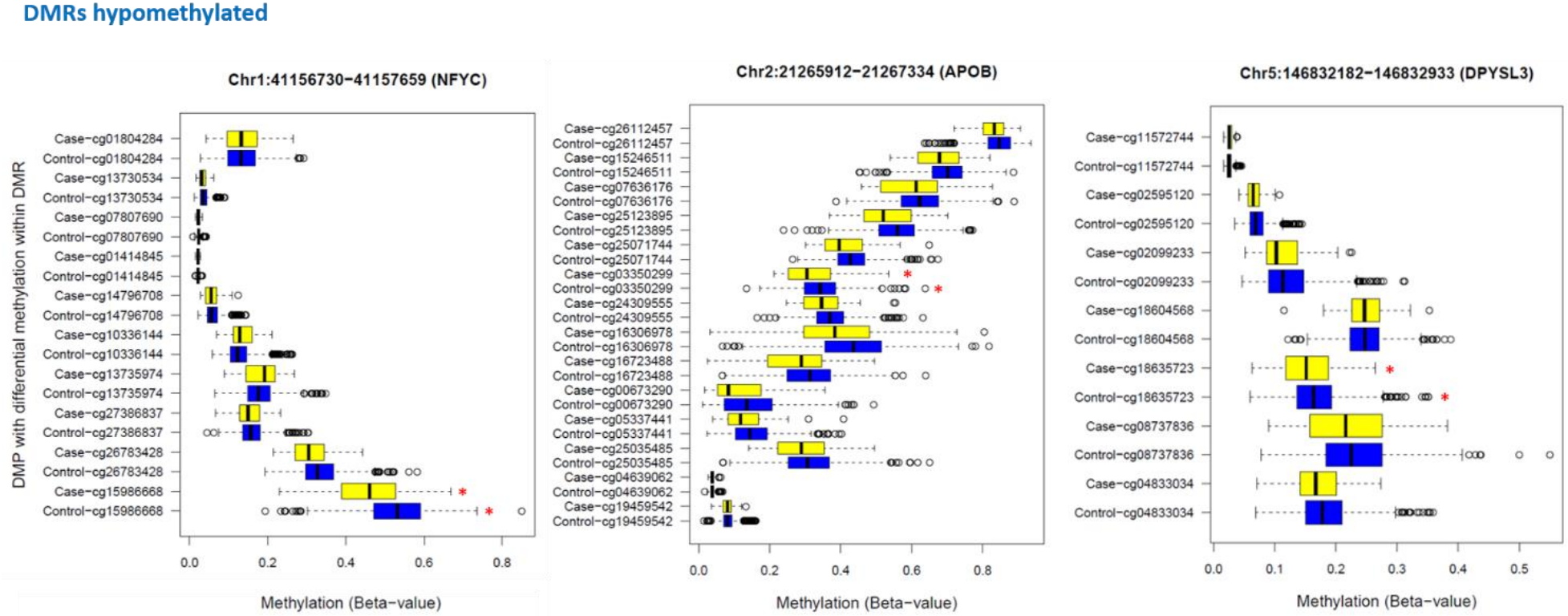
Table S8-5 List of 53 DMPs identified within 12 DMRs strongly associated with prevalent T2D. DMRs were generated by comb-p using summary data from a fully-adjusted EWAS model. Gene: closest gene to the DMR; DMP: CpG site identified with differential methylation within the DMR, Beta: effect-size identified in the, P: p-value reported in the EWAS, P-region: p-value for the region calculated by comb-p *using the Stouffer-Liptak correction, Sidak: level of significance to establish regions of interest. DMRs were selected as significant based on Sidak < 0.05.*

DMR	Gene	DMP	Beta (EWAS)	P (EWAS)	P-region	Sidak
Chr20:30,073,399-30,073,577	*NCRNA00028*	cg02991085	0.060	3.86E-05	5.82E-10	1.25E-06
		cg13159946	0.035	2.72E-04	5.82E-10	1.25E-06
		cg15537254	0.015	7.29E-02	5.82E-10	1.25E-06
		cg21846177	0.025	6.88E-05	5.82E-10	1.25E-06
		cg25502144	0.035	1.90E-03	5.82E-10	1.25E-06
Chr2:27,485,967-27,486,134	*SLC30A3*	cg01878321	-0.020	6.31E-03	1.01E-08	2.32E-05
		cg03023068	-0.011	4.51E-03	1.01E-08	2.32E-05
		cg10629682	-0.001	1.73E-02	1.01E-08	2.32E-05
		cg13174651	-0.002	7.35E-04	1.01E-08	2.32E-05
		cg23151303	-0.005	1.06E-04	1.01E-08	2.32E-05
Chr19:41,256,647-41,256,914	*C19orf54*	cg11481490	-0.004	2.38E-03	3.51E-08	5.04E-05
		cg13793525	-0.004	5.77E-03	3.51E-08	5.04E-05
		cg22745273	-0.002	2.87E-02	3.51E-08	5.04E-05
		cg23456263	-0.005	4.01E-03	3.51E-08	5.04E-05
		cg26015947	-0.005	1.26E-04	3.51E-08	5.04E-05
Chr16:3,507,460-3,507,583	*NAT15*	cg00484396	-0.025	7.77E-04	6.95E-08	2.17E-04
		cg05754148	-0.012	8.88E-04	6.95E-08	2.17E-04
		cg09873201	-0.001	1.36E-02	6.95E-08	2.17E-04
		cg21433313	-0.015	7.18E-03	6.95E-08	2.17E-04
		cg22508957	-0.018	5.74E-03	6.95E-08	2.17E-04
Chr2:21,266,947-21,267,114	*APOB*	cg03350299	-0.033	7.97E-04	1.19E-07	2.72E-04
		cg16306978	-0.041	1.80E-02	1.19E-07	2.72E-04
		cg16723488	-0.033	1.03E-02	1.19E-07	2.72E-04
		cg24309555	-0.022	1.20E-02	1.19E-07	2.72E-04
		cg25071744	-0.025	4.48E-03	1.19E-07	2.72E-04
		cg25123895	-0.032	4.07E-03	1.19E-07	2.72E-04
Chr5:146,832,182-146,832,357	*DPYSL3*	cg04833034	-0.016	7.11E-03	3.29E-06	7.17E-03
		cg08737836	-0.017	9.43E-03	3.29E-06	7.17E-03
		cg18604568	-0.011	1.15E-02	3.29E-06	7.17E-03
		cg18635723	-0.016	2.91E-03	3.29E-06	7.17E-03
Chr12:53,591,756-53,591,767	*ITGB7*	cg04972065	-0.034	2.67E-05	3.70E-07	1.28E-02
		cg23029655	-0.034	4.54E-04	3.70E-07	1.28E-02
Chr11:63,974,772-63,974,956	*FERMT3*	cg01447914	-0.002	2.85E-05	7.65E-06	1.58E-02
		cg01647936	-0.002	6.70E-02	7.65E-06	1.58E-02
		cg20136100	-0.003	9.00E-03	7.65E-06	1.58E-02
Chr7:5,609,731-5,609,898	*Unannotated*	cg05281338	-0.006	7.92E-04	8.71E-06	1.98E-02
		cg08253188	-0.002	1.18E-03	8.71E-06	1.98E-02
Chr20:57,427,274-57,427,504	*GNAS*	cg03010274	-0.014	1.12E-02	1.23E-05	2.02E-02
		cg06065549	-0.027	2.89E-04	1.23E-05	2.02E-02
		cg07105596	-0.036	3.14E-03	1.23E-05	2.02E-02
		cg08587534	-0.004	6.05E-01	1.23E-05	2.02E-02
		cg12321149	-0.017	1.29E-02	1.23E-05	2.02E-02
		cg25652859	-0.021	3.70E-03	1.23E-05	2.02E-02
		cg26534489	-0.006	4.20E-01	1.23E-05	2.02E-02
		cg26811638	-0.014	1.12E-01	1.23E-05	2.02E-02
		cg27304369	-0.021	5.12E-02	1.23E-05	2.02E-02

Continuation Table S8-5.

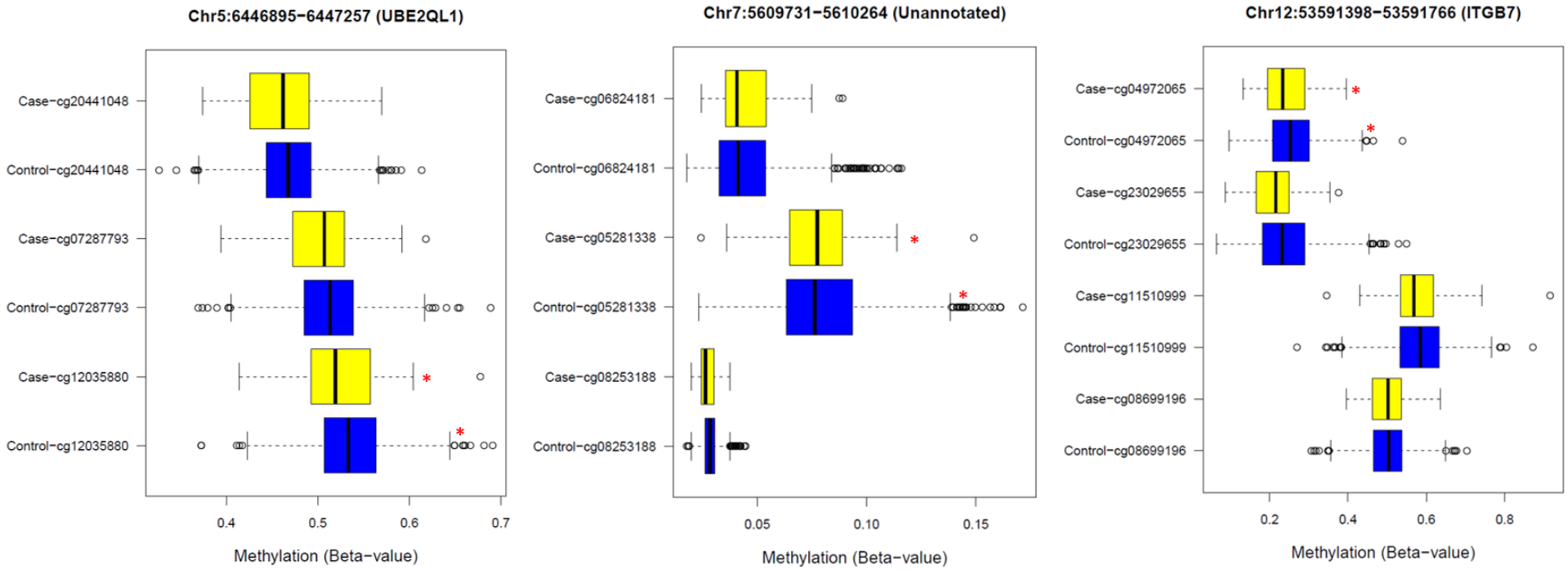
DMR	Gene	DMP	Beta (EWAS)	P (EWAS)	P-region	Sidak
Chr12:48,298,924-48,298,993	*VDR*	cg02522757	-0.003	1.96E-02	7.26E-06	3.95E-02
		cg13865595	-0.002	8.47E-04	7.26E-06	3.95E-02
		cg23654431	-0.004	1.55E-03	7.26E-06	3.95E-02
Chr5:6,447,235-6,447,258	*UBE2QL1*	cg07287793	-0.016	1.88E-03	3.01E-06	4.89E-02
		cg12035880	-0.019	5.27E-04	3.01E-06	4.89E-02
		cg20441048	-0.014	7.21E-03	3.01E-06	4.89E-02

Figure S8-6 Representation of methylation values for DMPs included in DMRs which were found in common between DMRcate and Comb-p. Distribution of methylation values was stratified betweenT2D cases (yellow-box) and controls (blue-box). Red asterisk demarks the strongest DMP within the region according to the EWAS. Average difference in methylation for the regions annotated to NFYC, APOB, and DPYSL3, UBE2QL1, ITGB7 and the unannotated DMR in Chr7, suggested that T2D cases were on average hypomethylated compared to controls. In contrast, average difference in methylation in the DMR annotated to NCRNA00028, suggested that cases were on average hypermethylated.



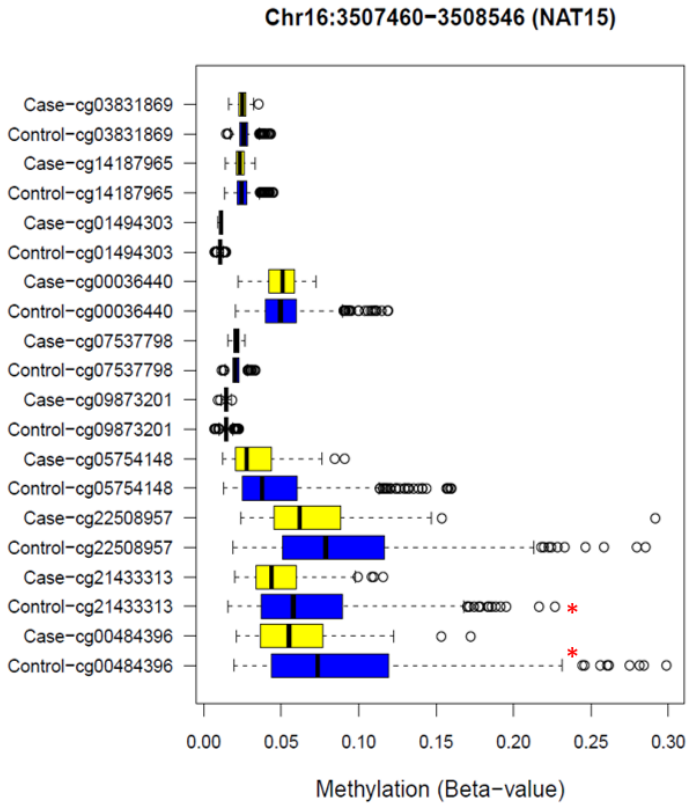
Continuation of DMR plots from Figure S8-6

DMRs hypomethylated



Continuation of DMR plots from Figure S8-6

DMR hypomethylated



DMR hypermethylated

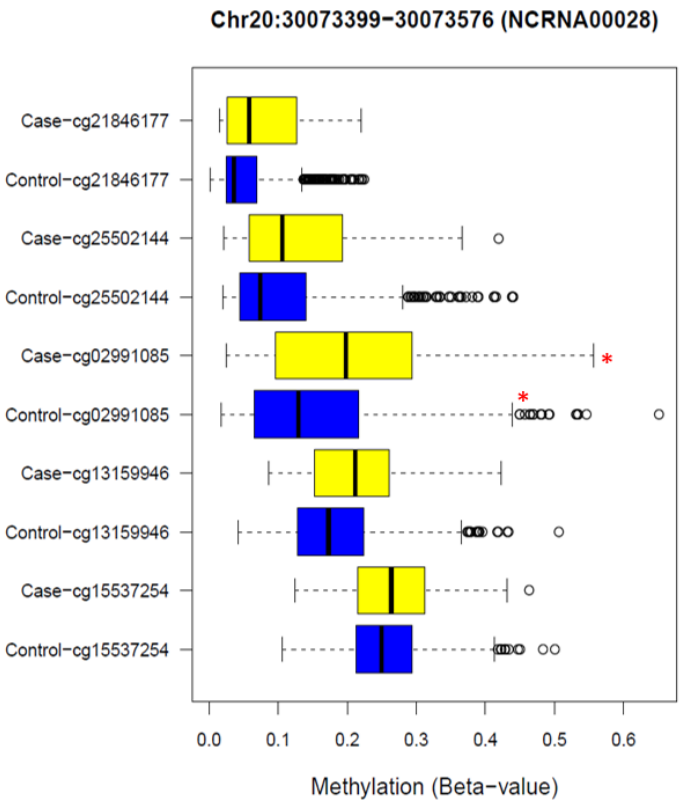


Figure S8-7 Q-Q plot (left) and Manhattan plot (right) showing results of the EWAS of T2D conducted in participants in KORA (n=1,719). A fully adjusted model was used adjusting for age, sex, 10 PCs, 6 predicted cell-counts, BMI and smoking (never, former and current-smoker). Q-Q plot shows the distribution of observed versus expected P-values, and the red line represents the distribution of P-values under the null hypothesis of no-associations. Deviation of observed P-values from the line of null associations, and outside the grey area demarking the 95% confidence intervals, suggested a signal in strong association with T2D. Manhattan plot shows the distribution of p-values across genomic coordinates for each CpG tested. The horizontal red line is the Bonferroni corrected p-threshold ($p<1.07x10^{-7}$). No association was detected with Bonferroni significance in this analysis, and top signal with borderline association was detected at DMP cg11696475 in GNG4 ($beta=-0.001$, $SE=1.7x10^{-4}$, $p=6.01x10^{-7}$).

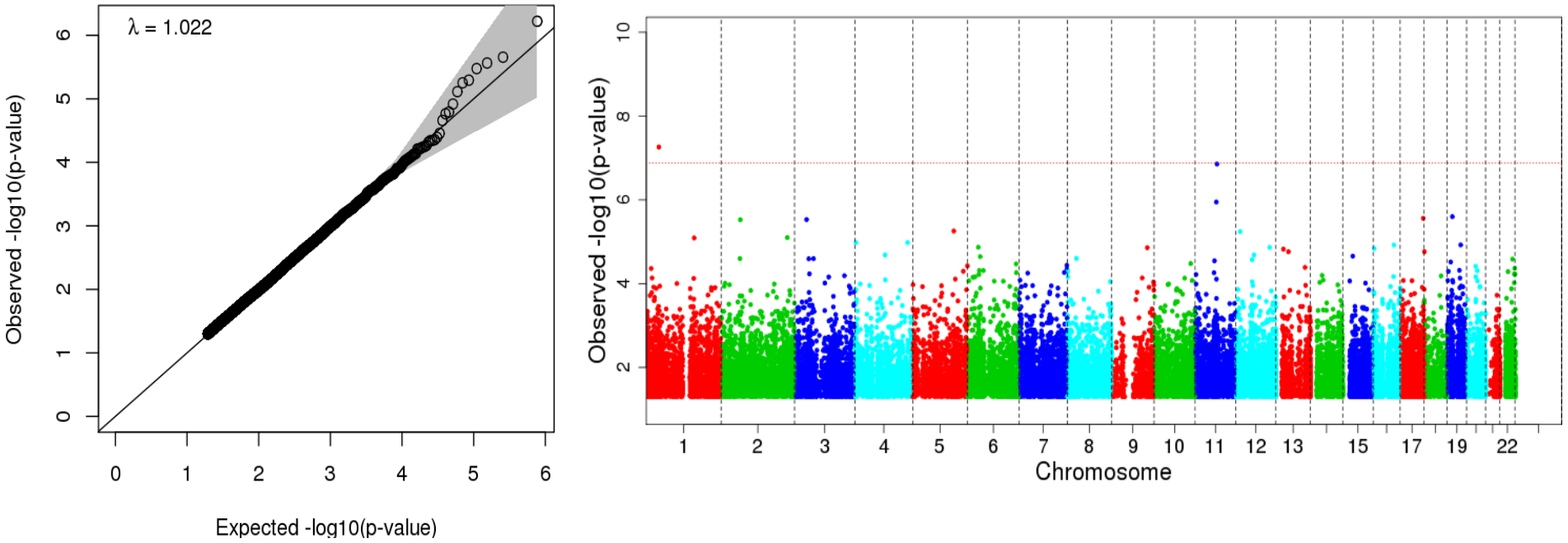


Table S8-6 Comparison of results of the EWAS in T2D between ALSPAC and KORA (n=1,719). Models considered were a minimally adjusted model (Model 1), adjusted for age, sex and SVs; a second model additionally adjusted for predicted cell-counts (Model 2), a third model additionally adjusted for smoking (Model 3), and a fully-adjusted model additionally adjusted for BMI (Model 4).

	ALSPAC/ARIES					KORA (F4)				
	CpG	**Loci**	**Beta**	**SE**	**P-value**	**CpG**	**Loci**	**Beta**	**SE**	**P-value**
Model 1	cg10870892	*CTTN*	-0.050	0.009	6.24E-08	cg11696475	*GNG4*	-0.001	1.63E-04	5.84E-07
	cg24605023	*CADPS*	-0.031	0.006	4.84E-07	cg10950524	*MAD1L1*	-0.034	0.007	2.30E-06
	cg15986668	*NFYC*	-0.064	0.013	7.00E-07	cg01622006	*Unannotated*	0.004	0.001	3.60E-06
	cg17749033	*Unannotated*	-0.019	0.004	1.29E-06	cg10780164	*CALY*	0.013	0.003	5.08E-06
	cg25341923	*KRTAP4-7*	-0.016	0.003	1.74E-06	cg14768946	*STAT1*	0.002	4.57E-04	5.21E-06
	cg19823491	*OTX1*	-0.006	0.001	1.79E-06	cg24377329	*SLC37A4*	-0.006	0.001	5.69E-06
	cg26353859	*SLC16A7*	0.031	0.007	2.99E-06	cg06596743	*MON1B*	0.004	0.001	7.79E-06
	cg04016326	*GRIN2B*	-0.054	0.012	3.44E-06	cg10937131	*MSRA*	-0.004	0.001	1.64E-05
	cg05575921	*AHRR*	-0.036	0.008	3.61E-06	cg00939432	*NIPSNAP1*	-0.010	0.002	2.04E-05
	cg03206717	*SLC25A38*	-0.003	0.001	4.38E-06	cg11949207	*C1orf66*	-0.005	0.001	2.09E-05
Model 2	cg14045803	*STARD10*	-0.011	0.002	2.70E-07	cg11696475	*GNG4*	-0.001	1.63E-04	4.79E-07
	cg15986668	*NFYC*	-0.065	0.013	4.61E-07	cg10780164	*CALY*	0.013	0.003	3.22E-06
	cg10870892	*CTTN*	-0.045	0.009	6.40E-07	cg11949207	*C1orf66*	-0.005	0.001	4.53E-06
	cg19823491	*OTX1*	-0.006	0.001	1.62E-06	cg10950524	*MAD1L1*	-0.031	0.007	4.73E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.71E-06	cg14768946	*STAT1*	0.002	4.57E-04	4.96E-06
	cg26652413	*CPAMD8*	-0.022	0.005	2.96E-06	cg01622006	*Unannotated*	0.004	0.001	5.48E-06
	cg00204249	*DNAH17*	-0.014	0.003	3.29E-06	cg16898425	*ITIH1*	-0.007	0.001	5.99E-06
	cg05575921	*AHRR*	-0.036	0.008	3.72E-06	cg24377329	*SLC37A4*	-0.006	0.001	6.05E-06
	cg14290451	*RPL10A*	-0.004	0.001	4.38E-06	cg02976539	*SLC9A3R1*	0.007	0.002	7.07E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.76E-06	cg06596743	*MON1B*	0.004	0.001	7.17E-06
Model 3	cg14045803	*STARD10*	-0.011	0.002	3.07E-07	cg11696475	*GNG4*	-0.001	1.63E-04	4.52E-07
	cg15986668	*NFYC*	-0.065	0.013	5.78E-07	cg10780164	*CALY*	0.013	0.003	3.44E-06
	cg10870892	*CTTN*	-0.045	0.009	7.58E-07	cg10950524	*MAD1L1*	-0.032	0.007	3.99E-06
	cg19823491	*OTX1*	-0.006	0.001	1.70E-06	cg11949207	*C1orf66*	-0.005	0.001	4.50E-06
	cg00204249	*DNAH17*	-0.014	0.003	2.47E-06	cg01622006	*Unannotated*	0.004	0.001	5.41E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.78E-06	cg14768946	*STAT1*	0.002	4.57E-04	5.66E-06
	cg26652413	*CPAMD8*	-0.022	0.005	3.49E-06	cg06596743	*MON1B*	0.004	0.001	7.24E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.84E-06	cg24377329	*SLC37A4*	-0.006	0.001	7.25E-06
	cg14290451	*RPL10A*	-0.004	0.001	5.20E-06	cg16898425	*ITIH1*	-0.006	0.001	7.88E-06
	cg24605023	*CADPS*	-0.028	0.006	6.06E-06	cg02976539	*SLC9A3R1*	0.007	0.002	9.32E-06
Model 4	cg15986668	*NFYC*	-0.071	0.013	5.48E-08	cg11696475	*GNG4*	-0.001	1.66E-04	6.01E-07
	cg14045803	*STARD10*	-0.012	0.002	1.39E-07	cg10950524	*MAD1L1*	-0.033	0.007	2.21E-06
	cg10870892	*CTTN*	-0.045	0.009	1.13E-06	cg02976539	*SLC9A3R1*	0.008	0.002	2.73E-06
	cg26652413	*CPAMD8*	-0.023	0.005	2.51E-06	cg24377329	*SLC37A4*	-0.007	0.001	3.34E-06
	cg00204249	*DNAH17*	-0.015	0.003	2.76E-06	cg10780164	*CALY*	0.013	0.003	5.09E-06
	cg03206717	*SLC25A38*	-0.003	0.001	2.95E-06	cg14768946	*STAT1*	0.002	4.64E-04	5.60E-06
	cg19823491	*OTX1*	-0.006	0.001	2.99E-06	cg11949207	*C1orf66*	-0.005	0.001	7.71E-06
	cg02307288	*TRPC7*	-0.038	0.008	5.54E-06	cg06596743	*MON1B*	0.004	0.001	1.21E-05
	cg04016326	*GRIN2B*	-0.054	0.012	5.71E-06	cg16898425	*ITIH1*	-0.006	0.001	1.61E-05
	cg04656330	*PNKD*	-0.002	0.000	7.96E-06	cg01622006	*Unannotated*	0.004	0.001	1.72E-05

Figure S8-8 Q-Q plot (left) and Manhattan plot (right) showing results of the EWAS in T2D conducted in participants in LBC1936 cohort (n=915). A fully adjusted model was used adjusting for age, sex, 8 SVs, 6 predicted cell-counts, BMI and smoking (non-smoker, former, current smoker). Q-Q plot shows the distribution of observed versus expected P-values, and the red line represents the distribution of P-values under the null hypothesis of no-associations Deviation of observed P-values from the line of null associations, and outside the grey area demarking the 95% confidence intervals, suggested signals in strong association with T2D. Manhattan plot shows the distribution of P-values across genomic coordinates for each CpG tested. The horizontal red line is the Bonferroni corrected p-threshold ($p<1.07x10^{-7}$). No association with Bonferroni significance was detected in this analysis, and top signal with borderline association was detected at DMP cg07051796 in ZFHX3 (beta=-0.01, SE=$3.0x10^{-3}$, p=$3.5x10^{-7}$).

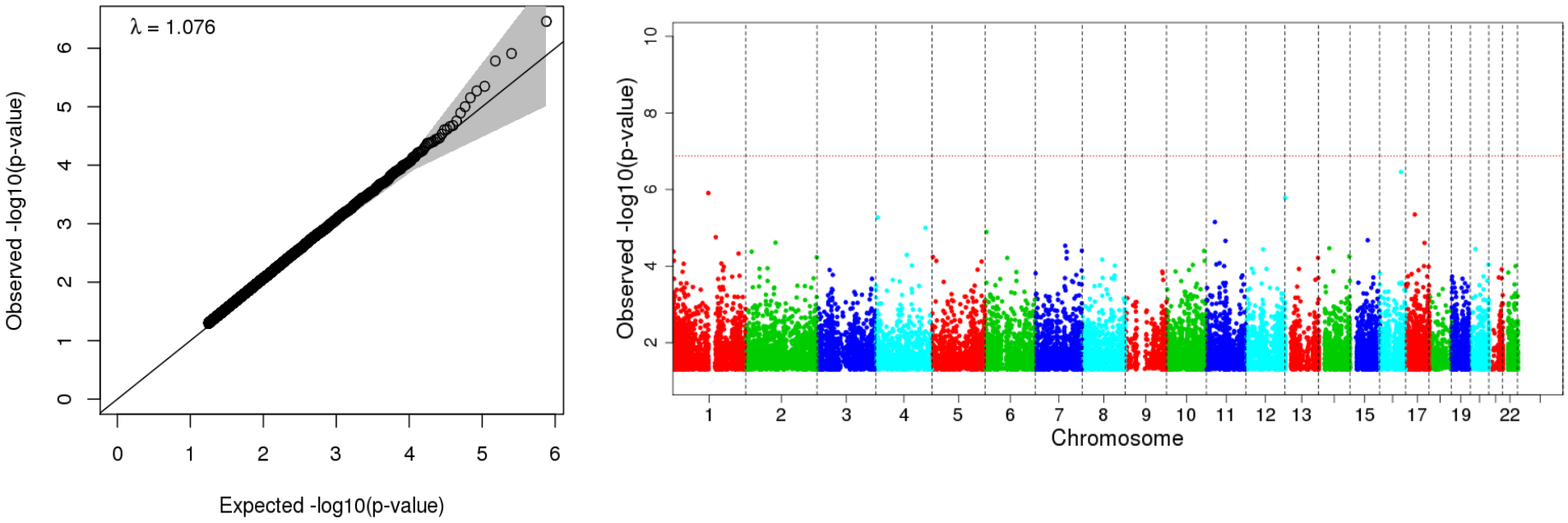


Table S8-7. Comparison of results of the EWAS in T2D between ALSPAC/ARIES and LBC1936. Models considered were a minimally adjusted model (Model 1), adjusted for age, sex and SVs; a second model additionally adjusted for predicted cell-counts (Model 2), a third model additionally adjusted for smoking (Model 3), and a fully-adjusted model additionally adjusted for BMI (Model 4).

	ALSPAC/ARIES					LBC1936				
	CpG	**Loci**	**Beta**	**SE**	**P-value**	**CpG**	**Loci**	**Beta**	**SE**	**P-value**
Model 1	cg10870892	*CTTN*	-0.050	0.009	6.24E-08	cg06500161	*ABCG1*	0.025	0.004	1.94E-08
	cg24605023	*CADPS*	-0.031	0.006	4.84E-07	cg27243685	*ABCG1*	0.017	0.004	1.07E-06
	cg15986668	*NFYC*	-0.064	0.013	7.00E-07	cg20068400	*Unannotated*	-0.036	0.007	1.38E-06
	cg17749033	*Unannotated*	-0.019	0.004	1.29E-06	cg17055821	*C17orf75*	0.030	0.006	1.64E-06
	cg25341923	*KRTAP4-7*	-0.016	0.003	1.74E-06	cg12194745	*BAHCC1*	0.018	0.004	2.06E-06
	cg19823491	*OTX1*	-0.006	0.001	1.79E-06	cg13555278	*EXTL1*	0.017	0.004	3.29E-06
	cg26353859	*SLC16A7*	0.031	0.007	2.99E-06	cg07051796	*ZFHX3*	-0.012	0.003	3.81E-06
	cg04016326	*GRIN2B*	-0.054	0.012	3.44E-06	cg03312117	*Unannotated*	-0.043	0.009	4.07E-06
	cg05575921	*AHRR*	-0.036	0.008	3.61E-06	cg09371351	*HSD3B2*	0.038	0.008	4.59E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.38E-06	cg08515811	*TBC1D16*	-0.013	0.003	4.67E-06
Model 2	cg14045803	*STARD10*	-0.011	0.002	2.70E-07	cg06500161	*ABCG1*	0.024	0.004	5.89E-08
	cg15986668	*NFYC*	-0.065	0.013	4.61E-07	cg19693031	*TXNIP*	-0.026	0.006	2.68E-06
	cg10870892	*CTTN*	-0.045	0.009	6.40E-07	cg09371351	*HSD3B2*	0.039	0.008	2.77E-06
	cg19823491	*OTX1*	-0.006	0.001	1.62E-06	cg07051796	*ZFHX3*	-0.012	0.003	3.51E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.71E-06	cg27243685	*ABCG1*	0.016	0.003	4.89E-06
	cg26652413	*CPAMD8*	-0.022	0.005	2.96E-06	cg17055821	*C17orf75*	0.026	0.006	4.97E-06
	cg00204249	*DNAH17*	-0.014	0.003	3.29E-06	cg20068400	*Unannotated*	-0.029	0.006	6.70E-06
	cg05575921	*AHRR*	-0.036	0.008	3.72E-06	cg15127702	*EMID2*	-0.019	0.004	7.35E-06
	cg14290451	*RPL10A*	-0.004	0.001	4.38E-06	cg13565670	*FBRSL1*	0.010	0.002	8.96E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.76E-06	cg03312117	*Unannotated*	-0.040	0.009	1.15E-05
Model 3	cg14045803	*STARD10*	-0.011	0.002	3.07E-07	cg06500161	*ABCG1*	0.024	0.004	5.18E-08
	cg15986668	*NFYC*	-0.065	0.013	5.78E-07	cg09371351	*HSD3B2*	0.040	0.008	2.31E-06
	cg10870892	*CTTN*	-0.045	0.009	7.58E-07	cg07051796	*ZFHX3*	-0.012	0.003	3.40E-06
	cg19823491	*OTX1*	-0.006	0.001	1.70E-06	cg19693031	*TXNIP*	-0.026	0.006	3.88E-06
	cg00204249	*DNAH17*	-0.014	0.003	2.47E-06	cg27243685	*ABCG1*	0.016	0.003	4.65E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.78E-06	cg17055821	*C17orf75*	0.025	0.006	6.22E-06
	cg26652413	*CPAMD8*	-0.022	0.005	3.49E-06	cg20068400	*Unannotated*	-0.029	0.006	7.41E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.84E-06	cg13565670	*FBRSL1*	0.010	0.002	9.33E-06
	cg14290451	*RPL10A*	-0.004	0.001	5.20E-06	cg03312117	*Unannotated*	-0.041	0.009	9.35E-06
	cg24605023	*CADPS*	-0.028	0.006	6.06E-06	cg15127702	*EMID2*	-0.019	0.004	9.62E-06
Model 4	cg15986668	*NFYC*	-0.071	0.013	5.48E-08	cg07051796	*ZFHX3*	-0.013	0.003	3.47E-07
	cg14045803	*STARD10*	-0.012	0.002	1.39E-07	cg09371351	*HSD3B2*	0.042	0.009	1.23E-06
	cg10870892	*CTTN*	-0.045	0.009	1.13E-06	cg13565670	*FBRSL1*	0.011	0.002	1.65E-06
	cg26652413	*CPAMD8*	-0.023	0.005	2.51E-06	cg17055821	*C17orf75*	0.027	0.006	4.47E-06
	cg00204249	*DNAH17*	-0.015	0.003	2.76E-06	cg22077313	*Unannotated*	0.022	0.005	5.36E-06
	cg03206717	*SLC25A38*	-0.003	0.001	2.95E-06	cg20068400	*Unannotated*	-0.030	0.007	7.02E-06
	cg19823491	*OTX1*	-0.006	0.001	2.99E-06	cg17384323	*Unannotated*	0.012	0.003	9.98E-06
	cg02307288	*TRPC7*	-0.038	0.008	5.54E-06	cg21740964	*FAM50B*	-0.033	0.007	1.28E-05
	cg04016326	*GRIN2B*	-0.054	0.012	5.71E-06	cg19693031	*TXNIP*	-0.025	0.006	1.75E-05
	cg04656330	*PNKD*	-0.002	0.000	7.96E-06	cg16611005	*FOXB1*	-0.003	0.001	2.10E-05

Figure S8-9 Q-Q plot (left) and Manhattan plot (right) showing results of the EWAS of T2D conducted in participants in the Rotterdam Study III-1 (n=723). A fully adjusted model was used adjusting for age, sex, 8 SVs, 6 predicted cell-counts, BMI and smoking (non-smoker, smoker). Q-Q plot shows the distribution of observed versus expected P-values, and the red line represents the distribution of P-values under the null hypothesis of no-associations. Deviation of observed P-values from the line of null associations, and outside the grey area demarking the 95% confidence intervals, suggested signals in strong association with T2D. Manhattan plot shows the distribution of P-values across genomic coordinates for each CpG tested. The horizontal red line is the Bonferroni corrected p-threshold ($p=1.07x10^{-7}$). No association with Bonferroni significance was detected in this analysis, and top signal with borderline association was detected at the DMP cg16330965 in SNAPC5 ($beta=-0.01$, $SE=3.0x10^{-3}$, $p=5.0x10^{-7}$).

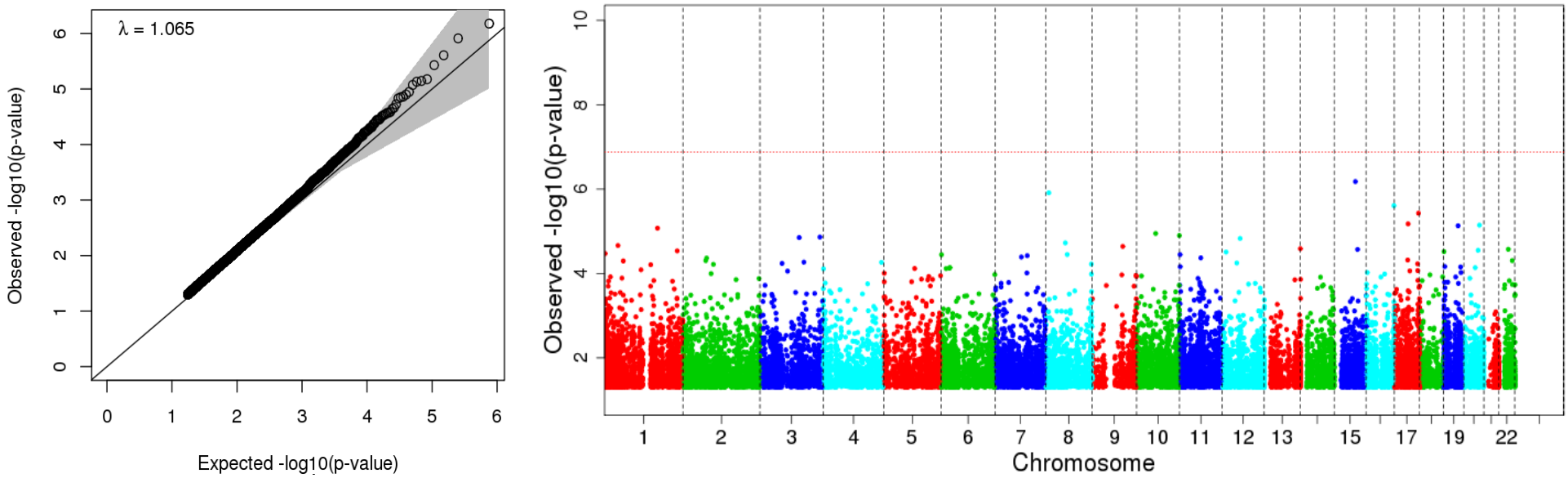


Table S8-8 Comparison of results of the EWAS in T2D between ALSPAC/ARIES and the Rotterdam study (RSIII-1, n=723). Models considered were the minimally-adjusted model (Model 1), adjusted for age, sex and SVs; a second model additionally adjusted for predicted cell-counts (Model 2), a third model additionally adjusted for smoking (Model 3), and a fully-adjusted model additionally adjusted for BMI (Model 4).

	ALSPAC/ARIES					Rotterdam Study-III-1				
	CpG	**Loci**	**Beta**	**SE**	**P-value**	**CpG**	**Loci**	**Beta**	**SE**	**P-value**
Model 1	cg10870892	*CTTN*	-0.050	0.009	6.24E-08	cg20052079	*JARID2*	-0.034	0.007	8.30E-07
	cg24605023	*CADPS*	-0.031	0.006	4.84E-07	cg05311626	*TOLLIP*	-0.009	0.002	1.03E-06
	cg15986668	*NFYC*	-0.064	0.013	7.00E-07	cg05887092	*PGS1*	-0.013	0.003	1.35E-06
	cg17749033	*Unannotated*	-0.019	0.004	1.29E-06	cg13212575	*MAEL*	-0.009	0.002	1.51E-06
	cg25341923	*KRTAP4-7*	-0.016	0.003	1.74E-06	cg16646600	*C1orf83*	-0.013	0.003	1.65E-06
	cg19823491	*OTX1*	-0.006	0.001	1.79E-06	cg26321644	*PLTP*	0.010	0.002	1.70E-06
	cg26353859	*SLC16A7*	0.031	0.007	2.99E-06	cg07746918	*SREBF2*	0.008	0.002	1.90E-06
	cg04016326	*GRIN2B*	-0.054	0.012	3.44E-06	cg08121984	*APOC1P1*	-0.013	0.003	2.63E-06
	cg05575921	*AHRR*	-0.036	0.008	3.61E-06	cg05375728	*DAB1*	0.009	0.002	3.57E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.38E-06	cg07264682	*Unannotated*	-0.023	0.005	6.89E-06
Model 2	cg14045803	*STARD10*	-0.011	0.002	2.70E-07	cg14278808	*LOC157627*	0.020	0.004	2.71E-07
	cg15986668	*NFYC*	-0.065	0.013	4.61E-07	cg16330965	*SNAPC5*	-0.012	0.002	5.29E-07
	cg10870892	*CTTN*	-0.045	0.009	6.40E-07	cg13212575	*MAEL*	-0.009	0.002	9.05E-07
	cg19823491	*OTX1*	-0.006	0.001	1.62E-06	cg02484673	*JPH3*	-0.012	0.003	1.30E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.71E-06	cg05887092	*PGS1*	-0.012	0.003	3.16E-06
	cg26652413	*CPAMD8*	-0.022	0.005	2.96E-06	cg08121984	*APOC1P1*	-0.012	0.003	3.90E-06
	cg00204249	*DNAH17*	-0.014	0.003	3.29E-06	cg04845819	*PRR23C*	-0.011	0.002	4.21E-06
	cg05575921	*AHRR*	-0.036	0.008	3.72E-06	cg21193660	*Unannotated*	-0.017	0.004	5.75E-06
	cg14290451	*RPL10A*	-0.004	0.001	4.38E-06	cg00574958	*CPT1A*	-0.018	0.004	6.16E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.76E-06	cg05084700	*Unannotated*	-0.011	0.002	6.23E-06
Model 3	cg14045803	*STARD10*	-0.011	0.002	3.07E-07	cg14278808	*LOC157627*	0.020	0.004	2.75E-07
	cg15986668	*NFYC*	-0.065	0.013	5.78E-07	cg16330965	*SNAPC5*	-0.012	0.002	4.79E-07
	cg10870892	*CTTN*	-0.045	0.009	7.58E-07	cg13212575	*MAEL*	-0.009	0.002	9.15E-07
	cg19823491	*OTX1*	-0.006	0.001	1.70E-06	cg02484673	*JPH3*	-0.012	0.003	1.32E-06
	cg00204249	*DNAH17*	-0.014	0.003	2.47E-06	cg05887092	*PGS1*	-0.012	0.003	3.15E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.78E-06	cg08121984	*APOC1P1*	-0.012	0.003	3.88E-06
	cg26652413	*CPAMD8*	-0.022	0.005	3.49E-06	cg04845819	*PRR23C*	-0.011	0.002	4.26E-06
	cg03206717	*SLC25A38*	-0.003	0.001	4.84E-06	cg21193660	*Unannotated*	-0.017	0.004	5.79E-06
	cg14290451	*RPL10A*	-0.004	0.001	5.20E-06	cg05084700	*Unannotated*	-0.011	0.002	6.18E-06
	cg24605023	*CADPS*	-0.028	0.006	6.06E-06	cg00574958	*CPT1A*	-0.018	0.004	6.19E-06
Model 4	cg15986668	*NFYC*	-0.071	0.013	5.48E-08	cg16330965	*SNAPC5*	-0.013	0.003	5.00E-07
	cg14045803	*STARD10*	-0.012	0.002	1.39E-07	cg14278808	*LOC157627*	0.020	0.004	1.04E-06
	cg10870892	*CTTN*	-0.045	0.009	1.13E-06	cg02484673	*JPH3*	-0.012	0.003	2.01E-06
	cg26652413	*CPAMD8*	-0.023	0.005	2.51E-06	cg05887092	*PGS1*	-0.013	0.003	3.20E-06
	cg00204249	*DNAH17*	-0.015	0.003	2.76E-06	cg21477861	*PLCD3*	-0.004	0.001	5.79E-06
	cg03206717	*SLC25A38*	-0.003	0.001	2.95E-06	cg08121984	*APOC1P1*	-0.012	0.003	6.82E-06
	cg19823491	*OTX1*	-0.006	0.001	2.99E-06	cg12986726	*CEBPB*	0.009	0.002	7.40E-06
	cg02307288	*TRPC7*	-0.038	0.008	5.54E-06	cg13212575	*MAEL*	-0.008	0.002	9.47E-06
	cg04016326	*GRIN2B*	-0.054	0.012	5.71E-06	cg07264682	*Unannotated*	-0.022	0.005	1.02E-05
	cg04656330	*PNKD*	-0.002	0.000	7.96E-06	cg07416844	*Unannotated*	-0.013	0.003	1.20E-05

Figure S8-10 Q-Q plot (left) and Manhattan plot (right) showing results of the EWAS of T2D conducted in participants in Rotterdam-Bios (n=723). A fully adjusted model was used adjusting for age, sex, 9 SVs, predicted cell-counts, BMI and smoking. Q-Q plot shows the distribution of observed versus expected P-values, and the red line represents the distribution of P-values under the null hypothesis of no-associations. Deviation of observed P-values from the line of null associations, and outside the grey area demarking the 95% confidence intervals, suggested signals in strong association with T2D. Manhattan plot shows the distribution of p-values across genomic coordinates for each CpG tested. The horizontal red line is the Bonferroni corrected p-threshold ($p=1.07x10^{-7}$). No association with Bonferroni significance was detected in this analysis, and top signal with borderline association was detected at the DMP cg16339915 in TIFAB (beta=-0.01, SE=0.002, $p=3.32x10^{-6}$).

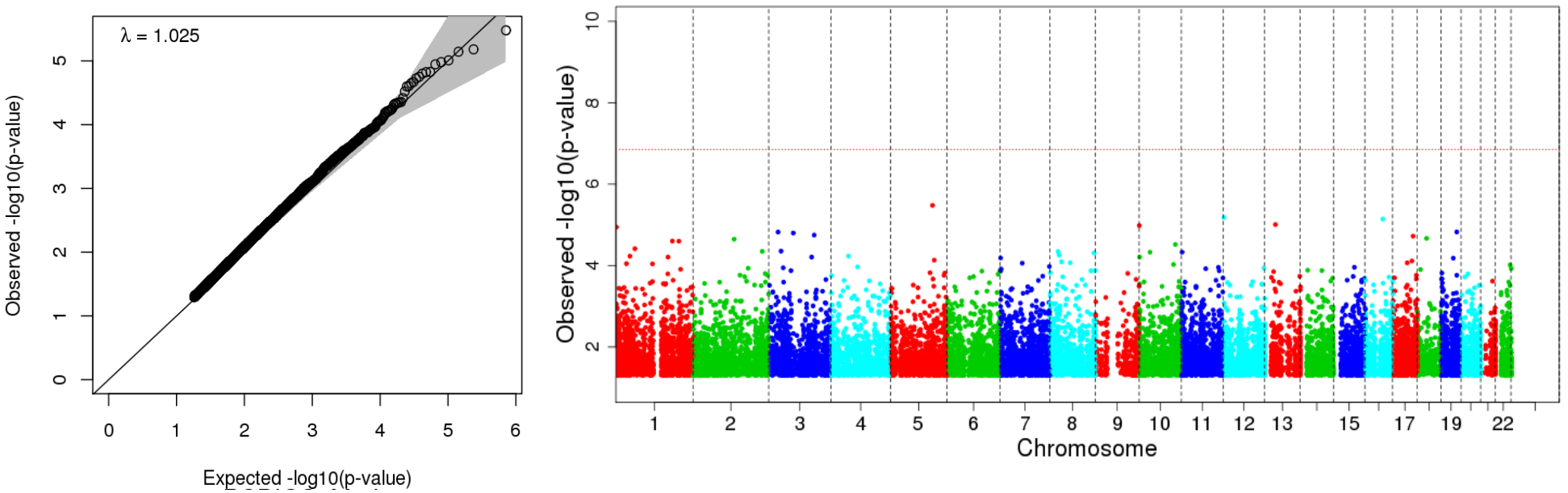


Table S8-9 Comparison of results of the EWAS in T2D between ALSPAC/ARIES and Rotterdam Bios (RS-Bios, n=723). Models considered were the minimally-adjusted model (Model 1), adjusted for age, sex and SVs; a second model additionally adjusted for predicted cell-counts (Model 2), a third model additionally adjusted for smoking (Model 3), and a fully-adjusted model additionally adjusted for BMI (Model 4).

	ALSPAC/ARIES					Rotterdam-Bios				
	CpG	**Loci**	**Beta**	**SE**	**P-value**	**CpG**	**Loci**	**Beta**	**SE**	**P-value**
Model 1	cg10870892	*CTTN*	-0.050	0.009	6.24E-08	cg24795867	*WNT5B*	-0.008	0.002	1.69E-06
	cg24605023	*CADPS*	-0.031	0.006	4.84E-07	cg05059607	*PITPNC1*	0.021	0.004	2.34E-06
	cg15986668	*NFYC*	-0.064	0.013	7.00E-07	cg02243386	*XIRP1*	-0.009	0.002	4.49E-06
	cg17749033	*Unannotated*	-0.019	0.004	1.29E-06	cg01316152	*Unannotated*	-0.008	0.002	5.59E-06
	cg25341923	*KRTAP4-7*	-0.016	0.003	1.74E-06	cg16929139	*RNF125*	0.016	0.004	6.31E-06
	cg19823491	*OTX1*	-0.006	0.001	1.79E-06	cg03328041	*FAM168B*	0.006	0.001	6.81E-06
	cg26353859	*SLC16A7*	0.031	0.007	2.99E-06	cg14491707	*CACNA1B*	-0.015	0.003	1.02E-05
	cg04016326	*GRIN2B*	-0.054	0.012	3.44E-06	cg11983038	*Unannotated*	-0.024	0.005	1.29E-05
	cg05575921	*AHRR*	-0.036	0.008	3.61E-06	cg02859537	*AKT1S1*	0.012	0.003	1.30E-05
	cg03206717	*SLC25A38*	-0.003	0.001	4.38E-06	cg24512093	*Unannotated*	-0.011	0.003	1.42E-05
Model 2	cg14045803	*STARD10*	-0.011	0.002	2.70E-07	cg05059607	*PITPNC1*	0.021	0.004	1.95E-06
	cg15986668	*NFYC*	-0.065	0.013	4.61E-07	cg24795867	*WNT5B*	-0.008	0.002	2.29E-06
	cg10870892	*CTTN*	-0.045	0.009	6.40E-07	cg16339915	*TIFAB*	-0.007	0.002	3.65E-06
	cg19823491	*OTX1*	-0.006	0.001	1.62E-06	cg14491707	*CACNA1B*	-0.016	0.003	4.00E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.71E-06	cg02243386	*XIRP1*	-0.008	0.002	7.44E-06
	cg26652413	*CPAMD8*	-0.022	0.005	2.96E-06	cg11983038	*Unannotated*	-0.025	0.006	8.82E-06
	cg00204249	*DNAH17*	-0.014	0.003	3.29E-06	cg24512093	*ROBO1*	-0.011	0.003	1.05E-05
	cg05575921	*AHRR*	-0.036	0.008	3.72E-06	cg01316152	*Unannotated*	-0.008	0.002	1.37E-05
	cg14290451	*RPL10A*	-0.004	0.001	4.38E-06	cg10500218	*IER5*	0.006	0.001	1.54E-05
	cg03206717	*SLC25A38*	-0.003	0.001	4.76E-06	cg18143317	*Unannotated*	-0.007	0.002	1.64E-05
Model 3	cg14045803	*STARD10*	-0.011	0.002	3.07E-07	cg05059607	*PITPNC1*	0.021	0.004	1.95E-06
	cg15986668	*NFYC*	-0.065	0.013	5.78E-07	cg24795867	*WNT5B*	-0.008	0.002	2.29E-06
	cg10870892	*CTTN*	-0.045	0.009	7.58E-07	cg16339915	*TIFAB*	-0.007	0.002	3.65E-06
	cg19823491	*OTX1*	-0.006	0.001	1.70E-06	cg14491707	*CACNA1B*	-0.016	0.003	4.00E-06
	cg00204249	*DNAH17*	-0.014	0.003	2.47E-06	cg02243386	*XIRP1*	-0.008	0.002	7.44E-06
	cg04016326	*GRIN2B*	-0.055	0.012	2.78E-06	cg11983038	*Unannotated*	-0.025	0.006	8.82E-06
	cg26652413	*CPAMD8*	-0.022	0.005	3.49E-06	cg24512093	*ROBO1*	-0.011	0.003	1.05E-05
	cg03206717	*SLC25A38*	-0.003	0.001	4.84E-06	cg01316152	*Unannotated*	-0.008	0.002	1.37E-05
	cg14290451	*RPL10A*	-0.004	0.001	5.20E-06	cg10500218	*IER5*	0.006	0.001	1.54E-05
	cg24605023	*CADPS*	-0.028	0.006	6.06E-06	cg18143317	*Unannotated*	-0.007	0.002	1.64E-05
Model 4	cg15986668	*NFYC*	-0.071	0.013	5.48E-08	cg16339915	*TIFAB*	-0.007	0.002	3.32E-06
	cg14045803	*STARD10*	-0.012	0.002	1.39E-07	cg24795867	*WNT5B*	-0.008	0.002	6.61E-06
	cg10870892	*CTTN*	-0.045	0.009	1.13E-06	cg16575444	*CX3CL1*	-0.008	0.002	7.19E-06
	cg26652413	*CPAMD8*	-0.023	0.005	2.51E-06	cg11983038	*Unannotated*	-0.025	0.006	9.82E-06
	cg00204249	*DNAH17*	-0.015	0.003	2.76E-06	cg14491707	*CACNA1B*	-0.015	0.003	1.05E-05
	cg03206717	*SLC25A38*	-0.003	0.001	2.95E-06	cg02635644	*AGRN*	-0.003	0.001	1.13E-05
	cg19823491	*OTX1*	-0.006	0.001	2.99E-06	cg02859537	*AKT1S1*	0.012	0.003	1.49E-05
	cg02307288	*TRPC7*	-0.038	0.008	5.54E-06	cg16565002	*RBMS3*	-0.011	0.003	1.50E-05
	cg04016326	*GRIN2B*	-0.054	0.012	5.71E-06	cg24512093	*ROBO1*	-0.011	0.003	1.59E-05
	cg04656330	*PNKD*	-0.002	0.000	7.96E-06	cg25250358	*PLOD2*	-0.015	0.004	1.78E-05

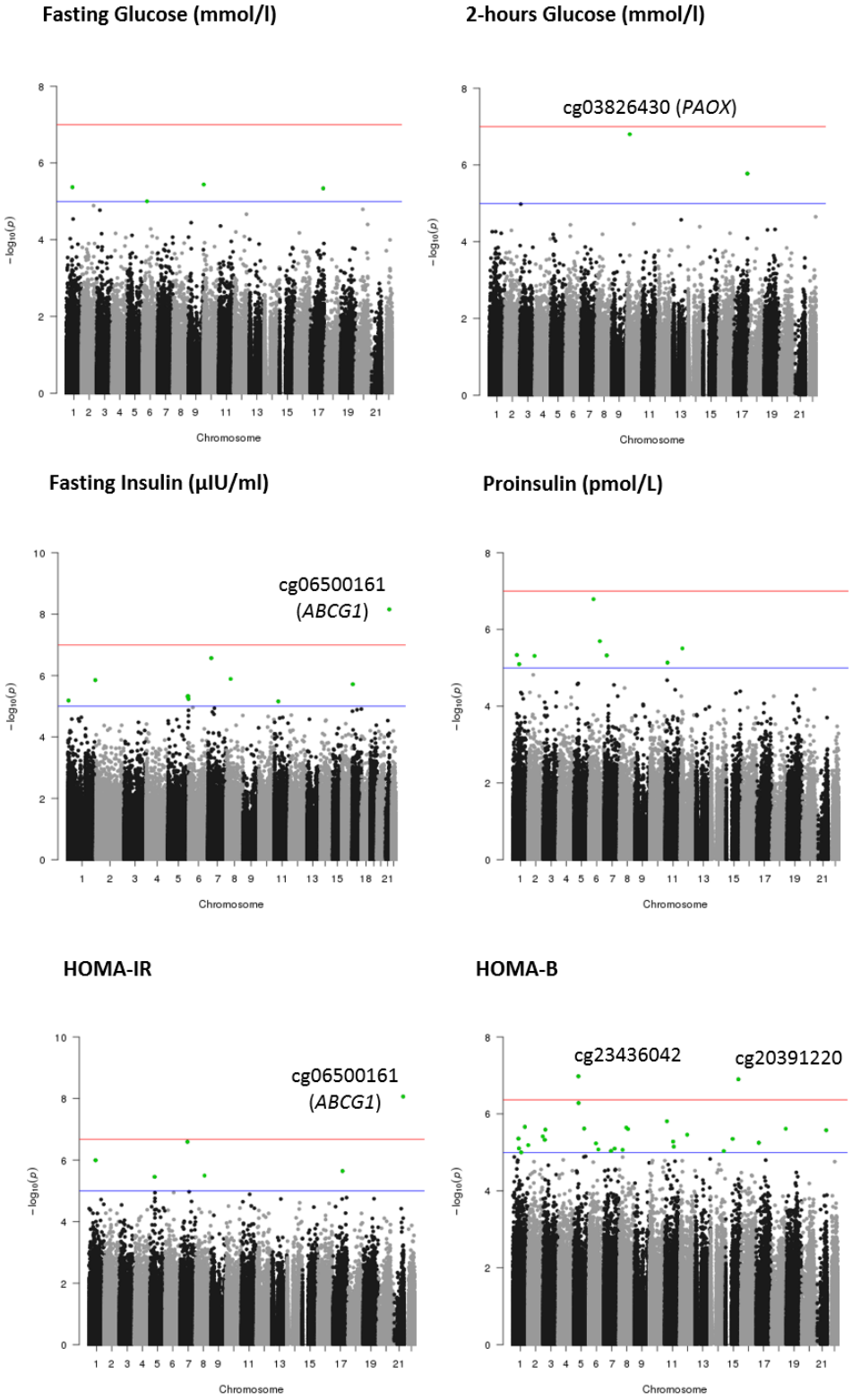
Table S8-10 Comparison of top-ten signals detected in the EWAS of prevalent T2D using participants in ALSPAC (n=1050, 48 cases and 1002 controls). Model 1 adjusted for age, sex, 3 SVs and 6 predicted cell counts (Houseman et al.[119]), Model 2 additionally adjusted for smoking (non-smoker, smoker) and Model 3 additionally adjusted for BMI. Genetic context: position of the probe with respect to the closest gene; CpG island: position of the CpG site with respect to the nearest CpG island (shore: within 2kb from the CpG island, shelf: >2kb from CpG island, open sea: >4kb from CpG island); OR: effect of 1% increase in methylation in the odds of T2D; SE: standard error (odds-scale); P: p-value of significance without adjustment for multiple testing; FDR: false-discovery rate. Associations were significant at $p<1.07x10^{-7}$ or FDR<0.05.

	CpG	Chr	Gene	Genetic context	CpG island	OR	SE	P	FDR
Model 1	cg15986668	1	*NFYC*	TSS1500	N_Shore	0.92	1.02	3.40X10-6	0.75
	cg05575921	5	*AHRR*	Body	N_Shore	0.90	1.02	5.16X10-6	0.75
	cg06500161	21	*ABCG1*	Body	S_Shore	1.16	1.03	6.99X10-6	0.75
	cg03734323	10	*GDF2*	Body	Island	0.76	1.07	1.65X10-5	0.75
	cg23867673	10	*CDH23*	Body	Open sea	0.89	1.03	1.69X10-5	0.75
	cg24605023	3	*CADPS*	Body	Open sea	0.88	1.03	2.63X10-5	0.75
	cg13888858	9	*ANKRD19*	Body	Island	1.29	1.06	2.78X10-5	0.75
	cg21597596	6	*CAP2*	1stExon	Open sea	1.60	1.12	2.92X10-5	0.75
	cg26353859	12	*SLC16A7*	TSS1500	Open sea	1.17	1.04	3.24X10-5	0.75
	cg10870892	11	*CTTN*	Body	Open sea	0.92	1.02	3.35X10-5	0.75
Model 2	cg15986668	1	*NFYC*	TSS1500	N_Shore	0.92	1.02	3.88X10-6	0.83
	cg06500161	21	*ABCG1*	Body	S_Shore	1.17	1.03	4.72X10-6	0.83
	cg03734323	10	*GDF2*	Body	Island	0.76	1.07	1.71X10-5	0.83
	cg03364130	22	*BRD1*	Body	Island	1.23	1.05	2.47X10-5	0.83
	cg21597596	6	*CAP2*	1stExon	Open sea	1.60	1.12	2.59X10-5	0.83
	cg23867673	10	*CDH23*	Body	Open sea	0.89	1.03	2.61X10-5	0.83
	cg26353859	12	*SLC16A7*	TSS1500	Open sea	1.17	1.04	3.11X10-5	0.83
	cg04656330	2	*PNKD*	Body	Island	0.04	2.16	3.11X10-5	0.83
	cg10870892	11	*CTTN*	Body	Open sea	0.92	1.02	3.31X10-5	0.83
	cg19545560	19	*Intergenic*	Intergenic	S_Shelf	0.84	1.04	3.66X10-5	0.83
Model 3	cg15986668	1	*NFYC*	TSS1500	N_Shore	0.92	1.02	1.86X10-6	0.87
	cg02307288	5	*TRPC7*	3'UTR	Open sea	0.88	1.03	1.12X10-5	0.87
	cg03364130	22	*BRD1*	Body	Island	1.25	1.05	1.66X10-5	0.87
	cg03923934	11	*OPCML*	5'UTR	Open sea	1.31	1.06	2.00X10-5	0.87
	cg04656330	2	*PNKD*	Body	Island	0.04	2.18	2.02X10-5	0.87
	cg13888858	9	*ANKRD19*	Body	Island	1.30	1.06	2.39X10-5	0.87
	cg26353859	12	*SLC16A7*	TSS1500	Open sea	1.18	1.04	3.35X10-5	0.87
	cg03734323	10	*GDF2*	Body	Island	0.76	1.07	3.64X10-5	0.87
	cg08870587	11	*SHANK2*	Body	Open sea	1.10	1.02	4.20X10-5	0.87
	cg21597596	6	*CAP2*	1stExon	Open sea	1.60	1.12	4.35X10-5	0.87

Table S8-11 Comparison of top-ten signals detected in the EWAS of fasting glucose in a subsample of normoglycemic middle-age participants in ALSPAC (n=1002). Model 1 adjusted for age, sex, 3SVs and 6-predicted cell counts (Houseman et al. [119]), Model 2 additionally adjusted for smoking (non-smoker, smoker) and Model 3 additionally adjusted for BMI. Genetic context: position of the probe with respect to the closest gene; CpG island: position of the CpG site with respect to the nearest CpG island (shore: within 2kb from the CpG island, shelf: >2kb from CpG island, open sea: >4kb from CpG island); Beta: regression coefficient showing the effect of 1% increase in methylation on one unit change in the levels of fasting glucose (mmol/l); SE: standard error; P: p of significance without adjustment for multiple testing; FDR: false-discovery rate. Associations were significant at p< $1.07x10^{-7}$ or FDR < 0.05.

	CpG	Chr	Gene	Genetic context	CpG island	Beta	SE	P	FDR
Model 1	cg01099300	10	*Intergenic*	Intergenic	Open sea	-0.028	0.006	3.49X10-6	0.71
	cg23274377	1	*BPNT1*	TSS200	Open sea	0.046	0.010	4.05X10-6	0.71
	cg17540765	17	*RECQL5*	Body	S_Shelf	-0.017	0.004	4.50X10-6	0.71
	cg17219086	6	*Intergenic*	Intergenic	Open sea	0.011	0.002	1.05X10-5	0.98
	cg26234543	2	*TMEM17*	Body	N_Shore	-0.012	0.003	1.25X10-5	0.98
	cg14099787	3	*Intergenic*	Intergenic	S_Shelf	0.018	0.004	1.64X10-5	1.00
	cg05857996	20	*EBF4*	Body	S_Shore	0.004	0.001	1.72X10-5	1.00
	cg23861120	12	*Intergenic*	Intergenic	Open sea	0.032	0.008	2.12X10-5	1.00
	cg22724847	1	*OR14C36*	1stExon	Open sea	0.008	0.002	2.88X10-5	1.00
	cg03693099	9	*CEL*	TSS1500	Open sea	-0.015	0.004	3.81X10-5	1.00
Model 2	cg01099300	10	*Intergenic*	Intergenic	Open sea	-0.028	0.006	3.63X10-6	0.72
	cg23274377	1	*BPNT1*	TSS200	Open sea	0.046	0.010	4.27X10-6	0.72
	cg17540765	17	*RECQL5*	Body	S_Shelf	-0.017	0.004	4.58X10-6	0.72
	cg17219086	6	*Intergenic*	Intergenic	Open sea	0.011	0.002	9.94X10-6	1.00
	cg26234543	2	*TMEM17*	Body	N_Shore	-0.012	0.003	1.29X10-5	1.00
	cg05857996	20	*EBF4*	Body	S_Shore	0.004	0.001	1.61X10-5	1.00
	cg14099787	3	*Intergenic*	Intergenic	S_Shelf	0.018	0.004	1.70X10-5	1.00
	cg23861120	12	*Intergenic*	Intergenic	Open sea	0.032	0.008	2.16X10-5	1.00
	cg22724847	1	*OR14C36*	1stExon	Open sea	0.008	0.002	2.88X10-5	1.00
	cg03693099	9	*CEL*	TSS1500	Open sea	-0.015	0.004	3.58X10-5	1.00
Model 3	cg17540765	17	*RECQL5*	Body	S_Shelf	-0.02	0.004	1.14X10-6	0.35
	cg01099300	10	*Intergenic*	Intergenic	Open sea	-0.03	0.006	2.15X10-6	0.35
	cg23274377	1	*BPNT1*	TSS200	Open sea	0.05	0.010	2.20X10-6	0.35
	cg26234543	2	*TMEM17*	Body	N_Shore	-0.01	0.003	5.46X10-6	0.64
	cg03693099	9	*CEL*	TSS1500	Open sea	-0.02	0.004	7.28X10-6	0.69
	cg04555287	20	*ARFGAP1*	3'UTR	Island	-0.02	0.004	1.77X10-5	1.00
	cg17219086	6	*Intergenic*	Intergenic	Open sea	0.01	0.002	2.52X10-5	1.00
	cg14099787	3	*Intergenic*	Intergenic	S_Shelf	0.02	0.004	3.24X10-5	1.00
	cg22724847	1	*OR14C36*	1stExon	Open sea	0.01	0.002	3.77X10-5	1.00
	cg10218733	16	*TEKT5*	Body	Island	0.01	0.002	4.81X10-5	1.00

Figure S8-11 Manhattan plot showing main results of the EWAS of glycaemic traits and prevalent T2D in two subsamples from ALSPAC. Methylation was considered the independent variable in all the analyses. Results correspond to a model adjusted for age, sex (not when including only females), SVs (batch effects), 6-Houseman cells and smoking (non-smoker, smoker). Blue line is the threshold of borderline significance in the EWAS at $p=1.0x10^{-5}$; red line is the threshold of EWAS significance at $p<1.07x10^{-7}$ or FDR < 0.05.



Continuation Figure S8-11

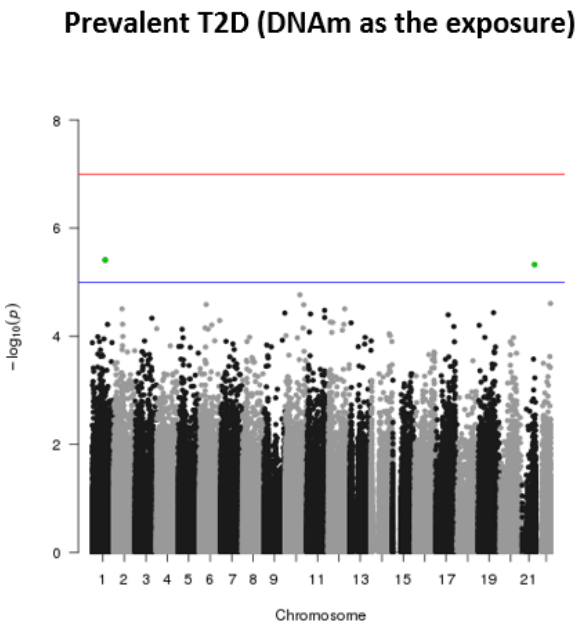
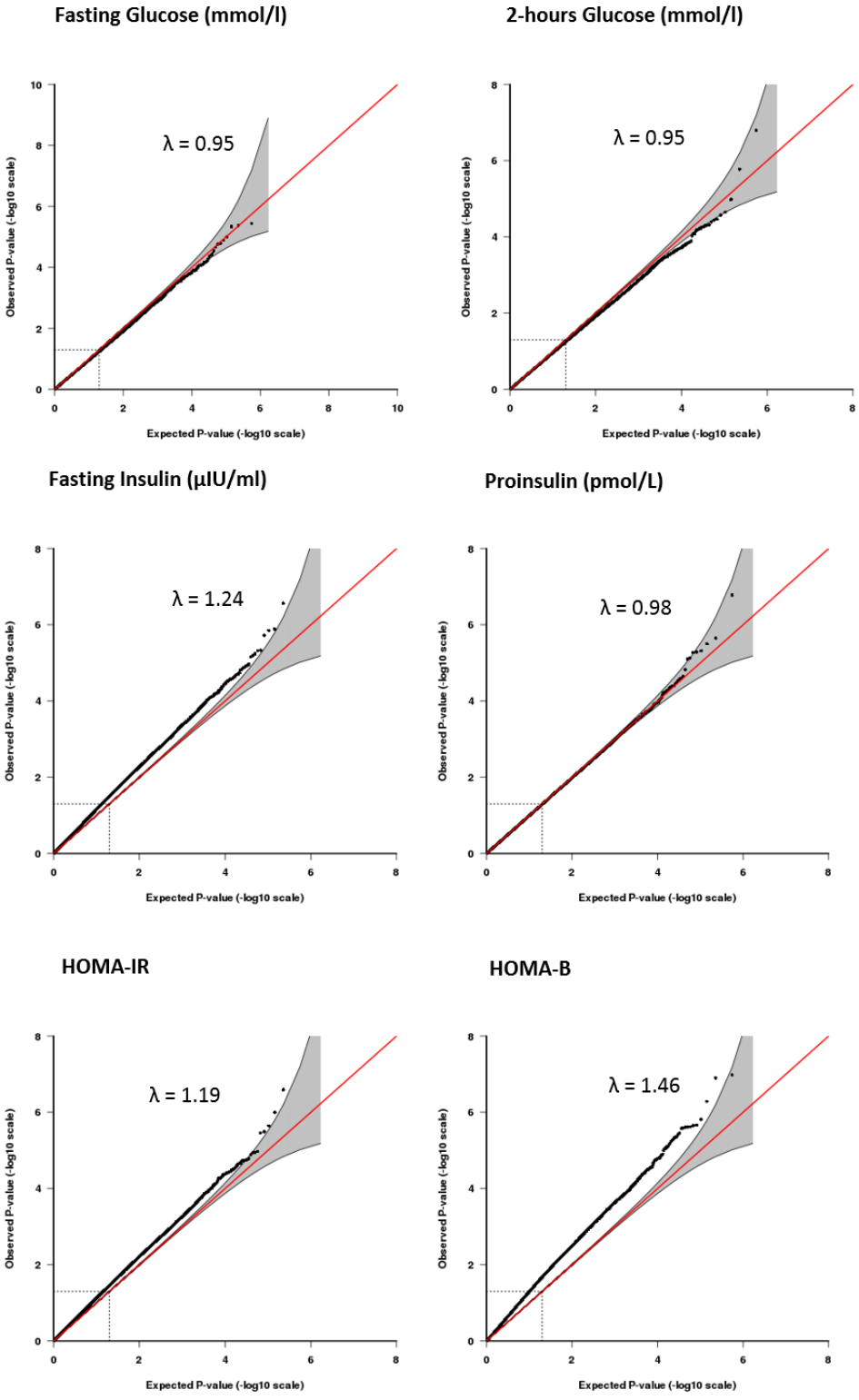


Figure S8-12 QQ-plot showing the distribution of observed versus expected p-values obtained in the EWAS of glycaemic traits (n=1002 and 622) and prevalent T2D (n=1050) in two subsamples of middle-age adults in ALSPAC. Red-line represents the distribution of observed p-values under the null hypothesis of no-associations, and deviation from this line represents potential significant associations. Grey area is the expected 95%CI distribution of -log10(p-values), and lambda (λ) is the genomic inflation factor; lambda > 1.0 indicates high genomic inflation, but this can also be due to analyses with small sample-size.



Continuation Figure S8-12

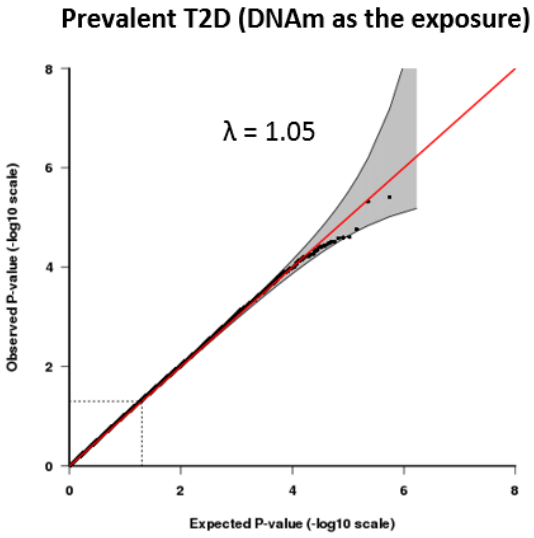
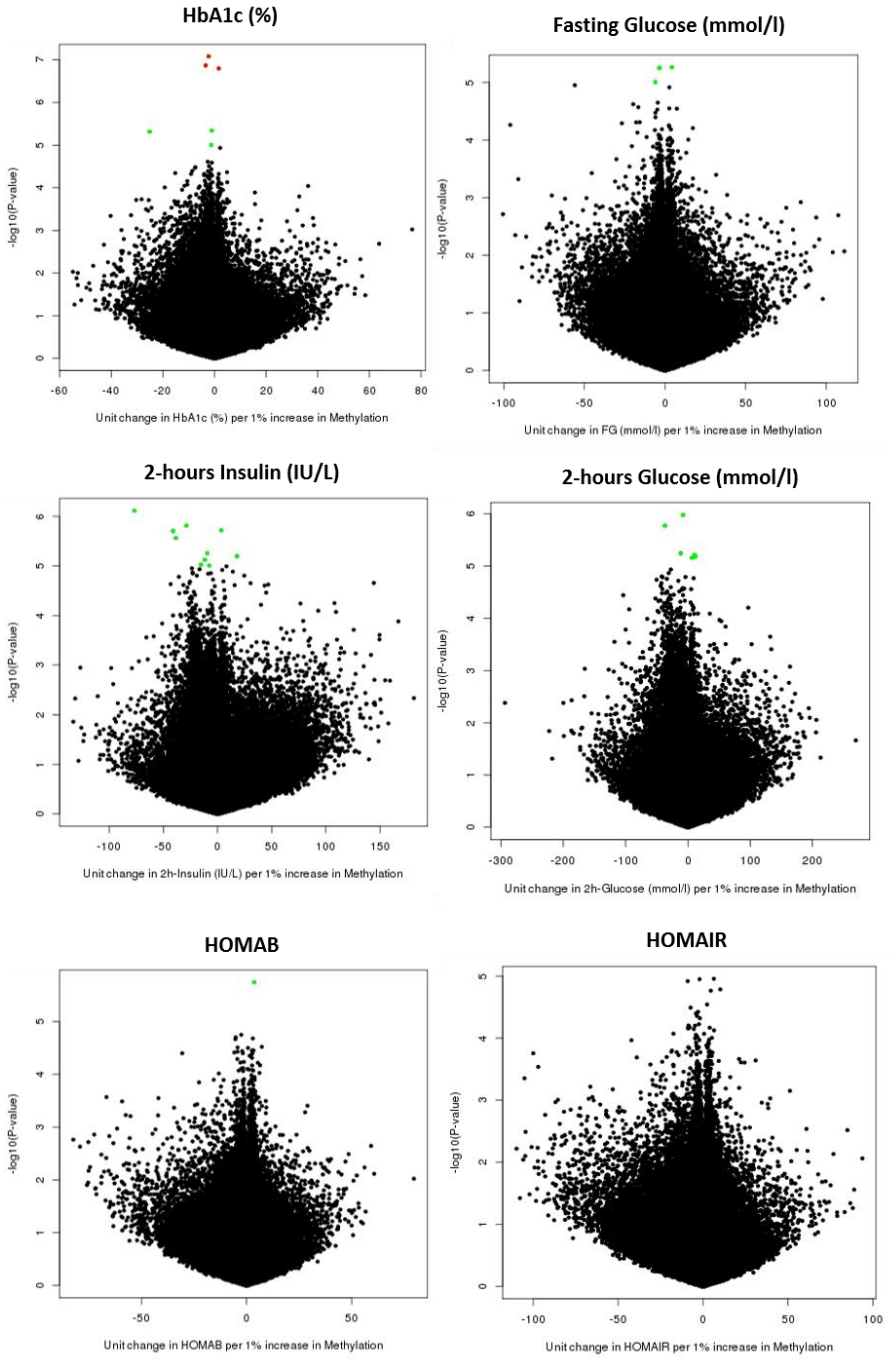


Figure S8-13 Distribution of effect estimates against the -Log10(p-value) for associations detected in the EWAS of glycaemic traits in SABRE (n=382 males). Results correspond to the model adjusted for age, SVs, 7 Houseman cells and smoking (never, former and current smokers). Effect estimates are interpreted as the effect of 1% increase in methylation on a unit change in the outcome. Effect estimates for variables log-transformed (i.e. fasting insulin, 2-hours insulin, HOMA-IR and HOMA-B), require back-transformation to the original units of the outcome using [exp(beta/100)], while for non-transformed outcomes (i.e. fasting glucose, 2-hours glucose and HbA1c), regression coefficients need to be interpreted after dividing coefficients by a hundred (beta/100). Associations were regarded borderline significant if -Log10(P-value) ≥ 5.0 or p < 1.0x10^{-5} (green-dots), and significant if -Log10(P-value) ≥ 7.0 or 1.06x10^{-7} (red-dots). An overrepresentation of probes with large positive effects was identified in the EWAS of 2-hours insulin and HbA1c.



Continuation Figure S8-13

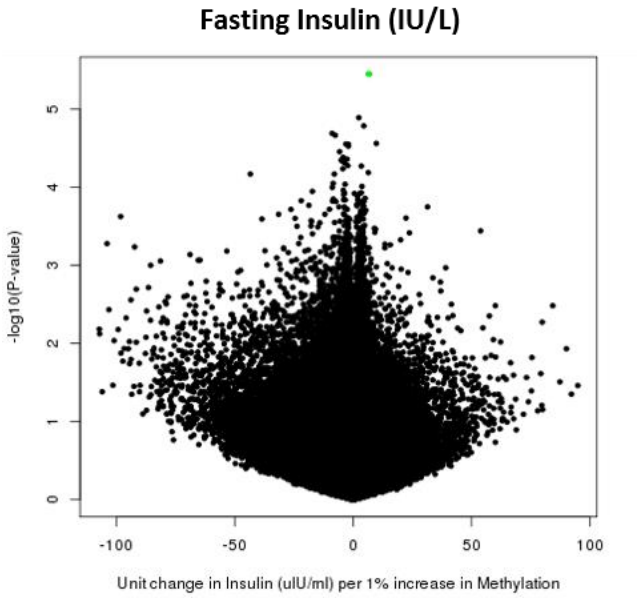
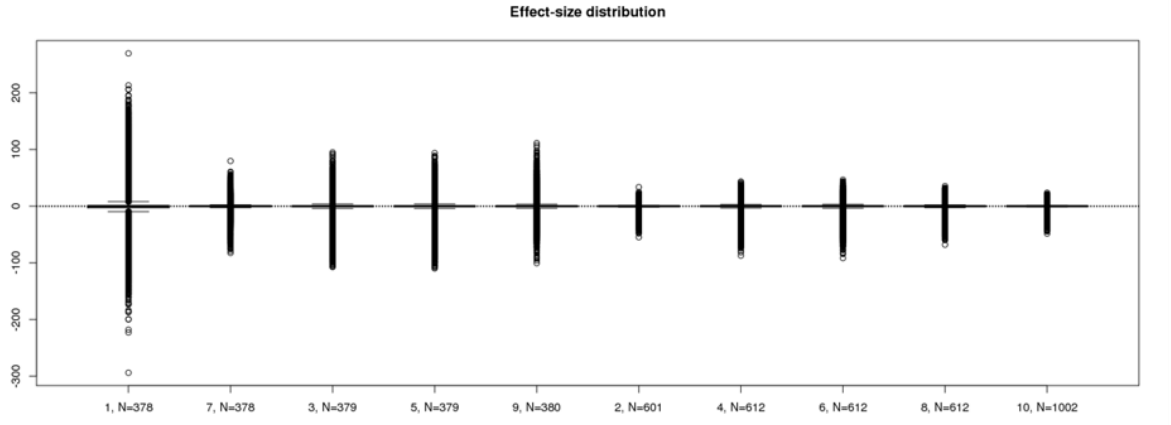
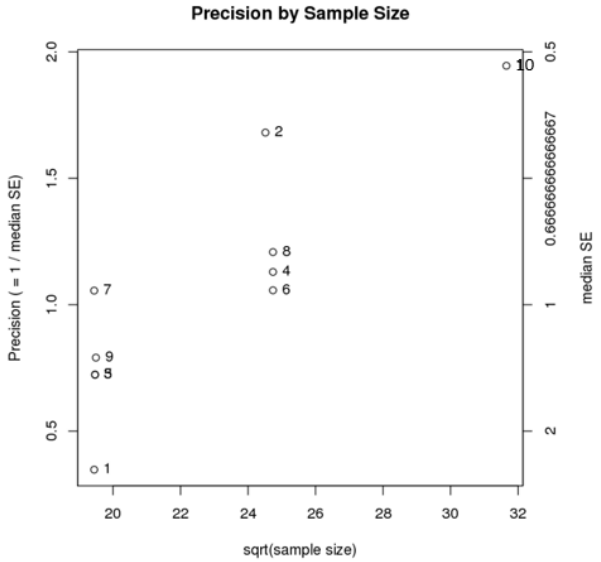


Figure S8-14 Distribution of the effect-size against sample-size (top-plot) and precision plot (bottom) reported by QCEWAS using results of the EWAS of glycaemic traits conducted in ALSPAC and SABRE. Results of the EWAS correspond to the model adjusted for age, SVs, predicted cells and smoking. A narrower distribution of effect-sizes indicates studies with a good control of outliers in the analysis, and this is a characteristic directly proportional to the sample size. In the precision plot, the different EWAS were arranged based on the magnitude of the variation in the standard error and the sample size. Studies with narrower distribution of the standard error and bigger sample-size have more precision in estimating the effect in the outcome (i.e. EWAS of fasting glucose in ALSPAC), compared to studies with larger distribution of the standard error and small sample-size (i.e. EWAS of 2-h glucose in SABRE).





Legend: 1= 2-h gluc (SABRE), 2= 2-h gluc(ALSPAC), 3= Fast. Insulin (SABRE), 4= Fast. Insulin (ALSPAC), 5= HOMA-IR (SABRE), 6= HOMA-IR (ALSPAC), 7= HOMA-B (SABRE), 8= HOMA-B (ALSPAC), 9= Fast. glucose (SABRE) and 10= Fast. glucose (ALSPAC).

Figure S8-15 Distribution of effect estimates against the -Log10(p-value) for the association between DNA methylation and glycaemic traits using results obtained in the meta-analysis between ALSPAC (n=1002 females and males and 622 females) and SABRE (n=382). Results were adjusted for age, SVs, Houseman cells and smoking. Effect estimates are interpreted as the effect of 1% increase in methylation on a unit change in the outcome; for fasting insulin, HOMA-IR and HOMA-B, effect estimates need to be transformed to the original units of the outcome using [exp(beta/100)], while for fasting glucose and 2-h glucose, effect estimates need to be transformed by applying beta/100. Associations were regarded borderline significant if -Log10(p-value) ≥ 5.0 or p < 1.0x10^{-5} (green-dots), and significant if -Log10(p-value) ≥ 7.0 or p<10x10^{-7} (red-dots).

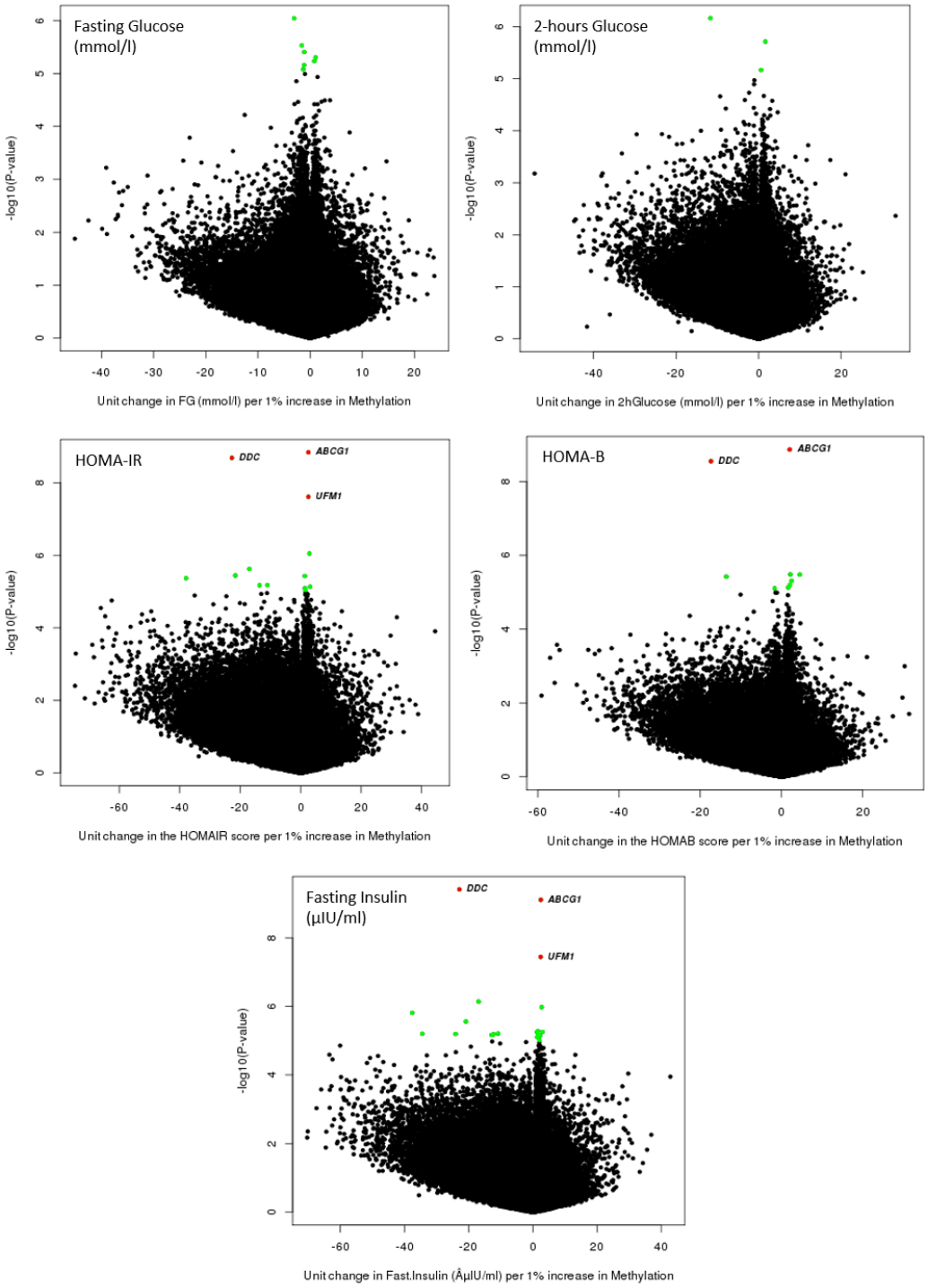


Table S8-12 Correlation estimates between glycaemic traits included in the meta-analysis. Correlation was estimated using the Spearman method considering the non-parametric distribution of regression coefficients across phenotypes for 20 CpG sites, which were initially identified as top signals in the meta-EWAS of fasting insulin.

Trait 1	Trait 2	r	P
Fasting insulin	HOMA-IR	1.00	< 0.001
HOMA-B	HOMA-IR	0.98	3.69E-14
Fasting insulin	HOMA-B	0.98	7.15E-14
Fasting insulin	2h glucose	0.88	2.59E-07
2h glucose	HOMA-IR	0.88	2.59E-07
2h glucose	HOMA-B	0.86	9.62E-07
Fasting insulin	Fasting glucose	0.66	1.39E-03
Fasting glucose	HOMA-IR	0.66	1.49E-03
Fasting glucose	2h glucose	0.63	2.65E-03
Fasting glucose	HOMA-B	0.60	5.16E-03

Table S8-13 Association between quintiles methylation at cg18232548 (DDC), and different clinical risk factors. Continuous variables were summarized using the mean and the standard deviation, while categorical variables were summarized using the proportion of samples per category per quintile. P-for-trend represents the Bonferroni adjusted-p, with p< 0.05 for evidence of a linear trend in the outcome per increase in the quintile of methylation.

	Quintile 1 (n=201)	Quintile 2 (n=200)	Quintile 3 (n=200)	Quintile 4 (n=200)	Quintile 5 (n=201)		
Continuous Phenotype	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P	P for trend (adjusted-p)
Age [years]	49.86(5.00)	49.95(5.19)	50.54(5.80)	50.17(4.92)	49.31(5.42)	0.21	1.05
BMI [kg/m2] [a]	26.46(4.26)	26.37(4.69)	27.16(5.07)	26.24(4.28)	25.97(4.12)	0.12	0.60
waist-circumference [cm]	88.17(12.27)	89.16(12.84)	90.9(13.03)	87.5(13.09)	85.85(13.12)	2.01E-03	0.01
Fasting Glucose [mmol/l]	5.26(0.45)	5.32(0.45)	5.29(0.5)	5.29(0.46)	5.20(0.46)	0.13	0.63
2-hours Glucose [mmol/l] [b]	4.31(0.37)	4.35(0.38)	4.34(0.41)	4.30(0.41)	4.24(0.39)	0.19	0.95
HbA1c [%] [c]	5.56(0.28)	5.5(0.26)	5.54(0.33)	5.58(0.32)	5.49(0.35)	0.31	1.55
C-reactive Protein [mg/l] [a]	2.06(2.78)	2.00(2.72)	2.48(3.16)	1.68(2.91)	1.56(2.22)	3.80E-04	1.90E-03
fasting Insulin [µIU/ml] [a,b]	6.33(4.48)	5.64(3.99)	5.84(4.08)	5.16(2.66)	4.43(3.51)	1.04E-04	5.22E-04
HOMA-IR [a,b]	1.48(1.09)	1.33(1.04)	1.35(0.95)	1.21(0.69)	1.03(0.93)	1.58E-04	7.90E-04
HOMA-B [a,b]	75.85(57.37)	65.97(39.73)	74.61(74.60)	61.22(28.58)	60.85(80.68)	4.80E-04	2.40E-03
Cholesterol [mmol/l]	4.86(0.88)	4.87(0.91)	4.91(1.00)	4.8(0.94)	4.73(0.85)	0.35	1.73
Triglycerides [mmol/l] [a]	1.15(0.52)	1.17(0.65)	1.21(0.61)	1.23(0.82)	1.00(0.49)	1.45E-03	0.01
HDL [mmol/l]	1.39(0.33)	1.40(0.35)	1.39(0.36)	1.40(0.34)	1.50(0.36)	0.01	0.03
LDL [mmol/l]	3.10(0.76)	3.09(0.84)	3.12(0.84)	3.04(0.82)	2.95(0.78)	0.28	1.38
Systolic Blood Pressure [mmHg]	121.15(13.67)	124.13(13.24)	124.42(14.19)	123.41(14.85)	120.94(13.74)	0.02	0.12
Diastolic Blood Pressure [mmHg] [a]	73.64(11.15)	75(10.83)	74.59(9.82)	74.49(11.69)	72.05(8.00)	0.04	0.18
$CD8^+$ T cells	0.17(0.05)	0.17(0.05)	0.16(0.05)	0.17(0.06)	0.18(0.06)	0.03	0.16
$CD4^+$ T cells	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.01(0.02)	0.22	1.12
Natural Killer Cells	0.20(0.05)	0.20(0.05)	0.20(0.06)	0.19(0.05)	0.20(0.05)	0.20	0.99
B cells	0.10(0.03)	0.09(0.03)	0.09(0.03)	0.09(0.03)	0.1(0.03)	0.04	0.22
Monocytes	0.07(0.03)	0.08(0.03)	0.07(0.03)	0.07(0.02)	0.07(0.03)	0.27	1.37
Granulocytes	0.51(0.09)	0.50(0.08)	0.52(0.09)	0.51(0.08)	0.50(0.08)	0.23	1.15
Categorical Phenotypes							
Sex [female/male]	135/66	104/96	105/95	134/66	144/57	0.03	0.14
Glucose tolerance [IFG/NGT] [b,d]	4/118	3/116	2/117	4/117	1/120	0.34	1.71
Glucose tolerance [IFG/IGT/NGT] [c]	7/2/66	6/1/69	7/1/67	9/1/65	2/0/74	0.24	1.18

[a] Variables log transformed to calculate the P-values. [b] Variables only available in a subset of 622 normoglycemic females in ALSPAC/ARIES, distribution between quintiles (125/124/124/124/125). [c] Variable only available in 382 normoglycemic male samples from the SABRE study, distribution between quintiles (77/76/76/76/77). IFG: impaired fasting glucose. IGT: impaired glucose tolerance. NGT: normal glucose tolerance. [d] IGT was not considered in the subsample of 622 females in ALSPAC/ARIES, as none of the measures of 2-hours glucose surpassed 5.0mmol/l. By contrast, IGT was reported in the subsample of males in SABRE.

Table S8-14 Association between quintiles of methylation at cg19750657 (UFM1), and different clinical risk factors. P-for-trend represents the Bonferroni adjusted-p, with p< 0.05 for evidence of a linear trend in the outcome per increase in quintile of methylation.

	Quintile 1 (n=201)	Quintile 2 (n=200)	Quintile 3 (n=200)	Quintile 4 (n=200)	Quintile 5 (n=201)		
Continuous Phenotype	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P	P for trend (adjusted-p)
Age [years]	49.45(5.51)	49.81(5.25)	49.39(5.06)	50.65(5.26)	50.53(5.25)	0.04	0.19
BMI [kg/m2] [a]	25.57(3.88)	26.45(4.57)	26.34(4.77)	26.92(4.35)	26.92(4.80)	0.01	0.07
waist-circumference [cm]	85.58(11.25)	88.68(13.5)	87.21(13.03)	90.23(13.33)	89.81(13.12)	1.47E-03	0.01
Fasting Glucose [mmol/l]	5.19(0.46)	5.25(0.48)	5.31(0.45)	5.32(0.49)	5.29(0.43)	0.02	0.11
2-hours Glucose [mmol/l] [b]	4.23(0.39)	4.31(0.39)	4.35(0.41)	4.32(0.36)	4.33(0.41)	0.14	0.72
HbA1c [%] [c]	5.49(0.31)	5.48(0.28)	5.55(0.33)	5.56(0.32)	5.59(0.31)	0.16	0.79
C-reactive Protein [mg/l] [a]	1.75(2.42)	2.20(2.94)	1.71(2.19)	1.93(3.27)	2.18(2.99)	0.21	1.03
fasting Insulin [µIU/ml] [a,b]	4.92(3.61)	4.80(2.81)	5.59(3.49)	5.62(4.2)	6.47(4.64)	1.48E-03	0.01
HOMA-IR [a,b]	1.14(0.94)	1.12(0.74)	1.31(0.88)	1.31(1.00)	1.52(1.15)	1.09E-03	0.01
HOMA-B [a,b]	67.03(80.93)	58.29(28.41)	66.1(39.33)	67.84(55.98)	79.12(74.89)	0.03	0.13
Cholesterol [mmol/l]	4.74(0.84)	4.86(0.99)	4.75(0.89)	4.91(0.98)	4.9(0.87)	0.17	0.87
Triglycerides [mmol/l] [a]	1.03(0.58)	1.17(0.65)	1.13(0.59)	1.23(0.64)	1.20(0.68)	0.01	0.04
HDL [mmol/l]	1.45(0.33)	1.43(0.39)	1.42(0.34)	1.41(0.34)	1.36(0.33)	0.10	0.50
LDL [mmol/l]	2.97(0.76)	3.08(0.86)	3.00(0.78)	3.07(0.80)	3.17(0.82)	0.10	0.51
Systolic Blood Pressure [mmHg]	121.39(13.77)	122.04(13.55)	123.24(14.77)	124.51(13.78)	122.85(14.04)	0.22	1.08
Diastolic Blood Pressure [mmHg] [a]	73.43(9.55)	73.81(9.82)	74.33(11.49)	74.87(11.77)	73.31(9.19)	0.64	3.18
$CD8^+$ T cells	0.15(0.05)	0.15(0.05)	0.17(0.05)	0.18(0.05)	0.2(0.06)	2.20E-16	1.10E-15
$CD4^+$ T cells	0.01(0.02)	0.02(0.02)	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.20	1.01
Natural Killer Cells	0.18(0.05)	0.19(0.05)	0.20(0.05)	0.21(0.05)	0.21(0.05)	1.59E-14	7.95E-14
B cells	0.09(0.03)	0.09(0.03)	0.09(0.03)	0.10(0.03)	0.10(0.03)	4.68E-04	2.34E-03
Monocytes	0.08(0.03)	0.08(0.03)	0.07(0.03)	0.07(0.03)	0.06(0.03)	2.58E-05	1.29E-04
Granulocytes	0.54(0.09)	0.53(0.08)	0.51(0.08)	0.48(0.08)	0.47(0.08)	2.20E-16	1.10E-15
Categorical Phenotypes							
Sex [female/male]	132/69	121/79	134/66	110/90	125/76	0.25	1.25
Glucose tolerance [IFG/NGT] [b,d]	2/117	2/119	2/119	4/115	4/118	0.26	1.30
Glucose tolerance [IFG/IGT/NGT] [c]	4/2/70	8/1/66	6/1/68	2/0/74	11/1/64	0.40	2.02

[a] Variables log transformed to calculate the P-values. [b] Variables only available in a subset of 622 normoglycemic females in ALSPAC/ARIES, distribution between quintiles (125/124/124/124/125). [c] Variable only available in 382 normoglycemic male samples from the SABRE study, distribution between quintiles (77/76/76/76/77). [d] IGT was not considered in the subsample of 622 females in ALSPAC, as none of the measures of 2-hours glucose surpassed 5.0mmol/l. In contrast, IGT was reported in the subsample of males in SABRE.

Table S8-15 Pathway analysis using genes annotated to top 1,000 CpG sites identified in the meta-EWAS of fasting insulin, HOMA-IR and HOMA-B. Pathway analysis was conducted in missMethyl using KEGG database. Total genes: total number of genes reported in a pathway with respect to genes included in the HumanMethylation 450K array; Differentially methylated: proportion of genes identified with differential methylation from the total number of genes in a pathway; P: p-value for pathway enrichment (unadjusted). None of the pathways were identified with FDR < 0.05 after multiple testing correction.

	Pathway	Total genes	Differentially Methylated	P
Fasting insulin	Terpenoid backbone biosynthesis	22	5	0.004
	Non-alcoholic fatty liver disease (NAFLD)	136	13	0.013
	AMPK signalling pathway	112	13	0.021
	Other types of O-glycan biosynthesis	19	4	0.022
	Insulin resistance	100	11	0.025
	Neurotrophin signalling pathway	110	11	0.048
	Th1 and Th2 cell differentiation	78	8	0.052
	T cell receptor signalling pathway	94	9	0.052
	B cell receptor signalling pathway	64	7	0.053
	Parkinson disease	119	10	0.055
	cGMP-PKG signalling pathway	152	14	0.057
	Thermogenesis	202	16	0.057
	Oocyte meiosis	112	10	0.058
	Oxidative phosphorylation	111	9	0.064
	mTOR signalling pathway	140	13	0.071
	Alzheimer disease	153	12	0.080
	Osteoclast differentiation	118	9	0.084
	Fc epsilon RI signalling pathway	63	6	0.090
	Adipocytokine signalling pathway	61	6	0.101
	Platelet activation	117	10	0.104
HOMA-IR	Sphingolipid signalling pathway	111	11	0.032
	mRNA surveillance pathway	76	7	0.095
	Glutathione metabolism	53	4	0.098
	Mineral absorption	44	4	0.099
	Carbohydrate digestion and absorption	38	4	0.111
HOMA-B	Hepatitis C	109	12	0.005
	Notch signalling pathway	47	7	0.018
	Endocytosis	222	19	0.032
	Hepatitis B	120	12	0.032
	Sulphur relay system	7	2	0.035
	Huntington disease	174	14	0.035
	Cell adhesion molecules (CAMs)	122	11	0.037
	PPAR signalling pathway	65	6	0.050
	Rheumatoid arthritis	70	6	0.058
	Bladder cancer	38	5	0.071
	ABC transporters	39	4	0.083
	Primary bile acid biosynthesis	16	2	0.098
	Caffeine metabolism	5	1	0.099
	Prostate cancer	93	9	0.106
	Cellular senescence	147	12	0.108
	MicroRNAs in cancer	261	16	0.109
	Peroxisome	79	6	0.116
	Aminoacyl-tRNA biosynthesis	42	4	0.119

Table S8-16 Regression analysis between HOMA-B and methylation score generated in ALSPAC (n=622 females), and replicated in SABRE (n=382 males). The score for HOMA-B was generated using two top CpG sites (in ABCG1 and DDC) identified in the meta-EWAS for this trait. Results are interpreted as a unit change in HOMA-B, per unit increase in the methylation score. P-value is presented for the score, and for the different adjustment model. Associations were considered significant at p<0.05.

Model [a]	ALSPAC/ARIES						SABRE					
	Score parameters		Model Parameters				Score parameters		Model Parameters			
	Effect [b]	P	P	R^2	P (LRT) [c]	RMSE	Effect	P	P	R^2	P (LRT)	RMSE
Crude	NA	NA	NA	0.79	NA	0.24	NA	NA	NA	0.84	NA	0.154
M1	-0.001	1.00	1.00	-1.62E-03	< 2.2e-16	0.54	-0.42	0.79	0.79	-2.46E-03	NA	0.388
M2	-0.11	0.94	1.97E-03	0.02	< 2.2e-16	0.53	-0.18	0.91	0.12	0.01	NA	0.385
M3	-0.15	0.92	7.41E-04	0.02	< 2.2e-16	0.53	-0.44	0.77	1.07E-05	0.06	NA	0.373
M4	-0.76	0.55	< 2.2e-16	0.23	< 2.2e-16	0.47	-1.96	0.18	9.62E-13	0.15	NA	0.355
M5	-0.90	0.18	< 2.2e-16	0.79	0.18	0.24	-0.75	0.24	< 2.2e-16	0.84	0.23	0.154

[a] Crude: model without the score but including as covariates age, smoking, BMI, fasting glucose and fasting insulin. M1: score, M2: score and age, M3: score, age and smoking, M4: score, age, smoking and BMI, M5: score, age, smoking, BMI, fasting insulin and fasting glucose. [b] Effect estimates are in log-units since HOMA-B was log-transformed before the regression analysis. [c] P-value of the likelihood ratio test was not calculated in SABRE as there was an imbalance in the number of samples between models for this study.

Table S8-17 Association between quartiles of methylation score and HOMA-B in ALSPAC and SABRE. Score was generated using two top CpG sites mapping to ABCG1 and DDC. Associations were additively adjusted for age, smoking, BMI, fasting glucose, and fasting insulin, and were considered significant at p< 0.05.

Adjustment	ALSPAC/ARIES						SABRE					
	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model	Mean diff. Q4 vs Q1	SE	95%CI	P-score	R^2	P-model
None	0.06	0.06	(-0.06, 0.18)	0.33	-1.71E-03	5.84E-01	-0.014	0.06	(-0.13, 0.10)	0.80	3.63E-03	2.25E-01
Age	0.05	0.06	(-0.065, 0.173)	0.37	0.02	5.59E-03	-0.005	0.06	(-0.12, 0.11)	0.93	0.01	5.95E-02
Smoking	0.05	0.06	(-0.065, 0.172)	0.37	0.02	1.72E-03	-0.02	0.05	(-0.13, 0.09)	0.70	0.07	2.11E-05
BMI	0.02	0.05	(-0.08, 0.13)	0.67	0.23	< 2.2e-16	-0.06	0.05	(-0.17, 0.04)	0.22	0.15	3.00E-12
Fasting Glucose + Fasting Insulin	-0.02	0.03	(-0.07, 0.04)	0.49	0.79	< 2.2e-16	-0.04	0.02	(-0.08, 0.01)	0.13	0.84	< 2.2e-16

Values of HOMA-B were log-transformed before the analysis, and the effect estimates should be interpreted as mean difference in HOMA-B for Q4 versus Q1 of methylation score after back-transformation of the coefficients using the exponential function [e^x]. Sample size per quartile in ALSPAC (156/155/155/156) and in SABRE (96/95/95/96).

Figure S8-16 Distribution of HbA1c per quartiles of methylation score calculated in SABRE (n=382). Significant differences in mean percent of HbA1c were detected between Q2 and Q1 (beta=-0.09, 95% CI=-0.18, -0.003, p=0.04) and between Q4 and Q1 (beta=-0.09, 95% CI=-0.18, -0.01, p=0.04), without suggestion of a linear decrease in HbA1c across quartiles of the score. Star (*) indicates the quartile where significant differences in mean percent of HbA1c were identified with the first quartile at p< 0.05. Mean percent of HbA1c per quartile of the score in SABRE (Q1=5.58%, Q2=5.49%, Q3=5.57% and Q4=5.49%).

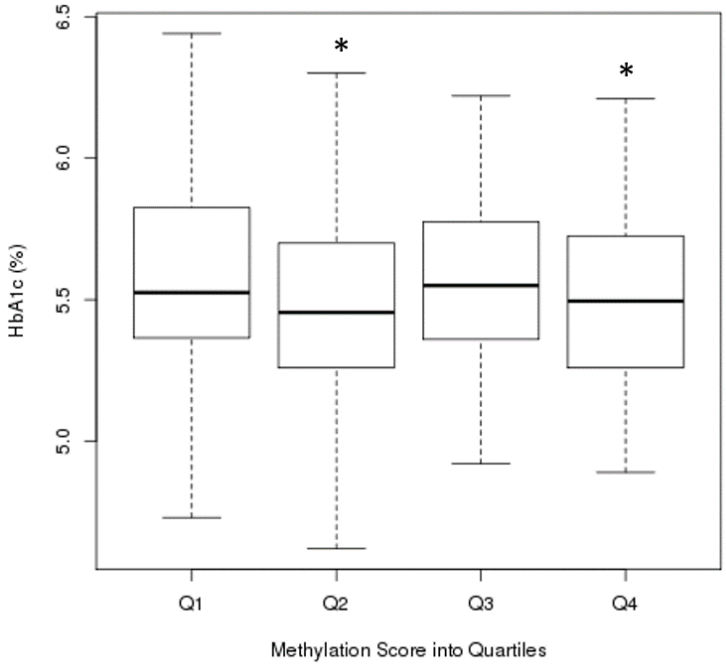


Table S 8-18 Baseline characteristics of participants in ALSPAC and KORA based on T2D status. Categorical variables were described using the percentage and (n), while continuous variables were described using the mean and/or (SD). Comparisons were considered statistically significant at p<0.05.

	ALSPAC (N=1,050)			KORA (N=1,719)		
	Controls (n=1,002)	**Cases (n=48)**	**P-value**	**Controls (n=1,564)**	**Cases (n=155)**	**P-value**
Age (yrs)	49.97 (5.28)	51.94 (7.40)	1.70E-01	60.48	66	<0.01
Ethnicity [% white]	94.61 (948)	93.75 (45)	2.50E-01	100 (1564)	100 (155)	---
Fasting Glucose (mmol/l)	5.27 (0.46)	8.45 (3.00)	<0.01	---	---	---
Body mass index (kg/m^2)	26.44 (4.51)	30.09 (5.52)	<0.01	27.81	31.1	<0.01
Waist circumference (cm)	88.31 (12.96)	99.63 (15.65)	<0.01	---	---	---
Waist-hip ratio (cm)	0.86 (0.09)	0.92 (0.10)	<0.01	---		---
Systolic BP (mmHg)	122.81 (13.99)	129.45 (15.93)	<0.01	123.78	134.3	1.79E-10
Diastolic BP (mmHg)	73.95 (10.41)	76.82 (12.48)	4.00E-02	75.98	76.69	4.24E-01
Serum Total Cholesterol (mmol/L)	4.84 (0.92)	4.28 (0.83)	<0.01	223.14	210.65	2.77E-04
Triglycerides (mmol/l)	1.15 (0.63)	1.62 (0.70)	<0.01	128.02	185.48	2.50E-07
HDL cholesterol (mmol/l)	1.42 (0.35)	1.17 (0.29)	<0.01	57.36	48.38	2.06E-18
LDL cholesterol (mmol/l)	3.06 (0.81)	2.57 (0.85)	<0.01	140.95	131.37	1.05E-03
Fasting C-reactive protein (mg/L)	1.96 (2.79)	2.52 (3.05)	<0.01			
Sex [% male]	37.92 (380)	18.50 (25)	5.00E-02	47.57 (744)	61.94 (96)	1.00E-03
Estimated white cell subset						
CD4T	0.17 (0.05)	0.16 (0.05)	6.10E-01	0.24	0.24	1.50E-01
CD8T	0.02 (0.03)	0.03 (0.04)	8.00E-02	0.10	0.10	4.70E-01
Natural Killer Cells	0.19 (0.05)	0.19 (0.05)	8.80E-01	0.06	0.06	1.30E-01
B-cells	0.09 (0.03)	0.09 (0.03)	7.00E-01	0.07	0.07	9.20E-01
Monocytes	0.07 (0.03)	0.08 (0.03)	5.00E-02	0.10	0.10	9.10E-01
Granulocytes	0.51 (0.08)	0.50 (0.09)	3.30E-01	0.45	0.44	3.30E-01
Smoking‡						
Never smoker	91.02 (912)	85.42 (41)	1.90E-01	43.86 (686)	41.94 (65)	9.70E-01
Former smoker	---	---		2.11 (33)	1.94 (3)	
Current smoker	8.98 (90)	14.58 (7)		54.02 (845)	56.13 (87)	
Physical activity [% < 4h/week]*	83.24 (586)	76.92 (20)	4.00E-01	---	---	---
Socioeconomic status						
High income	57.64 (479)	35.14 (13)	2.00E-02	---	---	---
Middle income	30.45 (253)	48.65 (18)		---	---	
Low income	11.91 (99)	16.22 (6)		---	---	

* Physical activity was defined as a binary variable in ALSPAC: < 4h/week or ≥ 4h/week. ‡in KORA, there were four categories of smoking (never/former/casual and current smoker). A casual smoker was defined as a participant without daily smoking. Casual smokers were combined with current smokers to match categories of smoking in KORA and in ALSPAC.

Table S 8-19 Baseline characteristics of participants in LBC1936 and RSIII-1 based on T2D status. Categorical variables were described using the percentage and (n), while continuous variables were described using the mean and/or (SD). Comparisons were considered statistically significant at p<0.05.

	LBC1936 (N=915)			RSIII-1 (N=728)		
	Controls (n=805)	**Cases (n=110)**	**P-value**	**Controls (n=654)**	**Cases (n=74)**	**P-value**
Age (yrs)	69.60	69.70	0.09	59.55	63.03	9.00E-04
Ethnicity [% white]	100 (805)	100 (110)	---	100 (654)	100 (74)	---
Fasting Glucose (mmol/l)	5.72*	7.37*	<0.01	5.33	8.01	<0.01
Body mass index (kg/m^2)	27.40	30.80	1.20E-09	27.19	30.59	9.54E-07
Waist circumference (cm)	---	---	---	92.77	102.74	1.29E-09
Waist-hip ratio (cm)	---	---	---	---	---	---
Systolic BP (mmHg)	148.70	146.70	2.80E-01	133.8	140.4	6.23E-03
Diastolic BP (mmHg)	83.50	80.30	1.20E-03	82.92	82.03	4.83E-01
Serum Total Cholesterol (mmol/L)	5.54	4.75	7.70E-12	5.61	5.1	3.45E-04
Triglycerides (mmol/l)	1.59	1.91	9.80E-06	1.44	1.96	2.01E-05
HDL cholesterol (mmol/l)	1.54	1.34	5.10E-07	1.42	1.21	1.45E-05
LDL cholesterol (mmol/l)	---	---	---	3.54	3.01	2.30E-05
Fasting C-reactive protein (mg/L)	---	---	---	---	---	---
Sex [male %]	50 (400)	56 (62)	1.90E-01	44.95 (294)	55.41 (41)	7.59E-02
Estimated white cell subset						
CD4T	0.15	0.14	9.00E-02	0.26	0.25	4.80E-02
CD8T	0.04	0.06	1.00E-02	0.09	0.09	4.30E-01
Natural Killer Cells	0.07	0.08	6.90E-01	0.14	0.14	3.36E-01
B-cells	0.07	0.07	4.70E-01	0.1	0.09	1.76E-01
Monocytes	0.07	0.07	9.80E-01	0.08	0.09	5.92E-01
Granulocytes	0.63	0.62	5.70E-01	0.37	0.4	1.32E-01
Smoking						
Never smoker	48 (386)	40 (44)	2.40E-01	30.12 (197)	20.27 (15)	8.72E-02
Former smoker	41 (332)	45 (50)		42.66 (279)	55.41 (41)	
Current smoker	11 (87)	15 (16)		27.22 (178)	24.32 (18)	
Physical activity**	---	---	---	58.98	55.69	3.71E-01
Socioeconomic status						
High income	---	---	---	26.60 (170)	16.44 (12)	1.63E-01
Middle income	---	---		61.66 (394)	71.23 (52)	
Low income	---	---		11.74 (75)	12.33 (9)	

*Mean values of HbA1c since FG were not available for participants in LBC1936. **Physical activity was measured in the Rotterdam studies as a continuous variable using metabolic equivalent (MET) hours/week.

Table S8-20 Baseline characteristics of participants in RS-Bios based on T2D status. Categorical variables were described using the percentage and (n), while continuous variables were described using the mean and/or (SD). Comparisons were considered statistically significant at p<0.05.

	RS-Bios (N=735)		
	Controls (n=627)	**Cases (n=108)**	**P-value**
Age (yrs)	67.51	68.79	1.14E-01
Ethnicity [% white]	100 (627)	100 (108)	---
Fasting Glucose (mmol/l)	5.40	7.39	<0.01
Body mass index (kg/m^2)	27.42	29.35	1.14E-05
Waist circumference (cm)	93.27	100.7	1.38E-09
Waist-hip ratio (cm)			
Systolic BP (mmHg)	144.30	148.10	2.32E-02
Diastolic BP (mmHg)	84.36	84.32	7.02E-01
Serum Total Cholesterol (mmol/L)	5.60	4.94	1.01E-08
Triglycerides (mmol/l)	1.42	1.75	4.61E-05
HDL cholesterol (mmol/l)	1.55	1.32	1.31E-07
LDL cholesterol (mmol/l)	3.41	2.83	3.11E-09
Fasting C-reactive protein (mg/L)	---	---	---
Sex [male %]	40.99 (257)	50.93 (55)	4.62E-02
Measured white cell subset†			
Lymphocytes	36.62	34.64	1.43E-02
Monocytes	7.14	7.16	5.55E-01
Granulocytes	6.75	7.35	5.66E-03
Smoking			
Never smoker	36.20 (227)	25.00 (27)	6.01E-02
Former smoker	53.91 (338)	61.11 (66)	
Current smoker	9.89 (62)	13.89 (15)	
Physical activity**	60.11	56.03	3.80E-01
Socioeconomic status			
High income	20.92 (128)	20.56 (22)	5.72E-01
Middle income	72.55 (444)	70.09 (75)	
Low income	6.54 (40)	9.35 (10)	

†Direct white-cell counts were available for samples in RS-Bios, where cells were categorized into three groups: lymphocytes, monocytes and granulocytes.
** Physical activity was measured in the Rotterdam studies as a continuous variable using metabolic equivalent (MET) hours/week.

Table S 8-21 Top associations identified in the meta-analysis of EWASs in T2D using a basic model adjusted for age, sex and SVs. I^2 is the heterogeneity test and the P-value for heterogeneity, significant at p<0.05. Association in the meta-analysis were regarded significant at p<1.033x10^{-7}. I^2 is the heterogeneity estimate and the P-value for heterogeneity (significant at p<0.05). Direction (+/-/?) represents the direction of the association based on results from each cohort (? If for one study the result for that probe is unknown). From left to right, direction is the effect detected in: ALSPAC, KORA, LBC1936, RSIII-1 and RS-Bios. I^2 is the heterogeneity estimate and the P value for heterogeneity.

CpG	Nearest Gene	ALSPAC (N=1050)		KORA (N=1719)		LBC1936 (N=915)		RSIII-1 (N=728)		RS-Bios (N=735)		Meta-analysis (N=5147)					
		Beta	P	Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Direction	I^2	P
cg19693031	*TXNIP*	-0.019	0.021	9.90E-05	0.980	-0.026	6.82E-06	-0.017	0.001	-0.016	3.15E-04	-0.013	0.002	1.13E-08	-+---	77.4	1.39E-03
cg06468695	*CCDC42*	0.003	0.661	0.006	0.001	0.013	3.99E-04	0.002	0.380	0.005	9.38E-03	0.006	0.001	4.72E-07	+++++	37.4	1.72E-01
cg07068382	*MTCH1*	0.004	0.611	0.005	0.275	0.023	3.81E-05	0.012	0.012	0.009	1.40E-02	0.010	0.002	1.38E-06	+++++	48.2	1.03E-01
cg06500161	*ABCG1*	0.027	0.000	-0.003	0.331	0.025	1.94E-08	0.008	0.027	0.008	6.46E-03	0.007	0.002	2.45E-06	+-+++	89.6	8.59E-08
cg01317029	*FAM131A*	0.015	0.001	0.005	0.039	0.012	4.69E-04	-0.003	0.553	0.004	6.28E-02	0.007	0.001	4.13E-06	+++-+	62.9	2.90E-02
cg00574958	*CPT1A*	-0.005	0.060	3.51E-04	0.869	-0.008	8.55E-05	-0.015	0.001	-0.004	4.99E-02	-0.005	0.001	5.64E-06	-+---	71.8	6.70E-03
cg19134130	*TMEM104*	0.009	0.028	0.002	0.295	0.009	6.60E-04	0.005	0.034	0.005	2.26E-02	0.005	0.001	7.02E-06	+++++	44.4	1.26E-01
cg00082384	*NISCH*	0.015	0.014	0.008	0.014	0.018	1.52E-04	-0.001	0.878	0.005	1.04E-01	0.008	0.002	8.73E-06	+++-+	57.8	5.02E-02
cg12727256	*C22orf45*	-0.003	0.668	0.004	0.013	0.014	1.22E-03	0.009	0.003	0.003	1.71E-01	0.005	0.001	9.30E-06	-++++	48.7	9.90E-02
cg24704287	*Intergenic*	-0.016	0.018	0.006	0.153	-0.015	9.21E-03	-0.013	0.008	-0.011	2.50E-04	-0.009	0.002	9.85E-06	-+---	74.4	3.53E-03

Table S 8-22 Top associations identified in the meta-analysis of EWASs in T2D using a model adjusted for age, sex, SVs and 6-Houseman cells. I^2 is the heterogeneity test and the P-value for heterogeneity, significant at p<0.05. Association in the meta-analysis were regarded significant at p<1.033x10^{-7}. I^2 is the heterogeneity estimate and the P-value for heterogeneity (significant at p<0.05). Direction (+ - ?) represents the direction of the association based on results from each cohort (? If for one study the result for that probe is unknown). From left to right, direction is the effect detected in: ALSPAC, KORA, LBC1936, RSIII-1 and RS-Bios. I^2 is the heterogeneity estimate and the P value for heterogeneity.

CpG	Nearest Gene	ALSPAC (N=1,050)		KORA (N=1,719)		LBC1936 (N=915)		RSIII-1 (N=735)		RS-Bios (N=723)		Meta-analysis (N=5147)					
		Beta	P	Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Direction	I^2	P
cg19693031	*TXNIP*	-0.023	0.005	0.002	0.616	-0.026	0.000	-0.019	0.000	-0.015	0.001	-0.013	2.13E-03	3.80E-09	-+---	83.9	5.46E-05
cg13826139	*Intergenic*	-0.004	0.416	-0.005	0.001	-0.005	0.048	-0.007	0.001	NA	NA	-0.006	1.06E-03	1.67E-07	----?	0.0	9.09E-01
cg00574958	*CPT1A*	-0.005	0.042	2.31E-05	0.991	-0.008	0.000	-0.018	0.000	-0.003	0.073	-0.005	1.01E-03	8.08E-07	-+---	79.8	5.37E-04
cg00082384	*NISCH*	0.014	0.025	0.008	0.012	0.015	0.001	0.006	0.187	0.005	0.106	0.008	1.79E-03	2.29E-06	+++++	1.7	3.97E-01
cg01317029	*FAM131A*	0.014	0.002	0.004	0.077	0.009	0.001	0.007	0.051	0.003	0.216	0.006	1.25E-03	3.65E-06	+++++	41.6	1.44E-01
cg15560632	*LRCH4*	-0.001	0.007	-2.32E-04	0.813	-0.001	0.003	-0.002	0.010	-0.001	0.160	-0.001	2.00E-04	4.49E-06	-----	0.0	4.54E-01
cg06500161	*ABCG1*	0.025	0.000	-0.003	0.240	0.024	0.000	0.010	0.005	0.008	0.013	0.007	1.53E-03	4.91E-06	+-+++	89.3	1.42E-07
cg22103637	*HS6ST3*	0.009	0.009	0.003	0.259	0.007	0.032	0.008	0.041	0.006	0.008	0.006	1.28E-03	5.14E-06	+++++	0.0	5.45E-01
cg06468695	*CCDC42*	0.003	0.622	0.006	0.001	0.012	0.001	0.001	0.571	0.004	0.028	0.005	1.11E-03	5.68E-06	+++++	34.2	1.93E-01
cg27237541	*MYO3A*	-0.019	0.010	-0.010	0.002	-0.011	0.031	-0.008	0.065	-0.003	0.275	-0.008	1.80E-03	6.70E-06	-----	21.3	2.79E-01
cg22628512	*Intergenic*	0.011	0.000	0.005	0.055	0.004	0.153	0.003	0.235	0.004	0.020	0.005	1.06E-03	6.95E-06	+++++	27.7	2.37E-01
cg06039489	*C20orf26*	0.008	0.348	0.005	0.491	0.028	0.002	0.010	0.141	0.018	0.000	0.014	3.01E-03	7.14E-06	+++++	31.7	2.10E-01
cg07184465	*SPZ1*	-0.012	0.040	-0.003	0.126	-0.006	0.051	-0.012	0.000	-0.004	0.102	-0.005	1.21E-03	7.14E-06	-----	46.1	1.15E-01
cg11851382	*PPAP2B*	-0.012	0.021	-0.003	0.187	-0.012	0.003	-0.005	0.214	-0.006	0.006	-0.006	1.38E-03	7.14E-06	-----	18.9	2.95E-01
cg07068382	*MTCH1*	0.001	0.937	0.005	0.256	0.021	0.000	0.011	0.014	0.007	0.036	0.009	2.07E-03	7.46E-06	+++++	47.1	1.09E-01
cg20154947	*PLEC1*	-0.002	0.000	0.001	0.641	-0.002	0.002	-0.002	0.147	0.000	0.735	-0.002	4.02E-04	7.62E-06	-+---	29.8	2.23E-01
cg17566334	*PACRG*	-0.004	0.701	0.006	0.007	-0.002	0.847	0.014	0.001	0.007	0.011	0.007	1.54E-03	8.48E-06	-+-++	28.1	2.34E-01
cg25741837	*SMYD5*	0.022	0.005	0.004	0.179	0.007	0.099	0.009	0.023	0.007	0.012	0.008	1.70E-03	9.32E-06	+++++	11.2	3.42E-01
cg09185884	*KCTD2*	0.012	0.024	0.002	0.726	0.016	0.058	0.013	0.042	0.010	0.001	0.009	2.07E-03	9.49E-06	+++++	0.0	4.30E-01
cg08127348	*GTPBP3*	-0.001	0.015	-0.002	0.027	-0.001	0.045	-0.003	0.074	-0.002	0.045	-0.001	3.22E-04	9.63E-06	-----	0.0	6.42E-01
cg08273233	*HTR1E*	-0.005	0.446	-0.004	0.083	-0.012	0.007	-0.005	0.197	-0.009	0.001	-0.006	1.45E-03	9.68E-06	-----	0.0	4.24E-01
cg10036510	*KLHL7*	-0.001	0.026	-0.001	0.118	-0.001	0.030	-0.004	0.007	-0.002	0.033	-0.001	2.81E-04	9.77E-06	-----	7.9	3.61E-01

Table S8-23 Top associations identified in the meta-analysis of EWAS in T2D using a model adjusted for age, sex, SVs, 6 Houseman cells, smoking and BMI. I^2 is the heterogeneity estimate and the P-value for heterogeneity (significant at p<0.05). Direction (+ - ?) represents the direction of the association based on results from each cohort (? If for one study the result for that probe is unknown). From left to right, direction is the effect detected in: ALSPAC, KORA, LBC1936, RSIII-1 and RS-Bios. I^2 is the heterogeneity estimate and the P value for heterogeneity.

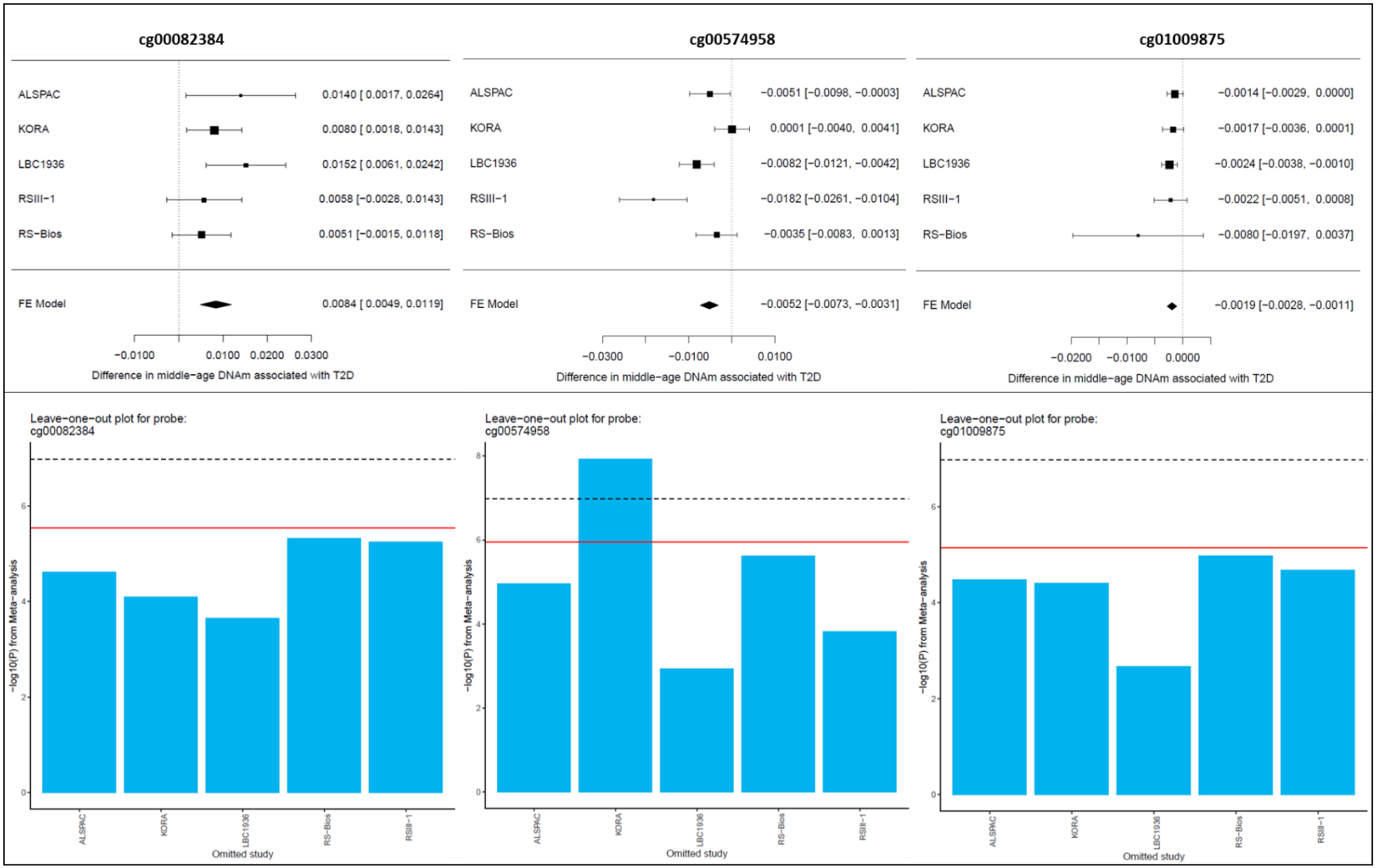
CpG	Nearest Gene	ALSPAC (N=1050)		KORA (N=1719)		LBC1936 (N=915)		RSIII-1 (N=728)		RS-Bios (N=735)		Meta-analysis (N=5147)					
		Beta	P	Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Direction	I^2	P
cg19693031	*TXNIP*	-0.019	0.018	0.001	0.730	-0.025	0.000	-0.018	0.000	-0.014	0.002	-0.012	2.17E-03	4.99E-08	-+---	80.10	4.81E-04
cg13826139	*Intergenic*	-0.004	0.464	-0.005	0.001	-0.004	0.119	-0.006	0.001	NA	NA	-0.005	1.08E-03	1.08E-06	----?	0.00	9.17E-01
cg17566334	*PACRG*	-0.005	0.586	0.006	0.003	-0.002	0.783	0.016	0.000	0.007	0.009	0.007	1.57E-03	2.59E-06	-+-++	41.70	1.43E-01
cg19008097	*CD14*	0.021	0.027	0.017	0.001	0.007	0.325	0.010	0.113	0.010	0.040	0.013	2.71E-03	3.61E-06	++++	0.00	6.22E-01
cg00082384	*NISCH*	0.015	0.021	0.007	0.029	0.017	0.000	0.005	0.302	0.006	0.094	0.008	1.83E-03	4.80E-06	+++++	28.80	2.30E-01
cg22628512	*Intergenic*	0.011	0.000	0.005	0.055	0.005	0.075	0.003	0.244	0.004	0.026	0.005	1.08E-03	5.43E-06	+++++	25.70	2.50E-01
cg14275576	*Intergenic*	-0.003	0.021	-0.003	0.008	-0.002	0.004	-0.004	0.031	0.000	0.583	-0.002	4.37E-04	8.79E-06	-----	30.80	2.16E-01

Table S8-24 Comparing results of the fixed effect and the random effect models for top associations identified in the meta-analysis adjusted for age, sex, SVs, 6 Houseman cells and smoking. The suffix FE refers to results of the fixed-effect model, and the suffix RE refers to those from the random effect model. I^2 is the heterogeneity estimate for the FE model, and Tau is the heterogeneity test for the RE model. Association were significantly heterogeneous at p<0.05 for the P(I^2) or at Tau ≠ 0.

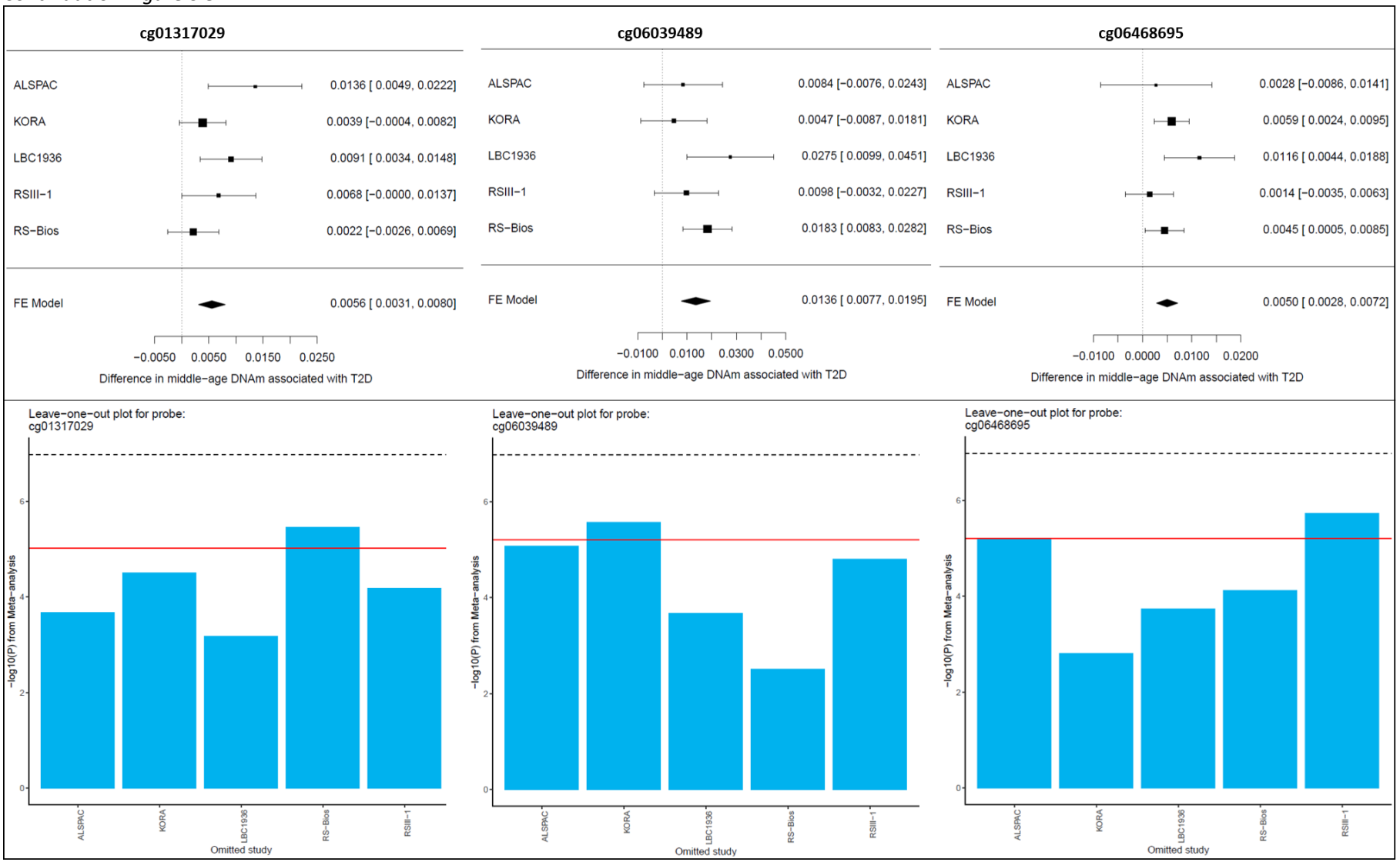
CpG	Beta_FE	SE_FE	P_FE	Direction	Beta_RE	SE_RE	P_RE	I^2	P(I^2)	Tau	Chr	Gene
cg19693031*	-0.013	2.13E-03	4.26E-09	-+---	-0.015	5.41E-03	4.66E-03	83.30	8.18E-05	1.18E-04	1	*TXNIP*
cg13826139**	-0.006	1.06E-03	1.27E-07	----?	-0.006	1.06E-03	1.27E-07	0.00	9.17E-01	0.00E+00	6	*Intergenic*
cg00574958	-0.005	1.07E-03	1.11E-06	-+---	-0.006	2.44E-03	9.56E-03	79.50	6.16E-04	2.31E-05	11	*CPT1A*
cg14275576	-0.002	5.05E-04	1.24E-06	-----	-0.002	5.05E-04	1.24E-06	0.00	9.36E-01	0.00E+00	20	*Intergenic*
cg27237541	-0.009	1.80E-03	1.99E-06	-----	-0.009	1.80E-03	1.99E-06	0.00	4.52E-01	0.00E+00	10	*MYO3A*
cg19611616	-0.003	7.03E-04	2.56E-06	-----	-0.003	7.27E-04	4.40E-06	3.90	3.85E-01	1.00E-07	12	*STK38L*
cg00082384	0.008	1.79E-03	2.86E-06	+++++	0.008	1.87E-03	5.92E-06	6.30	3.71E-01	1.10E-06	3	*NISCH*
cg06500161	0.007	1.53E-03	3.30E-06	+-+++	0.012	4.99E-03	1.81E-02	89.40	1.22E-07	1.07E-04	21	*ABCG1*
cg14186584	-0.002	3.47E-04	4.01E-06	-----	-0.002	3.47E-04	4.01E-06	0.00	9.37E-01	0.00E+00	5	*Intergenic*
cg25741837	0.008	1.71E-03	4.44E-06	+++++	0.008	1.86E-03	1.88E-05	12.50	3.34E-01	2.20E-06	2	*SMYD5*
cg15560632	-0.001	2.00E-04	4.58E-06	-----	-0.001	2.00E-04	4.58E-06	0.00	4.72E-01	0.00E+00	7	*LRCH4*
cg07400328	-0.003	5.48E-04	5.03E-06	----+	-0.003	5.48E-04	5.03E-06	0.00	6.94E-01	0.00E+00	6	*MUTED*
cg22628512	0.005	1.06E-03	5.66E-06	+++++	0.005	1.30E-03	1.18E-04	29.60	2.24E-01	2.50E-06	1	*Intergenic*
cg06468695	0.005	1.11E-03	6.19E-06	+++++	0.005	1.43E-03	3.97E-04	31.30	2.13E-01	3.10E-06	17	*CCDC42*
cg06039489	0.014	3.01E-03	6.27E-06	+++++	0.013	3.67E-03	2.55E-04	29.60	2.24E-01	1.99E-05	20	*C20orf26*
cg27374726	-0.007	1.49E-03	6.52E-06	-----	-0.007	2.25E-03	1.37E-03	52.70	7.63E-02	1.30E-05	10	*Intergenic*
cg01009875	-0.002	4.31E-04	7.17E-06	-----	-0.002	4.31E-04	7.17E-06	0.00	7.35E-01	0.00E+00	1	*TMCO1*
cg17566334	0.007	1.55E-03	7.52E-06	-+-++	0.007	2.05E-03	6.42E-04	26.80	2.43E-01	5.50E-06	6	*PACRG*
cg07184465	-0.005	1.21E-03	8.27E-06	-----	-0.006	1.78E-03	5.66E-04	46.30	1.14E-01	7.00E-06	5	*SPZ1*
cg11851382	-0.006	1.38E-03	8.81E-06	-----	-0.006	1.55E-03	4.37E-05	15.10	3.18E-01	1.90E-06	1	*PPAP2B*
cg08273233	-0.006	1.45E-03	8.85E-06	-----	-0.006	1.45E-03	8.85E-06	0.00	4.27E-01	0.00E+00	6	*HTR1E*
cg20154947	-0.002	4.02E-04	9.02E-06	-+---	-0.002	5.10E-04	9.30E-04	27.70	2.37E-01	4.00E-07	8	*PLEC1*
cg13927560	-0.002	4.64E-04	9.05E-06	-----	-0.002	4.64E-04	9.05E-06	0.00	7.82E-01	0.00E+00	4	*TMEM33*
cg01317029	0.006	1.26E-03	9.48E-06	+++++	0.006	1.79E-03	5.56E-04	46.60	1.12E-01	7.30E-06	3	*FAM131A*
cg17155612	-0.002	5.33E-04	9.55E-06	-----	-0.002	5.33E-04	9.55E-06	0.00	4.29E-01	0.00E+00	19	*LOC148189*

*CpG site identified in strong association with T2D only in the FE model. **CpG site identified in strong association with T2D in both, the FE and the RE models.

Figure S 8-17 Forest plot (top) and leave-one-out analysis (bottom) for top 25 CpG sites associated with T2D based on results of the meta-analysis adjusted for age, sex, SVs, 6-Houseman cells and smoking. In the forest plot, effect estimate (95%CI) of the EWAS in each cohort, and results of the fixed-effect meta-analysis at the bottom represented by the diamond symbol. From left to right, the study removed from the meta-analysis was: ALSPAC, KORA, LBC1936, RSIII-1 and RS-Bios.

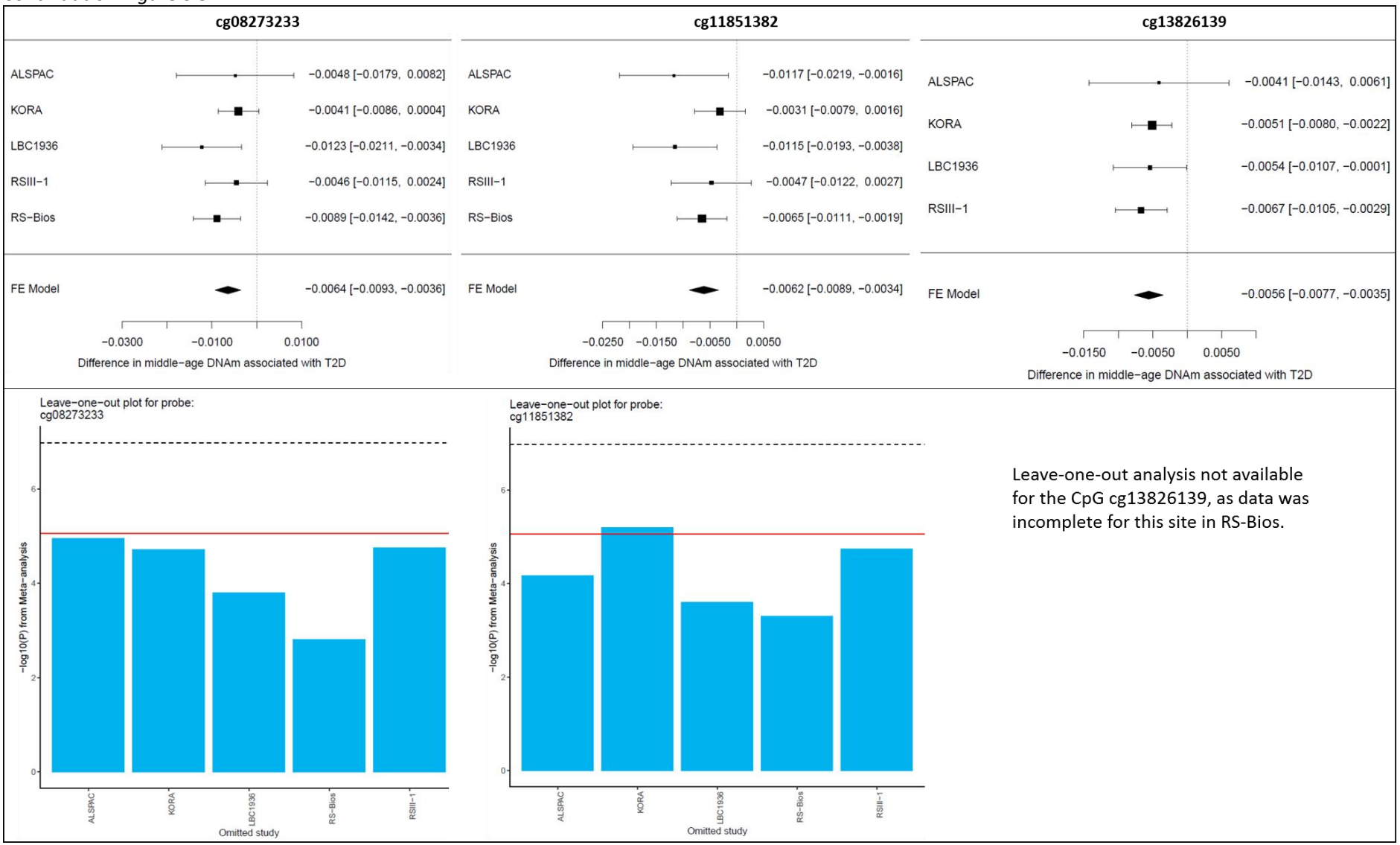


Continuation Figure S 8-17.



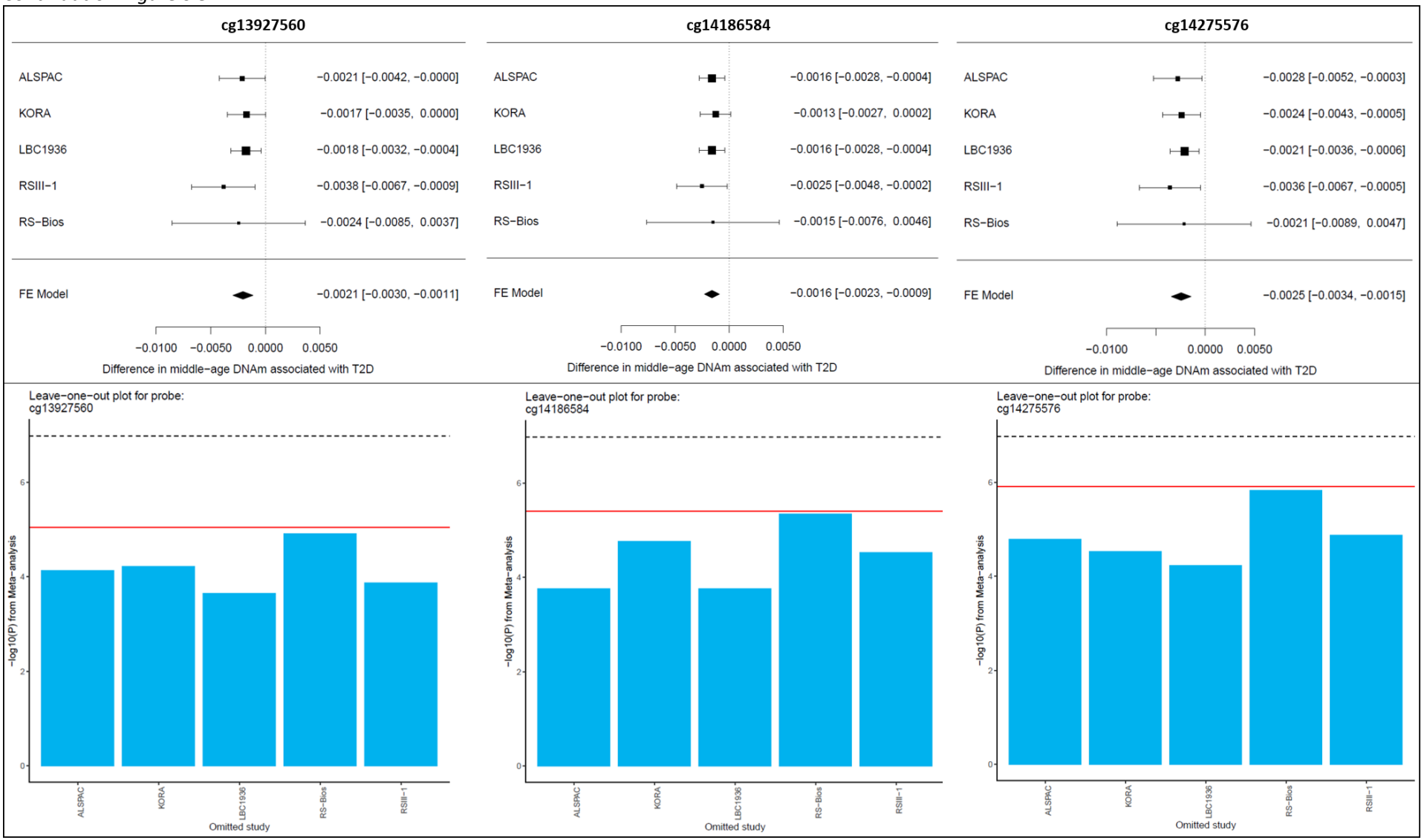
Continuation Figure S 8-17.

Continuation Figure S 8-17.

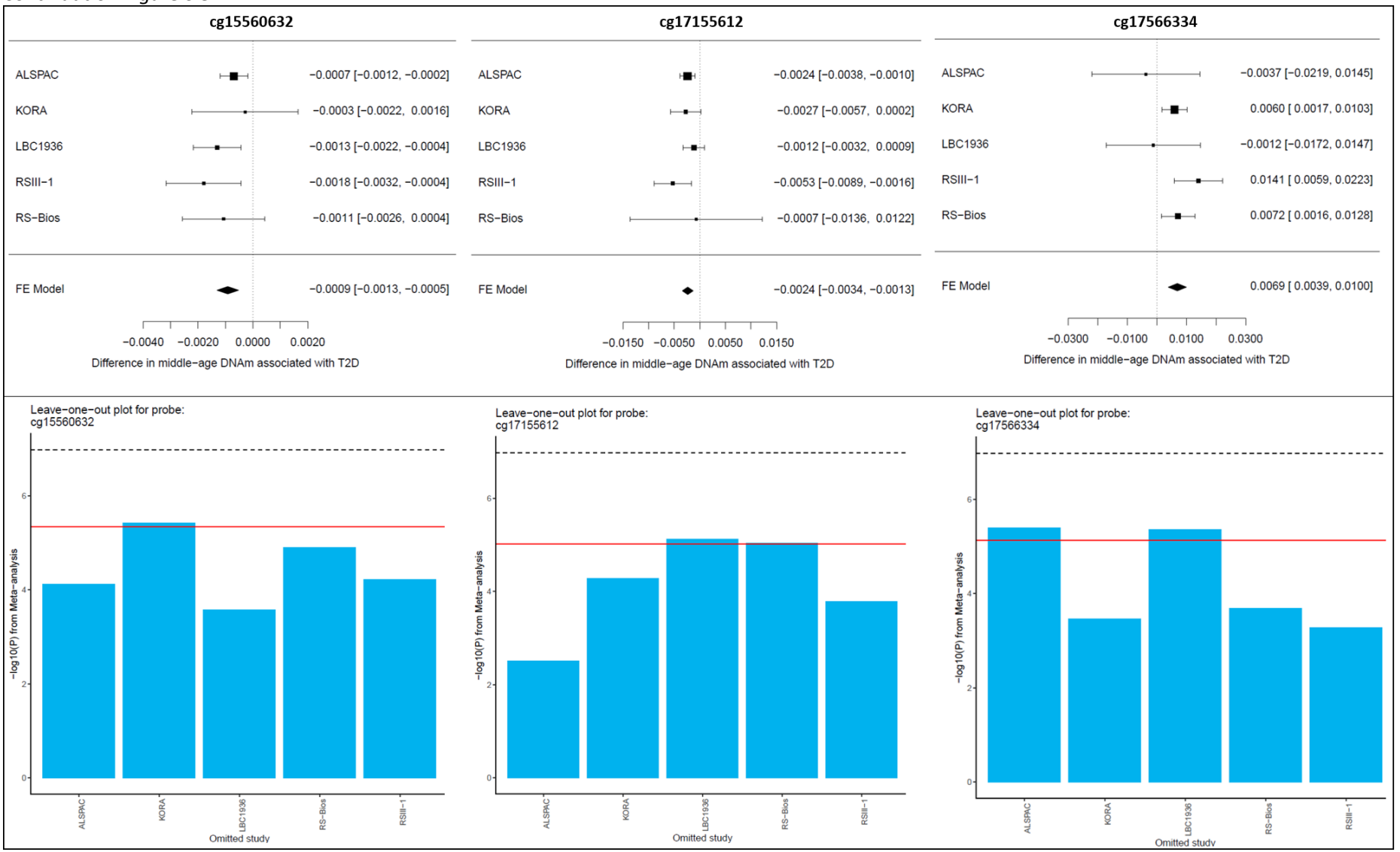


Leave-one-out analysis not available for the CpG cg13826139, as data was incomplete for this site in RS-Bios.

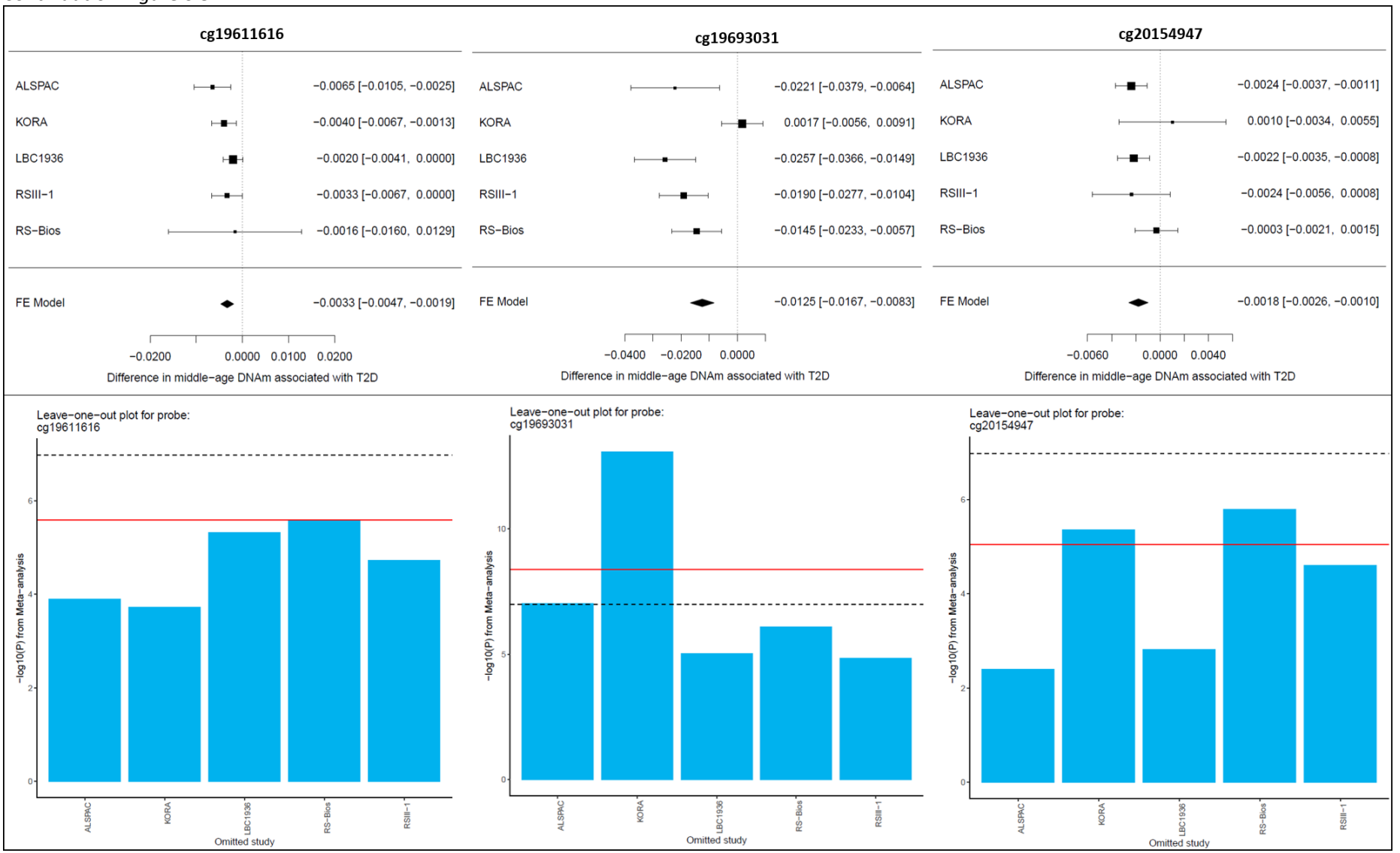
Continuation Figure S 8-17.



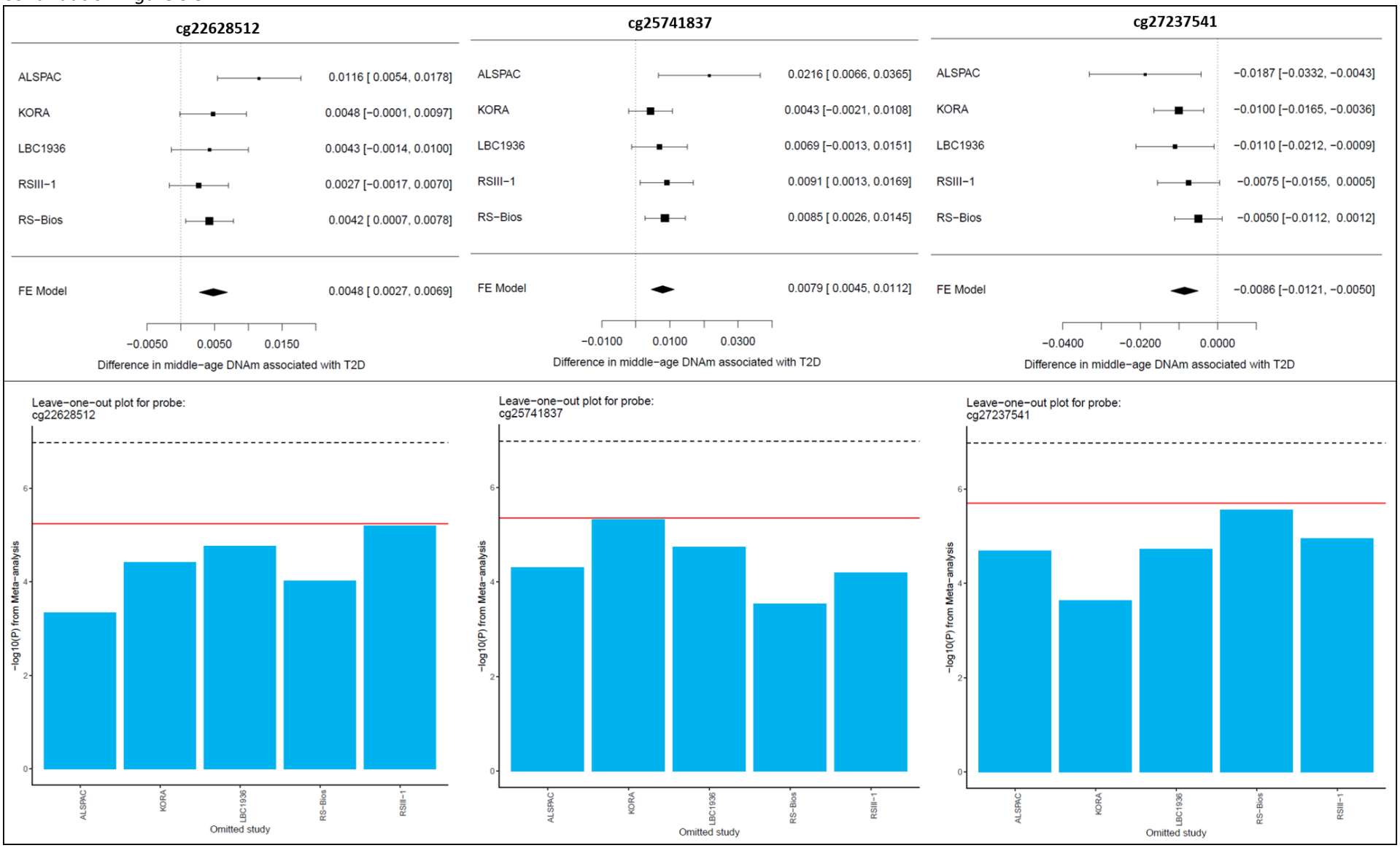
Continuation Figure S 8-17.



Continuation Figure S 8-17.



Continuation Figure S 8-17.



Continuation Figure S 8-17.

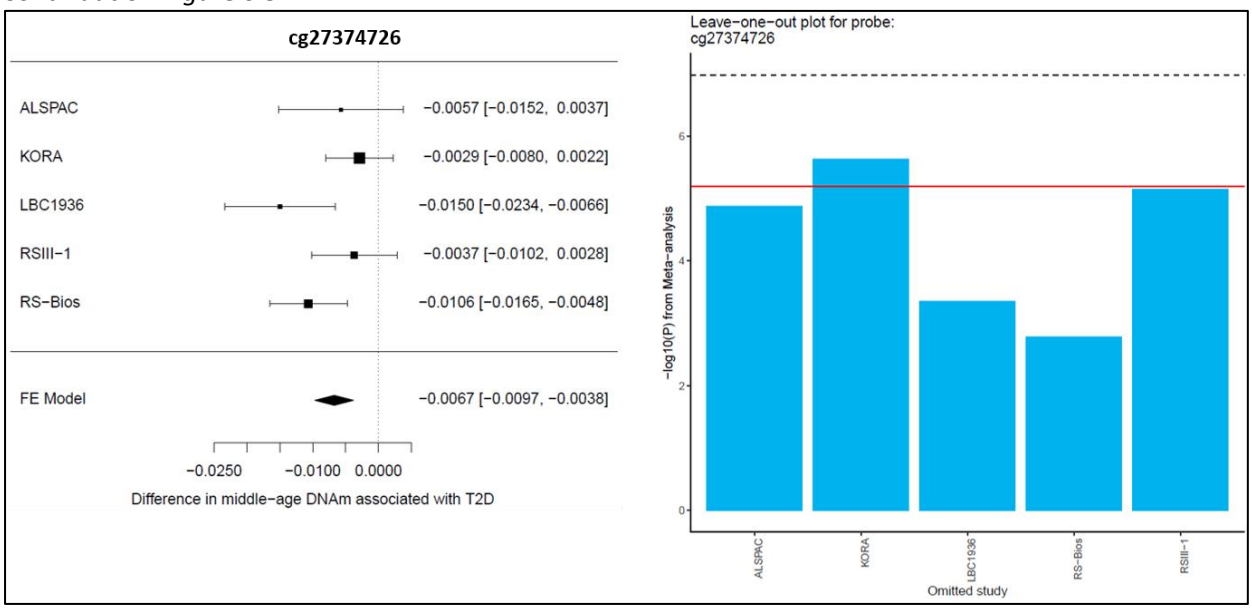


Table S 8-25 Top associations identified in the sensitivity meta-analysis of EWAS in T2D using a model adjusted for age, sex, SVs, 6-Houseman cells and smoking. Sensitivity analysis after excluding KORA from the main analysis. Adj. P is the Bonferroni corrected P-value, I^2 is the heterogeneity estimate, and the P-value for heterogeneity. Direction (+ - ?) represents the direction of the association based on results from each cohort (? If for one study the result for that probe is unknown). From left to right, direction is the effect detected in: ALSPAC, LBC1936, RSIII-1 and RS-Bios.

		ALSPAC (N=1,050)		LBC1936 (N=915)		RSIII-1 (N=728)		RS-Bios (N=735)		Meta-analysis (N=3,428)						
CpG site	Gene	Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Adj. P	Direction	I^2	P
cg19693031	*TXNIP*	-0.022	0.006	-0.026	3.88E-06	-0.019	1.86E-05	-0.015	1.27E-03	-0.019	2.59E-03	8.75E-14	3.29E-08	----	0.00	0.455
cg06500161	*ABCG1*	0.026	0.000	0.024	5.18E-08	0.010	5.06E-03	0.008	9.90E-03	0.013	1.92E-03	2.34E-11	8.81E-06	++++	77.70	0.004
cg16765088	*Intergenic*	-0.021	0.008	-0.014	4.20E-04	-0.009	2.20E-02	-0.009	5.43E-05	-0.011	1.75E-03	5.50E-10	2.07E-04	----	0.00	0.429
cg00574958	*CPT1A*	-0.005	0.036	-0.008	6.68E-05	-0.018	6.19E-06	-0.004	1.51E-01	-0.007	1.25E-03	1.20E-08	0.005	----	72.20	0.013
cg24704287	*Intergenic*	-0.009	0.141	-0.012	1.20E-02	-0.012	6.71E-04	-0.011	8.31E-04	-0.011	1.97E-03	2.34E-08	0.009	----	0.00	0.968
cg00144180	*HDAC4*	0.012	0.056	0.013	5.88E-04	0.019	3.67E-04	0.008	3.94E-02	0.012	2.23E-03	5.64E-08	0.021	++++	4.70	0.369
cg04567334	*CDH23*	-0.001	0.781	-0.005	5.49E-02	-0.007	1.89E-03	-0.007	9.38E-05	-0.006	1.18E-03	1.67E-07	0.063	----	0.00	0.670
cg10584271	*ITIH1*	-0.007	0.358	-0.019	8.08E-04	-0.017	2.92E-04	-0.010	2.43E-02	-0.014	2.63E-03	1.73E-07	0.065	----	5.60	0.365
cg26270261	*KRT4*	-0.004	0.435	-0.011	8.41E-05	-0.007	1.37E-02	-0.004	8.07E-03	-0.006	1.24E-03	5.68E-07	0.214	----	31.10	0.225
cg16575444	*CX3CL1*	0.001	0.790	-0.004	1.29E-01	-0.006	7.51E-03	-0.007	2.90E-05	-0.006	1.22E-03	6.83E-07	0.257	+---	0.00	0.454
cg24512093	*ROBO1*	-0.018	0.036	-0.009	7.87E-02	-0.004	3.28E-01	-0.011	1.06E-05	-0.010	1.92E-03	7.16E-07	0.269	----	16.50	0.309
cg11983038	*Intergenic*	-0.027	0.001	-0.003	7.69E-01	-0.007	2.44E-01	-0.025	7.05E-06	-0.017	3.38E-03	7.23E-07	0.272	----	65.10	0.035
cg25136644	*ATG9B*	-0.010	0.083	-0.007	4.43E-02	-0.005	2.25E-02	-0.012	4.52E-05	-0.007	1.44E-03	7.27E-07	0.274	----	34.80	0.204
cg20812370	*PBX1*	-0.017	0.025	-0.017	9.03E-05	-0.003	1.51E-01	-0.007	5.11E-04	-0.007	1.34E-03	7.40E-07	0.278	----	70.70	0.017
cg24686009	*RAP1B*	-0.002	0.000	-0.001	2.82E-02	-0.003	3.34E-03	0.000	8.43E-01	-0.002	3.83E-04	1.19E-06	0.446	---+	0.20	0.391
cg11024682	*SREBF1*	0.007	0.107	0.007	1.95E-02	0.011	1.08E-03	0.006	1.48E-02	0.008	1.59E-03	1.33E-06	0.501	++++	0.00	0.694
cg06114363	*ZNF683*	-0.016	0.023	-0.012	5.08E-04	-0.008	7.40E-03	na	na	-0.010	2.16E-03	1.37E-06	0.514	---?	0.00	0.530
cg01963618	*LOC285768*	-0.010	0.117	-0.011	5.08E-03	-0.007	4.91E-02	-0.007	1.77E-03	-0.008	1.58E-03	1.55E-06	0.584	----	0.00	0.801
cg22680424	*HCCA2*	0.009	0.238	0.013	6.20E-04	0.007	3.18E-02	0.006	8.16E-03	0.008	1.63E-03	2.16E-06	0.813	++++	0.00	0.492
cg19876302	*Intergenic*	-0.020	0.023	-0.007	1.06E-01	-0.010	2.21E-03	-0.006	5.08E-03	-0.008	1.70E-03	2.22E-06	0.836	----	0.00	0.430
cg08857797	*VPS25*	0.002	0.808	0.012	1.09E-02	0.008	6.17E-03	0.008	2.53E-03	0.009	1.84E-03	2.28E-06	0.856	++++	0.00	0.813
cg27374726	*Intergenic*	-0.006	0.235	-0.015	5.06E-04	-0.004	2.66E-01	-0.011	3.93E-04	-0.009	1.83E-03	2.32E-06	0.872	----	42.40	0.157
cg09185884	*KCTD2*	0.011	0.029	0.015	6.40E-02	0.013	4.18E-02	0.010	1.28E-03	0.011	2.31E-03	2.33E-06	0.875	++++	0.00	0.919
cg27115863	*Intergenic*	-0.023	0.003	-0.006	2.40E-01	-0.010	2.27E-02	-0.012	1.23E-03	-0.011	2.31E-03	2.41E-06	0.907	----	13.40	0.325
cg24795867	*WNT5B*	-0.007	0.359	-0.002	6.13E-01	-0.002	3.62E-01	-0.008	5.51E-07	-0.006	1.29E-03	2.47E-06	0.929	----	43.20	0.152
cg08945443	*ZMYND17*	-0.006	0.392	0.009	1.63E-01	0.015	6.67E-04	0.012	2.08E-04	0.010	2.23E-03	2.64E-06	0.994	-+++	51.70	0.102
cg06039489	*C20orf26*	0.008	0.303	0.028	2.22E-03	0.010	1.39E-01	0.018	3.43E-04	0.016	3.35E-03	2.71E-06	1.000	++++	16.80	0.308
cg08570691	*RPL13AP5*	-0.012	0.078	-0.010	3.93E-03	-0.014	3.64E-04	-0.004	8.95E-02	-0.008	1.76E-03	2.78E-06	1.000	----	42.60	0.156

Continuation Table S 8-25.

CpG site	Gene	ALSPAC (N=1,050)		LBC1936 (N=915)		RSIII-1 (N=728)		RS-Bios (N=735)		Meta-analysis (N=3,428)						
		Beta	P	Beta	P	Beta	P	Beta	P	Beta	SE	P	Adj. P	Direction	I^2	P
cg12593793	*Intergenic*	-0.012	0.029	-0.011	1.17E-03	-0.014	4.03E-04	-0.003	2.27E-01	-0.008	1.65E-03	2.90E-06	1.000	----	60.50	0.055
cg27037013	*Intergenic*	-0.027	0.024	-0.025	7.68E-04	-0.019	2.20E-03	-0.007	1.41E-01	-0.015	3.22E-03	2.90E-06	1.000	----	51.90	0.101
cg07212837	*Intergenic*	0.003	0.545	0.008	6.22E-03	0.007	2.72E-02	0.006	1.93E-03	0.006	1.38E-03	3.28E-06	1.000	++++	0.00	0.843
cg16192197	*Intergenic*	0.007	0.431	0.005	3.38E-01	0.009	1.75E-02	0.013	5.29E-05	0.010	2.06E-03	3.71E-06	1.000	++++	0.00	0.503
cg15560632	*LRCH4*	-0.001	0.007	-0.001	3.40E-03	-0.002	1.02E-02	-0.001	1.62E-01	-0.001	2.04E-04	3.83E-06	1.000	----	3.50	0.375
cg14003143	*SGK2*	-0.021	0.001	-0.007	1.49E-02	-0.007	8.21E-03	-0.004	3.85E-02	-0.006	1.27E-03	4.12E-06	1.000	----	62.60	0.045
cg20154947	*PLEC1*	-0.002	0.000	-0.002	1.49E-03	-0.002	1.48E-01	0.000	7.36E-01	-0.002	4.09E-04	4.34E-06	1.000	----	23.40	0.271
cg25741837	*SMYD5*	0.022	0.005	0.007	9.85E-02	0.009	2.33E-02	0.009	4.74E-03	0.009	2.01E-03	4.76E-06	1.000	++++	0.00	0.396
cg26766064	*MIR657*	-0.014	0.086	-0.010	7.14E-03	-0.008	1.59E-02	-0.005	6.10E-03	-0.007	1.43E-03	5.17E-06	1.000	----	0.00	0.412
cg25536676	*DHCR24*	-0.007	0.252	-0.012	1.48E-03	-0.014	3.67E-04	-0.004	8.16E-02	-0.008	1.68E-03	5.39E-06	1.000	----	54.80	0.084
cg11376147	*SLC43A1*	-0.008	0.010	-0.004	5.34E-02	-0.007	1.94E-02	-0.005	1.45E-02	-0.006	1.27E-03	5.43E-06	1.000	----	0.00	0.764
cg20316538	*RUFY4*	-0.008	0.115	-0.007	7.53E-03	-0.007	1.39E-02	-0.004	1.06E-02	-0.005	1.16E-03	6.11E-06	1.000	----	0.00	0.605
cg18181703	*SOCS3*	-0.024	0.007	-0.011	2.08E-02	-0.016	2.85E-04	-0.004	2.26E-01	-0.010	2.31E-03	6.20E-06	1.000	----	56.50	0.075
cg11851382	*PPAP2B*	-0.012	0.023	-0.012	3.79E-03	-0.005	2.13E-01	-0.006	5.87E-03	-0.008	1.69E-03	6.42E-06	1.000	----	0.00	0.491
cg00162348	*RNF40*	-0.002	0.000	-0.001	5.06E-02	-0.003	4.86E-02	-0.001	5.94E-01	-0.002	4.23E-04	6.64E-06	1.000	----	0.00	0.600
cg07184465	*SPZ1*	-0.012	0.038	-0.006	6.66E-02	-0.011	1.40E-04	-0.004	1.03E-01	-0.007	1.55E-03	7.18E-06	1.000	----	38.20	0.183
cg14284506	*Intergenic*	-0.009	0.001	-0.005	1.01E-03	-0.003	2.51E-01	-0.002	5.00E-01	-0.005	1.08E-03	7.31E-06	1.000	----	21.90	0.279
cg11252555	*RPL13AP5*	-0.013	0.047	-0.010	3.55E-03	-0.015	1.85E-04	-0.003	2.03E-01	-0.008	1.73E-03	7.44E-06	1.000	----	63.30	0.043
cg10082515	*Intergenic*	-0.022	0.003	-0.006	5.79E-01	-0.010	7.04E-02	-0.014	1.56E-03	-0.013	3.00E-03	7.46E-06	1.000	----	0.00	0.505
cg00896068	*Intergenic*	-0.003	0.650	-0.011	7.96E-03	-0.004	3.10E-01	-0.009	1.79E-04	-0.008	1.74E-03	7.58E-06	1.000	----	0.00	0.482
cg01577083	*Intergenic*	-0.018	0.005	-0.012	4.99E-02	-0.008	5.13E-02	-0.012	1.44E-02	-0.011	2.54E-03	7.93E-06	1.000	----	0.00	0.621
cg00320980	*Intergenic*	-0.006	0.344	-0.007	8.21E-02	-0.012	2.75E-03	-0.010	4.24E-03	-0.009	2.09E-03	7.97E-06	1.000	----	0.00	0.744
cg20231084	*Intergenic*	-0.011	0.112	-0.008	1.78E-02	-0.005	6.20E-02	-0.006	2.49E-03	-0.006	1.39E-03	8.36E-06	1.000	----	0.00	0.796
cg15832662	*RTN3*	-0.010	0.170	-0.012	6.91E-03	-0.012	5.16E-02	-0.008	5.65E-03	-0.009	2.08E-03	8.45E-06	1.000	----	0.00	0.829
cg13178597	*RGS17*	-0.018	0.014	-0.005	4.64E-01	-0.009	2.36E-02	-0.011	1.60E-03	-0.010	2.27E-03	8.57E-06	1.000	----	0.00	0.580
cg20116935	*SEMA3B*	-0.015	0.004	-0.005	5.45E-02	-0.006	6.60E-02	-0.005	6.69E-03	-0.006	1.38E-03	8.89E-06	1.000	----	4.80	0.369
cg00989505	*MIR299*	0.003	0.516	-0.004	1.00E-01	-0.006	4.00E-04	-0.004	6.73E-03	-0.004	9.41E-04	9.33E-06	1.000	+---	14.80	0.318
cg07068382	*MTCH1*	0.000	0.991	0.020	1.38E-04	0.011	1.32E-02	0.007	3.60E-02	0.010	2.37E-03	9.46E-06	1.000	-+++	46.70	0.131
cg14476101	*PHGDH*	-0.021	0.041	0.000	9.46E-01	-0.021	2.41E-04	-0.017	4.72E-03	-0.015	3.36E-03	9.46E-06	1.000	----	52.00	0.100
cg20456243	*SPEG*	-0.015	0.065	-0.004	2.74E-01	-0.010	3.31E-03	-0.007	3.41E-03	-0.007	1.65E-03	9.99E-06	1.000	----	0.00	0.490

Table S 8-26 Function of 4 top CpG sites identified in the meta-analysis of T2D conducted across five European cohorts. Functional description of genes extracted from the Human gene database Gene cards (https://www.genecards.org/).

TXNIP: This protein binds and inhibits Thioredoxin, an important regulator of cellular oxidative stress. Therefore, *TXNIP* can induce a state of oxidative stress in the cell by increasing the levels of reactive oxygen species. *TXNIP* is also a tumour suppressor gene and a regulator of cellular metabolism. *TXNIP* expression is upregulated under increased levels of glucose, and under these conditions, *TXNIP* downregulates the expression of GLUT1, a key transmembrane transporter of glucose across the cell. It has been reported that *TXNIP* is importantly expressed in human-islets, and when upregulated, it can induce β-cell death, but downregulation of *TXNIP* can prevent β-cell death and obesity-induced T2D.
ABCG1: this protein belongs to the superfamily of ATP-binding cassette (ABC) transporters, which comprises 7 sub-families. The *ABCG1* gene belongs to the White sub-family. These proteins are involved in the movement of different molecules across the membrane. In macrophages, *ABCG1* participates in the transport of phospholipids and cholesterol, and may also be involved in the maintenance of lipid homeostasis in other cell-types. *ABCG1* may participate also in insulin secretion. Pathways associated with *ABCG1* are the “transport of glucose and other sugars, biles, metal ions and amine compounds”, and a pathway related with “nuclear receptors in lipid's metabolism and toxicity”.
CPT1A: The carnitine palmitoyl-transferase 1A enzyme catalyses the transfer of the acyl group of long-chain fatty acid-CoA conjugates onto carnitine, an essential step for the mitochondrial uptake of long-chain fatty acids, and their subsequent beta-oxidation in the mitochondrion. Plays an important role in triglyceride metabolism.
HDAC4: Histone deacetylases (HDACs) are a group of enzymes closely related to sirtuins. They catalyse acetyl group removal from lysine residues in histones and non-histone proteins, causing transcriptional repression. The protein encoded by *HDAC4* belongs to class II of the histone deacetylase/acuc/apha family. It possesses histone deacetylase activity and represses transcription when tethered to a promoter. This protein does not bind DNA directly, but through transcription factors MEF2C and MEF2D. *HDAC4* seems to interact in a multiprotein complex with *RbAp48* and *HDAC3*.

Table S8-27 Top GO terms identified in enrichment for genes annotated to top 25 CpG sites in the meta-analysis of T2D (main analysis), and to top 58 CpG sites in a sensitivity analysis (excluding KORA from meta-analysis). Total genes: total number of genes pertaining to a term in relation to background genes included in the HM450K array. Differentially methylated: proportion of genes identified with differential methylation from the total number of genes in a term. P: p-value for enrichment (unadjusted). None of the terms identified by GO surpassed multiple testing correction at FDR<0.05.

	GO term	Total Genes	Differentially methylated	P
Main meta-analysis	intracellular lipid transport	27	2	6.17E-04
	ADP binding	30	2	9.41E-04
	ER-nucleus signalling pathway	41	2	1.44E-03
	canonical Wnt signalling pathway involved in positive regulation of cell-cell adhesion	1	1	1.78E-03
	canonical Wnt signalling pathway involved in positive regulation of endothelial cell migration	1	1	1.78E-03
	canonical Wnt signalling pathway involved in positive regulation of wound healing	1	1	1.78E-03
	regulation of sphingolipid mediated signalling pathway	1	1	1.78E-03
	actin binding	365	4	1.87E-03
	cellular response to tumour cell	1	1	1.87E-03
	detection of hormone stimulus	1	1	1.89E-03
	glycoprotein transporter activity	1	1	1.89E-03
	Palmitoleoyl-transferase activity	1	1	1.92E-03
	regulation of endoplasmic reticulum tubular network organization	3	1	3.45E-03
	glycoprotein transport	2	1	3.65E-03
	sterol-transporting ATPase activity	2	1	3.72E-03
	positive regulation of PERK-mediated unfolded protein response	4	1	3.78E-03
	radial spoke stalk	3	1	3.78E-03
	cellular process involved in reproduction in multicellular organism	252	3	3.85E-03
	lipid-transporting ATPase activity	3	1	3.97E-03
	response to high density lipoprotein particle	3	1	4.24E-03

Continuation Table S8-27.

	GO term	Total Genes	Differentially methylated	P
Sensitivity Analysis	Positive regulation of cholesterol biosynthetic process	9	2	3.63E-04
	Positive regulation of sterol biosynthetic process	9	2	3.63E-04
	Positive regulation of cholesterol metabolic process	10	2	4.42E-04
	Positive regulation of small molecule metabolic process	127	4	6.59E-04
	Cholesterol biosynthetic process	69	3	1.14E-03
	Secondary alcohol biosynthetic process	70	3	1.18E-03
	Positive regulation of steroid biosynthetic process	19	2	1.29E-03
	Pulmonary valve morphogenesis	15	2	1.33E-03
	Sterol biosynthetic process	75	3	1.46E-03
	Steroid biosynthetic process	177	4	1.67E-03
	Negative regulation of response to external stimulus	305	5	2.00E-03
	Pulmonary valve development	18	2	2.02E-03
	Positive regulation of steroid metabolic process	28	2	2.48E-03
	Cellular response to tumour cell	1	1	2.76E-03
	Phosphoglycerate dehydrogenase activity	1	1	2.77E-03
	Positive regulation of alcohol biosynthetic process	27	2	2.80E-03
	CX3C chemokine receptor binding	1	1	2.82E-03
	CXCR1 chemokine receptor binding	1	1	2.82E-03
	Positive regulation of calcium-independent cell-cell adhesion	1	1	2.82E-03
	Negative regulation of biological process	5008	25	2.94E-03

Table S 8-28 Top KEGG pathways identified in enrichment for genes annotated to 25 top CpG sites in the meta-analysis of T2D (main analysis), and for 58 top CpG sites detected in a sensitivity analysis (excluding KORA from meta-analysis). Total genes: total number of genes pertaining to a pathway in relation to background genes included in the HM450K array. Differentially methylated: proportion of genes identified with differential methylation from the total number of genes in a pathway. P: p-value for enrichment (unadjusted). None of the pathways identified in KEGG surpassed multiple testing correction at FDR<0.05. Highlighted in bold are pathways identified in common between the main and the sensitivity analysis.

	Pathway (KEGG)	Total Genes	Differentially methylated	P
Main meta-analysis	**Fat digestion and absorption**	35	1	2.34E-02
	ABC transporters	39	1	3.05E-02
	Ether lipid metabolism	46	1	3.49E-02
	Fatty acid degradation	42	1	3.64E-02
	Sphingolipid metabolism	41	1	3.95E-02
	Fatty acid metabolism	46	1	4.64E-02
	PPAR signalling pathway	65	1	4.96E-02
	Glycerolipids metabolism	53	1	5.01E-02
	Taste transduction	75	1	5.35E-02
	Adipocytokine signalling pathway	61	1	5.99E-02
	Glycerophospholipid metabolism	88	1	8.32E-02
	Glucagon signalling pathway	91	1	8.39E-02
	Fc gamma R-mediated phagocytosis	84	1	8.70E-02
	Serotonergic synapse	103	1	8.90E-02
	Choline metabolism in cancer	90	1	9.17E-02
	Insulin resistance	100	1	1.03E-01
	NOD-like receptor signalling pathway	142	1	1.12E-01
	Spliceosome	116	1	1.19E-01
	AMPK signalling pathway	112	1	1.21E-01
	Phospholipase D signalling pathway	133	1	1.35E-01
Sensitivity Analysis	Cushing syndrome	145	3	1.47E-02
	Axon guidance	168	3	2.62E-02
	TNF signalling pathway	100	2	3.54E-02
	Steroid biosynthesis	16	1	4.45E-02
	Arginine biosynthesis	17	1	4.94E-02
	Sphingolipid signalling pathway	111	2	5.02E-02
	Platelet activation	118	2	5.62E-02
	Autophagy - animal	115	2	5.73E-02
	Non-alcoholic fatty liver disease (NAFLD)	136	2	5.96E-02
	Oestrogen signalling pathway	127	2	6.21E-02
	Insulin signalling pathway	124	2	6.41E-02
	Apelin signalling pathway	130	2	6.91E-02
	Glycine, serine and threonine metabolism	31	1	7.43E-02
	Autophagy - other	29	1	9.11E-02
	Chemokine signalling pathway	177	2	9.50E-02

Table S 8-29 Summary of associations between DNA methylation at three of the top seven CpG sites identified in the meta-analysis, and phenotypes related with T2D. Results are based on unadjusted regressions between untransformed β-values of methylation as the exposure, and the phenotype as the outcome. Association analysis restricted to samples in ALSPAC (n=1,050). In bold are highlighted associations surpassing significance at p<0.05.

	cg13826139 (intergenic)			cg16765088 (intergenic)			cg24704287 (intergenic)		
	Estimate†	SE	P	Estimate	SE	P	Estimate	SE	P
2-hours Glucose	-1.57	0.70	**2.43E-02**	-1.48	0.45	**1.12E-03**	-0.65	0.49	0.19
Fasting glucose	-2.61	0.83	**1.79E-03**	-1.33	0.55	**1.58E-02**	-0.72	0.59	0.23
Fasting insulin*	-1.02	0.62	9.92E-02	-0.30	0.44	5.02E-01	-0.38	0.45	0.41
HOMA-IR*	-1.33	0.68	4.97E-02	-0.44	0.48	3.64E-01	-0.36	0.50	0.46
HOMA-B*	-0.27	0.57	6.33E-01	0.02	0.40	9.55E-01	-0.51	0.41	0.22
Glucose Tolerance	Mean (SD)	Effect size (%) ‡	P**	Mean (SD)	Effect size (%) ‡	P	Mean (SD)	Effect size (%) ‡	P
Normoglycaemic	0.79(0.04)	Ref	Ref	0.45(0.06)	Ref	Ref	0.34(0.05)	Ref	Ref
Prediabetes	0.78(0.04)	0.60	8.25E-02	0.45(0.06)	0.04	9.96E-01	0.34(0.05)	0.29	7.41E-01
T2D cases	0.77(0.04)	1.74	1.91E-02	0.42(0.07)	2.90	8.50E-03	0.32(0.06)	1.87	1.04E-01
P for trend‡‡			**4.58E-03**			**1.20E-02**			1.10E-01

†Estimates are interpreted as the effect of 1% increase in methylation on a unit change in the phenotype. ‡ Percentage of the absolute difference in DNA methylation between the reference category (normoglycaemic) and the comparison categories (prediabetes, diabetes). *Variables log-transformed before conducting the regression analysis, and available only for a subset of 645 females in ALSPAC. **Unadjusted p-value for the paired comparison between categories of glucose tolerance. ‡‡ Adjusted P-value for the comparison across categories. P< 0.05 indicates evidence of a linear trend in the association between methylation and glucose tolerance categories.

Table S8-30 Comparison of mean levels of β-values of methylation across quartiles for seven top CpG sites identified in the meta-analysis of T2D. Quartile analysis restricted to samples in ALSPAC (N=1,050). Difference in methylation across quartiles was calculated using an ANOVA. P-for-trend <0.05 indicates significant difference in methylation between Q1 and Q4.

CpG	Gene	Quartile	N	Mean	SD	P	P (Q4 vs Q1)
cg19693031	*TXNIP*	Q1	262	0.65	0.03	Ref*	**<0.001**
		Q2	262	0.72	0.01	<0.001	
		Q3	261	0.76	0.01	<0.001	
		Q4	262	0.80	0.02	<0.001	
cg13826139	*Intergenic*	Q1	263	0.79	0.04	Ref	0.01
		Q2	262	0.79	0.03	0.65	
		Q3	262	0.78	0.03	0.93	
		Q4	263	0.78	0.04	0.16	
cg06500161	*ABCG1*	Q1	262	0.57	0.05	Ref	0.89
		Q2	262	0.57	0.05	0.98	
		Q3	261	0.57	0.05	0.99	
		Q4	262	0.57	0.06	0.86	
cg16765088	*Intergenic*	Q1	259	0.44	0.06	Ref	0.31
		Q2	259	0.45	0.06	0.31	
		Q3	259	0.45	0.05	0.41	
		Q4	259	0.45	0.06	0.76	
cg00574958	*CPT1A*	Q1	262	0.05	0.02	Ref	0.90
		Q2	262	0.05	0.02	0.31	
		Q3	262	0.05	0.02	0.41	
		Q4	262	0.05	0.02	0.76	
cg24704287	*Intergenic*	Q1	263	0.27	0.04	Ref	**<0.001**
		Q2	262	0.33	0.01	<0.001	
		Q3	262	0.36	0.01	<0.001	
		Q4	262	0.40	0.02	<0.001	
cg00144180	*HDAC4*	Q1	263	0.79	0.06	Ref	0.05
		Q2	262	0.80	0.06	0.30	
		Q3	262	0.80	0.06	0.19	
		Q4	262	0.79	0.06	1.00	

*Quartile one is the reference quartile. Quartiles were generated to represent an increase in methylation from Q1 to Q4.

Table S8-31 Summary of associations between quartiles of methylation at six of the seven top CpG sites detected in the meta-analysis, and different sociodemographic, anthropometric and metabolic factors of relevance in T2D. Sensitivity analysis restricted to samples in the ALSPAC dataset (n=1,050). Continuous variables were summarized using the mean and the standard deviation, while for categorical variables the proportion of samples per category is shown. Results are interpreted as the change in the trait between Q1 (lower methylation) and Q4 (higher methylation) of methylation at cg13826139.

cg13826139 (*intergenic*)	Quartile 1 (n=263)	Quartile 2 (n=262)	Quartile 3 (n=262)	Quartile 4 (n=263)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	48.29(4.62)	49.56(4.36)	50.29(5.88)	52.09(5.91)	**4.31E-15**
BMI [kg/m^2]	26.09(4.68)	26.49(4.54)	27.17(5.04)	26.68(4.13)	6.02E-02
waist-circumference [cm]	85.55(13.43)	88.28(12.99)	90.43(14.08)	91.07(11.99)	**3.74E-06**
Fasting Glucose [mmol/l] *	5.24(0.48)	5.4(1.1)	5.52(1.35)	5.52(0.94)	**7.41E-04**
2-hours Glucose [mmol/l]	4.3(0.43)	4.37(0.97)	4.45(1.03)	4.37(0.76	3.11E-01
C-reactive Protein [mg/l] *	1.82(3.06)	1.94(2.65)	2.19(3.27)	1.98(2.09)	**2.85E-03**
Fasting Insulin [µIU/ml] [a]*	6.1(4.96)	5.65(4.57)	6.08(7.71)	5.54(3.93)	7.82E-01
HOMA-IR [a]*	1.59(2.02)	1.36(1.27)	1.5(2.68)	1.33(1.11)	7.28E-01
HOMA-B [a]*	66.52(41.57)	65.53(53.13)	73.06(77.59)	67.85(64.74)	6.82E-01
Cholesterol [mmol/l]	4.8(0.92)	4.82(0.92)	4.77(0.9)	4.86(0.95)	7.35E-01
Triglycerides [mmol/l] *	1.11(0.67)	1.11(0.55)	1.21(0.65)	1.26(0.68	**1.20E-03**
HDL [mmol/l]	1.48(0.36)	1.43(0.36)	1.37(0.35)	1.34(0.31)	**1.13E-05**
LDL [mmol/l]	3.04(0.86)	3.04(0.84)	3.01(0.79)	3.06(0.78	9.20E-01
Systolic Blood Pressure [mmHg]	120.37(12.81)	122.1(13.62)	122.97(14.45)	126.98(14.89)	**7.81E-07**
Diastolic Blood Pressure [mmHg] *	72.71(11.51)	73.88(8.73)	73.75(10.66)	75.97(10.79)	**9.99E-04**
CD8$^+$ T cells	0.01(0.03)	0.01(0.03)	0.02(0.03)	0.02(0.03)	**1.03E-03**
CD4$^+$ T cells	0.2(0.05)	0.18(0.05)	0.16(0.04)	0.13(0.05)	**<0.001**
Natural Killer Cells	0.21(0.05)	0.2(0.05)	0.2(0.05)	0.19(0.06)	**2.80E-04**
B cells	0.11(0.03)	0.1(0.03)	0.09(0.03)	0.08(0.03)	**<0.001**
Monocytes	0.06(0.03)	0.07(0.03	0.07(0.03)	0.08(0.03)	**4.09E-14**
Granulocytes	0.47(0.08)	0.5(0.08)	0.52(0.07)	0.54(0.09)	**<0.001**
Categorical Phenotypes					
Sex [female/male]	202/61	174/88	156/106	113/150	**6.99E-16**
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	213/47/3	203/52/7	190/60/12	173/76/14	**7.09E-06**

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (161/161/160/161).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Continuation Table S8-31.

***ABCG1* (cg06500161)**	Quartile 1 (n=262)	Quartile 2 (n=262)	Quartile 3 (n=261)	Quartile 4 (n=262)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	50.41(5.77)	50.75(5.23)	50.1(5.65)	48.97(4.83	**1.05E-03**
BMI [kg/m^2]	27.02(4.58)	26.65(4.6)	26.4(4.43)	26.32(4.86)	3.04E-01
waist-circumference [cm]	91.02(13.21)	89.87(13.63)	87.65(12.52)	86.8(13.51)	**7.54E-04**
Fasting Glucose [mmol/l] *	5.54(1.16)	5.43(0.96)	5.38(0.81)	5.32(1.13)	**1.83E-02**
2-hours Glucose [mmol/l]	4.45(0.98)	4.34(0.71)	4.36(0.67)	4.35(0.92)	3.50E-01
C-reactive Protein [mg/l] *	1.96(2.82)	2.04(2.82)	2.07(2.91)	1.87(2.67)	5.44E-01
Fasting Insulin [µIU/ml] [a]	4.84(3.57)	5.71(7.89)	5.98(3.71)	6.81(5.45)	**1.66E-04**
HOMA-IR [a]	1.11(0.86)	1.45(2.78)	1.45(1.11)	1.76(2.05)	**1.63E-04**
HOMA-B [a]	63.34(65.95)	61.12(41.92)	73.6(72.93)	74.7(56.48)	**1.87E-03**
Cholesterol [mmol/l]	4.89(1.04)	4.93(0.9)	4.76(0.85)	4.66(0.86)	**4.23E-03**
Triglycerides [mmol/l] *	1.24(0.7)	1.24(0.63)	1.12(0.61)	1.1(0.62)	**1.06E-03**
HDL [mmol/l]	1.36(0.34)	1.4(0.35)	1.41(0.34)	1.46(0.36)	**8.76E-03**
LDL [mmol/l]	3.05(0.87)	3.11(0.83)	3.04(0.77)	2.94(0.78)	1.24E-01
Systolic Blood Pressure [mmHg]	124.9(14.1)	124.71(15.33)	121.48(13.27)	121.37(13.58)	**1.89E-03**
Diastolic Blood Pressure [mmHg] *	74.72(8.83)	74.99(12.05)	73.33(9.8)	73.3(11.12)	8.81E-02
CD8$^+$ T cells	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.01(0.02)	**3.65E-03**
CD4$^+$ T cells	0.17(0.05)	0.17(0.05)	0.17(0.05)	0.17(0.05)	5.82E-01
Natural Killer Cells	0.2(0.06)	0.2(0.05)	0.2(0.05)	0.2(0.05)	1.99E-01
B cells	0.09(0.03)	0.09(0.03)	0.09(0.03)	0.1(0.03)	2.00E-01
Monocytes	0.08(0.03)	0.07(0.03)	0.07(0.03)	0.07(0.03)	**4.44E-02**
Granulocytes	0.5(0.09)	0.51(0.09)	0.51(0.08)	0.52(0.08)	**6.62E-03**
Categorical Phenotypes					
Sex [female/male]	138/124	137/125	173/88	195/67	**3.88E-09**
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	178/69/15	191/63/8	199/55/7	208/48/6	**8.67E-04**

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (162/161/161/161).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Continuation Table S8-31.

cg16765088 (*intergenic*)	Quartile 1 (n=259)	Quartile 2 (n=259)	Quartile 3 (n=259)	Quartile 4 (n=259)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	51.47(5.66)	49.91(5.56)	49.74(5.3)	49.06(4.91)	**4.51E-06**
BMI [kg/m^2]	27.43(4.8)	26.95(4.41)	26.6(4.66)	25.56(4.41)	**3.91E-05**
waist-circumference [cm]	93.07(13.26)	90.2(12.76)	88.69(13.49)	83.76(11.97)	**9.24E-15**
Fasting Glucose [mmol/l] *	5.51(0.9)	5.5(1.36)	5.39(0.95)	5.25(0.62)	**1.42E-03**
2-hours Glucose [mmol/l]	4.38(0.67)	4.42(1.02)	4.37(0.88)	4.29(0.57)	4.04E-01
C-reactive Protein [mg/l] *	2.17(2.88)	2.23(3.13)	2.03(2.9)	1.5(2.16)	**2.62E-05**
Fasting Insulin [µIU/ml] [a]*	6.24(5.32)	5.17(3.33)	6.09(7.85)	5.86(4.3)	4.44E-01
HOMA-IR [a]*	1.59(1.96)	1.26(1.13)	1.5(2.71)	1.43(1.27)	4.83E-01
HOMA-B [a]*	73.66(78.43)	60.73(29.92)	72.16(78)	66.27(39.19)	4.00E-01
Cholesterol [mmol/l]	4.96(0.95)	4.86(0.95)	4.77(0.9)	4.66(0.87)	**0.001761**
Triglycerides [mmol/l] *	1.37(0.66)	1.24(0.7)	1.13(0.68)	0.97(0.44)	**7.80E-15**
HDL [mmol/l]	1.31(0.32)	1.39(0.34)	1.44(0.34)	1.47(0.37)	**8.70E-07**
LDL [mmol/l]	3.12(0.84)	3.07(0.82)	3.03(0.8)	2.93(0.78)	5.30E-02
Systolic Blood Pressure [mmHg]	126.5(13.92)	123.83(12.84)	123.83(15.58)	118.48(13.04)	**1.19E-09**
Diastolic Blood Pressure [mmHg] *	75.16(9.6)	74.65(10.46)	74.51(11.08)	71.98(10.37)	**5.00E-04**
CD8$^+$ T cells	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.01(0.03)	5.45E-02
CD4$^+$ T cells	0.16(0.05)	0.17(0.05)	0.17(0.05)	0.18(0.05)	**8.69E-06**
Natural Killer Cells	0.19(0.06)	0.2(0.05)	0.2(0.05)	0.21(0.05)	6.06E-02
B cells	0.09(0.03)	0.09(0.03)	0.1(0.03)	0.1(0.03)	**2.51E-04**
Monocytes	0.08(0.03)	0.07(0.03)	0.07(0.03)	0.07(0.03)	**3.93E-04**
Granulocytes	0.52(0.09)	0.51(0.08)	0.5(0.08)	0.5(0.08)	**7.57E-03**
Categorical Phenotypes					
Sex [female/male]	109/150	151/108	167/92	205/54	**<0.001**
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	175/73/11	184/65/10	191/58/10	220/35/4	**1.51E-05**

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (161/161/161/161).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Continuation Table S8-31.

***CPT1A* (cg00574958)**	Quartile 1 (n=262)	Quartile 2 (n=262)	Quartile 3 (n=262)	Quartile 4 (n=262)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	51.05(5.83)	50.04(5.18)	49.58(5.00)	49.48(5.44)	**3.09E-03**
BMI [kg/m^2]	26.59(4.51)	26.22(4.16)	27.08(4.98)	26.53(4.77)	1.97E-01
waist-circumference [cm]	88.82(13.37)	87.86(12.74)	89.23(13.65)	89.41(13.49)	5.47E-01
Fasting Glucose [mmol/l] *	5.46(1.07)	5.5(1.46)	5.38(0.71)	5.34(0.64)	4.98E-01
2-hours Glucose [mmol/l]	4.44(0.94)	4.44(1.11)	4.32(0.54)	4.28(0.56)	6.76E-02
C-reactive Protein [mg/l] *	2.13(3.18)	2.04(3.14)	1.9(2.44)	1.87(2.35)	9.85E-01
Fasting Insulin [µIU/ml] [a]*	6.53(4.46)	5.67(4.23)	5.71(8.15)	5.29(3.54)	**2.32E-03**
HOMA-IR [a]*	1.57(1.25)	1.41(1.37)	1.44(2.84)	1.25(0.91)	**3.08E-03**
HOMA-B [a]*	76.76(66.99)	64.85(37.07)	69.62(87.34)	61.86(36.53)	**1.14E-02**
Cholesterol [mmol/l]	4.86(0.92)	4.79(0.95)	4.79(0.91)	4.80(0.91)	8.09E-01
Triglycerides [mmol/l] *	1.22(0.68)	1.18(0.69)	1.12(0.58)	1.18(0.62)	2.94E-01
HDL [mmol/l]	1.4(0.33)	1.45(0.37)	1.39(0.35)	1.38(0.34)	7.75E-02
LDL [mmol/l]	3.07(0.79)	2.98(0.86)	3.05(0.83)	3.05(0.78)	6.33E-01
Systolic Blood Pressure [mmHg]	121.94(13.39)	122.76(14.36)	123.82(14.16)	123.92(14.72	3.29E-01
Diastolic Blood Pressure [mmHg] *	73.39(9.33)	73.51(10.06)	74.63(10.2)	74.81(12.3)	3.28E-01
CD8$^+$ T cells	0.03(0.04)	0.02(0.03)	0.01(0.02)	0.01(0.02)	**2.79E-09**
CD4$^+$ T cells	0.18(0.06)	0.17(0.05)	0.17(0.06)	0.16(0.05)	2.90E-02
Natural Killer Cells	0.22(0.06)	0.2(0.05)	0.19(0.05)	0.19(0.05)	**1.24E-09**
B cells	0.10(0.03)	0.10(0.03)	0.09(0.03)	0.09(0.03)	2.01E-01
Monocytes	0.07(0.03)	0.07(0.03)	0.07(0.03)	0.08(0.03)	**1.11E-06**
Granulocytes	0.48(0.09)	0.51(0.09)	0.52(0.09)	0.52(0.07)	**4.81E-07**
Categorical Phenotypes					
Sex [female/male]	166/96	170/92	163/99	145/117	4.71E-02
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	187/63/12	200/52/10	197/58/7	194/61/7	3.44E-01

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (158/158/158/158).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Continuation Table S8-31.

cg24704287 (*intergenic*)	Quartile 1 (n=263)	Quartile 2 (n=262)	Quartile 3 (n=262)	Quartile 4 (n=262)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	50.18(5.75)	50.03(5.64)	50.03(5.04)	50.00(5.2)	9.82E-01
BMI [kg/m^2]	27.15(5.01)	26.08(4.17)	26.51(4.50)	26.62(4.64)	6.46E-02
waist-circumference [cm]	90.11(13.87)	87.7(12.41)	88.7(13.51)	88.69(13.24)	2.24E-01
Fasting Glucose [mmol/l] *	5.46(1.18)	5.47(1.22)	5.40(1.01)	5.34(0.52)	6.31E-01
2-hours Glucose [mmol/l]	4.41(0.92)	4.44(1.08)	4.34(0.76)	4.30(0.43)	3.73E-01
C-reactive Protein [mg/l] *	2.29(3.06)	1.68(2.31)	1.95(2.95)	1.99(2.81)	4.54E-02
Fasting Insulin [µIU/ml] [a]*	5.93(3.98)	5.46(4.08)	6.09(8.17)	5.82(4.50)	5.76E-01
HOMA-IR [a]*	1.39(1.01)	1.44(1.89)	1.56(2.88)	1.37(1.07)	7.18E-01
HOMA-B [a]*	75.64(75.99)	59.48(30.13)	65.74(46.14)	71.95(76.42)	1.89E-01
Cholesterol [mmol/l]	4.80(0.95)	4.75(0.94)	4.80(0.91)	4.90(0.89)	3.07E-01
Triglycerides [mmol/l] *	1.22(0.72)	1.10(0.56)	1.15(0.54)	1.22(0.73)	1.18E-01
HDL [mmol/l]	1.37(0.34)	1.43(0.35)	1.40(0.35)	1.42(0.36)	2.08E-01
LDL [mmol/l]	3.01(0.83)	3.01(0.82)	3.03(0.77)	3.09(0.83)	6.60E-01
Systolic Blood Pressure [mmHg]	125.32(14.99)	121.5(14.39)	122.38(13.09)	123.19(13.90)	**1.51E-02**
Diastolic Blood Pressure [mmHg] *	75.46(10.65)	73.06(9.93)	73.72(9.93)	74.05(11.42)	**4.16E-02**
CD8$^+$ T cells	0.02(0.03)	0.02(0.03)	0.02(0.03)	0.01(0.02)	**6.62E-06**
CD4$^+$ T cells	0.14(0.05)	0.16(0.05)	0.18(0.04)	0.20(0.05)	**<0.001**
Natural Killer Cells	0.19(0.06)	0.20(0.05)	0.20(0.05)	0.21(0.05)	**2.06E-07**
B cells	0.08(0.03)	0.09(0.02)	0.10(0.03)	0.11(0.03)	**<0.001**
Monocytes	0.08(0.03)	0.07(0.03)	0.07(0.03)	0.07(0.03)	**1.44E-05**
Granulocytes	0.55(0.09)	0.52(0.08)	0.50(0.08)	0.47(0.07)	**<0.001**
Categorical Phenotypes					
Sex [female/male]	154/109	172/90	158/104	160/102	8.69E-01
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	189/63/11	196/57/9	201/49/12	193/66/3	2.89E-01

*Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (161/161/161/161).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Continuation Table S8-31.

***HDAC4* (cg00144180)**	Quartile 1 (n=263)	Quartile 2 (n=262)	Quartile 3 (n=262)	Quartile 4 (n=262)	
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	P
Age [years]	48.89(4.71)	49.94(5.24)	50.43(5.7)	50.96(5.75)	**1.11E-04**
BMI [kg/m^2]	25.36(3.93)	26.30(4.55)	26.64(4.56)	28.15(4.95)	**6.24E-11**
waist-circumference [cm]	83.05(10.91)	87.91(12.76)	89.44(12.49)	94.94(14.16)	**<0.001**
Fasting Glucose [mmol/l] *	5.19(0.47)	5.39(1.06)	5.43(0.98)	5.67(1.34)	**3.18E-08**
2-hours Glucose [mmol/l]	4.30(0.43)	4.37(0.88)	4.34(0.74)	4.49(1.10)	7.00E-02
C-reactive Protein [mg/l] *	1.59(2.23)	1.84(2.64)	2.01(2.62)	2.50(3.51)	**3.36E-05**
Fasting Insulin [µIU/ml] [a]*	5.25(3.99)	5.40(4.15)	6.66(8.26)	6.06(4.29)	**2.31E-02**
HOMA-IR [a]*	1.22(0.99)	1.31(1.24)	1.74(3.11)	1.50(1.36)	**2.52E-02**
HOMA-B [a]*	70.22(90.77)	63.79(52.5)	71.43(48.46)	67.4(36.22)	8.24E-02
Cholesterol [mmol/l]	4.7(0.88)	4.76(0.91)	4.89(0.94)	4.89(0.95)	4.55E-02
Triglycerides [mmol/l] *	0.98(0.56)	1.13(0.53)	1.23(0.71)	1.36(0.69)	**3.74E-14**
HDL [mmol/l]	1.53(0.37)	1.42(0.34)	1.37(0.34)	1.29(0.31)	**5.33E-14**
LDL [mmol/l]	2.96(0.77)	3.04(0.80)	3.08(0.83)	3.08(0.86)	2.53E-01
Systolic Blood Pressure [mmHg]	119.46(12.62)	121.76(14.86)	123.37(13.33)	127.81(14.46)	**9.33E-11**
Diastolic Blood Pressure [mmHg] *	72.95(10.90)	73.45(10.95)	73.40(8.52)	76.49(11.18)	**8.66E-05**
CD8$^+$ T cells	0.01(0.02)	0.02(0.03)	0.02(0.03)	0.02(0.03)	**2.03E-02**
CD4$^+$ T cells	0.16(0.05)	0.17(0.05)	0.17(0.05)	0.18(0.06)	**2.30E-03**
Natural Killer Cells	0.20(0.05)	0.19(0.05)	0.20(0.05)	0.21(0.05)	**5.13E-05**
B cells	0.10(0.03)	0.09(0.03)	0.09(0.03)	0.10(0.03)	**6.72E-03**
Monocytes	0.07(0.03)	0.07(0.03)	0.08(0.03)	0.08(0.03)	**7.63E-05**
Granulocytes	0.53(0.08)	0.52(0.09)	0.50(0.09)	0.48(0.08)	**1.43E-11**
Categorical Phenotypes					
Sex [female/male]	215/48	179/83	151/111	100/161	**<0.001**
Glucose tolerance [Normoglycaemic/Prediabetes/Diabetes] †	217/45/1	196/59/7	197/53/12	168/78/16	**3.50E-07**

* Variables log transformed before the analysis. [a] Variables only available in a subset of 645 females from ALSPAC, distribution between quartiles (162/161/161/161).
†Categories of glucose tolerance based on ADA criteria (normoglycaemic if FG<5.6mmol/l, Prediabetes if FG ≥5.6mmol/l and <7.0mmol/l, and Diabetes if FG≥7.0mmol/l).

Table S 8-32 Regions identified using the Comb-p method as differentially methylated in association with T2D. Results obtained using CpG site summary statistics from the sensitivity meta-analysis (i.e. excluding KORA). Effect size measured as %Meth: median of the absolute percentage change in methylation between T2D cases and controls, calculated for all the DMPs within a region.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Lowest P	P_{region}	Sidak
1	Chr1:145441552-145441553	*TXNIP†*	1	1	1.93077	↓	3.29E-08	8.75E-14	3.29E-08
1	Chr1:28906210-28906539	*SNHG12*	329	5	0.50858	↓	9.67E-05	8.97E-09	1.03E-05
1	Chr1:120255941-120255993	*PHGDH*	52	2	1.26024	↓	7.25E-04	5.01E-08	3.62E-04
1	Chr1:28573736-28573941	*Intergenic*	205	4	0.26184	↓	1.72E-03	2.42E-07	4.44E-04
1	Chr1:223317364-223317549	*TLR5*	185	4	0.636595	↑	7.02E-03	2.26E-06	4.59E-03
1	Chr1:201708500-201708789	*NAV1*	289	5	0.7124	↓	8.84E-03	4.04E-06	5.25E-03
1	Chr1:36023025-36023429	*NCDN*	404	8	0.102505	↓	6.14E-03	7.02E-06	6.52E-03
1	Chr1:44457124-44457407	*CCDC24*	283	6	0.10145	↓	6.48E-03	5.70E-06	7.54E-03
1	Chr1:92012615-92012737	*Intergenic*	122	3	1.46487	↓	8.84E-03	4.03E-06	1.24E-02
1	Chr1:181074635-181074791	*Intergenic*	156	3	0.1512	↓	1.22E-02	7.79E-06	1.86E-02
1	Chr1:228890801-228891037	*Intergenic*	236	5	1.01745	↑	1.22E-02	1.19E-05	1.88E-02
1	Chr1:19600454-19600913	*AKR7L*	459	7	0.86912	↑	1.91E-02	2.37E-05	1.92E-02
1	Chr1:61548526-61549011	*NFIA*	485	4	0.098435	↓	2.53E-02	4.04E-05	3.09E-02
1	Chr1:108023249-108023483	*NTNG1*	234	5	1.06771	↑	2.69E-02	3.06E-05	4.80E-02
2	Chr2:239046879-239047337	*KLHL30*	458	5	0.50651	↓	5.99E-04	2.64E-08	2.17E-05
2	Chr2:65593761-65593934	*SPRED2*	173	3	0.93782	↓	5.62E-03	1.48E-06	3.21E-03
2	Chr2:11123476-11123617	*Intergenic*	141	3	0.88629	↑	5.76E-03	1.56E-06	4.16E-03
2	Chr2:120124292-120124678	*C2orf76*	386	10	0.070005	↑	6.56E-03	4.59E-06	4.47E-03
2	Chr2:32390673-32390938	*SLC30A6*	265	5	0.09833	↓	9.60E-03	4.64E-06	6.57E-03
2	Chr2:240294246-240294363	*HDAC4*	117	2	0.7153	↑	7.02E-03	2.11E-06	6.75E-03
2	Chr2:74669349-74669517	*RTKN*	168	5	0.79635	↓	6.84E-03	3.79E-06	8.45E-03
2	Chr2:231692812-231693071	*Intergenic*	259	3	0.75571	↓	1.28E-02	8.71E-06	1.26E-02
2	Chr2:68592345-68592395	*PLEK*	50	4	0.562945	↑	8.84E-03	4.06E-06	3.01E-02
2	Chr2:233284402-233284662	*Intergenic*	260	2	1.11487	↓	1.10E-02	3.53E-05	4.98E-02
3	Chr3:4534791-4535155	*ITPR1*	364	9	0.09694	↓	1.52E-02	1.17E-05	1.21E-02

Continuation Table S 8-32.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Lowest P	P_{region}	Sidak
4	Chr4:57333365-57333860	*SRP72*	495	6	0.112135	↓	2.32E-02	2.73E-05	2.06E-02
5	Chr5:143978290-143978421	*Intergenic*	131	4	1.223095	↑	3.11E-03	5.36E-07	1.54E-03
5	Chr5:139927110-139927472	*YIF1B*	362	10	0.13607	↓	3.35E-03	1.53E-06	1.59E-03
5	Chr5:139488181-139488624	*Intergenic*	443	3	0.59402	↓	6.48E-03	2.29E-06	1.94E-03
5	Chr5:136340060-136340208	*SPOCK1*	148	2	0.58565	↓	1.36E-02	9.33E-06	2.34E-02
6	Chr6:72130742-72131021	*C6orf155*	279	4	1.01404	↑	3.11E-03	5.24E-07	7.06E-04
6	Chr6:13574034-13574574	*SIRT5*	540	6	0.27097	↑	1.09E-02	1.41E-05	9.79E-03
6	Chr6:5261091-5261561	*LYRM4*	470	10	0.09588	↓	8.84E-03	2.08E-05	1.65E-02
6	Chr6:149806635-149806733	*ZC3H12D*	98	2	0.925515	↓	1.20E-02	5.13E-06	1.95E-02
6	Chr6:125283726-125283970	*STL*	244	3	0.41345	↓	2.23E-02	2.00E-05	3.04E-02
7	Chr7:129007902-129008408	*AHCYL2*	506	4	0.37632	↓	1.25E-02	9.32E-06	6.91E-03
7	Chr7:1961785-1961870	*MAD1L1*	85	2	0.84275	↑	8.77E-03	3.40E-06	1.49E-02
8	Chr8:145018010-145018301	*PLEC1*	291	5	0.11378	↓	1.09E-03	1.13E-07	1.46E-04
8	Chr8:48675647-48676055	*Intergenic*	408	6	1.05857	↑	1.55E-02	2.94E-05	2.68E-02
8	Chr8:41583321-41583524	*ANK1*	203	3	1.24942	↑	1.52E-02	2.74E-05	4.95E-02
10	Chr10:6214016-6214080	*PFKFB3*	64	3	0.92262	↓	4.80E-04	2.68E-08	1.58E-04
10	Chr10:74057705-74058093	*Intergenic*	388	5	0.34896	↓	7.02E-03	6.37E-06	6.16E-03
10	Chr10:121356513-121356866	*TIAL1*	353	9	0.10793	↓	1.68E-03	8.38E-06	8.90E-03
10	Chr10:123734658-123734890	*NSMCE4A*	232	4	0.05742	↓	2.62E-02	2.75E-05	4.36E-02
11	Chr11:68607622-68608226	*CPT1A*	604	4	0.48859	↓	3.56E-05	1.79E-12	1.11E-09
11	Chr11:1778524-1778628	*HCCA2*	104	3	0.7725	↑	3.41E-03	7.62E-07	2.75E-03
11	Chr11:1769289-1769523	*HCCA2*	234	7	0.9782	↑	8.69E-03	3.27E-06	5.25E-03
11	Chr11:1029029-1029337	*MUC6*	308	5	0.19199	↓	7.31E-03	8.50E-06	1.03E-02
11	Chr11:124294778-124295016	*OR8B4*	238	2	0.34566	↓	2.53E-02	2.59E-05	4.02E-02
11	Chr11:1036471-1036866	*MUC6*	395	8	0.69628	↑	1.91E-02	4.35E-05	4.06E-02
12	Chr12:6642229-6642355	*GAPDH*	126	2	0.496375	↓	3.12E-03	5.64E-07	1.68E-03

Continuation Table S 8-32.

Chr	DMR	Nearest gene	Size (bp)	CpG count	% Meth	Direction	Lowest P	P_{region}	Sidak
12	Chr12:6881997-6882084	*LAG3*	87	2	0.633275	↓	4.85E-03	1.15E-06	4.95E-03
12	Chr12:49463725-49464042	*RHEBL1*	317	9	0.08913	↓	8.73E-03	8.75E-06	1.03E-02
12	Chr12:14926744-14926987	*H2AFJ*	243	3	1.33454	↑	1.26E-02	3.02E-05	4.57E-02
16	Chr16:50321678-50322157	*ADCY7*	479	5	0.67212	↑	3.56E-05	6.38E-10	5.01E-07
16	Chr16:75150456-75150881	*LDHD*	425	7	1.04557	↓	5.99E-04	5.91E-08	5.24E-05
16	Chr16:3114948-3115287	*IL32*	339	6	0.478725	↓	7.58E-04	7.20E-08	7.99E-05
16	Chr16:87734816-87735078	*LOC100129637*	262	3	0.59359	↑	3.11E-03	4.71E-07	6.76E-04
16	Chr16:89044523-89044883	*CBFA2T3*	360	5	0.38832	↓	3.35E-03	9.14E-07	9.55E-04
16	Chr16:2203008-2203177	*RAB26*	169	2	0.53523	↓	1.20E-02	7.32E-06	1.62E-02
17	Chr17:27052676-27052829	*TLCD1*	153	2	0.958115	↓	7.45E-03	2.69E-06	6.60E-03
17	Chr17:7832680-7833164	*KCNAB3*	484	8	0.67294	↓	7.35E-03	1.42E-05	1.09E-02
17	Chr17:26662301-26662585	*TNFAIP1*	284	7	0.07301	↓	2.88E-02	2.88E-05	3.75E-02
19	Chr19:47287778-47288264	*SLC1A5*	486	6	0.52254	↓	3.87E-05	1.88E-09	1.46E-06
19	Chr19:49993865-49994150	*SNORD34*	285	4	0.52651	↓	7.64E-04	6.91E-08	9.12E-05
19	Chr19:38806746-38806875	*EIF4EBP3*	129	5	0.08273	↓	1.69E-03	5.45E-07	1.59E-03
19	Chr19:58790125-58790440	*ZNF8*	315	9	0.11752	↓	8.84E-03	5.93E-06	7.05E-03
19	Chr19:55549590-55549843	*GP6*	253	6	1.057215	↓	7.99E-03	5.45E-06	8.07E-03
19	Chr19:13951481-13951482	*Intergenic†*	1	1	1.10077	↓	4.80E-04	2.34E-08	8.75E-03
19	Chr19:39389915-39390199	*SIRT2*	284	2	0.10437	↓	1.06E-02	1.71E-05	2.25E-02
20	Chr20:32700182-32700555	*EIF2S2*	373	8	0.13329	↓	1.80E-02	1.84E-05	1.84E-02
20	Chr20:47897124-47897452	*C20orf199*	328	4	0.51384	↓	1.78E-02	3.39E-05	3.82E-02
21	Chr21:35320596-35320668	*Intergenic*	72	2	1.424175	↓	3.56E-05	8.16E-10	4.26E-06
21	Chr21:43656587-43656588	*ABCG1†*	1	1	1.28229	↑	4.41E-06	2.34E-11	8.81E-06
21	Chr21:35831871-35832165	*KCNE1*	294	8	1.224025	↓	7.35E-03	2.56E-06	3.27E-03
22	Chr22:22987059-22987127	*POM121L1P*	68	5	0.96734	↑	1.50E-03	1.83E-07	1.01E-03
22	Chr22:32598479-32598717	*RFPL2*	238	3	0.61468	↓	2.44E-02	2.33E-05	3.61E-02

†Gene regions in overlap between the meta-analysis and the DMR analysis. **DMR less informative as only 1 CpG site was identified within the region. CpG count: number of DMPs detected within a DMR. Direction: relative effect observed across CpG sites included in a DMR. Pregion: P value of the region calculated by *comb-p* using the Stouffer-Liptak correction. Sidak: significance of the DMR after multiple-testing correction. DMRs were considered significant at Sidak < 0.05

Table S 8-33 Top GO terms identified in enrichment for CpG sites included within 33 DMRs associated with T2D (main DMR analysis), and for sites included within 77 DMRs detected in a secondary analysis. Total genes: total number of genes within a term. Differentially methylated: proportion of genes identified with differential methylation from the total number of genes in a term. P: p-value for enrichment (unadjusted). None of the terms identified by GO surpassed multiple testing correction at FDR<0.05. In bold are GO terms detected in common across enrichment analyses.

	GO term	Total Genes	Differentially methylated	P
Main DMR analysis	Mitochondrion	1403	10	3.62E-04
	Mitochondrial part	916	8	3.68E-04
	Sarcoplasm	71	3	8.06E-04
	Cellular response to caloric restriction	1	1	2.07E-03
	NAD-dependent histone deacetylase activity (H4-K16 specific)	1	1	2.07E-03
	Negative regulation of oligodendrocyte progenitor proliferation	1	1	2.07E-03
	Tubulin deacetylase activity	1	1	2.07E-03
	Cellular response to tumour cell	1	1	2.21E-03
	D-lactate dehydrogenase activity	1	1	2.40E-03
	Lactate catabolic process	1	1	2.40E-03
	Regulation of thyroid-stimulating hormone secretion	1	1	3.68E-03
	Thyroid-stimulating hormone secretion	1	1	3.68E-03
	Positive regulation of oocyte maturation	2	1	4.09E-03
	Inositol 1,4,5-trisphosphate receptor activity involved in regulation of postsynaptic cytosolic calcium levels	1	1	4.14E-03
	Positive regulation of proteasomal ubiquitin-dependent protein catabolic process involved in cellular response to hypoxia	2	1	4.16E-03
	Palmitoleoyl-transferase activity	1	1	4.31E-03
	Histone deacetylase activity (H4-K16 specific)	2	1	4.37E-03
	3 iron, 4 sulphur cluster binding	2	1	4.41E-03
	Aconitate hydratase activity	2	1	4.58E-03
	Aryldialkylphosphatase activity	2	1	5.08E-03

Continuation Table S 8-33.

	GO term	Total Genes	Differentially Methylated	P
Secondary DMR analysis	NAD-dependent protein deacetylase activity	15	3	4.22E-05
	NAD binding	50	4	6.05E-05
	GAIT complex	4	2	1.09E-04
	Glycolytic process	67	4	3.02E-04
	ATP generation from ADP	68	4	3.28E-04
	Pyruvate biosynthetic process	71	4	3.49E-04
	Peptidyl-lysine deacetylation	6	2	3.49E-04
	Nucleolus	1155	13	4.11E-04
	ADP metabolic process	75	4	4.91E-04
	Main axon	65	4	5.23E-04
	Nucleoside diphosphate phosphorylation	83	4	5.53E-04
	Nucleotide phosphorylation	85	4	6.14E-04
	Negative regulation of cellular carbohydrate metabolic process	36	4	6.75E-04
	Regulation of small molecule metabolic process	326	3	7.31E-04
	Purine nucleoside diphosphate metabolic process	83	7	7.35E-04
	Purine ribonucleoside diphosphate metabolic process	83	4	7.36E-04
	Ribonucleoside diphosphate metabolic process	85	4	7.36E-04
	Regulation of glycolytic process	38	4	8.16E-04
	Histone H4 deacetylation	10	3	9.07E-04

Table S 8-34 Top KEGG pathways identified in enrichment for CpG sites included in 33 DMRs associated with T2D (main analysis) and for sites included in 77 DMRs identified in a secondary analysis. Total genes: total number of genes pertaining to a pathway. Differentially methylated: proportion of genes identified with differential methylation from the total number of genes in a pathway. P: p-value for enrichment (unadjusted). None of the pathways identified in KEGG surpassed multiple testing correction at FDR<0.05. Highlighted in bold are pathways identified in common between the main and the secondary enrichment analysis.

	Pathway (KEGG)	Total Genes	Differentially methylated	P
Main DMR analysis	**Thyroid hormone synthesis**	65	3	4.63E-04
	NOD-like receptor signalling pathway	149	3	1.74E-03
	Platelet activation	118	3	2.44E-03
	Adipocytokine signalling pathway	61	2	7.78E-03
	B cell receptor signalling pathway	64	2	8.79E-03
	Cortisol synthesis and secretion	57	2	9.38E-03
	Pancreatic secretion	91	2	1.15E-02
	Salivary secretion	83	2	1.20E-02
	Gastric acid secretion	72	2	1.37E-02
	GnRH signalling pathway	89	2	1.58E-02
	Glucagon signalling pathway	91	2	1.69E-02
	Gap junction	82	2	1.71E-02
	Phosphatidylinositol signalling system	91	2	2.05E-02
	Inflammatory mediator regulation of TRP channels	97	2	2.06E-02
	Aldosterone synthesis and secretion	90	2	2.24E-02
	Oocyte meiosis	113	2	2.35E-02
	Circadian entrainment	92	2	2.53E-02
	Parathyroid hormone synthesis, secretion and action	98	2	2.71E-02
	Vascular smooth muscle contraction	126	2	2.82E-02
	Oestrogen signalling pathway	127	2	3.00E-02
Secondary DMR analysis	Apelin signalling pathway	130	3	2.60E-02
	Biosynthesis of amino acids	60	2	2.66E-02
	PPAR signalling pathway	67	2	2.73E-02
	Carbon metabolism	97	2	6.35E-02
	Progesterone-mediated oocyte maturation	88	2	6.39E-02
	Protein export	21	1	7.69E-02
	HIF-1 signalling pathway	92	2	7.84E-02

Figure S 8-18 Overlap between CpG sites detected within T2D-associated DMRs, and regions enriched for H3K4m1 histone marks based on epigenomic data from the consolidated ROADMAP project using different cell-types tissues. Enrichment analysis conducted in eFORGE (v1.2). Tissues with significant enrichment for H3K4m1 marks among target CpG sites, are represented by red dots (q-value<0.01), and pink dots for tissues with borderline evidence of enrichment (q<0.05). Image provided by eFORGE[149].

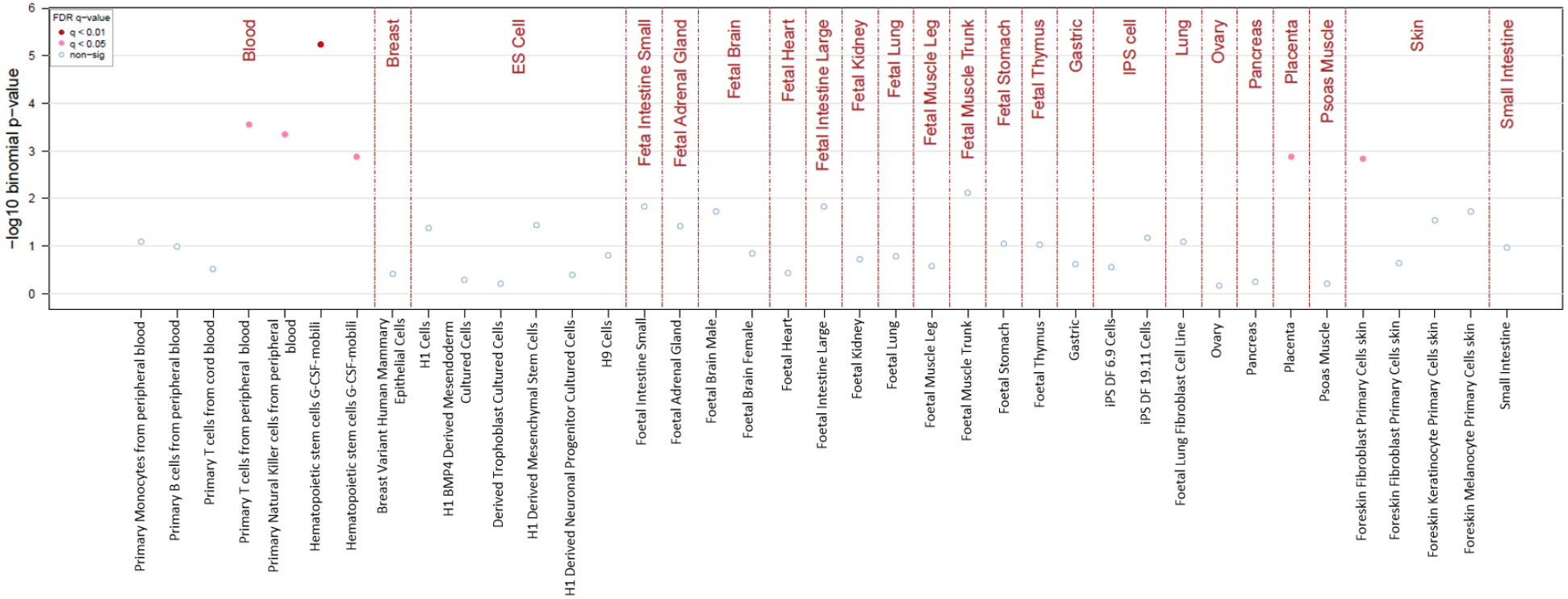


Figure S 8-19 Overlap between CpG sites detected within T2D-associated DMRs in a secondary analysis, and regions enriched for H3K4m1 and H3K4m3 histone marks based on epigenomic data from the consolidated ROADMAP project using different cell-types tissues. Enrichment analysis conducted in eFORGE (v1.2). Tissues with significant enrichment for H3K4m1 and H3K4m3 histone among T2D DMRs, are represented by red dots (q-value<0.01), and pink dots for tissues with borderline evidence of enrichment (q<0.05) for DMRs. Image provided by eFORGE[149].

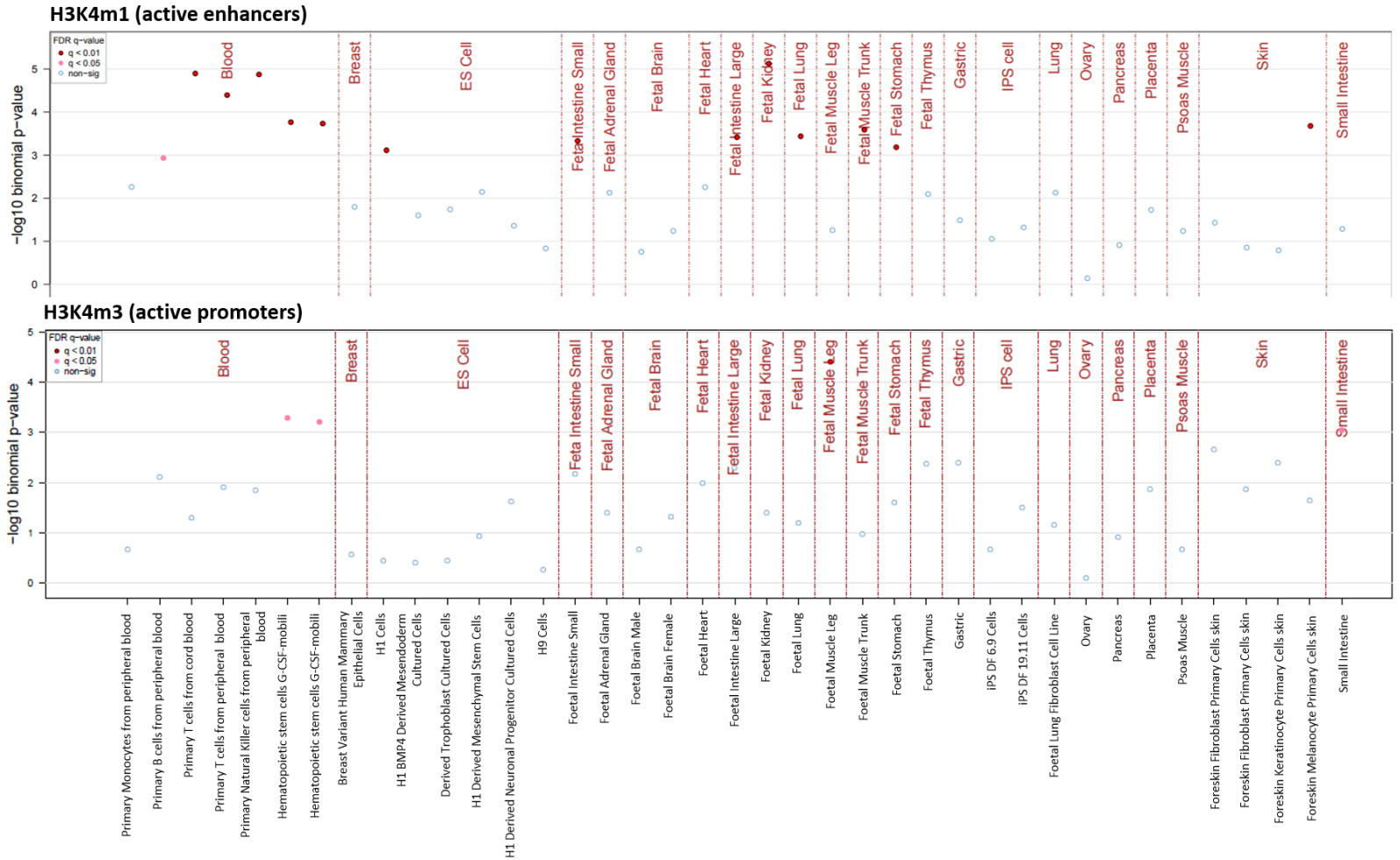
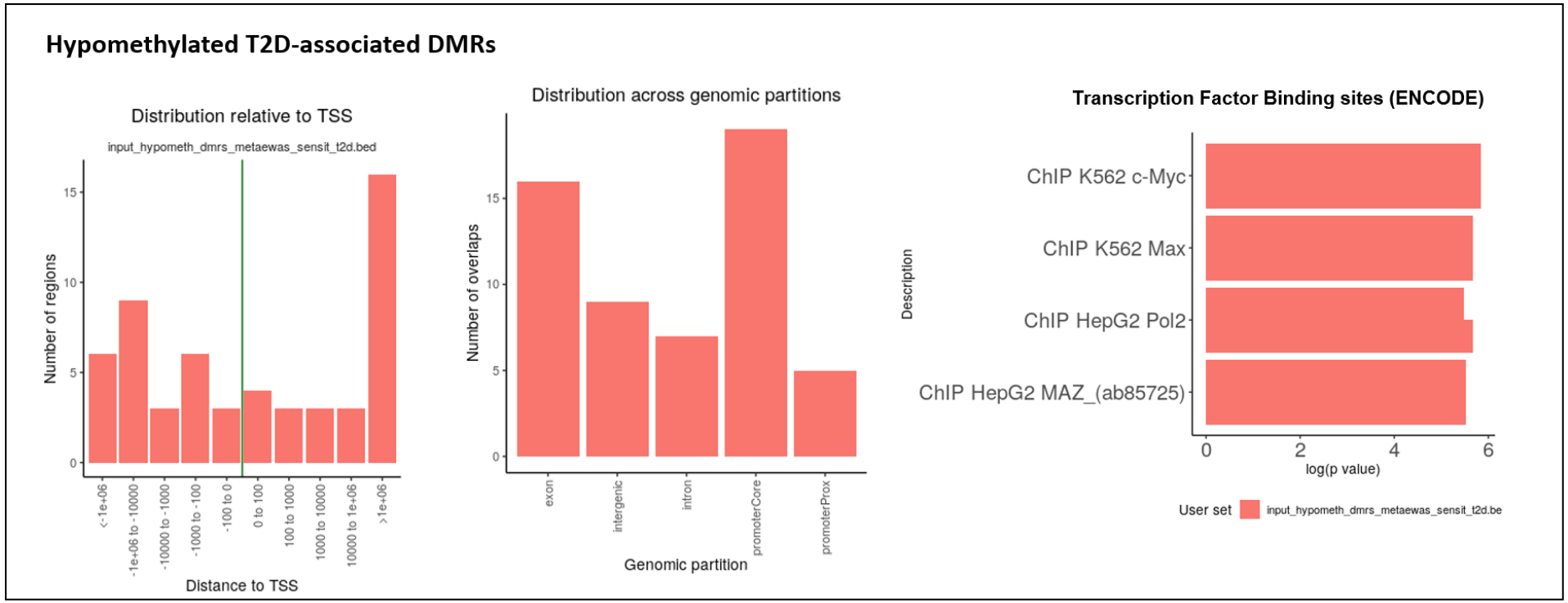


Figure S8-20 Genomic annotation of T2D-associated DMRs based on the analysis in LOLA web (http://lolaweb.databio.org). DMRs were annotated according to their distance from transcription start sites (TSS), distribution across genomic positions, and overlap with transcription factor binding sites based on ENCODE data. Results presented for hypomethylated and hypermethylated T2D-associated DMRs.



Continuation Figure S8-20.

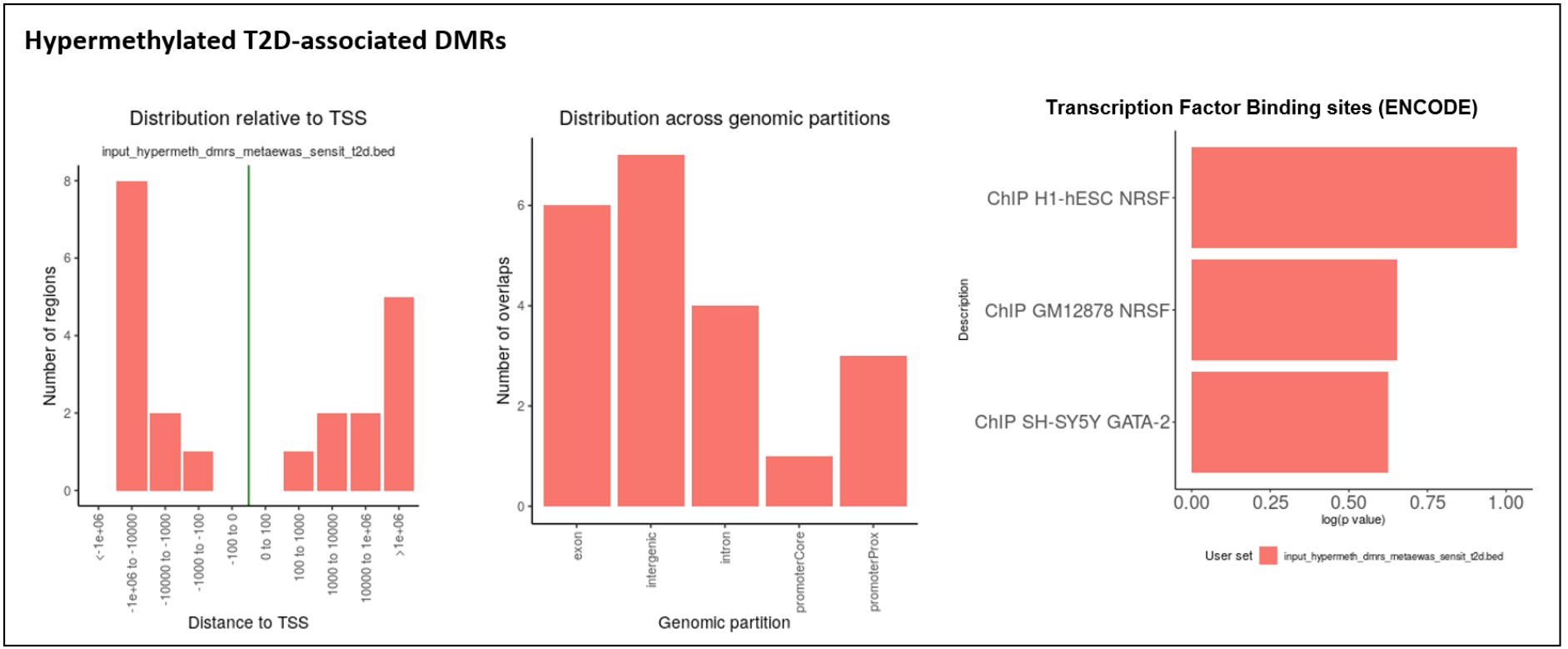


Table S8-35 Results of the regression between 75 independent SNPs and T2D using a subsample of middle-age adults in ALSPAC (n=1,252, cases=36 and controls=804). EA is the effect allele, EAF is the effect allele frequency calculated for cases and controls, and the effect estimate is reported for the unadjusted regression (OR_{unadj}), and for the regression adjusted for 10 genetic PCs and a batch-effect variable (batch1=male and batch2=female) ($OR_{adj.}$). Regressions were regarded significant at P < $6.67x10^{-4}$ after correction for multiple testing (0.05/75 tests). Highlighted in bold is the SNP with the strongest association with T2D identified in ALSPAC.

SNP	Chr	EA	FEA cases	FEA controls	OR (unadj)	P	OR (adj)	P
rs2820446	1	C	0.65	0.72	0.75	0.25	0.66	0.13
rs340874	1	C	0.61	0.57	1.19	0.49	1.14	0.63
rs17106184	1	G	0.89	0.91	0.81	0.58	NA	NA
rs35720761	2	T	0.08	0.13	0.62	0.27	0.51	0.17
rs1260326	2	T	0.36	0.41	0.81	0.41	0.85	0.55
rs243019	2	C	0.49	0.45	1.15	0.56	1.11	0.67
rs75297654	2	C	0.86	0.86	1.04	0.91	1.12	0.76
rs10190052	2	C	0.83	0.82	1.13	0.70	1.09	0.80
rs2972156	2	G	0.68	0.63	1.25	0.38	1.04	0.88
rs13389219	2	C	0.60	0.58	1.09	0.73	1.03	0.91
rs1496653	**3**	**A**	**0.68**	**0.78**	**0.59**	**0.04**	**0.54**	**0.03**
rs1470579	3	C	0.36	0.31	1.24	0.39	1.34	0.27
rs16861329	3	C	0.92	0.86	1.76	0.18	1.56	0.31
rs1801282	3	C	0.88	0.89	0.85	0.66	0.76	0.47
rs6808574	3	C	0.63	0.61	1.07	0.80	1.17	0.58
rs11717195	3	T	0.73	0.73	1.01	0.98	1.11	0.74
rs17676309	3	C	0.59	0.59	0.96	0.88	0.94	0.83
rs6813195	4	C	0.78	0.71	1.44	0.21	1.23	0.48
rs734312	4	A	0.56	0.52	1.18	0.52	1.14	0.64
rs35658696	5	A	0.91	0.96	0.49	0.10	0.45	0.09
rs6878122	5	G	0.36	0.31	1.30	0.32	1.55	0.12
rs702634	5	A	0.76	0.70	1.37	0.26	1.35	0.32
rs319598	5	C	0.56	0.58	0.91	0.71	0.85	0.56
rs459193	5	G	0.75	0.76	0.97	0.92	1.04	0.89
rs9505118	6	A	0.66	0.59	1.34	0.25	1.33	0.29
rs1535500	6	T	0.40	0.49	0.71	0.17	0.78	0.34
rs9472138	6	T	0.25	0.28	0.86	0.58	0.83	0.54
rs7756992	6	G	0.29	0.27	1.12	0.67	1.15	0.63
rs4273712	6	G	0.29	0.29	1.03	0.91	1.09	0.76
rs6937795	6	A	0.53	0.53	1.01	0.97	0.97	0.91
rs10276674	7	C	0.26	0.18	1.66	0.07	1.73	0.08
rs2284219	7	G	0.58	0.66	0.73	0.20	0.72	0.20
rs7795991	7	G	0.50	0.53	0.89	0.62	0.80	0.38
rs849135	7	G	0.43	0.50	0.76	0.25	0.83	0.48
rs1974620	7	T	0.56	0.54	1.07	0.79	1.06	0.84
rs878521	7	A	0.24	0.25	0.91	0.75	0.94	0.85
rs7845219	8	T	0.58	0.48	1.50	0.09	1.64	0.06
rs516946	8	C	0.74	0.75	0.94	0.83	0.81	0.47
rs3802177	8	G	0.73	0.68	1.25	0.42	1.23	0.49
rs944801	9	C	0.44	0.57	0.60	0.03	0.65	0.10
rs10757283	9	T	0.50	0.42	1.38	0.18	1.43	0.16
rs2796441	9	G	0.61	0.57	1.18	0.51	1.11	0.69
rs7041847	9	A	0.50	0.47	1.14	0.58	1.08	0.77
rs11257658	10	A	0.17	0.23	0.66	0.19	0.68	0.26
rs10788575	10	A	0.15	0.16	0.92	0.79	0.72	0.40

Continuation Table S8-35

SNP	Chr	EA	FEA cases	FEA controls	OR (unadj)	P	OR (adj)	P
rs7903146	10	T	0.36	0.30	1.34	0.24	1.23	0.44
rs12571751	10	A	0.51	0.55	0.86	0.54	0.90	0.68
rs2812533	10	C	0.87	0.86	1.07	0.85	0.90	0.78
rs10510110	10	C	0.49	0.48	1.03	0.90	1.02	0.94
rs5015480	10	C	0.58	0.58	1.01	0.97	1.01	0.96
rs1387153	11	T	0.36	0.27	1.50	0.11	1.47	0.15
rs1552224	11	A	0.89	0.82	1.72	0.15	1.49	0.31
rs2283220	11	A	0.65	0.71	0.78	0.35	0.76	0.36
rs231361	11	A	0.30	0.25	1.30	0.35	1.26	0.45
rs757110	11	A	0.61	0.63	0.92	0.74	0.99	0.98
rs7955901	12	C	0.51	0.42	1.45	0.12	1.61	0.08
rs7961581	12	C	0.21	0.26	0.73	0.30	0.66	0.20
rs1169288	12	C	0.34	0.32	1.09	0.73	1.23	0.46
rs2612035	12	G	0.11	0.10	1.12	0.77	1.30	0.49
rs10842994	12	C	0.81	0.80	1.01	0.97	0.86	0.64
rs4275659	12	C	0.68	0.71	0.88	0.63	0.91	0.73
rs12427353	12	G	0.79	0.81	0.88	0.67	1.02	0.95
rs1359790	13	G	0.81	0.72	1.60	0.12	1.54	0.18
rs2028299	15	C	0.36	0.29	1.35	0.23	1.60	0.09
rs7163757	15	C	0.65	0.57	1.44	0.14	1.45	0.17
rs12899811	15	G	0.31	0.28	1.11	0.68	1.16	0.60
rs7178572	15	G	0.71	0.71	0.99	0.97	1.14	0.65
rs7202877	16	T	0.97	0.89	4.11	0.03	3.17	0.11
rs9936385	16	C	0.46	0.37	1.45	0.13	1.52	0.11
rs4430796	17	G	0.48	0.47	1.07	0.79	1.08	0.77
rs12970134	18	A	0.26	0.26	1.05	0.87	0.87	0.64
rs2238689	19	C	0.46	0.41	1.23	0.41	1.38	0.23
rs58542926	19	T	0.08	0.08	1.07	0.88	0.99	0.98
rs4812829	20	A	0.18	0.14	1.34	0.35	1.28	0.47
rs41278853	22	A	0.93	0.89	1.61	0.31	1.43	0.47

Figure S8-21 Manhattan plot illustrating the association between two SNPs for T2D and HDL and PC9 in a subsample of middle-age adults in ALSPAC (n=1,252). Strong evidence of association was detected between SNP rs13389219 and HDL (GRB14, p=1.3x10^{-4}), with an increase in 0.95 mmol/L in the mean levels of HDL for each extra allele C. An increase in 0.003 in the genetic variation explained by PC9 was detected per extra allele C in the SNP rs516946 (ANK1, P=2.9x10^{-4}). Association were regarded significant at P<6.67x10^{-4} (0.05/75 tests).

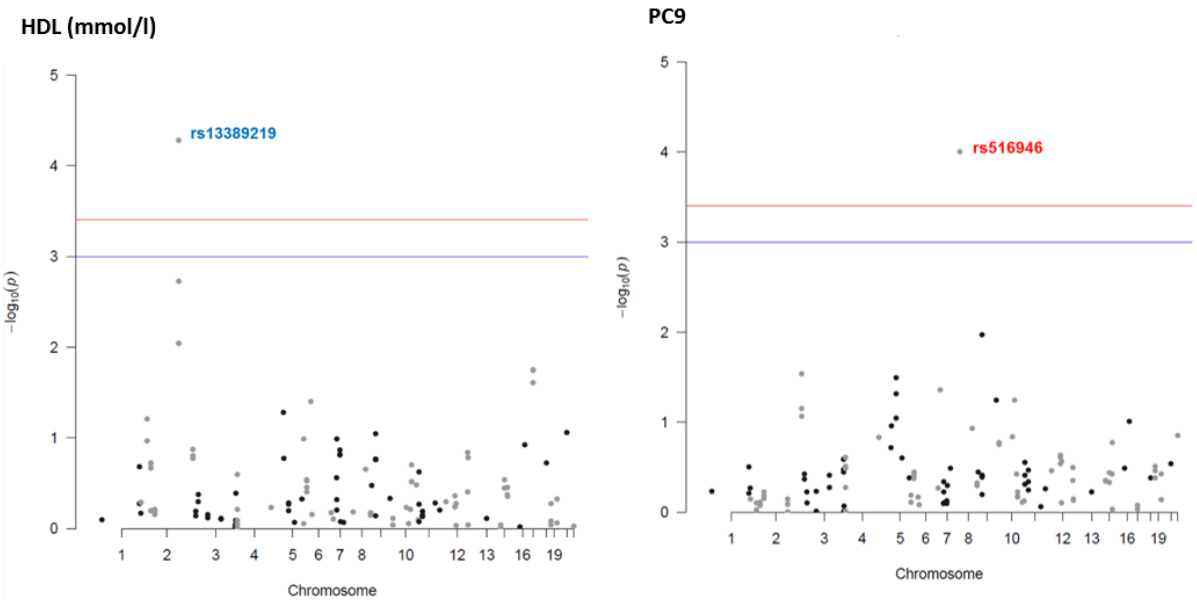


Table S8-36 Characteristics of 75 SNPs included in a Polygenic risk score for T2D. Detail for each SNP is provided according to the imputation performed in the subsample of adults in ALSPAC (n=1,252), and according to data reported in the DIAGRAM consortium. Minor allele frequency (MAF) and Hardy-Weinberg equilibrium P-value (P_{HW}) were calculated based on the subgroup of controls (n=804). Effect allele (EA), other allele (OA), effect estimate (OR), standard error (SE), and P-significance (P), were extracted from DIAGRAM studies. A1 and A2 are the minor and major allele, respectively, as reported from the genotyping in ALSPAC. Effect allele frequency (EAF) corresponds to the frequency of the risk allele for T2D based on data in ALSPAC.

SNP	Gene	Chr	ALSPAC (N=1,252)						DIAGRAM 2012-2016					
			A1	A2	$N_{controls}$	MAF	P_{HW}	EAF	EA	OA	OR	SE	P	Study
rs7903146	*TCF7L2*	10	T	C	804	0.30	1.00	0.30	T	C	1.39	0.02	1.20E-139	Morris_etal_2012
rs10811660	*NA*	9	A	G	790	0.18	0.04	0.82	G	A	1.27	0.02	1.10E-61	Gaulton_etal_2015
rs35261542	*CDKAL1*	6	A	C	795	0.26	1.00	0.26	A	C	1.17	0.01	1.50E-50	Gaulton_etal_2015
rs35510946	*IGF2BP2*	3	A	G	795	0.31	0.68	0.31	A	G	1.14	0.01	1.10E-39	Gaulton_etal_2015
rs11187140	*NA*	10	A	G	783	0.37	0.65	0.63	G	A	1.12	0.01	1.50E-31	Gaulton_etal_2015
rs13266634	*SLC30A8*	8	T	C	804	0.32	0.81	0.68	C	T	1.12	0.01	5.00E-28	Gaulton_etal_2015
rs1513272	*JAZF1*	7	C	T	796	0.50	0.83	0.50	C	T	1.10	0.01	7.80E-25	Gaulton_etal_2015
rs9936385	*FTO*	16	C	T	803	0.37	0.50	0.37	C	T	1.13	0.02	2.60E-23	Morris_etal_2012
rs11712037	*PPARG*	3	G	C	800	0.11	1.00	0.89	C	G	1.14	0.02	1.70E-20	Gaulton_etal_2015
rs2972156	*NA*	2	C	G	790	0.37	0.54	0.63	G	C	1.09	0.01	4.20E-20	Gaulton_etal_2015
rs10937721	*WFS1*	4	G	C	739	0.40	0.59	0.60	C	G	1.09	0.01	4.30E-18	Gaulton_etal_2015
rs4430796	*HNF1B*	17	G	A	752	0.47	0.77	0.47	G	A	1.09	0.01	7.80E-18	Gaulton_etal_2015
rs2238689	*GIPR*	19	C	T	746	0.41	0.88	0.41	C	T	1.08	0.01	8.30E-16	Gaulton_etal_2015
rs11257658	*NA*	10	A	G	787	0.23	0.09	0.23	A	G	1.09	0.01	1.20E-15	Gaulton_etal_2015
rs72999033	*HAPLN4*	19	T	C	786	0.07	0.14	0.07	T	C	1.16	0.02	1.80E-15	Gaulton_etal_2015
rs1169288	*HNF1A*	12	C	A	772	0.32	0.74	0.32	C	A	1.09	0.01	8.10E-15	Gaulton_etal_2015
rs7607980	*COBLL1*	2	C	T	803	0.14	0.20	0.86	T	C	1.15	0.02	8.30E-15	Fuchsberger_etal_2016
rs6813195	*NA*	4	T	C	803	0.29	0.49	0.71	C	T	1.08	0.01	4.10E-14	Mahajan_etal_2014
rs77981966	*THADA*	2	T	C	782	0.07	1.00	0.93	C	T	1.16	0.02	4.10E-14	Gaulton_etal_2015
rs11717195	*ADCY5*	3	C	T	802	0.27	0.72	0.73	T	C	1.11	0.02	6.50E-14	Morris_etal_2012
rs340874	*PROX1*	1	T	C	804	0.43	0.35	0.57	C	T	1.07	0.01	5.10E-13	Gaulton_etal_2015
rs7732130	*ZBED3-AS1*	5	G	A	784	0.31	0.61	0.31	G	A	1.08	0.01	2.40E-12	Gaulton_etal_2015
rs17676309	*ADAMTS9-AS2*	3	T	C	765	0.41	0.76	0.59	C	T	1.07	0.01	2.80E-12	Gaulton_etal_2015
rs1387153	*NA*	11	T	C	804	0.27	0.79	0.27	T	C	1.09	0.02	1.60E-11	Morris_etal_2012
rs2583941	*RPSAP52*	12	A	G	803	0.10	0.23	0.10	A	G	1.11	0.02	1.60E-11	Gaulton_etal_2015
rs10276674	*DGKB*	7	C	T	781	0.18	0.39	0.18	C	T	1.08	0.01	2.80E-11	Gaulton_etal_2015
rs3803563	*PRC1*	15	A	C	801	0.16	0.14	0.16	A	C	1.08	0.01	5.60E-11	Gaulton_etal_2015
rs12571751	*ZMIZ1*	10	G	A	794	0.45	0.89	0.55	A	G	1.08	0.01	1.00E-10	Morris_etal_2012
rs878521	*YKT6*	7	A	G	777	0.25	0.57	0.25	A	G	1.07	0.01	1.30E-10	Gaulton_etal_2015
rs1552224	*ARAP1*	11	C	A	804	0.18	0.72	0.82	A	C	1.11	0.02	1.80E-10	Morris_etal_2012
rs516946	*ANK1*	8	T	C	795	0.25	0.45	0.75	C	T	1.09	0.02	2.50E-10	Morris_etal_2012

Continuation Table S8-36

SNP	Gene	Chr	ALSPAC (N=1,252)						DIAGRAM 2012-2016					
			A1	A2	$N_{controls}$	MAF*	P_{HW}	EAF	EA	OA	OR	SE	P	Study
rs35720761	*THADA*	2	T	C	798	0.13	0.75	0.13	T	C	1.12	0.02	3.30E-10	Fuchsberger_etal_2016
rs780094	*GCKR*	2	T	C	804	0.40	0.61	0.60	C	T	1.06	0.01	3.40E-10	Gaulton_etal_2015
rs243020	*NA*	2	G	A	800	0.45	0.35	0.45	G	A	1.06	0.01	5.50E-10	Gaulton_etal_2015
rs35658696	*PAM*	5	G	A	773	0.04	1.00	0.96	A	G	1.17	0.03	5.70E-10	Fuchsberger_etal_2016
rs10842994	*NA*	12	T	C	800	0.20	0.58	0.80	C	T	1.10	0.02	6.10E-10	Morris_etal_2012
rs5215	*KCNJ11*	11	C	T	804	0.37	0.94	0.37	C	T	1.07	-0.01	8.50E-10	Morris_etal_2012
rs1974620	*NA*	7	C	T	804	0.46	0.62	0.54	T	C	1.06	0.01	1.00E-09	Gaulton_etal_2015
rs9502570	*NA*	6	T	C	759	0.26	0.71	0.74	C	T	1.06	0.01	1.00E-09	Mahajan_etal_2014
rs231361	*KCNQ1*	11	A	G	745	0.25	0.32	0.25	A	G	1.09	0.02	1.20E-09	Morris_etal_2012
rs944801	*CDKN2B-AS1*	9	G	C	790	0.43	0.88	0.57	C	G	1.08	0.01	2.40E-09	Morris_etal_2012
rs1496653	*UBE2E2*	3	G	A	804	0.22	0.21	0.78	A	G	1.09	0.02	3.60E-09	Morris_etal_2012
rs17106184	*FAF1*	1	A	G	799	0.09	0.02	0.91	G	A	1.10	0.02	4.10E-09	Mahajan_etal_2014
rs7177055	*NA*	15	G	A	804	0.29	0.67	0.71	A	G	1.08	0.01	4.60E-09	Morris_etal_2012
rs7161785	*NA*	15	C	G	801	0.43	0.77	0.57	G	C	1.06	0.01	4.90E-09	Gaulton_etal_2015
rs2796441	*BC036431*	9	A	G	763	0.43	0.88	0.57	G	A	1.07	0.01	5.40E-09	Morris_etal_2012
rs41278853	*MTMR3*	22	G	A	803	0.11	0.14	0.89	A	G	1.14	0.03	5.60E-09	Fuchsberger_etal_2016
rs6808574	*NA*	3	T	C	774	0.39	0.26	0.61	C	T	1.07	0.01	5.80E-09	Mahajan_etal_2014
rs459193	*NA*	5	A	G	741	0.24	1.00	0.76	G	A	1.08	0.02	6.00E-09	Morris_etal_2012
rs7955901	*NA*	12	C	T	803	0.42	0.61	0.42	C	T	1.07	0.01	6.50E-09	Morris_etal_2012
rs702634	*ARL15*	5	G	A	785	0.30	0.73	0.70	A	G	1.06	0.01	6.90E-09	Mahajan_etal_2014
rs4275659	*ABCB9*	12	T	C	791	0.29	0.20	0.71	C	T	1.06	0.01	9.50E-09	Mahajan_etal_2014
rs12970134	*NA*	18	A	G	804	0.26	0.64	0.26	A	G	1.08	0.02	1.20E-08	Morris_etal_2012
rs1359790	*NA*	13	A	G	784	0.28	0.25	0.72	G	A	1.08	0.01	1.40E-08	Morris_etal_2012
rs7202877	*NA*	16	G	T	804	0.11	0.06	0.89	T	G	1.12	0.02	3.50E-08	Morris_etal_2012
rs4812829	*HNF4A*	20	A	G	804	0.14	0.24	0.14	A	G	1.07	0.03	5.00E-08	Mahajan_etal_2014
rs7845219	*TP53INP1*	8	T	C	797	0.48	0.36	0.48	T	C	1.08	0.02	6.00E-08	Mahajan_etal_2014
rs10510110	*PLEKHA1*	10	C	T	800	0.48	0.78	0.48	C	C	1.05	0.01	1.00E-07	Mahajan_etal_2014
rs7961581	*TSPAN8*	12	C	T	802	0.26	0.36	0.26	C	T	1.06	0.01	1.80E-07	Gaulton_etal_2015
rs10190052	*NA*	2	T	C	804	0.18	0.41	0.82	C	T	1.07	0.02	2.00E-07	Mahajan_etal_2014
rs9472138	*AK097853*	6	T	C	804	0.28	0.34	0.28	T	C	1.06	0.01	2.00E-07	Mahajan_etal_2014
rs2283220	*KCNQ1*	11	G	A	767	0.29	0.86	0.71	A	G	1.06	0.01	2.40E-07	Gaulton_etal_2015
rs2028299	*AP3S2*	15	C	A	798	0.29	0.73	0.29	C	A	1.04	0.02	5.00E-07	Mahajan_etal_2014
rs7795991	*NA*	7	A	G	751	0.47	0.83	0.53	G	A	1.05	0.01	7.00E-07	Mahajan_etal_2014

Continuation Table S8-36

SNP	Gene	Chr	ALSPAC (N=1,252)						DIAGRAM 2012-2016					
			A1	A2	$N_{controls}$	MAF	P_{HW}	EAF	EA	OA	OR	SE	P	Study
rs2820446	*NA*	1	G	C	802	0.28	0.44	0.72	C	G	1.05	0.01	2.00E-06	Mahajan_etal_2014
rs319598	*PCBD2*	5	T	C	801	0.42	0.43	0.58	C	T	1.05	0.01	2.00E-06	Mahajan_etal_2014
rs4273712	*AK127472*	6	G	A	804	0.29	0.55	0.29	G	A	1.05	0.01	3.00E-06	Mahajan_etal_2014
rs12427353	*HNF1A*	12	C	G	778	0.19	1.00	0.81	G	C	1.12	0.03	4.00E-06	Mahajan_etal_2014
rs2812533	*NA*	10	T	C	765	0.14	0.05	0.86	C	C	1.07	0.01	5.00E-06	Mahajan_etal_2014
rs7041847	*GLIS3*	9	A	G	799	0.47	0.72	0.47	A	G	1.05	0.02	5.00E-06	Mahajan_etal_2014
rs6937795	*NA*	6	C	A	792	0.47	0.72	0.53	A	C	1.04	0.01	7.00E-06	Mahajan_etal_2014
rs1535500	*KCNK17*	6	T	G	793	0.49	0.12	0.49	T	G	1.13	0.03	8.00E-06	Mahajan_etal_2014
rs2284219	*CRHR2*	7	A	G	794	0.34	0.81	0.66	G	A	1.05	0.01	8.00E-06	Mahajan_etal_2014
rs10788575	*NA*	10	A	G	799	0.16	0.80	0.16	A	G	1.06	0.01	9.00E-06	Mahajan_etal_2014
rs16861329	*ST6GAL1*	3	T	C	804	0.14	0.66	0.86	C	T	1.03	0.04	9.00E-06	Mahajan_etal_2014

Figure S8-22 Association between T2D and two polygenic scores validated in a subsample of adults in ALSPAC (n=1,252). Top panel: log(odds) of T2D per SD increase in the score. Regression was assessed using an unadjusted model (T2D~PRS), and a model adjusted for sex and 10 genetic PCs. The x-axis corresponds to the number of SNPs included in each score, and the y-axis is the effect estimate between the score and T2D. Values of the score were standardised to Z-values. Bottom panel: total variation in T2D (R^2) explained by the scores according to the regression model implemented. Variation was calculated using the Nagelkerke's R^2 as an approximation of the R^2 reported in a linear regression.

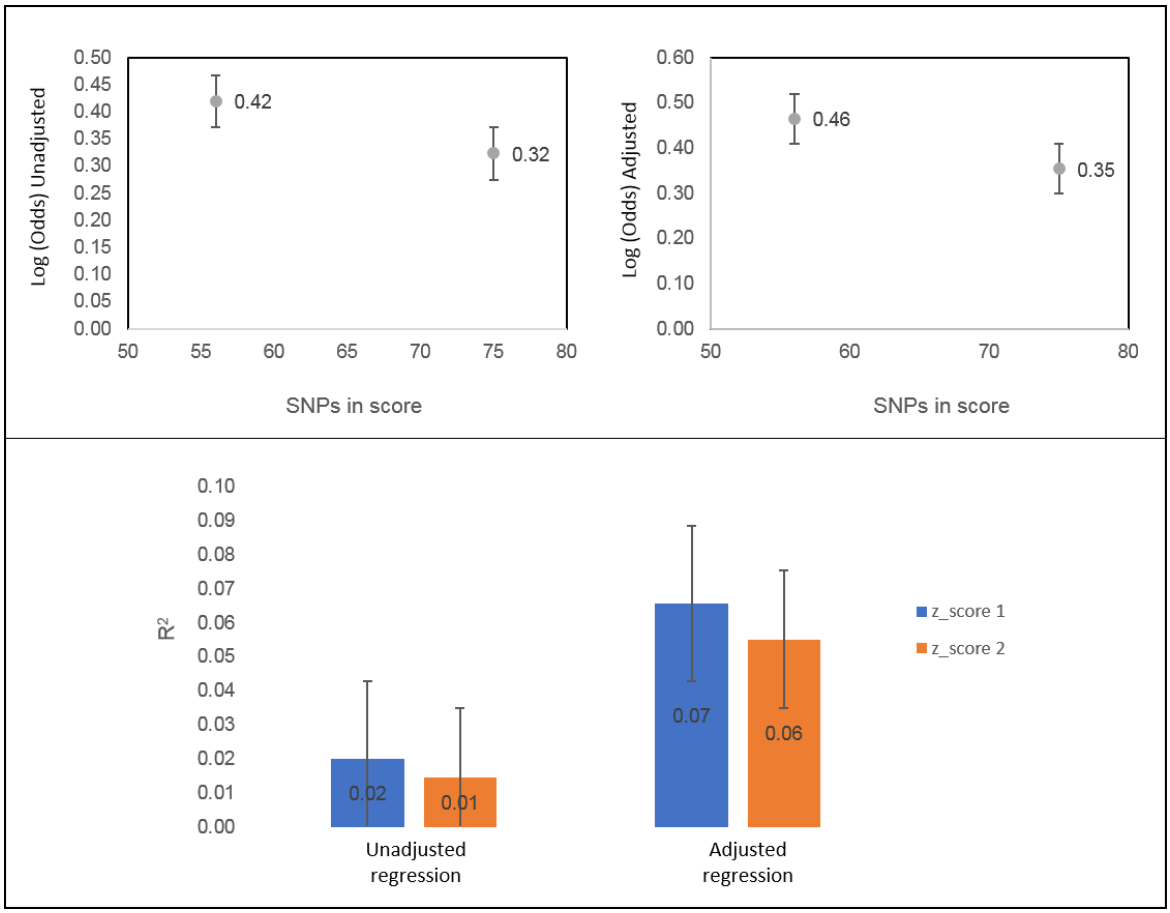


Table S8-37 Comparison of association statistics between the case control study (T2D vs Meth) and a polygenic risk score regression (IV vs Meth) for top DMPs identified in two observational analyses. Results of the PRS-outcome association were adjusted for age, sex, 10 genetic PCs and Houseman cells.

Table S8-37.1 Association statistics of the IV-outcome regression for 58 top DMPs identified in a sensitivity meta-EWAS of T2D (excluding KORA samples) at $p< 1.0x10^{-5}$ and $p\geq1.07x10^{-7}$, using a model adjusted for age, sex, SVs, smoking and Houseman cells. In bold, associations with opposite direction of effect across analyses.

CpG	Chr	Gene	T2D vs Meth†				PRS vs Meth‡			
			Estimate	SE	P	*Bonf	Estimate	SE	P	*Bonf
cg19693031	1	*TXNIP*	-0.019	2.59E-03	8.75E-14	3.29E-08	-1.85E-03	1.43E-03	0.20	1.00
cg06500161	21	*ABCG1*	0.013	1.92E-03	2.34E-11	8.81E-06	9.34E-04	1.40E-03	0.50	1.00
cg16765088	15	*Unannotated*	-0.011	1.75E-03	5.50E-10	2.07E-04	-1.87E-03	1.25E-03	0.13	1.00
cg00574958	11	*CPT1A*	-0.007	1.25E-03	1.20E-08	4.53E-03	-4.79E-04	3.93E-04	0.22	1.00
cg24704287	19	*Unannotated*	-0.011	1.97E-03	2.34E-08	8.79E-03	-6.90E-04	1.07E-03	0.52	1.00
cg00144180	2	*HDAC4*	0.012	2.23E-03	5.64E-08	0.02	7.89E-05	1.13E-03	0.94	1.00
cg04567334	**10**	***CDH23***	**-0.006**	1.18E-03	1.67E-07	0.06	**2.73E-04**	7.52E-04	0.72	1.00
cg10584271	3	*ITIH1*	-0.014	2.63E-03	1.73E-07	0.07	-5.87E-05	1.10E-03	0.96	1.00
cg26270261	**12**	***KRT4***	**-0.006**	1.24E-03	5.68E-07	0.21	**5.55E-04**	9.47E-04	0.56	1.00
cg16575444	**16**	***CX3CL1***	**-0.006**	1.22E-03	6.83E-07	0.26	**6.10E-04**	1.06E-03	0.56	1.00
cg24512093	**3**	***ROBO1***	**-0.010**	1.92E-03	7.16E-07	0.27	**2.79E-04**	1.55E-03	0.86	1.00
cg11983038	**13**	***Unannotated***	**-0.017**	3.38E-03	7.23E-07	0.27	**4.29E-04**	1.55E-03	0.78	1.00
cg25136644	7	*ATG9B*	-0.007	1.44E-03	7.27E-07	0.27	**-7.24E-04**	9.22E-04	0.43	1.00
cg20812370	1	*PBX1*	-0.007	1.34E-03	7.40E-07	0.28	-3.04E-03	1.37E-03	0.03	1.00
cg24686009	12	*RAP1B*	-0.002	3.83E-04	1.19E-06	0.45	-2.27E-04	1.07E-04	0.03	1.00
cg11024682	17	*SREBF1*	0.008	1.59E-03	1.33E-06	0.50	5.53E-05	7.45E-04	0.94	1.00
cg06114363	1	*ZNF683*	-0.010	2.16E-03	1.37E-06	0.51	-5.98E-04	1.30E-03	0.65	1.00
cg01963618	**6**	***LOC285768***	**-0.008**	1.58E-03	1.55E-06	0.58	**1.17E-03**	1.11E-03	0.29	1.00
cg22680424	**11**	***HCCA2***	**0.008**	1.63E-03	2.16E-06	0.81	**-1.22E-03**	1.38E-03	0.38	1.00
cg19876302	**10**	***Unannotated***	**-0.008**	1.70E-03	2.22E-06	0.84	**4.34E-04**	1.52E-03	0.78	1.00
cg08857797	**17**	***VPS25***	**0.009**	1.84E-03	2.28E-06	0.86	**1.20E-03**	1.69E-03	0.48	1.00
cg27374726	10	*Unannotated*	-0.009	1.83E-03	2.32E-06	0.87	-1.72E-03	8.27E-04	0.04	1.00
cg09185884	17	*KCTD2*	0.011	2.31E-03	2.33E-06	0.88	7.48E-04	9.50E-04	0.43	1.00
cg27115863	22	*Unannotated*	-0.011	2.31E-03	2.41E-06	0.91	-6.26E-04	1.50E-03	0.68	1.00
cg24795867	12	*WNT5B*	-0.006	1.29E-03	2.47E-06	0.93	-1.31E-03	1.49E-03	0.38	1.00
cg08945443	10	*ZMYND17*	0.010	2.23E-03	2.64E-06	0.99	1.73E-03	1.43E-03	0.23	1.00
cg06039489	**20**	***C20orf26***	**0.016**	3.35E-03	2.71E-06	1.00	**-2.40E-03**	1.64E-03	0.14	1.00
cg08570691	19	*RPL13AP5*	-0.008	1.76E-03	2.78E-06	1.00	-9.32E-04	1.21E-03	0.44	1.00

Continuation Table S8-37.1

CpG	Chr	Gene	T2D vs Meth†				PRS vs Meth‡			
			Estimate	SE	P	*Bonf	Estimate	SE	P	*Bonf
cg12593793	**1**	***Unannotated***	**-0.008**	1.65E-03	2.90E-06	1.00	**1.44E-03**	9.56E-04	0.13	1.00
cg27037013	21	*Unannotated*	-0.015	3.22E-03	2.90E-06	1.00	-1.93E-03	2.25E-03	0.39	1.00
cg07212837	**8**	***Unannotated***	**0.006**	1.38E-03	3.28E-06	1.00	**-8.63E-04**	8.59E-04	0.32	1.00
cg16192197	**6**	***Unannotated***	**0.010**	2.06E-03	3.71E-06	1.00	**-2.77E-03**	1.70E-03	0.10	1.00
cg15560632	7	*LRCH4*	-0.001	2.04E-04	3.83E-06	1.00	-5.95E-05	5.24E-05	0.26	1.00
cg14003143	20	*SGK2*	-0.006	1.27E-03	4.12E-06	1.00	-1.41E-05	1.13E-03	0.99	1.00
cg20154947	**8**	***PLEC1***	**-0.002**	4.09E-04	4.34E-06	1.00	**6.80E-05**	9.62E-05	0.48	1.00
cg25741837	**2**	***SMYD5***	**0.009**	2.01E-03	4.76E-06	1.00	**-1.46E-05**	1.45E-03	0.99	1.00
cg26766064	17	*MIR657*	-0.007	1.43E-03	5.17E-06	1.00	-1.78E-03	1.45E-03	0.22	1.00
cg25536676	1	*DHCR24*	-0.008	1.68E-03	5.39E-06	1.00	-5.52E-04	1.04E-03	0.60	1.00
cg11376147	11	*SLC43A1*	-0.006	1.27E-03	5.43E-06	1.00	-8.20E-04	5.36E-04	0.13	1.00
cg20316538	**2**	***RUFY4***	**-0.005**	1.16E-03	6.11E-06	1.00	**4.60E-04**	8.81E-04	0.60	1.00
cg18181703	**17**	***SOCS3***	**-0.010**	2.31E-03	6.20E-06	1.00	**1.50E-03**	1.58E-03	0.34	1.00
cg11851382	1	*PPAP2B*	-0.008	1.69E-03	6.42E-06	1.00	-2.31E-03	9.03E-04	0.01	0.62
cg00162348	16	*RNF40*	-0.002	4.23E-04	6.64E-06	1.00	-5.97E-05	1.07E-04	0.58	1.00
cg07184465	5	*SPZ1*	-0.007	1.55E-03	7.18E-06	1.00	-1.67E-03	9.88E-04	0.09	1.00
cg14284506	**17**	***Unannotated***	**-0.005**	1.08E-03	7.31E-06	1.00	**2.43E-04**	4.32E-04	0.57	1.00
cg11252555	19	*RPL13AP5*	-0.008	1.73E-03	7.44E-06	1.00	-1.16E-03	1.13E-03	0.30	1.00
cg10082515	**7**	***Unannotated***	**-0.013**	3.00E-03	7.46E-06	1.00	**3.88E-04**	1.46E-03	0.79	1.00
cg00896068	13	*Unannotated*	-0.008	1.74E-03	7.58E-06	1.00	-1.43E-03	1.37E-03	0.30	1.00
cg01577083	16	*Unannotated*	-0.011	2.54E-03	7.93E-06	1.00	-1.77E-03	1.05E-03	0.09	1.00
cg00320980	10	*Unannotated*	-0.009	2.09E-03	7.97E-06	1.00	-1.03E-03	1.01E-03	0.31	1.00
cg20231084	**11**	***Unannotated***	**-0.006**	1.39E-03	8.36E-06	1.00	**2.13E-03**	1.20E-03	0.07	1.00
cg15832662	11	*RTN3*	-0.009	2.08E-03	8.45E-06	1.00	-8.55E-05	1.22E-03	0.94	1.00
cg13178597	6	*RGS17*	-0.010	2.27E-03	8.57E-06	1.00	-2.62E-03	1.33E-03	0.05	1.00
cg20116935	3	*SEMA3B*	-0.006	1.38E-03	8.89E-06	1.00	-8.77E-04	8.84E-04	0.32	1.00
cg00989505	14	*MIR299*	-0.004	9.41E-04	9.33E-06	1.00	-1.19E-03	8.39E-04	0.16	1.00
cg07068382	**6**	***MTCH1***	**0.010**	2.37E-03	9.46E-06	1.00	**-8.28E-04**	1.67E-03	0.62	1.00
cg14476101	1	*PHGDH*	-0.015	3.36E-03	9.46E-06	1.00	-2.03E-03	2.14E-03	0.34	1.00
cg20456243	2	*SPEG*	-0.007	1.65E-03	9.99E-06	1.00	-5.07E-04	1.48E-03	0.73	1.00

†Observed exposure versus outcome association. ‡Observed instrumental-variable versus outcome association (observed IV-outcome). * Bonferroni-adjusted P-value.

Table S8-37.2 Association statistics of the IV-outcome regression for 11 top DMPs identified in the EWAS of T2D in ALSPAC ($p < 1.0x10^{-5}$) using a model adjusted for age, sex, SVs, BMI, smoking and Houseman cells. In bold, associations with opposite direction of effect across analyses.

CpG	Chr	Gene	T2D vs Meth†				PRS vs Meth‡			
			Estimate	SE	P	Bonf*	Estimate	SE	P	Bonf*
cg15986668	1	*NFYC*	-0.071	1.30E-02	5.48E-08	0.02	-3.94E-03	2.39E-03	0.10	1.00
cg14045803	11	*STARD10*	-0.012	2.00E-03	1.39E-07	0.05	-1.43E-04	3.67E-04	0.70	1.00
cg10870892	**11**	***CTTN***	**-0.045**	9.00E-03	1.13E-06	0.43	**9.97E-04**	1.59E-03	0.53	1.00
cg26652413	19	*CPAMD8*	-0.023	5.00E-03	2.51E-06	0.96	-1.34E-03	9.61E-04	0.16	1.00
cg00204249	**17**	***DNAH17***	**-0.015**	3.00E-03	2.76E-06	1.00	**3.14E-05**	5.25E-04	0.95	1.00
cg03206717	3	*SLC25A38*	-0.003	1.00E-03	2.95E-06	1.00	-4.20E-05	1.21E-04	0.73	1.00
cg19823491	2	*OTX1*	-0.006	1.00E-03	2.99E-06	1.00	-2.75E-04	1.92E-04	0.15	1.00
cg02307288	5	*TRPC7*	-0.038	8.00E-03	5.54E-06	1.00	-1.64E-03	1.69E-03	0.33	1.00
cg04016326	12	*GRIN2B*	-0.054	1.20E-02	5.71E-06	1.00	-2.84E-03	2.13E-03	0.18	1.00
cg04656330	2	*PNKD*	-0.002	3.52E-04	7.96E-06	1.00	-1.38E-04	6.79E-05	0.04	0.46
cg07251197	**1**	***Unannotated***	**-0.008**	1.76E-03	8.11E-06	1.00	**2.21E-04**	2.69E-04	0.41	1.00

†Observed exposure versus outcome association. ‡Observed instrumental-variable versus outcome association (observed IV-outcome). * Bonferroni-adjusted P-value.

Table S8-38 Robust linear regression (RLR) between T2D and methylation for top DMPs detected in three observational analyses at $p < 1.0x10^{-5}$. Results based on a basic model adjusted for age, sex, 8 SVs, 6 Houseman cells, BMI and smoking, and a second model additionally adjusted for the polygenic risk score (PRS 56 SNPs) using ALSPAC samples (cases=36 and controls=826). Associations were regarded significant at Bonferroni corrected $p < 0.05$.

Table S8-38.1 RLR estimates for 25 top DMPs identified in the meta-EWAS of T2D (5 cohorts).

CpG	Chr	Gene	RLR: Meth ~ T2D + Covariates					RLR: Meth ~ T2D + PRS + Covariates				
			Estimate	SE	Bonf	Cases	Controls	Estimate	SE	Bonf	Cases	Controls
cg00574958*	11	*CPT1A*	-5.66E-03	1.45E-03	2.48E-03	35	825	-5.64E-03	1.50E-03	4.16E-03	35	825
cg15560632*	7	*LRCH4*	-8.73E-04	2.51E-04	0.01	36	826	-8.40E-04	2.52E-04	0.02	36	826
cg22628512	1	*Unannotated*	1.13E-02	3.28E-03	0.01	36	825	1.13E-02	3.29E-03	0.02	36	825
cg06500161*	21	*ABCG1*	2.48E-02	7.49E-03	0.02	35	808	2.54E-02	7.56E-03	0.02	35	808
cg17155612*	19	*LOC148189*	-2.23E-03	7.15E-04	0.05	36	810	-2.29E-03	7.20E-04	0.04	36	810
cg20154947	8	*PLEC1*	-1.63E-03	5.59E-04	0.09	36	822	-1.63E-03	5.61E-04	0.09	36	822
cg01317029	3	*FAM131A*	1.39E-02	4.93E-03	0.12	36	824	1.41E-02	5.00E-03	0.12	36	824
cg25741837	2	*SMYD5*	1.93E-02	8.56E-03	0.60	35	823	1.91E-02	8.56E-03	0.64	35	823
cg19693031	1	*TXNIP*	-2.27E-02	1.03E-02	0.70	35	816	-2.17E-02	1.07E-02	1.00	35	816
cg00082384	3	*NISCH*	1.72E-02	8.38E-03	1.00	34	767	1.81E-02	8.49E-03	0.84	34	767
cg01009875	1	*TMCO1*	-1.05E-03	8.12E-04	1.00	35	825	-9.93E-04	8.17E-04	1.00	35	825
cg06039489	20	*C20orf26*	8.66E-03	7.12E-03	1.00	36	822	1.01E-02	7.08E-03	1.00	36	822
cg06468695	17	*CCDC42*	1.24E-03	7.26E-03	1.00	36	826	1.29E-03	7.29E-03	1.00	36	826
cg07184465	5	*SPZ1*	-9.52E-03	7.02E-03	1.00	36	819	-8.47E-03	7.42E-03	1.00	36	819
cg07400328	6	*MUTED*	-1.60E-03	1.51E-03	1.00	36	823	-1.53E-03	1.51E-03	1.00	36	823
cg08273233	6	*HTR1E*	-3.24E-03	5.23E-03	1.00	36	824	-3.62E-03	5.29E-03	1.00	36	824
cg11851382	1	*PPAP2B*	-7.00E-03	5.59E-03	1.00	36	823	-5.93E-03	5.38E-03	1.00	36	823
cg13826139	6	*Unannotated*	-1.59E-03	5.61E-03	1.00	36	813	-1.36E-03	5.62E-03	1.00	36	813
cg13927560	4	*TMEM33*	1.33E-05	1.00E-03	1.00	36	825	1.39E-05	1.00E-03	1.00	36	825
cg14186584	5	*Unannotated*	-1.45E-03	7.33E-04	1.00	36	826	-1.40E-03	7.35E-04	1.00	36	826
cg14275576	20	*Unannotated*	-4.25E-04	1.17E-03	1.00	36	826	-5.59E-04	1.16E-03	1.00	36	826
cg17566334	6	*PACRG*	-3.79E-03	7.89E-03	1.00	36	826	-3.54E-03	7.85E-03	1.00	36	826
cg19611616	12	*STK38L*	-2.72E-03	1.98E-03	1.00	36	826	-2.52E-03	1.97E-03	1.00	36	826
cg27237541	10	*MYO3A*	-8.48E-03	7.69E-03	1.00	36	816	-8.24E-03	7.74E-03	1.00	36	816
cg27374726	10	*Unannotated*	-1.77E-03	5.43E-03	1.00	36	826	-9.23E-04	5.75E-03	1.00	36	826

Table S8-38.2 RLR estimates for 58 top DMPs identified in a sensitivity Meta-EWAS of T2D (4 cohorts).

CpG	Chr	Gene	RLR: Meth ~ T2D + Covariates					RLR: Meth ~ T2D + PRS + Covariates				
			Estimate	SE	Bonf	Cases	Controls	Estimate	SE	Bonf	Cases	Controls
cg00574958	11	*CPT1A*	-5.66E-03	1.45E-03	0.01	36	826	-5.64E-03	1.50E-03	0.01	36	826
cg15560632	7	*LRCH4*	-8.73E-04	2.51E-04	0.03	36	826	-8.40E-04	2.52E-04	0.05	36	826
cg00162348	16	*RNF40*	-1.89E-03	5.69E-04	0.05	36	826	-1.89E-03	5.73E-04	0.06	36	826
cg06500161	21	*ABCG1*	2.48E-02	7.49E-03	0.06	36	825	2.54E-02	7.56E-03	0.05	36	825
cg14284506	17	*Unannotated*	-7.21E-03	2.29E-03	0.10	36	826	-7.38E-03	2.29E-03	0.08	36	826
cg20154947	8	*PLEC1*	-1.63E-03	5.59E-04	0.21	36	824	-1.63E-03	5.61E-04	0.22	36	824
cg14003143	20	*SGK2*	-1.55E-02	5.37E-03	0.22	36	824	-1.56E-02	5.35E-03	0.21	36	824
cg11024682	17	*SREBF1*	1.17E-02	4.14E-03	0.29	36	826	1.20E-02	4.15E-03	0.22	36	826
cg11983038	13	*Unannotated*	-2.43E-02	8.73E-03	0.31	36	825	-2.48E-02	8.75E-03	0.26	36	825
cg00144180	2	*HDAC4*	8.79E-03	5.93E-03	1.00	36	813	8.43E-03	5.95E-03	1.00	36	813
cg00320980	10	*Unannotated*	-3.05E-03	5.18E-03	1.00	36	824	-2.29E-03	5.25E-03	1.00	36	824
cg00896068	13	*Unannotated*	2.89E-03	7.47E-03	1.00	36	820	3.96E-03	7.65E-03	1.00	36	820
cg00989505	14	*MIR299*	5.70E-03	4.54E-03	1.00	36	826	6.12E-03	4.54E-03	1.00	36	826
cg01577083	16	*Unannotated*	-1.47E-02	8.14E-03	1.00	36	818	-1.39E-02	8.03E-03	1.00	36	818
cg01963618	6	*LOC285768*	-2.99E-03	7.32E-03	1.00	36	824	-3.29E-03	7.26E-03	1.00	36	824
cg04567334	10	*CDH23*	5.64E-04	3.79E-03	1.00	36	825	6.54E-04	3.80E-03	1.00	36	825
cg06039489	20	*C20orf26*	8.66E-03	7.12E-03	1.00	36	824	1.01E-02	7.08E-03	1.00	36	824
cg06114363	1	*ZNF683*	-9.62E-03	6.83E-03	1.00	35	825	-8.81E-03	6.77E-03	1.00	35	825
cg07068382	6	*MTCH1*	-2.55E-03	8.57E-03	1.00	36	826	-2.14E-03	8.66E-03	1.00	36	826
cg07184465	5	*SPZ1*	-9.52E-03	7.02E-03	1.00	35	816	-8.47E-03	7.42E-03	1.00	35	816
cg07212837	8	*Unannotated*	4.27E-03	4.54E-03	1.00	36	825	5.09E-03	4.58E-03	1.00	36	825
cg08570691	19	*RPL13AP5*	-7.35E-03	8.34E-03	1.00	36	826	-7.02E-03	7.98E-03	1.00	36	826
cg08857797	17	*VPS25*	1.55E-02	9.48E-03	1.00	36	819	1.47E-02	9.57E-03	1.00	36	819
cg08945443	10	*ZMYND17*	-9.35E-03	6.62E-03	1.00	36	826	-9.56E-03	6.69E-03	1.00	36	826
cg09185884	17	*KCTD2*	1.55E-02	7.71E-03	1.00	36	826	1.52E-02	7.77E-03	1.00	36	826
cg10082515	7	*Unannotated*	-1.44E-02	9.01E-03	1.00	36	826	-1.50E-02	8.83E-03	1.00	36	826
cg10584271	3	*ITIH1*	-2.39E-03	5.60E-03	1.00	36	826	-1.92E-03	5.63E-03	1.00	36	826
cg11252555	19	*RPL13AP5*	-6.64E-03	6.60E-03	1.00	36	824	-6.14E-03	6.59E-03	1.00	36	824
cg11376147	11	*SLC43A1*	-3.97E-03	3.09E-03	1.00	36	825	-3.50E-03	3.09E-03	1.00	36	825

* Associations surpassing Bonferroni adjustment in the Robust Linear regression between Methylation and T2D in ALSPAC, but not in the Ordinary Least Squared Regression used to compute the Meta-EWAS of T2D across five European cohorts.

Continuation Table S8-38.2

CpG	Chr	Gene	RLR: Meth ~ T2D + Covariates					RLR: Meth ~ T2D + PRS + Covariates				
			Estimate	SE	Bonf	Cases	Controls	Estimate	SE	Bonf	Cases	Controls
cg11851382	1	*PPAP2B*	-7.00E-03	5.59E-03	1	36	825	-5.93E-03	5.38E-03	1	36	825
cg12593793	1	*Unannotated*	-4.42E-03	4.98E-03	1	36	824	-4.74E-03	5.01E-03	1	36	824
cg13178597	6	*RGS17*	-1.44E-02	6.91E-03	1	36	826	-1.33E-02	7.00E-03	1	36	826
cg14476101	1	*PHGDH*	-2.51E-02	1.12E-02	1	35	816	-2.44E-02	1.13E-02	1	35	816
cg15832662	11	*RTN3*	-9.04E-03	6.84E-03	1	36	824	-8.69E-03	6.90E-03	1	36	824
cg16192197	6	*Unannotated*	1.78E-02	8.40E-03	1	36	826	1.87E-02	8.29E-03	1	36	826
cg16575444	16	*CX3CL1*	8.29E-05	5.37E-03	1	35	825	6.44E-04	5.39E-03	1	35	825
cg16765088	15	*Unannotated*	-6.58E-03	5.51E-03	1	36	826	-6.38E-03	5.67E-03	1	36	826
cg18181703	17	*SOCS3*	-2.28E-02	1.02E-02	1	34	820	-2.41E-02	1.03E-02	1	34	820
cg19693031	1	*TXNIP*	-2.27E-02	1.03E-02	1	35	826	-2.17E-02	1.07E-02	1	35	826
cg19876302	10	*Unannotated*	-2.19E-02	1.27E-02	1	36	826	-2.18E-02	1.27E-02	1	36	826
cg20116935	3	*SEMA3B*	-4.65E-03	4.38E-03	1	36	825	-3.97E-03	4.51E-03	1	36	825
cg20231084	11	*Unannotated*	-7.06E-03	5.73E-03	1	36	823	-7.58E-03	5.68E-03	1	36	823
cg20316538	2	*RUFY4*	-5.29E-03	4.83E-03	1	36	826	-5.11E-03	4.71E-03	1	36	826
cg20456243	2	*SPEG*	-1.09E-02	8.78E-03	1	36	823	-1.02E-02	8.55E-03	1	36	823
cg20812370	1	*PBX1*	-1.54E-02	1.03E-02	1	36	826	-1.37E-02	1.06E-02	1	36	826
cg22680424	11	*HCCA2*	8.09E-03	6.08E-03	1	36	824	8.37E-03	6.12E-03	1	36	824
cg24512093	3	*ROBO1*	-1.77E-02	8.62E-03	1	36	812	-1.75E-02	8.59E-03	1	36	812
cg24686009	12	*RAP1B*	-1.54E-03	9.14E-04	1	36	810	-1.51E-03	9.34E-04	1	36	810
cg24704287	19	*Unannotated*	1.47E-03	6.36E-03	1	35	819	2.17E-03	6.42E-03	1	35	819
cg24795867	12	*WNT5B*	2.18E-03	1.02E-02	1	36	826	3.19E-03	1.03E-02	1	36	826
cg25136644	7	*ATG9B*	-8.51E-03	5.83E-03	1	36	822	-7.97E-03	5.90E-03	1	36	822
cg25536676	1	*DHCR24*	-3.32E-03	5.95E-03	1	36	826	-2.97E-03	6.33E-03	1	36	826
cg25741837	2	*SMYD5*	1.93E-02	8.56E-03	1	36	826	1.91E-02	8.56E-03	1	36	826
cg26270261	12	*KRT4*	-3.59E-03	5.39E-03	1	36	826	-3.86E-03	5.43E-03	1	36	826
cg26766064	17	*MIR657*	-6.72E-03	7.72E-03	1	36	825	-5.51E-03	7.98E-03	1	36	825
cg27037013	21	*Unannotated*	-1.79E-02	1.47E-02	1	36	826	-1.70E-02	1.45E-02	1	36	826
cg27115863	22	*Unannotated*	-1.88E-02	8.32E-03	1	36	826	-1.83E-02	8.60E-03	1	36	826
cg27374726	10	*Unannotated*	-1.77E-03	5.43E-03	1	36	826	-9.23E-04	5.75E-03	1	36	826

Table S8-38.3 RLR estimates for 11 top DMPs identified in the EWAS of T2D in ALSPAC.

CpG	Chr	Gene	RLR: Meth ~ T2D + Covariates					RLR: Meth ~ T2D + PRS + Covariates				
			Estimate	SE	Bonf	Cases	Controls	Estimate	SE	Bonf	Cases	Controls
cg14045803*	**11**	***STARD10***	**-0.010**	**1.52E-03**	**1.23E-09**	36	825	**-0.010**	**0.002**	**1.27E-09**	32	819
cg26652413*	**19**	***CPAMD8***	**-0.023**	**4.22E-03**	**9.92E-07**	36	824	**-0.022**	**0.004**	**2.70E-06**	36	826
cg15986668	**1**	***NFYC***	**-0.076**	**1.76E-02**	**1.753E-04**	32	819	**-0.074**	**0.018**	**4.26E-04**	36	826
cg03206717*	**3**	***SLC25A38***	**-0.003**	**8.02E-04**	**2.420E-03**	36	818	**-0.003**	**0.001**	**3.50E-03**	34	812
cg04656330*	**2**	***PNKD***	**-0.002**	**5.00E-04**	**5.646E-03**	36	824	**-0.002**	**0.000**	**5.52E-03**	36	822
cg02307288*	**5**	***TRPC7***	**-0.031**	**9.80E-03**	**1.775E-02**	34	812	**-0.030**	**0.010**	**2.09E-02**	36	825
cg10870892*	**11**	***CTTN***	**-0.038**	**1.25E-02**	**2.679E-02**	36	816	**-0.039**	**0.012**	**2.16E-02**	36	824
cg07251197*	**1**	***Unannotated***	**-0.004**	**1.48E-03**	**2.718E-02**	36	826	**-0.005**	**0.001**	**2.36E-02**	36	824
cg19823491*	**2**	***OTX1***	**-0.004**	**1.19E-03**	**3.218E-02**	36	826	**-0.003**	**0.001**	**3.01E-02**	36	816
cg04016326	12	*GRIN2B*	-0.048	1.80E-02	9.008E-02	36	825	-0.047	0.018	1.14E-01	36	825
cg00204249	17	*DNAH17*	-0.011	4.96E-03	3.399E-01	36	822	-0.011	0.005	2.82E-01	36	818

* Associations surpassing Bonferroni adjustment in the Robust Linear regression between Methylation and T2D, but not in the Ordinary Least Squared Regression used to compute the EWAS of T2D in ALSPAC.

Figure S8-23 Summary plots of the EWAS with the Polygenic risk score for T2D (56 SNPs). Top panel shows plots for the basic model adjusted for age, sex and the first ten genetic PCs, while the bottom panel shows results of the model additionally adjusted for 6 Houseman cells. Top-ranked DMPs in the basic model in suggestive association with the PRS were DMPs cg26799188, cg01554963 (NXN), cg03676624 (AQP11) and cg11362770 (ST3GAL6). The first three of these DMPs were also identified in the model adjusted for cells. There was no evidence of genomic inflation based on a λ of 1.01 and 0.99 for the two models. The volcano plot shows top-ranked signals at the top of the plot, but none of them reached the threshold of genome-wide significance at P<1.31x10^{-7}.

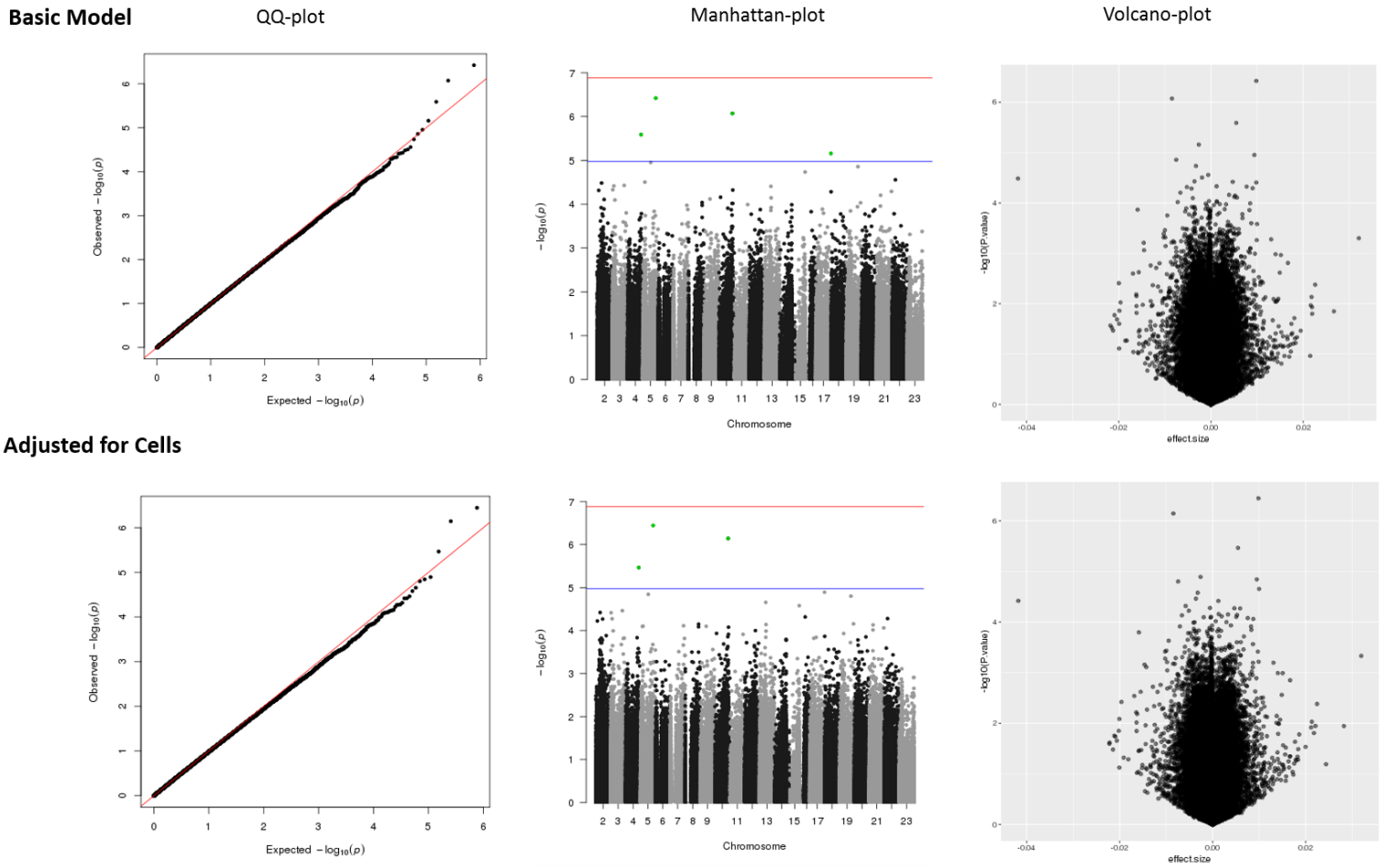


Table S8-39 Results of the matrix-eQTL analysis for 110 significant SNP-CpG associations identified in the EWAS of T2D-SNPs using the subsample of middle-age adults in ALSPAC (n=1248). Associations were considered significant at $p<1.60x10^{-9}$(α=0.05/31,330,975 tests). Effect estimate is interpreted as a unit change in inverse-normal transformed residuals of methylation per increase in the risk allele (MA).

											Matrix EQTL results						
SNP	CPG	SNP Chr	SNP Position	SNP Gene	CPG Chr	CPG Position	CPG Gene	A1	A2	MAF	Estimate	SE	P	R^2	FDR	N	meQTL
rs7845219	cg16049864	8	95937502	*TP53INP1*	8	95962084	*TP53INP1*	T	C	0.51	0.94	0.03	2.79E-163	0.45	8.75E-156	1248	Cis
rs2284219	cg22109827	7	30714436	*CRHR2*	7	30727326	*NA*	A	G	0.65	0.98	0.03	8.23E-160	0.44	1.29E-152	1248	Cis
rs7845219	cg20039814	8	95937502	*TP53INP1*	8	95962383	*TP53INP1*	T	C	0.51	0.92	0.03	6.89E-155	0.43	7.20E-148	1248	Cis
rs4275659	cg09084244	12	123447928	*NA*	12	1.24E+08	*CDK2AP1*	T	C	0.70	1.04	0.03	3.51E-153	0.43	2.75E-146	1248	Cis
rs7845219	cg18059933	8	95937502	*TP53INP1*	8	95962463	*TP53INP1*	T	C	0.51	0.90	0.03	1.67E-148	0.42	1.04E-141	1248	Cis
rs1535500	cg00012638	6	39284050	*KCNK16*	6	39280541	*KCNK17*	T	G	0.52	0.88	0.03	7.02E-142	0.40	3.66E-135	1248	Cis
rs4275659	cg00376283	12	123447928	*NA*	12	1.23E+08	*ABCB9*	T	C	0.70	-1.01	0.03	6.64E-140	0.40	2.97E-133	1248	Cis
rs7845219	cg13393036	8	95937502	*TP53INP1*	8	95962371	*TP53INP1*	T	C	0.51	0.88	0.03	1.93E-139	0.40	7.58E-133	1248	Cis
rs7845219	cg09323728	8	95937502	*TP53INP1*	8	95962352	*TP53INP1*	T	C	0.51	0.87	0.03	4.06E-135	0.39	1.41E-128	1248	Cis
rs1974620	cg19272540	7	15065467	*NA*	7	15055459	*NA*	C	T	0.53	0.81	0.03	2.82E-112	0.33	8.84E-106	1248	Cis
rs1535500	cg24018148	6	39284050	*KCNK16*	6	39290715	*KCNK16*	T	G	0.52	-0.74	0.03	7.63E-94	0.29	2.17E-87	1248	Cis
rs7845219	cg23172400	8	95937502	*TP53INP1*	8	95962367	*TP53INP1*	T	C	0.51	0.75	0.03	2.33E-93	0.29	6.07E-87	1248	Cis
rs4275659	cg07644039	12	123447928	*NA*	12	1.23E+08	*ARL6IP4*	T	C	0.70	0.83	0.04	2.72E-88	0.27	6.08E-82	1248	Cis
rs7177055	cg04480376	15	77832762	*NA*	15	77896785	*NA*	G	A	0.72	-0.76	0.04	4.11E-73	0.23	8.58E-67	1248	Cis
rs4275659	cg21745287	12	123447928	*NA*	12	1.23E+08	*ARL6IP4*	T	C	0.70	0.76	0.04	6.08E-73	0.23	1.19E-66	1248	Cis
rs4275659	cg13010344	12	123447928	*NA*	12	1.23E+08	*ARL6IP4*	T	C	0.70	0.73	0.04	1.05E-65	0.21	1.73E-59	1248	Cis
rs1535500	cg18601596	6	39284050	*KCNK16*	6	39283313	*KCNK16*	T	G	0.52	0.63	0.03	7.03E-65	0.21	1.10E-58	1248	Cis
rs2820446	cg19373347	1	219748818	*LYPLAL1*	1	2.2E+08	*NA*	G	C	0.70	0.68	0.04	1.56E-58	0.19	2.23E-52	1248	Cis
rs780094	cg04845466	2	27741237	*GCKR*	2	27665079	*KRTCAP3*	T	C	0.59	0.61	0.04	2.03E-57	0.19	2.76E-51	1248	Cis
rs4275659	cg10169515	12	123447928	*NA*	12	1.24E+08	*MPHOSPH9*	T	C	0.70	0.67	0.04	9.16E-55	0.18	1.20E-48	1248	Cis
rs780094	cg02592271	2	27741237	*GCKR*	2	27665507	*KRTCAP3*	T	C	0.59	0.59	0.04	9.83E-53	0.17	1.23E-46	1248	Cis
rs340874	cg01631319	1	214159256	*PROX1*	1	2.14E+08	*NA*	T	C	0.58	-0.60	0.04	2.12E-52	0.17	2.55E-46	1248	Cis
rs780094	cg17158414	2	27741237	*GCKR*	2	27665306	*KRTCAP3*	T	C	0.59	0.58	0.04	3.21E-51	0.17	3.73E-45	1248	Cis
rs1535500	cg04321126	6	39284050	*KCNK16*	6	39393690	*KIF6*	T	G	0.52	-0.56	0.04	5.41E-51	0.17	6.05E-45	1248	Cis
rs4275659	cg05973401	12	123447928	*NA*	12	1.23E+08	*ABCB9*	T	C	0.70	-0.64	0.04	3.22E-49	0.16	3.48E-43	1248	Cis
rs780094	cg21248554	2	27741237	*GCKR*	2	27665150	*KRTCAP3*	T	C	0.59	0.57	0.04	1.63E-48	0.16	1.65E-42	1248	Cis
rs780094	cg11618577	2	27741237	*GCKR*	2	27665543	*KRTCAP3*	T	C	0.59	0.56	0.04	1.49E-46	0.15	1.41E-40	1248	Cis
rs780094	cg12648201	2	27741237	*GCKR*	2	27665141	*KRTCAP3*	T	C	0.59	0.55	0.04	2.83E-46	0.15	2.61E-40	1248	Cis

Continuation Table S8-39

											Matrix EQTL results						
SNP	CPG	SNP Chr	SNP Position	SNP Gene	CPG Chr	CPG Position	CPG Gene	A1	A2	MAF	Estimate	SE	P	R^2	FDR	N	meQTL
rs243020	cg16665442	2	60585028	NA	2	60585866	NA	G	A	0.54	-0.54	0.04	7.71E-46	0.15	6.90E-40	1248	Cis
rs780094	cg20102877	2	27741237	GCKR	2	27665638	KRTCAP3	T	C	0.59	0.55	0.04	1.13E-45	0.15	9.85E-40	1248	Cis
rs780094	cg12000995	2	27741237	GCKR	2	27665139	KRTCAP3	T	C	0.59	0.55	0.04	2.61E-45	0.15	2.21E-39	1248	Cis
rs780094	cg18428193	2	27741237	GCKR	2	27665017	KRTCAP3	T	C	0.59	0.54	0.04	4.21E-44	0.14	3.47E-38	1248	Cis
rs4275659	cg12026538	12	123447928	NA	2	91935310	NA	T	C	0.70	0.59	0.04	1.62E-42	0.14	1.27E-36	1248	Trans
rs340874	cg24083324	1	214159256	PROX1	1	2.14E+08	PROX1	T	C	0.58	-0.53	0.04	1.07E-40	0.13	7.78E-35	1248	Cis
rs4275659	cg01687878	12	123447928	NA	12	1.24E+08	NA	T	C	0.70	0.58	0.04	1.96E-40	0.13	1.39E-34	1248	Cis
rs780094	cg24768116	2	27741237	GCKR	2	27665128	KRTCAP3	T	C	0.59	0.52	0.04	3.40E-40	0.13	2.37E-34	1248	Cis
rs12571751	cg20744163	10	80942631	ZMIZ1	10	80999841	ZMIZ1	G	A	0.54	0.52	0.04	5.10E-40	0.13	3.47E-34	1248	Cis
rs7177055	cg23627990	15	77832762	NA	15	77364890	TSPAN3	G	A	0.72	0.57	0.04	1.11E-39	0.13	7.38E-34	1248	Cis
rs780094	cg22903471	2	27741237	GCKR	2	27725779	GCKR	T	C	0.59	0.51	0.04	4.80E-39	0.13	3.07E-33	1248	Cis
rs11187140	cg27639046	10	94466910	NA	2	1.72E+08	NA	A	G	0.63	0.52	0.04	6.60E-37	0.12	4.05E-31	1248	Trans
rs2028299	cg23731826	15	90374257	AP3S2	15	90371692	NA	C	A	0.70	-0.54	0.04	9.23E-37	0.12	5.56E-31	1248	Cis
rs11717195	cg04890266	3	123082398	ADCY5	3	1.23E+08	ADCY5	C	T	0.74	0.55	0.04	3.98E-36	0.12	2.35E-30	1248	Cis
rs7845219	cg26343298	8	95937502	TP53INP1	8	95960752	TP53INP1	T	C	0.51	-0.47	0.04	3.90E-35	0.12	2.26E-29	1248	Cis
rs5215	cg01251548	11	17408630	KCNJ11	11	17372745	DKFZp686O24166	C	T	0.63	0.49	0.04	1.64E-34	0.11	9.34E-29	1248	Cis
rs1535500	cg06347083	6	39284050	KCNK16	6	39282316	KCNK17	T	G	0.52	0.46	0.04	5.53E-34	0.11	3.09E-28	1248	Cis
rs7177055	cg20380897	15	77832762	NA	15	77876657	NA	G	A	0.72	0.53	0.04	8.98E-34	0.11	4.93E-28	1248	Cis
rs5215	cg15432903	11	17408630	KCNJ11	11	17409602	KCNJ11	C	T	0.63	-0.49	0.04	1.78E-33	0.11	9.46E-28	1248	Cis
rs12571751	cg18737081	10	80942631	ZMIZ1	10	80999807	ZMIZ1	G	A	0.54	0.48	0.04	3.16E-33	0.11	1.65E-27	1248	Cis
rs7177055	cg15453836	15	77832762	NA	15	77711506	NA	G	A	0.72	0.52	0.04	1.52E-32	0.11	7.80E-27	1248	Cis
rs10510110	cg19863426	10	124192430	PLEKHA1	10	1.24E+08	PLEKHA1	C	T	0.53	0.44	0.04	1.93E-29	0.10	9.63E-24	1248	Cis
rs11187140	cg15757802	10	94466910	NA	10	94429530	NA	A	G	0.63	-0.47	0.04	1.94E-29	0.10	9.63E-24	1248	Cis
rs5215	cg03864215	11	17408630	KCNJ11	11	17408437	KCNJ11	C	T	0.63	0.45	0.04	5.17E-29	0.10	2.53E-23	1248	Cis
rs4275659	cg19016782	12	123447928	NA	12	1.24E+08	C12orf65	T	C	0.70	0.48	0.04	2.57E-27	0.09	1.22E-21	1248	Cis
rs7955901	cg19871235	12	71433293	NA	12	71552569	TSPAN8	C	T	0.57	-0.43	0.04	4.19E-27	0.09	1.96E-21	1248	Cis
rs780094	cg14021192	2	27741237	GCKR	2	27616791	PPM1G	T	C	0.59	0.42	0.04	5.31E-27	0.09	2.45E-21	1248	Cis
rs7955901	cg13355032	12	71433293	NA	12	71524030	TSPAN8	C	T	0.57	-0.41	0.04	5.00E-25	0.08	2.27E-19	1248	Cis
rs5215	cg08548044	11	17408630	KCNJ11	11	17249958	NA	C	T	0.63	-0.40	0.04	9.34E-23	0.07	4.01E-17	1248	Cis
rs7177055	cg12131826	15	77832762	NA	15	77904385	NA	G	A	0.72	0.43	0.04	1.35E-22	0.07	5.72E-17	1248	Cis
rs340874	cg10288510	1	214159256	PROX1	1	2.14E+08	NA	T	C	0.58	-0.39	0.04	9.82E-22	0.07	4.05E-16	1248	Cis
rs5215	cg01153817	11	17408630	KCNJ11	11	17409509	KCNJ11	C	T	0.63	-0.38	0.04	5.38E-21	0.07	2.19E-15	1248	Cis

Continuation Table S8-39

										Matrix EQTL results						
SNP	CPG	SNP Chr	SNP Position	SNP Gene	CPG Chr	CPG Position	CPG Gene	A1	A2	MAF	Estimate	SE	P	R^2	FDR	N	meQTL
rs7177055	cg12131826	15	77832762	*NA*	15	77904385	*NA*	G	A	0.72	0.43	0.04	1.35E-22	0.07	5.72E-17	1248	Cis
rs340874	cg10288510	1	214159256	*PROX1*	1	2.14E+08	*NA*	T	C	0.58	-0.39	0.04	9.82E-22	0.07	4.05E-16	1248	Cis
rs5215	cg01153817	11	17408630	*KCNJ11*	11	17409509	*KCNJ11*	C	T	0.63	-0.38	0.04	5.38E-21	0.07	2.19E-15	1248	Cis
rs9472138	cg11772020	6	43811762	*VEGFA*	6	43806470	*NA*	T	C	0.71	-0.41	0.04	1.35E-20	0.07	5.43E-15	1248	Cis
rs1535500	cg13075951	6	39284050	*KCNK16*	6	39282393	*KCNK17*	T	G	0.52	0.35	0.04	8.50E-20	0.06	3.33E-14	1248	Cis
rs4275659	cg19120225	12	123447928	*NA*	12	1.24E+08	*PITPNM2*	T	C	0.70	0.40	0.04	4.90E-19	0.06	1.90E-13	1248	Cis
rs1535500	cg13855924	6	39284050	*KCNK16*	6	39281183	*KCNK17*	T	G	0.52	0.33	0.04	1.41E-17	0.06	5.01E-12	1248	Cis
rs1535500	cg01429075	6	39284050	*KCNK16*	6	39290711	*KCNK16*	T	G	0.52	-0.33	0.04	2.06E-17	0.06	7.26E-12	1248	Cis
rs35510946	cg16570885	3	185518910	*IGF2BP2*	3	1.85E+08	*IGF2BP2*	A	G	0.69	0.36	0.04	3.94E-17	0.06	1.36E-11	1248	Cis
rs11187140	cg16060189	10	94466910	*NA*	10	94350577	*NA*	A	G	0.63	0.35	0.04	1.02E-16	0.05	3.39E-11	1248	Cis
rs35510946	cg23956648	3	185518910	*IGF2BP2*	3	1.85E+08	*IGF2BP2*	A	G	0.69	-0.35	0.04	1.32E-16	0.05	4.35E-11	1248	Cis
rs11187140	cg20554832	10	94466910	*NA*	10	94429526	*NA*	A	G	0.63	-0.35	0.04	1.39E-16	0.05	4.53E-11	1248	Cis
rs780094	cg26034919	2	27741237	*GCKR*	2	27665711	*KRTCAP3*	T	C	0.59	0.33	0.04	3.51E-16	0.05	1.11E-10	1248	Cis
rs780094	cg21747090	2	27741237	*GCKR*	2	27597821	*SNX17*	T	C	0.59	0.32	0.04	3.68E-16	0.05	1.14E-10	1248	Cis
rs2284219	cg07186765	7	30714436	*CRHR2*	7	30633504	*GARS*	A	G	0.65	0.34	0.04	5.18E-16	0.05	1.59E-10	1248	Cis
rs2028299	cg22710306	15	90374257	*AP3S2*	15	90358103	*ANPEP*	C	A	0.70	-0.35	0.04	8.60E-16	0.05	2.62E-10	1248	Cis
rs10510110	cg25446361	10	124192430	*PLEKHA1*	10	1.24E+08	*HTRA1*	C	T	0.53	0.32	0.04	1.66E-15	0.05	5.00E-10	1248	Cis
rs7177055	cg17467968	15	77832762	*NA*	15	77711092	*NA*	G	A	0.72	0.35	0.04	1.95E-15	0.05	5.83E-10	1248	Cis
rs7845219	cg00807342	8	95937502	*TP53INP1*	3	10182884	*VHL*	T	C	0.51	0.31	0.04	2.74E-15	0.05	8.11E-10	1248	Trans
rs5215	cg09575421	11	17408630	*KCNJ11*	11	17415270	*ABCC8*	C	T	0.63	0.32	0.04	1.11E-14	0.05	3.14E-09	1248	Cis
rs35261542	cg03523917	6	20675792	*CDKAL1*	6	20662519	*CDKAL1*	A	C	0.74	-0.35	0.05	2.23E-14	0.05	6.19E-09	1248	Cis
rs7845219	cg05266843	8	95937502	*TP53INP1*	8	95961618	*TP53INP1*	T	C	0.51	0.30	0.04	3.32E-14	0.05	9.13E-09	1248	Cis
rs780094	cg05385684	2	27741237	*GCKR*	2	27651135	*NRBP1*	T	C	0.59	-0.30	0.04	3.61E-14	0.05	9.84E-09	1248	Cis
rs4275659	cg08176694	12	123447928	*NA*	12	1.24E+08	*PITPNM2*	T	C	0.70	0.34	0.04	3.90E-14	0.04	1.05E-08	1248	Cis
rs1513272	cg01883759	7	28200097	*NA*	7	28220576	*JAZF1*	C	T	0.51	0.30	0.04	5.26E-14	0.04	1.41E-08	1248	Cis
rs1359790	cg20100768	13	80717156	*SPRY2*	13	80706924	*NA*	A	G	0.73	0.34	0.05	1.05E-13	0.04	2.74E-08	1248	Cis
rs1535500	cg06543101	6	39284050	*KCNK16*	6	39290342	*KCNK16*	T	G	0.52	0.29	0.04	1.05E-13	0.04	2.74E-08	1248	Cis
rs319598	cg10942914	5	134240235	*PCBD2*	5	1.34E+08	*NA*	T	C	0.58	-0.30	0.04	1.99E-13	0.04	4.99E-08	1248	Cis
rs780094	cg18947209	2	27741237	*GCKR*	2	27960869	*NA*	T	C	0.59	0.29	0.04	2.06E-13	0.04	5.12E-08	1248	Cis
rs4275659	cg10672416	12	123447928	*NA*	12	1.24E+08	*C12orf65*	T	C	0.70	-0.32	0.04	3.28E-13	0.04	8.10E-08	1248	Cis
rs4275659	cg03328639	12	123447928	*NA*	12	1.23E+08	*ABCB9*	T	C	0.70	-0.32	0.04	4.13E-13	0.04	1.00E-07	1248	Cis
rs2972156	cg05293897	2	227117778	*NA*	2	2.27E+08	*NA*	C	G	0.63	0.30	0.04	6.08E-13	0.04	1.45E-07	1248	Cis
rs1513272	cg26102728	7	28200097	*NA*	7	28218933	*JAZF1*	C	T	0.51	-0.29	0.04	6.54E-13	0.04	1.54E-07	1248	Cis

Continuation Table S8-39

SNP	CPG	SNP Chr	SNP Position	SNP Gene	CPG Chr	CPG Position	CPG Gene	A1	A2	MAF	Estimate	SE	P	R^2	FDR	N	meQTL
rs35261542	cg14546115	6	20675792	*CDKAL1*	6	20817980	*CDKAL1*	A	C	0.74	-0.33	0.05	1.07E-12	0.04	2.51E-07	1248	Cis
rs319598	cg25134345	5	134240235	*PCBD2*	5	1.34E+08	*PCBD2*	T	C	0.58	0.29	0.04	3.05E-12	0.04	6.92E-07	1248	Cis
rs340874	cg20692998	1	214159256	*PROX1*	1	2.14E+08	*NA*	T	C	0.58	-0.28	0.04	3.43E-12	0.04	7.68E-07	1248	Cis
rs780094	cg03023068	2	27741237	*GCKR*	2	27486122	*SLC30A3*	T	C	0.59	0.28	0.04	3.62E-12	0.04	8.04E-07	1248	Cis
rs4273712	cg19875578	6	126964510	*C6orf173*	6	1.27E+08	*C6orf173*	G	A	0.72	0.31	0.04	3.66E-12	0.04	8.07E-07	1248	Cis
rs4275659	cg22341471	12	123447928	*NA*	12	1.23E+08	*ABCB9*	T	C	0.70	-0.31	0.04	3.76E-12	0.04	8.23E-07	1248	Cis
rs7177055	cg06616081	15	77832762	*NA*	15	77336938	*TSPAN3*	G	A	0.72	0.30	0.04	1.57E-11	0.04	3.40E-06	1248	Cis
rs2284219	cg03667083	7	30714436	*CRHR2*	7	30721010	*CRHR2*	A	G	0.65	0.28	0.04	2.23E-11	0.04	4.79E-06	1248	Cis
rs1535500	cg23372795	6	39284050	*KCNK16*	6	39284679	*KCNK16*	T	G	0.52	-0.26	0.04	2.84E-11	0.03	6.02E-06	1248	Cis
rs2284219	cg18669823	7	30714436	*CRHR2*	7	30725669	*NA*	A	G	0.65	-0.27	0.04	7.13E-11	0.03	1.48E-05	1248	Cis
rs319598	cg05713859	5	134240235	*PCBD2*	5	1.34E+08	*PCBD2*	T	C	0.58	0.27	0.04	7.90E-11	0.03	1.62E-05	1248	Cis
rs1535500	cg19475903	6	39284050	*KCNK16*	6	39271655	*KCNK17*	T	G	0.52	0.25	0.04	1.44E-10	0.03	2.90E-05	1248	Cis
rs4273712	cg21089903	6	126964510	*C6orf173*	6	1.27E+08	*C6orf173*	G	A	0.72	0.28	0.04	1.82E-10	0.03	3.64E-05	1248	Cis
rs7177055	cg27398640	15	77832762	*NA*	15	77910606	*LINGO1*	G	A	0.72	0.28	0.04	2.26E-10	0.03	4.49E-05	1248	Cis
rs340874	cg26293546	1	214159256	*PROX1*	1	2.14E+08	*PROX1*	T	C	0.58	-0.25	0.04	4.01E-10	0.03	7.81E-05	1248	Cis
rs780094	cg14242246	2	27741237	*GCKR*	2	27434262	*C2orf28*	T	C	0.59	-0.25	0.04	4.98E-10	0.03	9.63E-05	1248	Cis
rs10510110	cg25542438	10	124192430	*PLEKHA1*	10	1.24E+08	*ARMS2*	C	T	0.53	0.25	0.04	5.29E-10	0.03	1.02E-04	1248	Cis
rs6937795	cg25543459	6	137291281	*IL20RA*	6	1.37E+08	*NHEG1*	C	A	0.52	0.25	0.04	8.11E-10	0.03	1.55E-04	1248	Cis
rs10510110	cg09588434	10	124192430	*PLEKHA1*	10	1.24E+08	*PLEKHA1*	C	T	0.53	0.24	0.04	9.39E-10	0.03	1.78E-04	1248	Cis
rs4275659	cg22931309	12	123447928	*ABCB9*	12	1.24E+08	*C12orf65*	T	C	0.70	0.27	0.04	1.18E-09	0.03	2.21E-04	1248	Cis

Summary of main results obtained in the EWAS of T2D-SNPs

The strongest association was identified in *cis* at the SNP-CpG pair rs7845219-cg16049864 (estimate=0.94, SE=0.03, $p=2.79x10^{-163}$), mapping to the gene *TP53INP1*, while the weakest association was identified in *cis* at the SNP-CpG pair rs4275659-cg22931309 (estimate=0.27, SE=0.04, $p=1.18x10^{-9}$), encompassing the genes *ABCB9* and *C12orf65*. Regarding distance from the CpG, 97.2% of the meQTL identified were in *cis* and another 2.7% were in *trans*. The most representative SNPs in terms of the number of associations where they were detected, were the SNPs rs7800094 (*GCKR*, n=18 SNP-CpG pairs), rs4275659 (*ABCB9*, n=16 SNP-CpG pairs) and rs1535500 (*KCNK16*, n=11 SNP-CpG pairs). The average variation in methylation explained by T2D-SNPs was 0.12 (R^2 range 0.03 to 0.45). For the strongest meQTL, the SNPs were predominantly located within introns, intergenic regions, and upstream coding regions, while the CpG's were predominantly located within the body of the gene, in intergenic regions, or 1500bp from the TSS. With respect to their distance from the nearest CpG island, most of the CpG's were in the open sea or inside CpG islands.

Figure S8-24 One-to-many forest plot summarizing results of the 2SMR for five DMPs detected in borderline association with T2D across different MR methods. DMPs included in the outcome were detected observationally in the sensitivity analysis of the meta-EWAS of T2D (excluding KORA). P-value is the unadjusted P, and results were considered significant at $p < 0.001$ after Bonferroni correction ($\alpha = 0.05/48$ DMPs tested). Stronger evidence of causality was detected at the DMP cg20812370 (PBX1) using the IVW estimate.

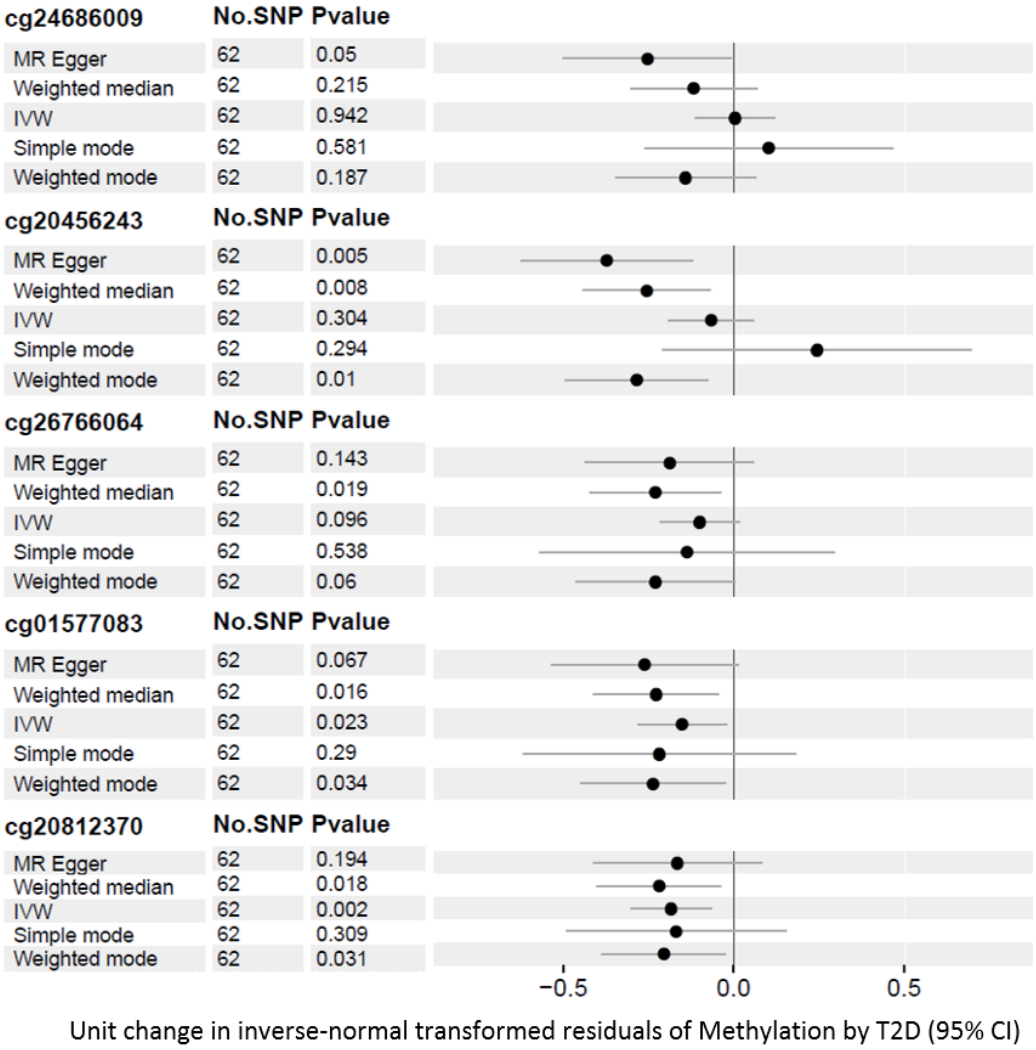


Figure S8-25 Forest plot (left) and funnel plot (right) showing results of the 2SMR for the effect of T2D on variation in methylation at the DMP cg10584271 in the ITHI1 gene. Results of the combined causal effect across different methods (red lines at the bottom of the forest plot), indicated no strong evidence of a causal association between T2D and this DMP. In addition, the SNP rs4275659 was shown to have an extreme negative effect on methylation, and it was previously demonstrated that this effect was strong, but independent of T2D. The funnel plot was asymmetric, suggesting evidence of horizontal pleiotropy, and this was attributed to the effect of the outlier SNP rs4275659.

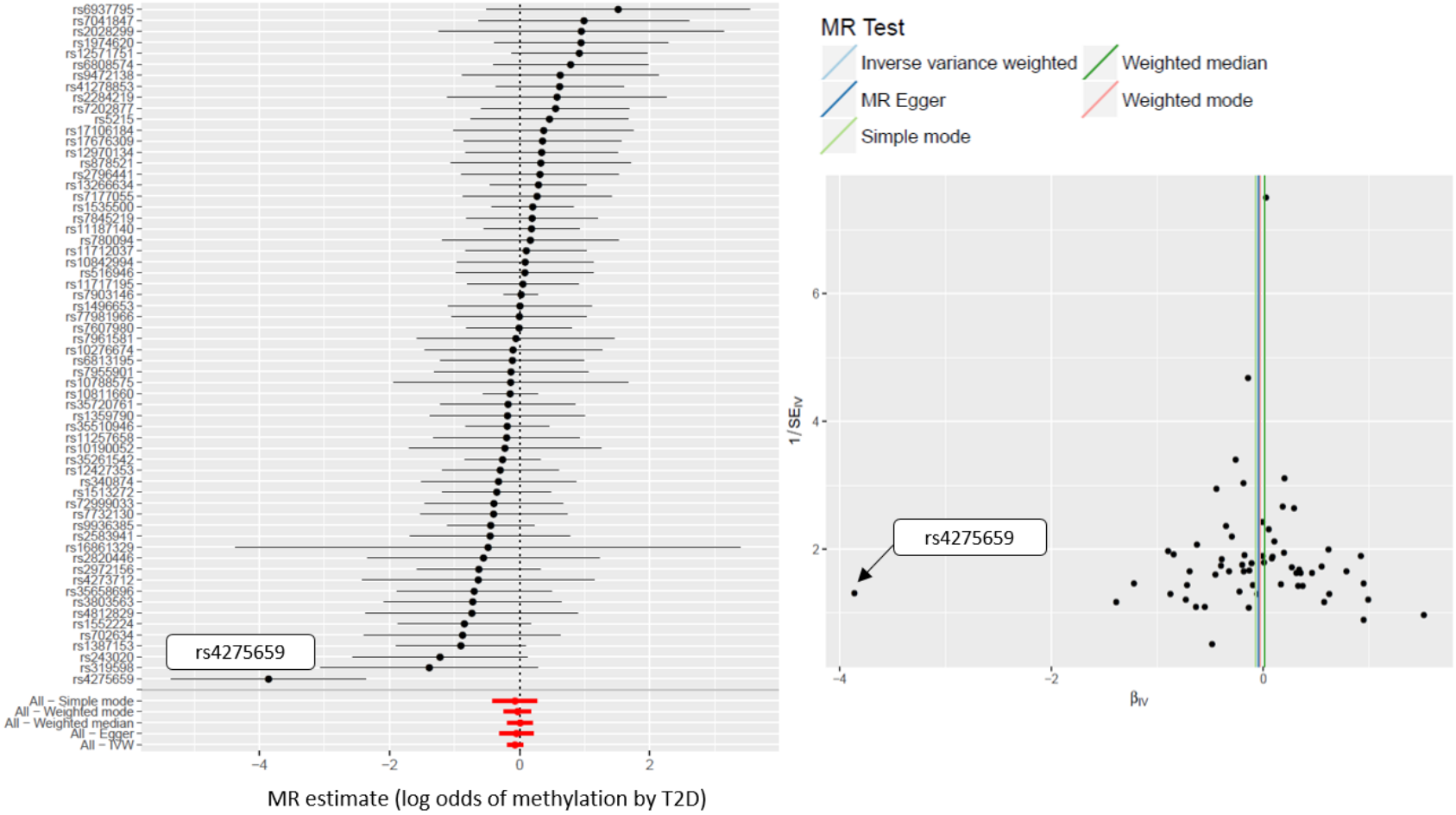


Table S8-40 Harmonized datasets for the reverse 2SMR. Harmonization was applied on MR-Base to guarantee using the same effect allele across the genotype-exposure and the genotype-outcome datasets.

Table S8-40.1 Harmonized dataset for DMPs identified in the meta-EWAS of T2D that were successfully instrumented by an meQTL in GoDMC. In total, 14 SNPs and 10 DMPs remained in the MR analysis after data harmonization.

Exposure	Gene	Outcome	Study‡	Proxy SNP	EA	OA	EAF Exp	EAF Out	Estimate Exposure	Estimate Outcome	N Exposure	N Outcome	P Exposure	P Outcome
cg06039489	*C20orf26*	T2D	UK Biobank	rs6081870	A	G	0.27	0.27	-0.14	0.00	24,388	113,116	7.27E-44	0.30
cg06039489	*C20orf26*	T2D	DIAGRAM	rs6081870	A	G	0.27	NA	-0.14	0.01	24,388	106,232	7.27E-44	0.50
cg06500161	*ABCG1*	T2D	UK Biobank	rs220182	T	C	0.55	0.54	0.06	0.00	24,474	114,976	1.00E-200	1.00
cg06500161	*ABCG1*	T2D	DIAGRAM	rs220182	T	C	0.55	NA	0.06	-0.01	24,474	91,284	1.00E-200	0.62
cg06468695	*CCDC42*	T2D	UK Biobank	rs140180165	T	C	0.97	0.97	-0.39	-0.18	15,366	115,743	1.00E-200	0.60
cg17155612	*LOC148189*	T2D	UK Biobank	rs56293553	A	G	0.81	0.80	-0.17	-0.10	26,699	117,370	4.67E-51	0.14
cg17155612	*LOC148189*	T2D	DIAGRAM	rs56293553†	A	G	0.81	NA	-0.17	0.00	26,699	104,275	4.67E-51	0.97
cg11851382	*PPAP2B*	T2D	UK Biobank	rs7535757	A	G	0.50	0.49	-0.08	0.10	26,658	115,683	1.00E-200	0.10
cg11851382	*PPAP2B*	T2D	DIAGRAM	rs7535757	A	G	0.50	NA	-0.08	0.02	26,658	102,682	1.00E-200	0.17
cg25741837	*SMYD5*	T2D	UK Biobank	rs62148128	A	G	0.06	0.06	0.71	0.10	20,444	116,700	1.00E-200	0.75
cg25741837	*SMYD5*	T2D	UK Biobank	rs6732515	A	C	0.98	0.98	0.54	0.29	22,690	116,210	1.08E-64	0.36
cg25741837	*SMYD5*	T2D	DIAGRAM	rs6732515†	A	C	0.98	NA	0.54	0.06	22,690	63,390	1.08E-64	0.35
cg01317029	*FAM131A*	T2D	UK Biobank	rs28421035	T	C	0.09	0.08	0.22	-0.13	25,851	117,708	2.42E-53	0.21
cg01317029	*FAM131A*	T2D	UK Biobank	rs35668024	A	C	0.97	0.98	-1.04	-0.59	25,970	117,261	1.00E-200	0.24
cg01317029	*FAM131A*	T2D	DIAGRAM	rs28421035	T	C	0.09	NA	0.22	0.07	25,851	10,854	2.42E-53	0.22
cg19693031	*TXNIP*	T2D	UK Biobank	rs6657798	C	G	0.80	0.82	-0.46	0.00	27,212	116,594	1.00E-200	0.71
cg07184465	*SPZ1*	T2D	UK Biobank	rs1500138	T	C	0.35	0.35	-0.15	-0.06	25,936	117,682	1.78E-58	0.07
cg07184465	*SPZ1*	T2D	DIAGRAM	rs1500138	T	C	0.35	NA	-0.15	0.01	25,936	104,490	1.78E-58	0.64
cg00082384	*NISCH*	T2D	UK Biobank	rs11716756	T	C	0.14	0.15	-0.24	-0.05	21,355	116,799	1.0E-200	0.36
cg00082384	*NISCH*	T2D	UK Biobank	rs35911561	T	C	0.89	0.89	0.12	0.12	22,455	117,393	6.63E-16	0.14
cg00082384	*NISCH*	T2D	DIAGRAM	rs11716756†	T	C	0.14	NA	-0.24	0.00	21,355	89,160	1.00E-200	0.95
cg00082384	*NISCH*	T2D	DIAGRAM	rs35911561	T	C	0.89	NA	0.12	0.01	22,455	10,853	6.63E-16	0.91

‡GWAS used to extract summary outcome data. From DIAGRAM it was the study by Mahajan et al. 2014, and from the UK-Biobank was the study by Wood et al. 2016.
†Proxied SNPs using effect estimates and allele frequencies from a second SNP identified in high correlation (LD > 0.8) with the target SNP in the outcome dataset.

Table S8-40.2 Harmonized dataset for DMPs identified in the sensitivity analysis of the meta-EWAS of T2D, which were successfully instrumented by an meQTL in GoDMC. In total, 32 SNPs and 24 DMPs remained in the MR analysis after data harmonization.

Exposure	Gene	Outcome	Study	Proxy SNP	EA	OA	EAF Exp	EAF Out	Estimate Exposure	Estimate Outcome	N Exposure	N Outcome	P Exposure	P Outcome
cg18181703	*SOCS3*	T2D	UK Biobank	rs4383852	A	G	0.52	0.52	-0.130	3.05E-02	27,746	117,641	2.3E-56	0.42
cg18181703	*SOCS3*	T2D	DIAGRAM	rs4383852†	A	G	0.52	NA	-0.130	0.00E+00	27,746	104,379	2.3E-56	0.98
cg27037013	*Unannotated*	T2D	UK Biobank	rs13051329	T	C	0.15	0.15	0.170	-5.13E-02	26,837	117,504	1.0E-200	0.38
cg27037013	*Unannotated*	T2D	DIAGRAM	rs13051329	T	C	0.15	NA	0.170	0.00E+00	26,837	104,328	1.0E-200	0.80
cg25536676	*DHCR24*	T2D	UK Biobank	rs174551	T	C	0.66	0.65	0.110	0.00E+00	24,653	117,520	5.1E-30	0.40
cg25536676	*DHCR24*	T2D	UK Biobank	rs6681644	C	G	0.42	0.42	0.260	9.53E-02	27,714	117,094	1.0E-200	0.05
cg25536676	*DHCR24*	T2D	UK Biobank	rs79365581	T	C	0.98	0.99	0.350	7.34E-01	5,834	117,420	7.8E-07	0.18
cg25536676	*DHCR24*	T2D	DIAGRAM	rs174551†	T	C	0.66	NA	0.110	3.92E-02	24,653	104,539	5.1E-30	0.00
cg01577083	*Unannotated*	T2D	UK Biobank	rs1107095	T	C	0.51	0.51	-0.200	0.00E+00	23,764	110,804	7.5E-109	0.41
cg14476101	*PHGDH*	T2D	UK Biobank	rs347903	T	C	0.67	0.65	0.230	0.00E+00	24,554	111,060	1.0E-200	0.42
cg14476101	*PHGDH*	T2D	UK Biobank	rs608358	A	C	0.28	0.28	-0.280	0.00E+00	25,566	115,429	5.8E-180	0.78
cg14476101	*PHGDH*	T2D	DIAGRAM	rs347903†	T	C	0.67	NA	0.230	-2.96E-02	24,554	107,656	1.0E-200	0.06
cg14476101	*PHGDH*	T2D	DIAGRAM	rs608358	A	C	0.28	NA	-0.280	0.00E+00	25,566	101,903	5.8E-180	0.97
cg10082515	*Unannotated*	T2D	UK Biobank	rs1525502	T	C	0.41	0.39	-0.250	9.53E-02	24,607	115,383	1.1E-172	0.10
cg10082515	*Unannotated*	T2D	DIAGRAM	rs1525502	T	C	0.41	NA	-0.250	1.98E-02	24,607	105,669	1.1E-172	0.12
cg08857797	*VPS25*	T2D	UK Biobank	rs1047891	A	C	0.32	0.32	-0.140	-7.26E-02	24,138	117,775	8.8E-47	0.04
cg08857797	*VPS25*	T2D	DIAGRAM	rs1047891†	A	C	0.32	NA	-0.140	-3.92E-02	24,138	24,243	8.8E-47	0.15
cg20231084	*Unannotated*	T2D	UK Biobank	rs750129	A	G	0.47	0.47	0.140	-1.01E-02	23,944	96,646	1.0E-200	0.80
cg20231084	*Unannotated*	T2D	DIAGRAM	rs750129	A	G	0.47	NA	0.140	0.00E+00	23,944	106,230	1.0E-200	0.77
cg07212837	*Unannotated*	T2D	UK Biobank	rs56261297	T	C	0.41	0.41	-0.340	0.00E+00	27,738	117,775	1.0E-200	0.27
cg07212837	*Unannotated*	T2D	DIAGRAM	rs5626129†	T	C	0.41	NA	-0.340	-2.96E-02	27,738	93,701	1.0E-200	0.07

† Proxied SNPs using effect estimates and allele frequencies from a second SNP identified in high correlation (LD > 0.8) with the target SNP in the outcome dataset.

Continuation Table S8-40.2 Harmonized dataset for DMPs identified in the sensitivity analysis of the meta-EWAS of T2D.

Exposure	Gene	Outcome	Study	Proxy SNP	EA	OA	EAF Exp	EAF Out	Estimate Exposure	Estimate Outcome	N Exposure	N Outcome	P Exposure	P Outcome
cg13178597	RGS17	T2D	UK Biobank	rs540908	A	G	0.82	0.81	-0.100	0.00E+00	26,442	116,958	6.4E-18	0.47
cg12593793	*Unannotated*	T2D	UK Biobank	rs11584621	A	T	0.21	0.21	-0.060	-1.01E-02	25,084	113,096	1.0E-200	0.86
cg24512093	*ROBO1*	T2D	UK Biobank	rs9309801	T	C	0.34	0.33	-0.110	0.00E+00	27,235	116,463	1.0E-200	0.42
cg24512093	*ROBO1*	T2D	UK Biobank	rs9831014	C	G	0.42	0.42	0.120	-2.02E-02	24,994	115,128	1.0E-200	0.58
cg24512093	*ROBO1*	T2D	DIAGRAM	rs9309801	T	C	0.34	NA	-0.110	0.00E+00	27,235	104,489	1.0E-200	0.81
cg11376147	*SLC43A1*	T2D	UK Biobank	rs2848634	A	G	0.75	0.75	0.090	0.00E+00	27,749	117,587	7.9E-18	0.68
cg11376147	*SLC43A1*	T2D	DIAGRAM	rs2848634	A	G	0.75	NA	0.090	-1.98E-02	27,749	100,365	7.9E-18	0.21
cg16765088	*Unannotated*	T2D	UK Biobank	rs7496161	A	G	0.11	0.10	-0.320	-1.74E-01	25,984	116,171	4.7E-122	0.03
cg16765088	*Unannotated*	T2D	DIAGRAM	rs7496161	A	G	0.11	NA	-0.320	9.95E-03	25,984	98,254	4.7E-122	0.51
cg00144180	*HDAC4*	T2D	UK Biobank	rs11693641	A	C	0.50	0.49	-0.150	-4.08E-02	23,360	117,061	1.0E-200	0.18
cg00144180	*HDAC4*	T2D	UK Biobank	rs1872614	A	T	0.58	0.58	-0.050	-9.53E-02	16,512	104,810	2.2E-06	0.04
cg00144180	*HDAC4*	T2D	DIAGRAM	rs11693641†	A	C	0.50	NA	-0.150	-1.98E-02	23,360	102,563	1.0E-200	0.18
cg11024682	*SREBF1*	T2D	UK Biobank	rs11652574	A	G	0.04	0.03	1.100	-1.98E-01	19,085	117,291	1.0E-200	0.42
cg10584271	*ITIH1*	T2D	UK Biobank	rs115738369	T	C	0.02	0.02	-1.740	-3.05E-02	22,070	117,430	1.0E-200	0.95
cg10584271	*ITIH1*	T2D	UK Biobank	rs62250760	T	C	0.35	0.35	0.050	-3.05E-02	25,857	116,430	9.1E-08	0.42
cg10584271	*ITIH1*	T2D	DIAGRAM	rs62250760†	T	C	0.35	NA	0.050	-1.98E-02	25,857	104,476	9.1E-08	0.15
cg01963618	*LOC285768*	T2D	UK Biobank	rs6596785	A	G	0.89	0.90	0.220	1.01E-02	25,226	115,281	2.2E-50	0.95
cg01963618	*LOC285768*	T2D	DIAGRAM	rs6596785	A	G	0.89	NA	0.220	9.95E-03	25,226	104,567	2.2E-50	0.54
cg00896068	*Unannotated*	T2D	UK Biobank	rs113786621	T	C	0.08	0.08	-0.340	9.53E-02	25,956	116,849	1.0E-200	0.30
cg00896068	*Unannotated*	T2D	UK Biobank	rs9525281	C	G	0.76	0.75	0.140	6.19E-02	18,956	109,280	3.2E-30	0.16
cg20456243	*SPEG*	T2D	UK Biobank	rs55760516	A	G	0.67	0.68	-0.120	0.00E+00	27,242	117,775	1.0E-200	0.40
cg20456243	*SPEG*	T2D	DIAGRAM	rs55760516†	A	G	0.67	NA	-0.120	-2.96E-02	27,242	103,923	1.0E-200	0.04
cg09185884	*KCTD2*	T2D	UK Biobank	rs71380866	C	G	0.97	0.97	0.210	3.05E-02	25,599	117,447	8.7E-15	0.92
cg11252555	*RPL13AP5*	T2D	UK Biobank	rs10421294	A	G	0.10	0.09	-0.140	9.53E-02	25,481	117,526	2.8E-20	0.37
cg11252555	*RPL13AP5*	T2D	DIAGRAM	rs10421294†	A	G	0.10	NA	-0.140	3.92E-02	25,481	97,947	2.8E-20	0.09
cg27115863	*Unannotated*	T2D	UK Biobank	rs6000773	C	G	0.75	0.74	-0.120	0.00E+00	24,389	116,130	1.0E-200	0.28
cg27115863	*Unannotated*	T2D	UK Biobank	rs7602568	T	C	0.22	0.23	0.200	-3.05E-02	27,625	117,775	1.0E-85	0.45
cg27115863	*Unannotated*	T2D	DIAGRAM	rs7602568	T	C	0.22	NA	0.200	0.00E+00	27,625	110,135	1.0E-85	0.91

† Proxied SNPs using effect estimates and allele frequencies from a second SNP identified in high correlation (LD > 0.8) with the target SNP in the outcome dataset.

Table S8-40.3 Harmonized dataset DMPs identified in the EWAS of T2D in ALSPAC, which were successfully instrumented by an meQTL in GoDMC. In total, 5 SNPs and 4 DMPs remained in the MR analysis after data harmonization.

Exposure	Gene	Outcome	Study‡	Proxy SNP	EA	OA	EAF Exp	EAF Out	Estimate Exposure	Estimate Outcome	N Exposure	N Outcome	P Exposure	P Outcome
cg26652413	*CPAMD8*	T2D	UK Biobank	rs773865	A	G	0.24	0.24	0.432	0.000	23,231	116,634	1.00E-200	0.92
cg26652413	*CPAMD8*	T2D	DIAGRAM	rs773865†	A	G	0.24	NA	0.432	-0.010	23,231	110,198	1.00E-200	0.29
cg00204249	*DNAH17*	T2D	UK Biobank	rs72903323	A	G	0.17	0.17	-0.150	-0.030	26,469	117,039	8.81E-38	0.60
cg00204249	*DNAH17*	T2D	DIAGRAM	rs72903323†	A	G	0.17	NA	-0.150	0.010	26,469	100,493	8.81E-38	0.70
cg14045803	*STARD10*	T2D	UK Biobank	rs4459332	T	C	0.53	0.52	0.185	0.073	26,453	117,504	1.55E-102	0.02
cg14045803	*STARD10*	T2D	DIAGRAM	rs4459332†	T	C	0.53	NA	0.185	-0.010	26,453	100,586	1.55E-102	0.64
cg15986668	*NFYC*	T2D	UK Biobank	rs115582802	T	C	0.10	0.10	-0.191	-0.062	21,291	115,690	1.00E-200	0.48
cg15986668	*NFYC*	T2D	UK Biobank	rs3767953	C	G	0.90	0.89	-0.246	0.010	18,890	116,493	1.32E-46	0.88
cg15986668	*NFYC*	T2D	DIAGRAM	rs115582802†	T	C	0.10	NA	-0.191	-0.010	21,291	92,569	1.00E-200	0.78

† Proxied SNPs using effect estimates and allele frequencies from a second SNP identified in high correlation (LD > 0.8) with the target SNP in the outcome dataset.

Figure S8-26 Scatterplot and forest plot illustrating results of the reverse Mendelian randomization for the effect of methylation on T2D at three DMPs detected in the Meta-EWAS of T2D. These DMPs were instrumented by two SNPs using meQTL data from the GoDMC consortium. Results correspond to the IVW-estimate.

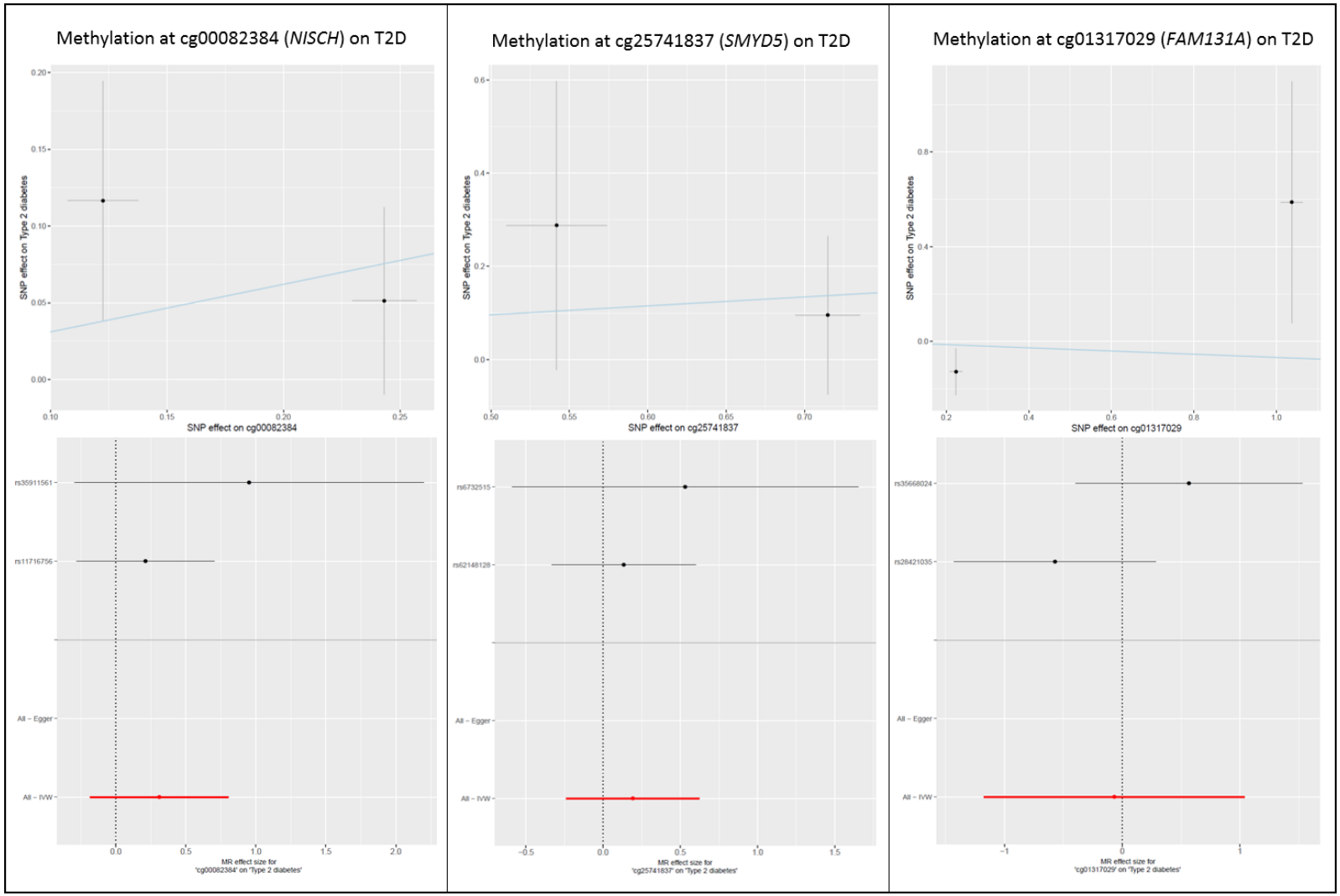
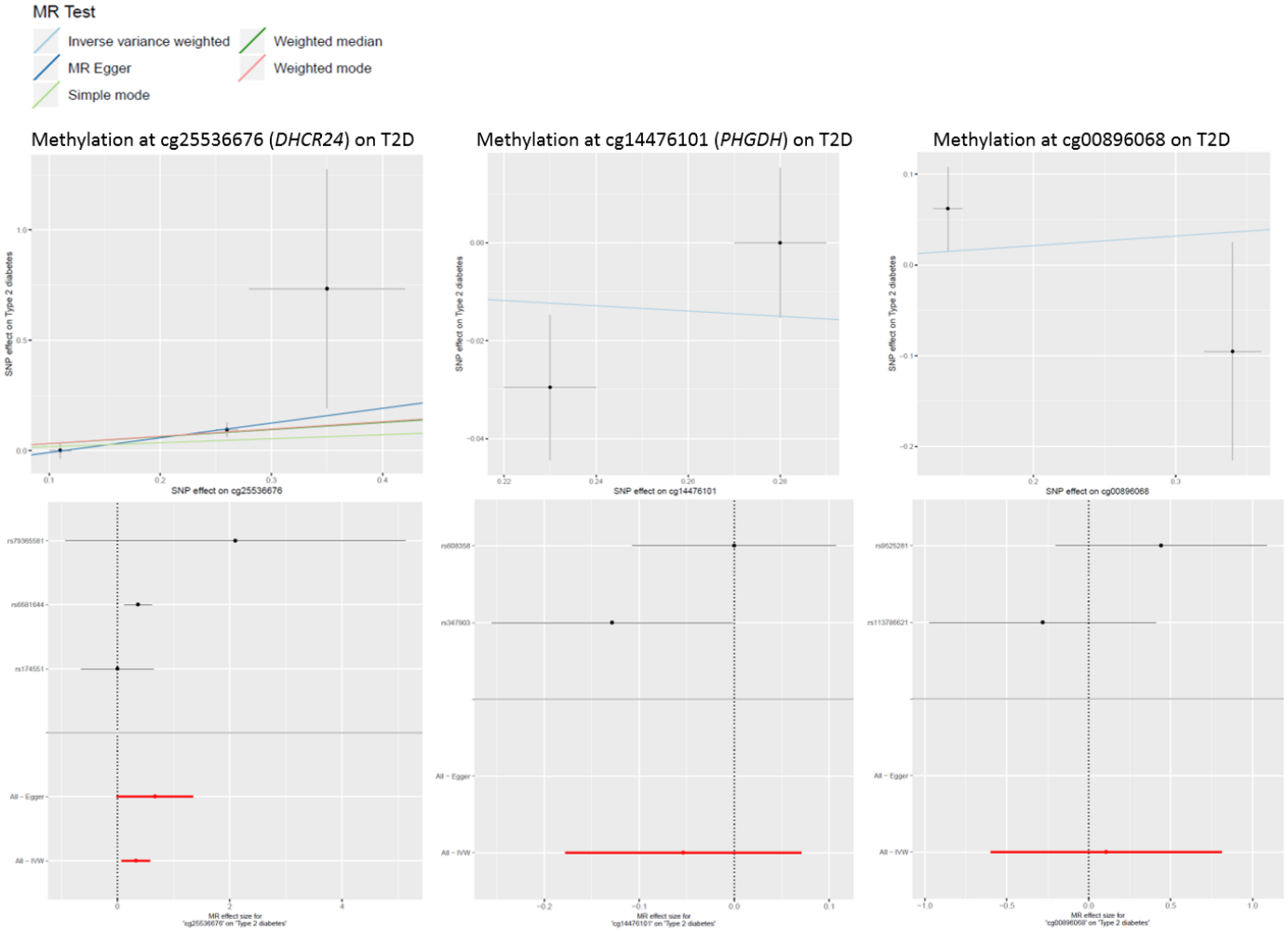


Figure S8-27 Scatterplot (top line) and forest plot (bottom line) illustrating the effect of methylation on T2D for three DMPs identified in the Sensitivity analysis of the Meta-EWAS of T2D. These DMPs were instrumented by two or three SNPs (cg25536676 in DHCR24) reported by GoDMC. For the DMP in DHCR24, additional methods were implemented and are represented by the different regression lines in the scatterplot. Forest plot shows results of the individual SNPs, and the combined causal effect using the MR-Egger estimate or the IVW estimate.



Continuation Figure S8-27. Forest plot illustrating the effect of methylation on T2D for DMPs instrumented by at least two SNPs. Results were obtained using the IVW estimate.

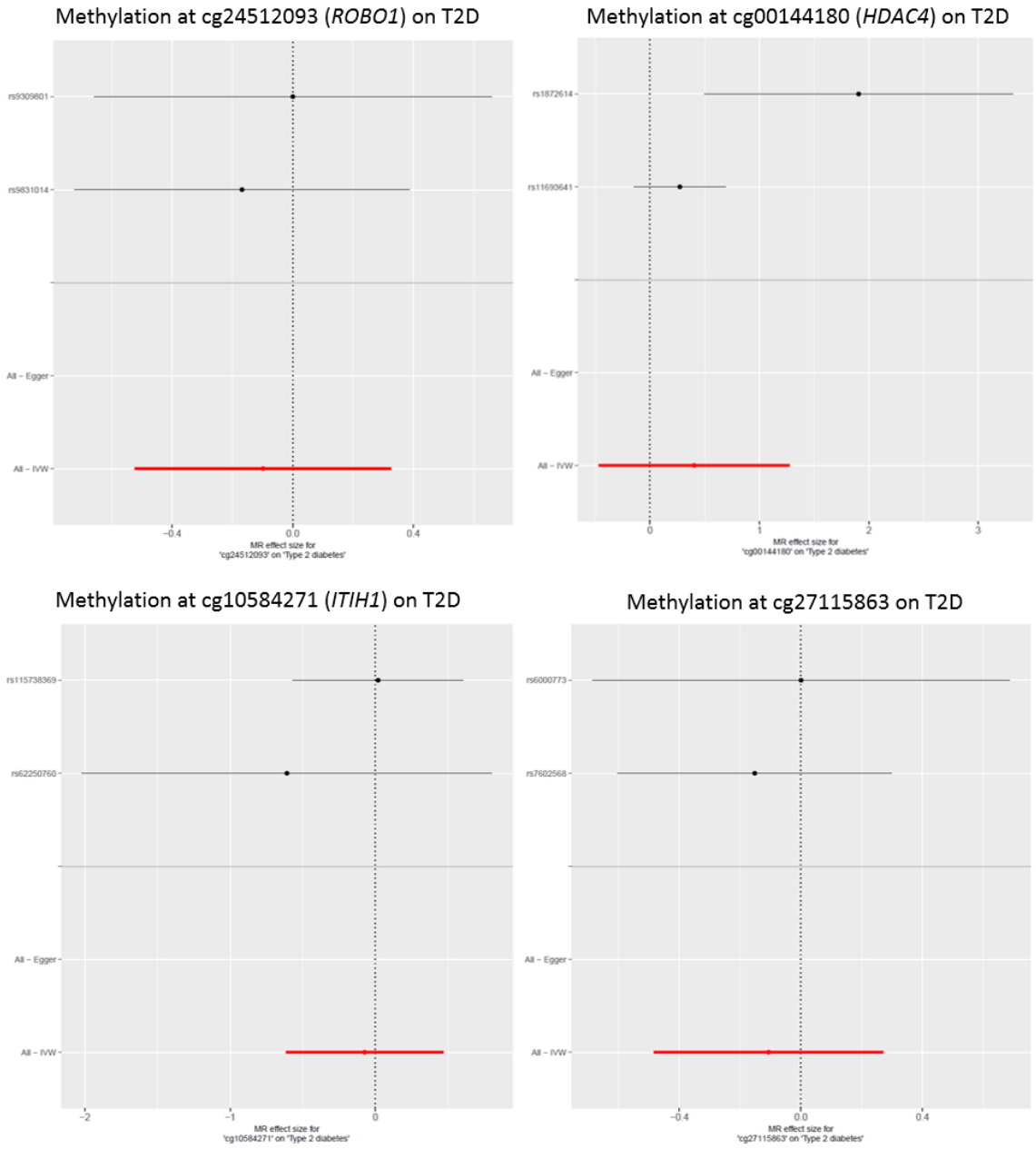


Table S8-41 Results of the reverse 2SMR for the effect of methylation at the DMP cg25536676 (DHCR24) on T2D using additional sensitivity methods. The DMP in DHCR24 was observationally identified in the sensitivity meta-EWAS of T2D. Results were considered significant at p<0.002 after Bonferroni correction

Other Method	OR (95%CI)	P
MR-Egger†	1.95(0.99,3.85)	0.305
Weighted Median	**1.38(1.09,1.74)**	**0.008**
Weighted Mode	1.39(1.08,1.79)	0.125

†Egger Intercept=-0.08, p=0.48. Ruckers' Q test=1.15, p=0.28.

Figure S8-28 Scatterplot and forest plot illustrating results of the reverse 2SMR for the effect of methylation at the DMP cg1598668 (NFYC) on T2D. Results were obtained using the IVW estimate, and no association was identified with nominal causality at the DMP in NYFC.

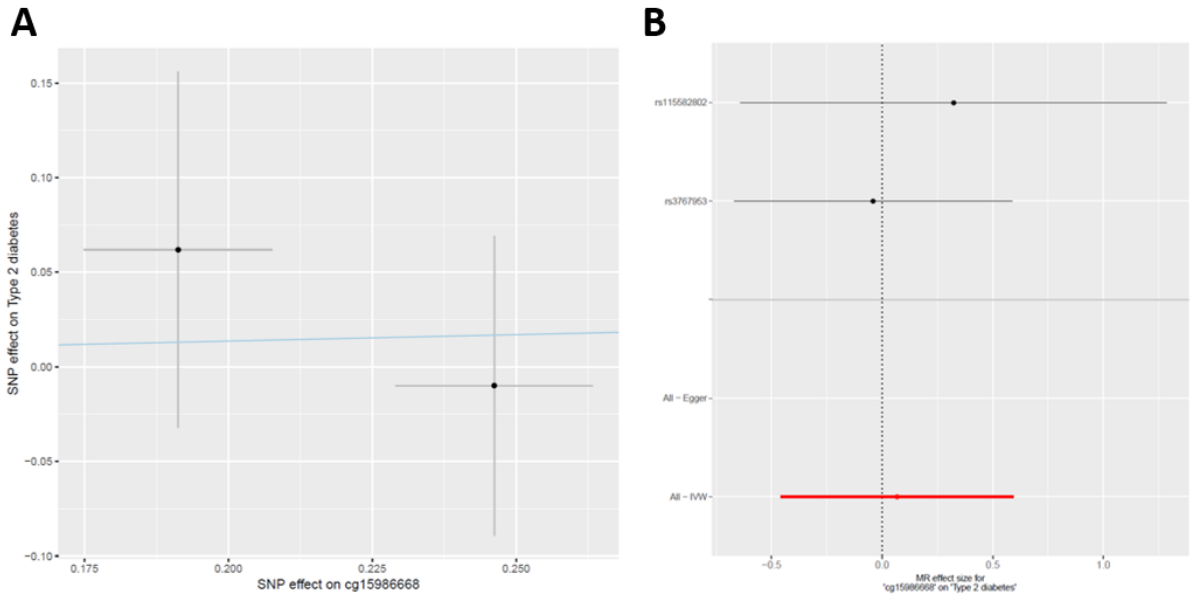


Table S8-42 Description of top 20 pathways reported by KEGG in relation to DMPs identified in the bidirectional MR. Results are presented separately for DMPs detected in the forward and in the reverse MR.N: total number of genes reported for a pathway with respect to those included as background genes in the analysis, DE: number of query genes matching to genes reported for a pathway, P: unadjusted P-value of enrichment. Enrichment was regarded significant at FDR < 0.05. Pathways shown here were nominally enriched.

DMPs Forward MR				**DMPs reverse MR**			
Description	N	DE	P	Description	N	DE	P
Cushing syndrome	3	2	0.06	Metabolic pathways	4	2	0.02
Bacterial invasion of epithelial cells	1	1	0.17	Choline metabolism in cancer	1	1	0.09
Pathogenic Escherichia coli infection	1	1	0.17	Ether lipid metabolism	1	1	0.09
Shigellosis	1	1	0.17	Fat digestion and absorption	1	1	0.09
Tight junction	1	1	0.17	Fc gamma R-mediated phagocytosis	1	1	0.09
Focal adhesion	1	1	0.17	Glycerolipid metabolism	1	1	0.09
Leukocyte trans-endothelial migration	1	1	0.17	Glycerophospholipid metabolism	1	1	0.09
MAPK signalling pathway	1	1	0.17	Phospholipase D signalling pathway	1	1	0.09
Neurotrophin signalling pathway	1	1	0.17	Sphingolipid metabolism	1	1	0.09
Pancreatic secretion	1	1	0.17	Steroid biosynthesis	1	1	0.10
Renal cell carcinoma	1	1	0.17	ABC transporters	1	0	1.00
Cortisol synthesis and secretion	1	1	0.17	AGE-RAGE signalling pathway in diabetic complications	1	0	1.00
Transcriptional mis regulation in cancer	1	1	0.17	Amphetamine addiction	1	0	1.00
Spliceosome	1	1	0.18	Amyotrophic lateral sclerosis (ALS)	1	0	1.00
NOD-like receptor signalling pathway	1	1	0.23	Antigen processing and presentation	1	0	1.00
Proteoglycans in cancer	2	1	0.27	Apoptosis	1	0	1.00
Long-term potentiation	2	1	0.29	Arginine and proline metabolism	1	0	1.00
Rap1 signalling pathway	2	1	0.29	Arginine biosynthesis	1	0	1.00
Ras signalling pathway	2	1	0.29	Autophagy - other	1	0	1.00
Chemokine signalling pathway	2	1	0.30	Bacterial invasion of epithelial cells	1	0	1.00

Table S8-43 Traits identified in association with meQTL using a SNP lookup in the PheWAS application tool (MR-Base). MeQTL identified in association with other traits were identified as proxies for five DMPs detected in the bidirectional MR. Reporting only the top 39 associations identified at genome-wide significance level ($p<5.0x10^{-8}$).

Trait	Estimate	SE	P	N
Other polyunsaturated fatty acids than 18:2	-0.383	0.012	1.19E-210	13,549
Ratio of bisallylic groups to double bonds	-0.373	0.012	4.81E-198	13,524
Ratio of bisallylic groups to total fatty acids	-0.348	0.013	5.30E-167	13,171
Average number of methylene groups per double bond	0.309	0.012	2.44E-135	13,532
Average number of double bonds in a fatty acid chain	-0.251	0.012	3.58E-102	15,728
Average number of methylene groups in a fatty acid chain	0.213	0.011	7.25E-83	19,021
Omega-3 fatty acids	-0.148	0.013	7.28E-31	13,544
18:2, linoleic acid (LA)	0.145	0.013	9.44E-30	13,527
Standing height	-0.020	0.002	1.94E-29	336,474
22:6, docosahexaenoic acid	-0.122	0.013	2.26E-21	13,499
Ankle spacing width	0.021	0.002	7.40E-20	463,010
Mean diameter for HDL particles	-0.095	0.011	1.25E-18	19,273
Comparative height size at age 10	-0.014	0.002	2.93E-17	332,021
Sitting height	-0.016	0.002	4.32E-17	336,172
Phospholipids in very large HDL	-0.090	0.011	4.72E-17	19,273
Total lipids in very large HDL	-0.090	0.011	5.51E-17	19,273
Impedance of arm (right)	0.015	0.002	5.71E-17	331,279
Mean diameter for VLDL particles	0.089	0.011	1.12E-16	19,273
Impedance of arm (left)	0.015	0.002	2.64E-16	331,292
Free cholesterol in very large HDL	-0.085	0.010	3.02E-16	21,542
Concentration of large HDL particles	-0.087	0.011	8.87E-16	19,273
Cholesterol esters in very large HDL	-0.086	0.011	9.92E-16	19,273
Total cholesterol in very large HDL	-0.082	0.010	1.89E-15	21,540
Total lipids in large HDL	-0.086	0.011	2.18E-15	19,273
Phospholipids in large HDL	-0.084	0.011	7.65E-15	19,273
Free cholesterol in large HDL	-0.081	0.010	9.84E-15	21,559
Concentration of very large HDL particles	-0.083	0.011	1.75E-14	19,273
Total cholesterol in large HDL	-0.077	0.010	1.82E-13	21,558
Cholesterol esters in large HDL	-0.079	0.011	2.26E-13	19,273
Impedance of whole body	0.014	0.002	3.39E-13	331,284
Triglycerides in very large HDL	-0.076	0.010	3.84E-13	21,536
Pulse rate automated reading	0.019	0.003	1.01E-12	317,756
Total cholesterol in HDL	-0.067	0.010	1.87E-10	21,555
Apolipoprotein A-I	-0.067	0.011	3.92E-10	20,687
Trunk fat-free mass	-0.009	0.002	3.82E-09	331,030
Trunk predicted mass	-0.009	0.002	7.61E-09	330,995
Arm fat-free mass (right)	-0.008	0.002	3.36E-08	331,221
Operation code: cholecystectomy/gall bladder removal	-0.002	0.000	3.80E-08	462,933
Ankle spacing width (left)	0.017	0.003	4.50E-08	463,010