Investigating the Role of Nuclear Encoded Mitochondrial Genes in the Onset of Type 2 Diabetes

Hannah Elizabeth Billaux Maude

December 9, 2020

Imperial College London, Section of Genetics & Genomics

Department of Metabolism, Digestion and Reproduction

PhD in Clinical Medical Research (Medicine)

Declaration of Originality

I declare that this thesis was composed by myself and that the work presented is entirely my own, unless otherwise stated. I further declare that this work has not been submitted for any other degree or previous qualification.

Copyright Declaration

Acknowledgements

Firstly, I would like to sincerely thank my primary Ph.D. Supervisor, Dr. Toby Andrew, to whom I will be forever grateful for the past five years of mentorship. I could not have had a more supportive supervisor. The hours of scientific conversation are not only fond memories, but the reason I can call myself a geneticist.

Secondly, to my Ph.D. supervisors, Dr. Derek Huntley, Dr. Filippo Tamanini and Professor Marjo-Riitta Järvelin. To Dr. Huntley, who has been a part of my academic journey for the longest, thank you for the years of support and guidance; I will miss the conversations while assisting in your classes. I credit my pursuit of Bioinformatics to your enthusiasm. Thank you to Dr. Tamanini for the years of support and feedback and to Professor Marjo-Riitta Järvelin for supporting my Ph.D. I am additionally grateful to both Dr. Andrew and Dr. Huntley for providing me with countless opportunities to both teach and supervise students. I learnt as much as I taught and hope to use this experience to mentor future students as well as you mentored me.

My sincere thanks also goes to Professor Nikolas Maniatis, under whose mentorship I began this project. Thank you for welcoming me into your group, for providing invaluable feedback and for sharing your enthusiasm for science. I greatly admire your work, on which this project is founded.

I am grateful to Dr. Kourosh Ahmadi and Dr. Ines Cebola, who have offered support both through collaborations and in my pursuit of a scientific career. Thank you for giving your time and guidance. Thank you to my Ph.D. examiners, Professor Richard Festenstein and Professor Elizabeth Shephard and to Professor Laki Buluwela and Nousheen Tariq, who run the Imperial College MRC DTP (Medical Research Council Doctoral Training Partnership) Research Studentships.

Finally, to my family and friends. Thank you for filling my life with joy, which has always been reflected in my approach to my Ph.D. To my parents, none of my achievements would have been possible without your selfless support and encouragement. Thank you.

This work was supported by the Medical Research Council, United Kingdom (Imperial MRC DTP Ph.D. Studentship 2016-2020).

Abstract

Mitochondrial dysfunction has long been implicated in Type 2 diabetes (T2D). This relationship appears to be bidirectional, with evidence that mitochondrial dysfunction is both caused by and causal of T2D-related phenotypes. A potential causal role in T2D onset would be supported by evidence of a genetic predisposition to mitochondrial dysfunction, since inherited genetic risk factors precede and contribute to disease onset. Here, a genetic study design is used to investigate the potential role of T2D-associated genetic risk loci (T2D loci) in disrupting mitochondrial function through the altered expression of nuclear-encoded mitochondrial genes (NEMGs). The mitochondria are targeted by multiple T2D drugs and therefore such loci may be informative for effective treatment and prevention measures. The functional *cis*–genes regulated by T2D loci were identified based on the co-location of T2D loci with adipose tissue expression quantitative trait (eQTL) within a genetic distance of 1 LDU. T2D loci and eQTL were previously mapped using LDU-based gene mapping, which is compared and contrasted in this thesis to other popular tests of association. 50 of the identified T2D *cis*–genes were NEMGs and implicated a number of pathways in the inherited risk of T2D, including the relevant pathway of branched-chain amino acid catabolism. These same 50 genes were enriched for decreased expression in T2D cases compared to controls in independent gene expression datasets. Compared to the total known NEMGs, the 50 *cis*-NEMGs showed further enrichment for decreased expression, suggesting that T2D-eQTL co-location may identify specific subsets of causal genes. Finally, a candidate T2D locus associated with the *cis*–NEMG *ACAD11* was fine-mapped using targeted sequence data for 94 T2D cases and 94 controls. Several candidate causal variants were identified, including two low-frequency haplotypes, one of which contained both an *ACAD11* splicing mutation and a mutation predicted to disrupt the observed binding of HNF4A and COUP-TFII within the *ACAD11* promoter region.

Contents

List of Figures

List of Tables

Abbreviations

AA = African American

ATP = adenosine triphosphate

BAT = brown adipose tissue

BCAA = branched chain amino acid

BMI = body mass index

CADD = Combined Annotation Dependent Depletion (score)

CD/CV = common disease, common variant (hypothesis)

CD/RV = common disease, rare variant (hypothesis)

ChiP = chromatin immunoprecipitation

cM = centimorgan

CNV = copy number variant

DAG = diacylglycerol

DGE = differential gene expression

DGI = Diabetes Genetics Initiative

DIAGRAM = Diabetes Genetics Replication and Meta-analysis (Consortium)

DNA = deoxyribonucleic acid

ER = endoplasmic reticulum

eSNP = expression single nucleotide polymorphism

ETC = electron transport chain

eQTL = expression quantitative trait loci/locus

FA = fatty acid

FAO = fatty acid oxidation

FDR = false discovery rate

FFA = free fatty acid

FH = family history

FUSION = Finland-United States Investigation of NIDDM Genetics

FWER = family wise error rate

GEO = Gene Expression Omnibus (database)

GoT2D = Genetics of T2D (Consortium)

GRS = genetic risk score

GSEA = gene set enrichment analysis

GTEx = Genotype-Tissue Expression (Project)

GWAS = genome-wide association study

HapMap = Haplotype Map (The International HapMap Consortium)

HepG2 = liver hepatocellular cells

HRC = Haplotype Reference Consortium

HUGO = HUGO Gene Nomenclature Committee

HWE = Hardy-Weinberg equilibrium

IGT = impaired glucose tolerance

INDEL = insertion or deletion

IR = insulin resistant

KB = kilobase(s)

LCFA = long-chain fatty acid

LD = linkage disequilibrium

LDU = linkage disequilibrium unit(s)

MAF = minor allele frequency

MODY = maturity onset diabetes of the young

mRNA = messenger RNA

MTC = metabochip

mtDNA = mitochondrial DNA

NEMG = nuclear encoded mitochondrial gene(s)

NIDDK = National Institute of Diabetes and Digestive and Kidney Diseases

NIDDM = non-insulin-dependent diabetes mellitus

REM = random effects model

RMA = robust multi-array averaging

ROS = reactive oxygen species

RNA = ribonucleic acid

$\hat{S}$ = estimated causal variant location

$\hat{S}_{T2D}$ = estimated location of T2D-associated variant

$\hat{S}_{eQTL}$ = estimated location of variant associated with gene expression

SIFT = sorting intolerant from tolerant (algorithm)

SNP = single nucleotide polymorphism

SNV = single nucleotide variant

SMR = single marker regression

SV = structural variant

TCA = tricarboxylic acid cycle

TFBS = transcription factor binding site

TG = triglyceride

Treg = T regulatory (cell)

tRNA = transfer RNA

T2D = type 2 diabetes

T2D-GENES = T2D Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples

VCF = variant call file

VLCFA = very long chain fatty acid

WAT = white adipose tissue

WES = whole exome sequence(ing)

WGS = whole genome sequence(ing)

WTCCC/WTC = Wellcome Trust Case Control Consortium

1 Introduction

1.1 Preface

This thesis will present research regarding the relationship between the human genome, the mitochondria and Type 2 diabetes (T2D). The following Chapter 1 will introduce T2D, along with the relevant literature which motivates our hypothesis that the dysfunction of the mitochondria is a heritable component of T2D aetiology[1]. The Introduction concludes by presenting the hypothesis, along with aims and a research plan.

The following study aims to identify genetic mechanisms which increase the risk of T2D through altered mitochondrial function. A genetic study design will be used to map genetic loci associated with the risk of T2D (T2D loci) and the (*cis*–)genes which they may regulate, as a follow-up to the study published by Lau et al. (2017). The genetic design used here and by Lau et al. will be presented in Chapter 2 and will be compared and contrasted with other gene mapping methods. As such, Chapter 2 will review published gene mapping studies of T2D, including the methods used and how they have influenced and contributed to our current understanding of T2D genetics. The Chapter concludes by reviewing the Lau et al. (2017) study and work completed prior to this project.

Subsequently, this thesis follows the discovery of genetic associations (Chapter 2) by identifying the downstream effects of T2D loci at a single-gene and pathway level (Chapter 3), complimented by independent validation (Chapter 4) and fine-mapping (Chapter 5). A breakdown of the aims for each Chapter is provided in **Section 1.7: Hypothesis, Aims and Research Plan**.

This thesis complements ongoing efforts to understand in depth the genetic factors which predispose to T2D. This goal is becoming increasingly important in light of growing evidence that genetic heterogeneity may underpin T2D phenotypic heterogeneity, including distinct aetiologies as well as the risk of diverse complications and treatment efficacies (discussed in **Section 1.4: Type 2 diabetes: a spectrum, a palette or a cluster?**).

[1] Aetiology - the cause, or causes of a disease or condition.

With the continued development of high-throughput tools and large datasets, pooling data and applying informative study designs will be crucial steps towards developing a detailed understanding of T2D risk. The methods used in this study can be widely applied to study complex disease genetics and their biological consequences.

1.2 Type 2 diabetes: a modern disease

Diabetes, scientifically known as *diabetes mellitus* is a collection of conditions recognised by high levels of the sugar, glucose, in the blood. The prevalence and financial impact of diabetes has been rising over the past decades, making it of critical importance to better understand the underlying causes and risk factors in order to treat and prevent this devastating disease. In 2019, the International Diabetes Federation (IDF) estimated that 463 million people were affected by diabetes worldwide (Saeedi et al., 2019). Diabetes prevalence was estimated to rise from 151 million to 463 million between the years 2000 and 2019 (Atlas, 2015). The prevalence of diabetes is higher in urban areas and in high-income countries (Saeedi et al., 2019), which effectively translates to a high health expenditure (Williams et al., 2020). In 2019, it was estimated that the global expenditure on diabetes was 760 billion US dollars. The UK annual expenditure is estimated at 10% of the annual NHS budget (Diabetes, 2014) and for the 2010/2011 year was estimated at £1 billion for Type 1 and £8.8 billion for Type 2 diabetes, with an additional £0.9 billion and £13 billion in indirect costs, respectively (Hex et al., 2012).

To briefly describe the history of diabetes, it was first reported as a condition of excessive sweet-tasting urine (polyuria) and excessive thirst. It is believed that the first mention of diabetes was around 1500 BC in the Ancient Egyptian medical text, the Ebers Papyrus (Marwood, 1973) and it was later described by the Greek Physician, Aretaeus of Cappadocia, as a "melting down of the flesh and bones into urine" (Adams et al., 1856). It is to Aretaeus that the name *diabetes* is attributed, which translates roughly from Greek as 'to pass through' or 'siphon' (Gemmill, 1972). In 1675, the celebrated Physician Thomas Willis described diabetes in his book Pharmaceutice rationalis, in a Chapter famously named 'The Pissing Evil' (Allan, 1953). He added the term *mellitus* after describing the

sweetness of the urine, which translates from Latin as 'honey' or 'sweetened with honey'. It was the English Physician Matthew Dobson who, in 1776, eventually recognised the residue from boiled urine as sugar. Dobson also noted the sweetness of blood serum, opposing the previously held opinion that diabetes was a disease of the kidneys (von Engelhardt, 1989). Diabetes *insipidus* describes a separate rare condition caused by a deficiency of the antidiuretic hormone, but which is also characterised by polyuria and excess thirst.

The metabolism of glucose is controlled largely by the hormone insulin, which is released by the pancreas in response to high blood glucose levels. Insulin release stimulates the storage of glucose in the liver and muscles and signals for the liver to stop the endogenous production of glucose. Diabetes may result when sufficient insulin is not produced (insulin deficiency), or when it is not effective (insulin resistance). Since its discovery in 1922[2], insulin has been used to effectively treat insulin deficient forms of diabetes, largely characterised as Type 1 diabetes (insulin-dependent) (Vecchio et al., 2018). Type 1 diabetes is an autoimmune condition in which the immune system attacks and destroys the pancreatic β-cells which produce insulin. In comparison, Type 2 diabetes (also known as non-insulin-dependent diabetes mellitus, NIDDM) is characterised by resistance to the action of insulin.

The American Diabetes Association classifies diabetes into four categories: (1) Type 1 diabetes (T1D), (2) Type 2 diabetes (T2D), (3) gestational diabetes mellitus (GDM)[3] and (4) other specific types including monogenic, syndromic and induced forms, such as neonatal diabetes, maturity onset diabetes of the young (MODY), latent onset autoimmune diabetes (LADA), Wolfram syndrome, Alström syndrome, cystic fibrosis-related diabetes (CFRD) and maternally inherited diabetes and deafness (MIDD) (Association et al., 2014, 2015). $\sim$90% of diabetes cases have T2D (Saeedi et al., 2019).

T2D, on which this study focusses, is largely characterised by a resistance to the effects of

[2]The discovery is accredited to Fred Banting and Charles Best, as well as J. J. R. Macleod and Bert Collip, however the Romanian Nicholai Paulescu described a pancreatic extract in 1919 which cured the Diabetic symptoms of dogs with their pancreases removed.

[3]Gestational diabetes mellitus (GDM) is diagnosed when women develop diabetes during pregnancy.

insulin accompanied by defective insulin production by pancreatic β-cells. This may occur when the increased insulin production is no longer sufficient to compensate for increased peripheral insulin resistance (Lyssenko et al., 2008). A diagnosis of 'pre-diabetes' is given to individuals with insulin resistance; the development of β-cell dysfunction may facilitate the transition to overt diabetes (Kolb and Martin, 2017). Pre-diabetes is a risk factor for T2D (Fletcher et al., 2002; Edwards and Cusi, 2016). The current rates of T2D, which have been likened to a 'global pandemic' (Unnikrishnan et al., 2017; Ali et al., 2017), have been attributed to an ageing population and modern lifestyle with increased calorific intake and decreased physical activity, since these are major risk factors for T2D (discussed in the next sections).

1.3 Type 2 diabetes: a complex disease

T2D is a complex disease, which is defined as being influenced by both environmental and genetic factors. Several models have been used to describe this genetic contribution, including the polygenic model in which common genetic variants shared across the entire population additively increase risk, or the genetic heterogeneity model in which closely related individuals have unique genetic risk factors. Environmental risk factors for T2D include obesity, age and physical inactivity, as well as diet, short or disturbed sleep, smoking, stress, depression and low socioeconomic status (Kolb and Martin, 2017). More recent studies have provided evidence of T2D subtypes, which may have their own unique risk factors such as age or BMI (Ahlqvist et al., 2018; Udler et al., 2018; Udler, 2019). These subtypes include those characterised predominantly by insulin resistance or insulin deficiency[4]. Additional risk factors for T2D include ethnic background and a positive family history; these are discussed in more detail below.

1.3.1 T2D heritability

The current study aims to better understand the genetic risk factors which predispose to T2D. However, there must be evidence that a trait or disease is influenced by genetic

[4]Type 1 diabetes is the classical 'insulin-deficient' phenotype, however this is diagnosed by the autoimmune destruction of the pancreatic β-cells and therefore the presence of autoantibodies. β-cell dysfunction in the absence of autoantibodies may be diagnosed as T2D.

factors before a genetic study can be carried out. For T2D, this evidence is extensive. Firstly, T2D clusters in families (Zimmet, 1982; Harlan et al., 1987; Morris et al., 1989). The lifetime risk of developing T2D increases to ∼40% if one parent is affected and to ∼70% if both parents are affected (Köbberling and Tattersall, 1982) compared to a population risk of ∼10% (Saeedi et al., 2019). Similarly, the odds ratio[5] for developing T2D with one affected parent has been estimated at ∼3.5 and with both affected parents at ∼6 (Meigs et al., 2000).

The sibling relative risk (RR)[5] λ_s, which is the risk of developing T2D if one sibling is affected compared to the general population, has been estimated at ∼3 (Köbberling and Tattersall, 1982; Hemminki et al., 2010). This number dramatically increases depending on the number of affected first-degree relatives (Hemminki et al., 2010). Crucially, adopted individuals showed no increased risk from adoptive parents, suggesting that T2D risk is driven by genetic, rather than environmental factors (Hemminki et al., 2010). A more recent study showed that individuals with a family history of T2D were less likely to have an average BMI and waist-hip ratio, despite being more likely to partake in regular exercise and to have healthy diets (Choi et al., 2019). These results highlight that the risk from a positive family history may even outweigh preventative lifestyle changes, emphasising the need to understand the underlying genetic mechanisms.

Another study design used to interrogate the extent of genetic risk is twin studies. Monozygotic twins, who share all of their DNA, are compared to dizygotic twins who share on average half of their DNA. Estimates for T2D proband concordance, which reflects the proportion of affected individuals with an affected twin (concordant twin pairs)[6], have ranged from 34-58% in monozygotic twins and 16-37% for dizygotic twins (Newman et al., 1987; Kaprio et al., 1992; Poulsen et al., 1999). In a 15 year follow-up, Medici et al. (1999) reported that 76% of originally discordant monozygotic twin pairs, i.e. in

[5]Several different measures quoted here include the odds ratio (OR) and relative risk (RR). OR is the odds of observing disease given a particular exposure (an affected parent, for example), compared to the odds of disease in the absence of the exposure (no affected parent). RR is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group.

[6]Probandwise concordance is calculated as 2C/(2C+D), where C is the number of concordant pairs and D is the number of discordant pairs.

which only one sibling was affected, had both developed diabetes after 15 years. When considering impaired glucose tolerance, the concordance after 15 years was 96%. The higher probability that monozygotic twins will share the same diagnosis is presumed to reflect their shared DNA.

Twin studies and other study designs have also been used to estimate the proportion of trait phenotypic variance which is explained by genetic factors; this is termed the heritability. Heritability estimates are founded on the assumption that T2D and other complex traits are polygenic, in that risk results from the additive effects of independent, multifactorial risk factors. These additive effects are modelled as a normal distribution known as liability (Falconer, 1967; Pearson and Lee, 1901), at which some threshold results in disease onset if crossed (if an individual accumulates a critical number of additive risk factors). This is shown in Figure 1. The "heritability" describes the proportion of the total latent liability which is attributed to genetic risk factors.

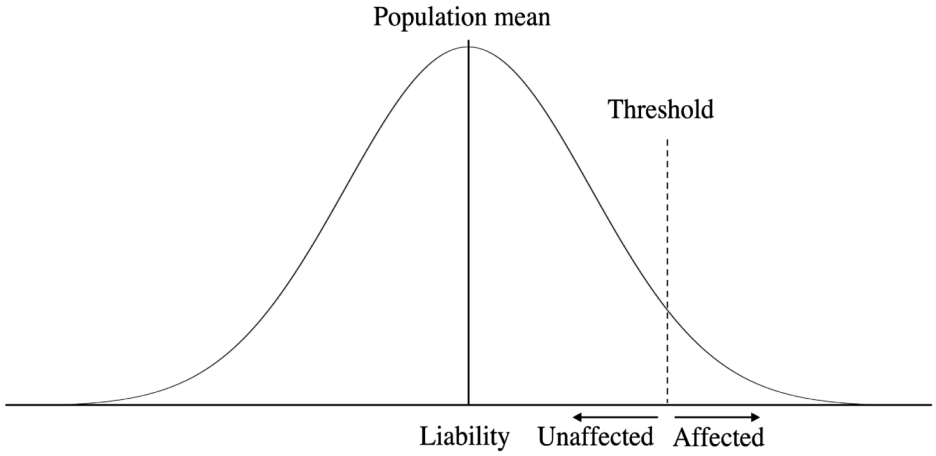


Figure 1: Liability, which represents the cumulative effects of additive risk factors is modelled as a normal distribution. The threshold represents the liability at which disease onset occurs. For a binary trait, such as T2D, the threshold corresponds to the population prevalence.

Estimates of the heritability of T2D liability (T2D heritability) range from ∼25% to ∼80% (Prasad and Groop, 2015). A recent meta-analysis of nearly 35,000 twin pairs estimated the T2D heritability at 72% (95% confidence interval 61-78%) (Willemsen et al., 2015); this is similar to an estimate of 69% by Almgren et al. (2011) for individuals with age-of-onset between 35 and 60 years. Almgren et al. (2011) further showed that the

estimated heritability for T2D differed depending on the age-of-onset, consistent with previous reports (Falconer, 1967; Simpson, 1969). There have been several criticisms, however, about using twin and family data to estimate heritability. These include the potential overestimation of risk for outbred population based on families in whom diabetes clusters (Prasad and Groop, 2015) and are discussed further in Chapter 2, **Section 2.6: Missing Heritability**.

In light of this evidence, many genetic studies have aimed to identify genetic variants which increase the risk of T2D; these are the focus of the following Chapter 2. Large-scale genetic studies may also estimate the proportion of the total estimated trait heritability explained by the identified variants, assuming the same liability scale (see Mahajan et al. (2018) and Vujkovic et al. (2020) for T2D). This approach is used to indicate how many of the additive genetic risk factors have been identified, on the assumption that all relevant variants will have been identified when the total heritability is explained.

1.3.2 The debated evolution of T2D

There have been several hypotheses which address the increasing rates of insulin resistance (IR) and T2D. One popular proposal is that of the 'thrifty gene', in which evolution favoured the selection of fat-storing genotypes in times of limited food and malnutrition, although these have a negative impact in the modern energy-rich and high-calorie environment (Neel, 1962). There is evidence that such selection may have influenced lipid and carbohydrate metabolism, selecting for thrifty (fat storage) mechanisms (Voight et al., 2006; Blekhman et al., 2008; Luca et al., 2010; Rubio-Ruiz et al., 2015; James et al., 2019) and metabolic rate is itself highly heritable (Pettersen et al., 2018). Following criticism of the 'thrifty gene', Speakman (2008) proposed the 'drifty gene' hypothesis, which explained the genetic propensity to obesity as a result of genetic drift following the removal of *Homo sapiens* predators and the resulting selective pressures. Alternatively, the 'thrifty phenotype' hypothesis proposed that T2D risk was driven by environmental factors early in life, including poor fetal growth and malnutrition (Hales and Barker, 1992, 2001). This hypothesis is of particular interest since there is strong evidence in support

and early-life interventions may prove preventative (Vaag et al., 2012).

IR may itself have had an evolutionary advantage. For example, IR may promote glucose availability for the inflammatory response during trauma and prevent harmful protein loss during starvation (Soeters and Soeters, 2012), or may act as a socioeconomic mechanism, diverting glucose to the brain and placenta to improve cognitive ability, invest in offspring (Watve and Yajnik, 2007)[7] and suppress physical aggression (Belsare et al., 2010). The 'carnivore connection' hypothesis stated that IR may have evolved when transitioning from a high-glucose diet to a low-carbohydrate diet in the ice ages, to which the adaptive response is IR (Miller and Colagiuri, 1994). Regardless of potential evolutionary explanations, the modern prevalence of T2D has been attributed to an overwhelmingly sedentary lifestyle (Freese et al., 2017).

1.4 Type 2 diabetes: a spectrum, a palette or a cluster?

More recently, the nature of T2D itself has been debated. T2D is often the diagnosis if patients do not fit into any other diabetes category, such as Type 1 or monogenic forms (Association et al., 2014). T2D is classically associated with obesity and other risk factors including age and a sedentary lifestyle, however recent studies have revealed a new level of complexity and have identified distinct subtypes of T2D. This complexity and heterogeneity of T2D is likely to impact any study aiming to identify the underlying causes of disease, including gene mapping.

One such study was reported by Ahlqvist et al. (2018), in which five distinct, replicable subgroups of T2D patients were identified based on the clustering of six clinical variables: glutamate decarboxylase antibodies, age at diagnosis, BMI, HbA_{1c}, and homoeostatic model assessment 2 (HOMA2) estimates of β-cell function and insulin resistance. These clusters were assigned (1) severe autoimmune diabetes, (2) severe insulin-deficient diabetes, (3) severe insulin-resistant diabetes, (4) mild obesity-related diabetes and (5) mild age-related diabetes. Importantly, the authors showed that the patients in each subgroup

[7]Watve and Yajnik (2007) described a 'behavioural switch' hypothesis in which insulin resistance facilitates a switch from 'soldier' to 'diplomat' by stimulating low birth but fast growth rates, with more glucose diverted to the brain.

presented different risks of micro and macro-vascular complications and distinct genetic associations. These clusters were effectively replicated by Dennis et al. (2019), however the authors questioned their clinical utility since simple clinical features such as age of diagnosis were as or more effective at predicting clinical outcomes and selecting optimal therapies. Van Smeden et al. (2018) also questioned the use of data-driven clustering to define subgroups, warning against over-interpreting clusters of somewhat correlated clinical measures. Both Dennis et al. and van Smeden et al. advocated the use of continuous clinical measures, rather than defining distinct categories which may still contain heterogeneity. Despite this, the clusters reported by Ahlqvist et al. (2018) support the idea that distinct risk factors may contribute to distinct prognoses.

T2D clusters can also be identified directly from genotypes. Udler et al. (2018) clustered T2D-variant associations for variants known to be associated with T2D (n = 94) and diabetes-related traits (n = 47). The authors reported five clusters, of which two contained trait-associations characteristic of reduced β-cell function and three were associated within insulin resistance, assigned as obesity-mediated, lipodystrophy-like and disrupted liver lipid metabolism. Similarly to the analysis of Ahlqvist et al. (2018), the clusters were associated with distinct clinical outcomes. The authors found that around a third of individuals with T2D could be assigned uniquely to one cluster based on genetic risk score and that these individuals had somewhat distinct phenotypes. This approach is of particular interest, since genetic risk precedes disease onset and may be of more use for preventative medicine, while the implicated genotypes can give direct insight into underlying mechanisms (Udler, 2019).

Alternatively, the 'palette model' of T2D proposed by McCarthy (2017) notes that distinct pathophysiological processes related to T2D risk, such as fat distribution, β-cell function and insulin resistance etc, may contribute disproportionately to risk in different individuals. The model illustrates this as a multidimensional space where individual risk is made up of the distinct component parts. Similarly to Udler et al. (2018), McCarthy suggests that initial efforts might benefit from identifying individuals whose risk is dominated by a restricted set of processes. However, a debate remains as to the current clinical utility of

the proposed T2D subtypes. On the other hand, deconstructing T2D heterogeneity has profound implications for discovery, for example by increasing the power of gene mapping by studying homogenous groups of patients.

Given the evidence that genetic heterogeneity may underlie phenotypic heterogeneity, it is of upmost importance to continue the effort to map and characterise genetic risk factors associated with T2D. The vision provided by McCarthy (2017); Udler et al. (2018); Ahlqvist et al. (2018) and others will continue to be realised as the biological pathways perturbed by distinct genetic risk factors are revealed.

1.5 Mitochondrial dysfunction in T2D

With this in mind, the current study aims to investigate and better understand the genetic mechanisms which underlie a specific biological process associated with T2D risk: mitochondrial dysfunction.

This study follows the observation by Lau et al. (2017) that many of the putative functional genes regulated by T2D loci were involved in mitochondrial function, potentially implicating inherited changes in gene expression and mitochondrial function in T2D risk. The following sections will review mitochondrial function and its role in diabetes, with emphasis on particular mechanisms which have been functionally implicated in diabetes onset and as such might be expected to be identified in further genetic analyses. The data obtained from Lau et al. for the current study included measures of gene expression in subcutaneous adipose. Therefore, the role of mitochondrial function in adipose and T2D will be discussed in detail.

1.5.1 Mitochondria: the powerhouse of the cell

The mitochondria are maternally inherited organelles present in eukaryotic cells. They are enclosed by an outer lipid membrane and an inner lipid membrane which contorts into folds called cristae. Mitochondria contain their own DNA (mtDNA); a circular genome of 15,569bp which contains 13 protein-coding genes, 22 tRNAs, and 2 rRNAs. It is thought that the mitochondria likely originated from an endosymbiotic relationship between an α-Proteobacteria contained by a larger, essentially eukaryotic host (Lane and Martin,

2010; Gray, 2012). Each mitochondrion can have between one and ten copies of the circular mtDNA molecule (Cole, 2016), which was originally much larger prior to the transfer of mitochondrial genes into the nuclear genome over time[8] (Martin, 2003). The mitochondrial proteome is now predominantly encoded by the nucleus, with the human nuclear genome containing up to 1,500 genes which encode proteins imported to the mitochondria (Calvo et al., 2015b); these are known as nuclear-encoded mitochondrial genes (NEMGs). The nuclear genome may also regulate transcription of the mtDNA (Ali et al., 2019).

The mitochondria have multiple functions, of which the most well-known is the production of energy as molecules of ATP (adenosine triphosphate). The numbers of mitochondria in each human cell can vary dramatically depending on the energy requirements, from hundreds to hundreds of thousands (Cole, 2016). The mitochondria are dynamic and undergo fusion, fission and degradation and can form greater fused networks (Rafelski, 2013; Zamponi et al., 2018). These dynamics can be regulated by metabolic processes (Mishra and Chan, 2016). Aside from their classic role in energy production, the mitochondria also contain various metabolic reactions including β-oxidation of fatty acids, branched-chain amino acid catabolism and production of steroids (McBride et al., 2006), making them particularly pertinent to metabolic disorders including T2D.

1.5.2 Mitochondrial function and T2D

A potential relationship between diabetes and mitochondrial function was first reported by Applegarth and Koneff (1946), with Yamada et al. (1975) describing a relationship between glucose intolerance and impaired function of the liver mitochondria. One of the clearest links between the mitochondria and diabetes can be seen in the manifestation of maternally-inherited forms of diabetes resulting from mutations in the mitochondrial genome (mtDNA) (Maassen et al., 2004). One example is maternally inherited diabetes and deafness (MIDD) (van den Ouweland et al., 1994) which results from mutations

[8]The colocation of gene and gene product for redox regulation of gene expression (CoRR) hypothesis proposes that the protein-coding genes retained in the mtDNA, which encode subunits of the electron transport chain, allow for swift response to environmental changes by rapid changes in the expression of these genes and subsequent energy production (Allen, 2015).

in one of three mitochondrial genes: MT-TL1, MT-TK, or MT-TE. These three genes encode transfer RNAs (tRNAs), which transport the amino acids leucine, lysine and glutamic acid, respectively, to translating polypeptide chains within the mitochondria. Mitochondrial dysfunction[9] has, as a result, been highly researched as a potential cause of common T2D. This is the focus of a large number of review papers (see Morino et al. (2006); Kim et al. (2008); Sivitz and Yorek (2010); Newsholme et al. (2012); Szendroedi et al. (2012); Montgomery and Turner (2015); Wada and Nakatsuka (2016); Gonzalez-Franquesa and Patti (2017); Sergi et al. (2019) for some examples of general reviews). The evidence can be broadly described as observational and functional, with some examples provided below.

Observational studies: it is very well established that perturbed mitochondrial function can be observed in individuals with T2D. This includes lower activity of mitochondrial oxidation and enzymes (Kelley et al., 2002; Ritov et al., 2005; Heilbronn et al., 2007), decreased expression of genes involved in maintaining mitochondrial function and decreased mitochondrial size (Kelley et al., 2002; Patti et al., 2003; Morino et al., 2005; Heilbronn et al., 2007; Zorzano et al., 2009), as well as alterations in the mitochondrial proteome (Chae et al., 2018). However, mitochondrial dysfunction is not observed in all cases of T2D (Holloway et al., 2007; Boushel et al., 2007). Furthermore, the observed mitochondrial dysfunction may be induced by diabetes itself (Hoeks et al., 2010; Fujimaki and Kuwabara, 2017; Haythorne et al., 2019), leading many to ask whether mitochondrial dysfunction is a cause or consequence of T2D (Turner and Heilbronn, 2008; Dumas et al., 2009), even in recent reviews[10]. While some studies observe mitochondrial dysfunction only in long-standing T2D patients (Van Tienen et al., 2012), others have observed mitochondrial dysfunction in the healthy offspring of individuals with T2D (Mootha et al., 2003; Petersen et al., 2004, 2005; Morino et al., 2005).

[9]Mitochondrial dysfunction has been used to refer to the altered expression of mitochondrial genes, altered content of mitochondria, mtDNA or mitochondrial proteins, altered enzymatic activity of key mitochondrial proteins, changes in mitochondrial size or shape, a change in ATP production and different rates of substrate oxidation (Montgomery and Turner, 2015).

[10]A quote from Sergi et al. (2019): "However, whether these mitochondrial defects represent a cause or a consequence of insulin resistance in skeletal muscle remains to be fully elucidated".

FUNCTIONAL STUDIES: in contrast to observational studies, functional studies in rodent models have demonstrated that T2D-related phenotypes such as insulin resistance or deficiency can result from directly perturbing mitochondrial function. Examples include reduced glucose uptake caused by the direct perturbation of oxidative phosphorylation electron chain subunits (Lim et al., 2006) and that mtDNA substitutions altered glucose tolerance (Pravenec et al., 2007). Increased reactive oxygen species (ROS) may enhance insulin signalling (Loh et al., 2009). However, mitochondrial oxidative stress has also been shown to impair insulin signalling and drive insulin resistance in rodent models (Hurrle and Hsu, 2017; Fazakerley et al., 2018) and may impact pancreatic β-cell function by mediating glucose toxicity and the death of β-cells (reviewed by Kaneto et al. (2010); Ma et al. (2012)). The mitochondria of both mice and humans receiving a high-fat diet release higher levels of ROS (Anderson et al., 2009). However, treating mice with mitochondria-targeted antioxidants was shown to both preserve insulin sensitivity (Anderson et al., 2009; Lee et al., 2010) but also cause chronic hyperinsulinaemia (Wang et al., 2008). Oxidative stress may be present with insulin resistance, even when other measures of mitochondrial function are normal (Samocha-Bonet et al., 2012). Altered fatty acid β-oxidation has been shown to induce insulin resistance (this is discussed in detail in Chapter 5). Inhibiting mitochondrial fission improved the insulin sensitivity of obese mice (Jheng et al., 2012) as well as insulin signalling in diabetes-susceptible cybrid cells (Lin et al., 2018a). Furthermore, inhibiting mitophagy (the removal of dysfunctional mitochondria) can cause insulin resistance in mice (Drew et al., 2014) (reviewed by Su et al. (2019)). Proper mitochondrial oxidation, ATP production, calcium release and mitochondrial dynamics are crucial for the functioning of pancreatic β-cells, including glucose-stimulated insulin release and the prevention of apoptosis (Soejima et al., 1996; Tsuruzoe et al., 1998; Kennedy et al., 1998; Zhang et al., 2001; Noda et al., 2002; Joseph et al., 2006; Gauthier and Wollhejm, 2006; Molina et al., 2009; Fridlyand and Philipson, 2010) (reviewed by Wiederkehr and Wollheim (2008)).

CONCLUSIONS and GENETIC STUDIES: the above cites just some studies of mitochondrial function in relation to T2D aetiology. Based on this evidence, the most

plausible scenario is that of a bidirectional relationship in which mitochondrial dysfunction, depending upon the context, can both contribute to and be induced by T2D. While mitochondrial dysfunction may not be necessary for insulin resistance or β-cell dysfunction, it may be sufficient to cause it. A further question may be whether the conflicting observations can be explained by the heterogenous nature of T2D, such that mitochondrial dysfunction may contribute at varying extents to T2D in different individuals (see the 'palette model' proposed by McCarthy (2017)). This may reflect variation in the predisposition to mitochondrial dysfunction, its penetrance, or in lifestyle factors such as nutrient intake or energy requirements.

An additional question is whether a genetic predisposition to mitochondrial dysfunction exists. Interestingly, NEMGs are enriched in genes found to be regulated by genetic variation (*cis*-genes), highlighting a potential important role of genetic variation in regulating mitochondrial function (Sajuthi et al., 2016). While the observation of mitochondrial dysfunction in the healthy offspring of T2D patients hints at a heritable mechanism, this could feasibly result from the early, asymptomatic onset of disease or be induced by other shared familial risk factors. Functional studies are also limited to the conditions in which they were carried out and may not always reflect physiological effects. Alternatively, genetic studies can demonstrate inherited mechanisms which precede and causally contribute to disease onset in the human population.

DNA mutations are known to cause monogenic and maternally-inherited forms of diabetes. In common T2D, genetic variants may, for example, mimic the effects of the functional perturbations described above such as increased ROS production, decreased fat oxidation, increased mitochondrial fission and may reduce the natural compensation of mitochondrial function in response to high-fat diets or insulin resistance (Katic et al., 2007; Sergi et al., 2019). The mitochondria are already of therapeutic importance since several current T2D drugs directly improve mitochondrial function (Yaribeygi et al., 2019) and studies in rodents have shown that increasing mitochondrial capacity can improve insulin sensitivity (Wright et al., 2011; Henstridge et al., 2014). Evidence has suggested that mitochondrial function may be suppressed particularly in T2D patients who are resistant

to exercise-induced metabolic improvements (Stephens et al., 2015), suggesting that stratified patients groups may benefit from mitochondrial-targeted therapies. Indeed, previous genetic studies have hinted at a role for genetic regulation of mitochondrial function in insulin metabolism, glucose metabolism and diabetes (Kraja et al., 2019).

To better understand the role of mitochondrial dysfunction in the genetic risk of T2D in humans, the following Chapters present a genetic study with the aim of identifying genetic risk factors for T2D which directly alter mitochondrial function. This will be achieved by systematically assigning functional *cis*–regulated genes to T2D genetic risk loci. The study design presented over the following Chapters also has the potential to build on patient stratification discussed in **Section 1.4: Type 2 diabetes: a spectrum, a palette or a cluster?**, by prioritising specific biological processes for further study.

1.6 Adipose, diabetes and the mitochondria

The current study, which is a follow-up to that of Lau et al. (2017), uses gene expression data from subcutaneous adipose to systematically identify the target genes of T2D-associated genetic risk loci (these methods are described in detail in the following Chapter 2, see **Section 2.7.3: LDU-based gene mapping in T2D**). The relationship between adipose and T2D is discussed here, with a particular focus on the mitochondria.

The mitochondria are central to adipose metabolism, playing a key role in adipocyte differentiation, lipid metabolism (lipogenesis and lipolysis), endocrine signalling and thermogenesis (Boudina and Graham, 2014; Cedikova et al., 2016). Effective mitochondrial function is required for adipose formation (adipogenesis) (Trifunovic et al., 2004; De Pauw et al., 2009). White adipose tissue (WAT) can be divided into subcutaneous adipose, which is located under the skin and visceral adipose, which surrounds organs (Bjørndal et al., 2011). An additional type of adipose is brown adipose tissue (BAT) which generates heat and contributes to thermogenesis through the release of energy from mitochondrial fatty acid β-oxidation. BAT decreases in mass as humans age, although there is evidence that some may remain active in adult humans (Nedergaard et al., 2007; Virtanen et al., 2009; van Marken Lichtenbelt et al., 2009; Cypess et al., 2009) and that this may protect

against diabetes (Chondronikola et al., 2014). BAT thermogenesis and maintenance requires mitochondrial oxidation and the mitochondrial uncoupling protein 1 (Heaton et al., 1978; Lee et al., 2015b; Cedikova et al., 2016; Gonzalez-Hurtado et al., 2018).

WAT stores fat as triglycerides, which are released when required such as in times of reduced food intake or increased energy requirements (Sethi and Vidal-Puig, 2007). Mitochondria are critical for this process, producing the glycerol 3-phosphate and acetyl-CoA necessary for triglyceride formation, and also contribute to the regulation of lipolysis (De Pauw et al., 2009). As such, adipose have high levels of mitochondrial fatty acid β-oxidation (this pathway and the evidence relating it to T2D is discussed in detail in Chapter 5). Reduced storage of fats as triglycerides, potentially due to adipose inflammation, can cause increased circulating fatty acid levels and may cause insulin resistance and ectopic fat accumulation in other organs, such as the liver and muscle (Guilherme et al., 2008; Lê et al., 2011). WAT is also an important endocrine organ and produces adipokines and other signalling molecules to regulate multiple processes including appetite, inflammation, fat and glucose metabolism. These include leptin which regulates appetite and food intake, prostaglandins which play a role in the repair of injuries and adiponectin which stimulates fatty acid and glucose uptake (Sethi and Vidal-Puig, 2007). Signalling molecules secreted by adipose tissue may regulate the insulin sensitivity of peripheral tissues (Van der Kolk et al., 2019), these include the inflammatory cytokines tumor necrosis factor-α (TNFα) and interleukin (IL)-6 (Rytka et al., 2011). The mitochondria play an important role in this endocrine activity (Koh et al., 2007), with their dysfunction related to the development of insulin resistance and diabetes (Medina-Gómez, 2012). Visceral adipose, which has a high mitochondrial content (Deveaud et al., 2004), has been particularly implicated in T2D and increased visceral fat deposits are associated with insulin resistance and diabetes (Bjørndal et al., 2011; Direk et al., 2013, 2014). Disruption of the mitochondrial electron transport chain may be one cause of visceral adipose insulin resistance (Ngo et al., 2019).

Several studies have implicated adipose function in T2D genetic risk. Small et al. (2018) showed that T2D-associated genetic variants altered the expression of *KLF14* in adipose,

resulting in larger, preferentially visceral adipocytes and insulin resistance. T2D is associated with missense mutations in *ADIPOQ*, which encodes the adipokine adiponectin, and *PPARG*, which encodes the nuclear receptor peroxisome proliferator-activated receptor-γ (PPARγ) and is highly expressed in adipose tissue (Stumvoll et al., 2002; Hivert et al., 2008; Gao et al., 2013; Ahmadian et al., 2013; Vergotine et al., 2014; Majithia et al., 2014). It is also worth noting that there are both 'healthy' and 'unhealthy' forms of adiposity, which may involve increased adipose inflammation (Aires et al., 2019; Smith et al., 2019). Further study of the genetic mechanisms regulating adipose gene expression may provide additional insight into the underlying processes of T2D risk. The mechanisms of insulin resistance in adipose subtypes are further reviewed by Czech (2020).

The following Chapters describe the mapping of genes regulated by T2D-associated variants using adipose gene expression data, however mitochondrial dysfunction may have significant implications in other tissues. Briefly, these include in the insulin response of muscle (see reviews Muoio and Neufer (2012); Hesselink et al. (2016); Devarshi et al. (2017)) and liver (Su et al., 2019; Bassot et al., 2019) and in maintaining insulin production from pancreatic β-cells (Fex et al., 2018; Las et al., 2020). Altered mitochondrial fatty acid β-oxidation may also increase the release of pro-inflammatory cytokines from T cells in T2D patients and influence adipose metabolism (Wang and Wu, 2018; Nicholas et al., 2019); increased production of pro-inflammatory cytokines from adipose-resident T cells is associated with insulin resistance (Hardy et al., 2011). Mitochondrial dysfunction may also contribute to T2D-associated kidney disease (Kang et al., 2015; Sharma, 2017). The following Chapter 3 describes the assigning of functional genes to T2D loci using subcutaneous adipose gene expression. However, the methods described can be applied to the gene expression data of other tissues.

1.7 Hypothesis, aims and research plan

1.7.1 Hypothesis

The hypothesis for this study is that mitochondrial dysfunction is a heritable risk factor for Type 2 diabetes (T2D) and can causally contribute to disease onset. As a result, we

anticipate that some genetic risk factors for T2D will exert their effects through altering the expression of nuclear-encoded mitochondrial genes (NEMGs).

1.7.2 Aims and research plan

CHAPTER 2: Gene mapping in T2D: a literature review

Aim: to carry out a literature review of gene mapping studies for T2D. The gene mapping method used by Lau et al. (2017) in generating the data analysed in this study (see Chapter 3 aims, below) will be compared and contrasted to other gene mapping methods.

CHAPTER 3: T2D loci regulate nuclear-encoded mitochondrial *cis*-genes

Aim: to investigate published and unpublished genetic loci associated with risk of Type 2 diabetes (T2D loci), previously mapped by Lau et al. (2017), for evidence of regulating neighbouring genes (*cis*-genes) including NEMGs in adipose tissue.

Research plan:

1. Identify T2D *cis*–genes by integrating independent location estimates corresponding to genetic variants associated with (1) T2D risk (T2D loci) and (2) neighbouring gene expression (eQTL), provided by Lau et al.
2. Cross-reference the identified T2D *cis*-genes with curated databases of known NEMGs to identify T2D *cis*-NEMGs.
3. Test the total T2D *cis*-genes for enrichment of mitochondrial pathways.

CHAPTER 4: T2D *cis*–gene expression in cases vs controls

Aim: to validate the identified *cis*-genes by demonstrating that the same genes show evidence of differential expression in T2D cases compared to controls for independent gene expression datasets.

Research plan:

1. Identify independent gene expression datasets with data for T2D or insulin resistant

cases and controls from the Gene Expression Omnibus (GEO) database.

2. Test for differential gene expression and carry out meta-analysis of the T2D *cis*-genes, *cis*-NEMGs and mitochondrial pathways using gene set enrichment analysis (GSEA).

CHAPTER 5: Fine-mapping a candidate locus

Aim: to investigate a candidate T2D locus using targeted next generation sequencing data, with an aim to identify the putative causal variant(s).

1. Identify single nucleotide variants, small insertions and deletions and structural variants from targeted next generation sequencing data for 94 T2D cases and 94 healthy controls (French).

2. Test the variants for significant differences in allele frequencies between T2D cases and controls.

3. Identify potentially functional variants by integrating variant risk prediction and chromatin modification data from publicly available sources.

2 Chapter 2: Gene Mapping in T2D: A Literature Review

2.1 Overview

This project aims to investigate whether an inherited predisposition for mitochondrial dysfunction increases the risk for Type 2 diabetes (T2D). As such, a genetic study design will be used to investigate whether genetic risk factors for T2D alter mitochondrial function. The starting point is therefore gene mapping; the mapping of genetic loci associated with T2D[11], followed by the study of their potential downstream effects on mitochondrial function.

Hundreds of T2D loci have been identified to date. These have been mapped mostly through genome-wide association studies (GWAS), which test single nucleotide polymorphisms (SNPs) for evidence of association by regressing genotype against case-control status. Conversely, this project investigates T2D loci mapped previously by Lau et al. (2017) using an alternative method referred to as LDU-based gene mapping. In order to appreciate the strengths and weaknesses of this method and hence the novelty of the results presented, Chapter 2 will review and compare the gene mapping methods previously applied to T2D. The efficiency of different gene mapping methods will be explored in light of new insights into complex disease genetic architecture[12]. Following this, Chapter 3 will explore the functional interpretation of published and unpublished T2D loci from Lau et al. All results presented in Chapter 2 for the literature review were generated previously, excluding the analysis on page 76 (Figure 10) which was carried out specifically to compare the genetic architecture of T2D loci identified by Lau et al. and lead SNPs mapped by Mahajan et al. (2018) using the more conventional single-SNP GWAS.

[11] These will mostly be referred to as T2D loci, but may also be called disease loci or risk loci.

[12] Genetic architecture has been defined as types of variation, allele frequency distribution, allele effect sizes and new mutation rates (Lupski et al., 2011). Allelic architecture may be the number of alleles which impact a phenotype at a given locus, their frequencies and penetrance (Pritchard and Cox, 2002)

2.2 Introduction

Gene mapping[13] aims to identify the causal genetic variants and the implicated functional gene(s) underlying a disease or trait. The importance of gene mapping is evident in the case of rare monogenic diseases, since mapping the single causal mutation can provide an accurate diagnosis, enable family genetic counselling and reveal causal mechanisms which can be targeted during treatment. An important example is in neonatal diabetes for which mutations in the potassium channel (*KCNJ11* or *ABCC8* genes) make up 50% of cases (Hattersley and Patel, 2017). These patients are now specifically treated with oral sulfonylureas to close the potassium channel and stimulate insulin release from pancreatic β-cells, demonstrating significantly improved outcomes compared with the traditional insulin injections (Hattersley and Patel, 2017). However, there is a longer road between gene mapping and clinical translation when concerned with complex diseases.

The risk or 'liability' of developing a complex traits is observed to be influenced by hundreds of genetic variants of predominantly small effect (here referred to as risk variants). It is important to identify these risk variants not only for clinical translation, but also to achieve a greater understanding of complex disease. Identifying the functional disease-associated genes can reveal novel functions for those genes and implicate novel biological pathways in disease onset. The findings from gene mapping can also be used to inform the development of improved analytical methods. However, complex disease risk variants are typically located in intronic or intergenic non-coding DNA rather than within coding genes, challenging their interpretation (discussed in detail in Chapter 3).

In terms of clinical utility, there is increasing evidence that genetic risk may underlie T2D phenotypic heterogeneity (discussed in Chapter 1, **Section 1.4: Type 2 diabetes: a spectrum, a palette or a cluster?**). This evidence further motivates gene mapping studies, since distinct underlying mechanisms may translate to targeted therapies and guided disease management (Udler, 2019). Stratifying patient groups is likely to improve

[13]The mapping of a trait associated locus has also been described as positional cloning (Collins, 1992). The terms 'positional cloning' and 'gene mapping' are now often used interchangeably. 'Gene mapping' will be used hereafter in this thesis.

the power of gene mapping. For example, Guan et al. (2016) mapped novel genetic loci in a study stratified for T2D patients with end-stage kidney disease vs healthy controls with no kidney disease (discussed in Chapter 5). Genetic risk scores (GRSs)[14] can also be correlated with other traits to investigate shared genetic mechanisms and co-morbities (Damkondwar et al., 2012; Hackinger and Zeggini, 2017; Zheutlin et al., 2019; Chasman et al., 2020) and to identify novel causal pathways (Monnereau et al., 2016; Ahmad et al., 2018; Mallet et al., 2020; Caspers et al., 2020). For example, Vujkovic et al. (2020) reported that T2D genetically correlated with waist circumference, overall health, BMI and fat mass, hypertension, coronary artery disease, dyslipidemia, alcohol intake, wheezing and smoking, while negatively correlating with years of educational attainment.

The most widely used gene-mapping method to date is the genome-wide association study (GWAS). This Chapter will review the published T2D GWAS prior to discussing their strengths and limitations in comparison to LDU-based gene mapping. The next section will first describe the fundamental concepts of linkage and linkage disequilibrium, in addition to briefly reviewing gene mapping studies for T2D prior to GWAS.

2.3 Linkage, LD and pre-GWAS gene mapping in T2D

This section introduces the concepts of linkage and linkage disequilibrium (LD), as well as linkage and association studies. Linkage and association studies exploit the related properties of linkage and LD, which correlate genotypes at two loci within and between defined pedigree structures, respectively (Terwilliger, 2001). They are described below, along with examples of their application to T2D gene mapping. Detailed reviews of T2D loci mapped prior to GWAS are also given by McCarthy (2004); McCarthy and Menzel (2001) and van Tilburg et al. (2001)[15].

Prior to the popularisation of GWAS from 2005 onwards, T2D genetic risk loci were identified using linkage analysis and then further investigated in larger cohorts using as-

[14]There is a large literature on the use of genetic risk scores which will not be discussed here. For several recent reviews, see: Torkamani et al. (2018); Lewis and Vassos (2020) and McCarthy and Mahajan (2018) for diabetes specifically

[15]See also Chapter 16 of Collins (2007), 'Identifying susceptibility variants for Type 2 diabetes' by Zeggini and McCarthy.

sociation analysis. Association studies were also used to investigate biologically plausible candidate genes identified through their involvement in monogenic diabetes, functionally-relevant pathways such as insulin production, or the known mechanisms of diabetes drug targets and later through the inclusion of additional -omics data (McCarthy et al., 2003). Although subject to certain limitations (discussed more below), linkage and candidate gene studies were successfully used to map several T2D risk loci.

2.3.1 Linkage and linkage mapping

The mapping of disease loci exploits the fundamental concept of linkage. Linkage refers to the co-inheritance of alleles at polymorphisms which are present on the same chromosome, unlike those on different chromosomes which are inherited independently[16]. Linkage breaks down when meiotic recombinations cause crossover between paired chromosomes and creates new combinations of alleles (see Figure 2). A haplotype represents the combination of alleles which occur together on the same chromatid and new haplotypes are formed as a result of recombination.

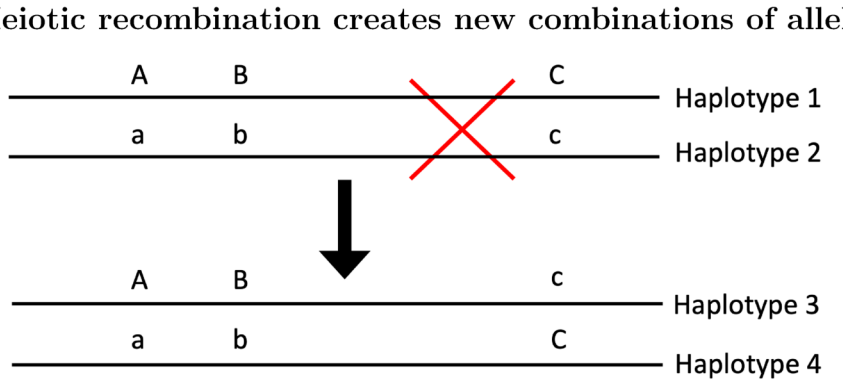


Figure 2: Meiotic recombination produces new combinations of alleles. The red cross shows the location of a recombination event, forming new haplotypes. In this scenario, the A and B are in complete linkage, as are the a and b alleles, since they are inherited together.

The linkage between two genetic variants can be measured as the frequency with which their alleles are co-inherited. This is determined by the frequency of recombination between them and is therefore closely related to physical distance[17], since the probability

[16] as stated by Mendel's law of independent assortment.

[17] Here, physical distance between two genetic variants is defined as the number of DNA bases between

of observing a recombination is smaller when a smaller region is considered. Prior to the availability of high-throughput genotyping, or indeed, the availability of any genotype data[18], linkage could be exploited to map the locations of disease-causing variants. Genetic polymorphisms of known location, also known as markers, were genotyped in related individuals (family data). Due to linkage, markers nearby a causal variant would be inherited together on the sample haplotype and would therefore also co-segregate through the pedigrees with affected individuals. Multiple generations were included so that a larger number of recombinations could be directly observed to break down linkage, resulting in the co-segregation of a smaller genetic region with disease and greater resolution for inferring the disease gene location. The resolution of linkage analysis is limited to the number of observed recombinations, with a 1% probability of observing a recombination equating to approximately 1Mb (Khil and Camerini-Otero, 2009). As a result, linkage studies typically implicated large chromosomal segments and could be carried out by genotyping relatively disperse markers. A limitation of family-based linkage studies is the need to recruit large pedigrees of related individuals.

Importantly, the probability of recombination is not entirely random. This may be influenced by the binding of transcription factors (Tiemann-Boege et al., 2017; Coop and Przeworski, 2007) and interfering DNA structures such as chiasmata, for example. Recombination events are also observed to cluster in 'hot spots' first observed by Chakravarti et al. (1984), the theory of which is reviewed elsewhere (Tiemann-Boege et al., 2017; Coop and Przeworski, 2007). As a result, linkage is non-linear with respect to physical distance and the number markers linked with a causal mutation will differ depending on the genomic location. These patterns can be captured by genetic linkage maps which can (1) inform where best to place markers to capture the inheritance of surrounding variants and (2) the physical region surrounding a significant marker in which the linked causal

them.

[18]The concept of linkage was first explained by Thomas Morgan in 1911, based on the observed inheritance of simple traits in Drosophila fruit flies such as eye colour and wing mutations that appeared to co-segregate together, violating Mendel's second law of independent assortment. Morgan announced the following profound conclusion: 'Instead of random segregation in Mendel's sense we find "associations of factors" that are located near together in the chromosomes.' (Morgan, 1911).

mutation is likely to reside. Linkage maps are discussed in **Section 2.7.1: Genetic and LDU maps** and a review of the various linkage analysis methods can be found in Chapter 7 of Neale et al. (2007) and Terwilliger and Ott (1994).

Variations of the linkage design were used to detect several T2D loci of large effect (McCarthy, 2003). These included model-free, non-parametric[19] methods such as affected sib-pair or affected-pedigree-member and analysis of 'extreme discordant sib pairs'[20] (Weeks and Lathrop, 1995; Risch and Zhang, 1995; Freimer and Sabatti, 2004). Examples of T2D loci first identified through linkage analysis include the *TCF7L2* (Reynisdottir et al., 2003; Duggirala et al., 1999) and *ADIPOQ* loci (Vionnet et al., 2000; Mori et al., 2002; Busfield et al., 2002). The *TCF7L2* locus analysis has since been detected as the most significant genome-wide signal of T2D association across multiple populations (Cauchi et al., 2007; Sladek et al., 2007; Tabassum et al., 2013; Ng et al., 2014; Qi et al., 2017; Mahajan et al., 2018; Chen et al., 2019a) and *ADIPOQ*, which encodes the cytokine adiponectin has accumulated a large literature regarding its role in T2D (Achari and Jain, 2017). The application of linkage methods to complex diseases were more recently discussed by Flaquer and Strauch (2012).

While linkage studies had great success in mapping genes which caused early-onset, familial forms of diabetes known as maturity-onset diabetes of the young (MODY) (Botstein and Risch, 2003), they were underpowered to detect variants which did not strictly co-segregate with disease. By 2001, genome-wide linkage studies had largely failed to detect loci which could be reproducibly associated with complex diseases (Altmüller et al., 2001). This supported the polygenic model of complex disease, in that liability resulted from the accumulation of multiple small effect variants[21] (see Figure 1). Advocated by Risch and Merikangas (1996); Lander and Schork (1994) and others, association studies of candidate genes largely supplanted linkage analysis in the study of complex diseases.

[19]Non-parametric studies make no assumption regarding the mode of inheritance, whereas parametric studies test a pre-defined mode of inheritance.

[20]Alternative designs include studies of parent-offspring trios (Huxtable et al., 2000) and the transmission-disequilibrium test (Spielman et al., 1993).

[21]The mapping of several large-effect loci is notably consistent with the modern view of T2D as having significant genetic heterogeneity (Udler, 2019).

2.3.2 Linkage disequilibrium and association mapping

Association analysis exploits linkage disequilibrium (LD), which is a measure of allelic association. LD measures the frequency with which alleles are found together across a population and large families. A population can be considered as an extended pedigree in which historical, unobserved recombinations have created new combinations of alleles. LD therefore measures linkage with a much higher resolution, since many more recombination events occur within the history of a population compared to a pedigree of several generations. That being said, LD is also influenced by population-parameters such as population admixture and inbreeding, as well as the number of founding individuals and the number of generations within a population, natural selection, genetic drift and mutation rate (Collins, 2007). Patterns of LD across the genome can be seen following the construction of genetic LDU maps described in **Section 2.7.1: Genetic and LDU maps**.

Various metrics used to measure LD include r^2, D' and ρ, which are all normalised measures of the covariance, D (Falconer and Mackay, 1996; Lynch et al., 1998). The calculation of these are shown below for two nearby SNPs with major alleles A and B and minor alleles a and b, where the frequency of each haplotype is denoted p_{AB}, p_{Ab}, p_{aB} and p_{ab} and the allele frequencies are denoted p_A, p_B, p_a and p_b:

		Locus B	
		B	b
Locus A	A	p_{AB}	p_{Ab}
	a	p_{aB}	p_{ab}

$$D_{AB} = p_{AB} - p_A p_B \qquad r^2_{AB} = \frac{D^2}{(p_A p_a p_B p_b)} \qquad \rho_{AB} = \frac{D_{AB}}{p_A p_b}$$

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

r^2 and ρ range between 0 and 1 with a value of 1 indicating complete LD; the alleles are found together 100% of the time. D' ranges between -1 and 1 and a D' of ± 1 is obtained when at least one of the haplotypes is not observed. The properties of ρ, D', r^2 and other LD metrics are compared by Morton et al. (2001) and Kang and Rosenberg (2019).

Association studies investigate cohorts of unrelated individuals to identify genetic variants which have significantly different frequencies in individuals with the disease, compared to healthy controls (higher if deleterious or lower if protective). Since LD is less extensive than linkage, LD can be exploited to genotype densely placed marker variants which are co-inherited with, or 'tag', a smaller number of variants in high LD[22]. This concept of indirect genotyping is illustrated in Figure 3. A significantly associated marker is either the causal variant itself (unlikely if an array is used), or is in high LD with one or more causal variant.

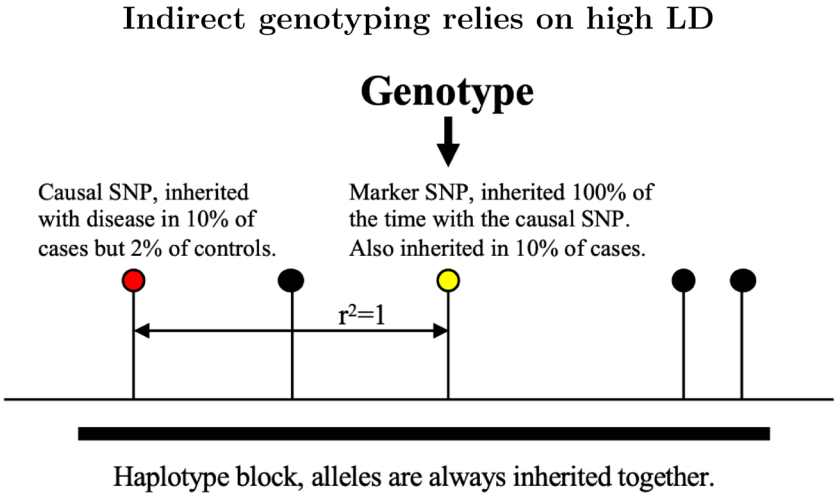


Figure 3: Indirect genotyping. In this example, a haplotype block contains five SNPs. The causal variant, shown in red, is inherited in 10% of cases but only 2% of controls. The other four SNPs are in complete LD and are always inherited together with the causal variant. Hence, they show the same level of association with disease. Rather than genotyping all five SNPs, the marker SNP shown in yellow is genotyped.

Association mapping has significantly higher resolution than linkage mapping, since smaller sets of variants are consistently co-inherited throughout the multiple generations of a population. Association mapping was therefore used to fine-map loci implicated by linkage

[22]For more information on the history of association mapping and LD, see Chapter 2 of Collins (2007) and Sved and Hill (2018)

analysis, for example both linkage signals at *TCF7L2* and *ADIPOQ* were fine-mapped to the respective genes using association analysis (Grant et al., 2006; Zhang et al., 2006; Vasseur et al., 2002). Other studies genotyped markers at candidate genes implicated in the onset of MODY, including the insulin promoter factor 1, *IPF-1/PDX1* (Stoffers et al., 1999; Macfarlane et al., 2000), the sulfonylurea receptor, *SUR* (Inoue et al., 1996; De Knijff et al., 1999), insulin receptor substrate 1, *IRS-1* (Almind et al., 1993) and the hepatocyte nuclear factor-4-α, *HNF4A* (Love-Gregory et al., 2004; Silander et al., 2004). Other genes were prioritised as known targets of T2D drugs, including *PPARG*, which encodes the ligand-activated transcription factor PPARγ (Altshuler et al., 2000) and *KCNJ11* and *ABCC8*, which encode the pancreatic potassium channel.

Despite offering a greater resolution and the freedom to recruit unrelated participants, the few successful association studies were outweighed by a general lack of consistency and reproducibility (Hirschhorn et al., 2002; Hirschhorn and Altshuler, 2002; Ioannidis et al., 2001), leading to arguments for (Burgner and Hull, 2000) and against their utility (Gambano et al., 2000). Various explanations addressing both study design and the nature of complex diseases were offered and calls for consistent study designs were made (Hirschhorn et al., 2002; Colhoun et al., 2003; Hattersley and McCarthy, 2005). Potential explanations included cohort heterogeneity, fluctuations in allele frequency, gene-gene or gene-environment interaction effects, variations in LD[23], population stratification[24] and too liberal statistical thresholds. Furthermore, existing methods had limited power to prioritise candidate genes (McCarthy et al., 2003), especially given that complex disease mechanisms were largely unknown (hence the requirement for gene mapping) (Hattersley and McCarthy, 2005). Hypothesis-free linkage scans also required prior evidence to prioritise genes, since their low resolution typically implicated large regions.

As high-throughput studies made available genome-wide maps of common variants and the

[23]Populations-specific LD and the relative difficulties in reproducing trans-ethnic association signals are the focus of Section 2.4.4: Trans-ethnic GWAS and defining replication.

[24]Population stratification arises when there is a systematic difference in allele frequencies between subpopulations in a population, which in the context of association study may cause confounding (e.g. false positives and false negatives) where the prevalence of disease also differs between subpopulations.

costs of genotyping decreased, hypothesis-free scanning of the genome using association analysis became feasible. Genome-wide association studies (GWAS) subsequently became one of the most popular methods with which to carry out gene mapping for complex disease; these are discussed below.

2.4 Genome-wide association studies (GWAS)

2.4.1 Overview

As mentioned previously, the GWAS design has to date been the most widely used method to identify variants associated with complex disease. The conventional GWAS tests individual SNPs, genome-wide, for evidence of association; this is shortened to single-SNP GWAS in the following text. The alternative method of LDU-based gene mapping was used to generate data for this current study. To compare the strengths and limitations of these two methods, the assumptions and design of the single-SNP GWAS are described below, including the common disease, common variant (CD/CV) hypothesis, indirect genotyping and imputation. Following this, **Section 2.4.3: GWAS of T2D** will review the published T2D GWAS, which is followed by **Section 2.4.4: Trans-ethnic GWAS and defining replication**. The limitations of single-SNP GWAS are discussed later in **Section 2.6: Missing heritability**.

2.4.2 GWAS: the design

GWAS are typically carried out by genotyping genetic markers using genotyping arrays for large case-control cohorts. The indirect genotyping approach illustrated in Figure 3 is used such that an informative subset of markers can be genotyped in order to capture information about unobserved genotypes in high LD (Collins et al., 1997; Johnson et al., 2001; Carlson et al., 2004a,b). Thus, exploiting LD to reduce the number of genotyped variants allowed for larger sample sizes and greater power[25] (Gabriel et al., 2002; Consortium et al., 2005). Early predictions for the number of genotyped SNPs required to capture the inheritance of all common variants ranged between 300,000 and 1 million,

[25]Crucially, the increase in power applies *only* to variants which are in high LD with markers.

depending on the population (Kruglyak, 1999; Gabriel et al., 2002; Hapmap, 2003). At this time, 1.42 million SNPs were reported (Sachidanandam et al., 2001) and this was predicted to rise to 10-15 million SNPs (minor allele frequency, MAF >1%) (Botstein and Risch, 2003; Kruglyak and Nickerson, 2001). By 2005, ∼1.5 million SNPs were published by the International HapMap Consortium (Hinds et al., 2005) and this increased to ∼8 million SNPs with a MAF >5% and ∼12 million SNPs with a MAF between 0.5% and 5% reported by the 1000 Genomes Project (Consortium et al., 2015a).

Marker SNPs are required to be in high LD with a causal variant in order to capture its association with disease. The placement of markers was highly influenced by both the common disease, common variant (CD/CV) hypothesis and the observation of haplotype blocks. The CD/CV hypothesis states that complex traits result from the additive small effects of common variants[26] found frequently across the population (Risch and Merikangas, 1996; Cargill et al., 1999; Chakravarti, 1999; Reich and Lander, 2001). As a result, markers were selected to be in high LD with common SNPs[27]. The debates and criticisms surrounding this hypothesis, including the contribution of rare and low-frequency variants, are discussed in **Section 2.6: Missing Heritability**. Haplotype blocks are extended blocks of variants in high LD which persist through a population due to little to no historical recombination (Daly et al., 2001); blocks of high LD typically extend for 10s of kilobases (kb) (Reich et al., 2001). In theory, a marker will capture or 'tag' all other common variants on the same haplotype. The placement of markers on early GWAS arrays was informed by the International HapMap Project (Gabriel et al., 2002; Hapmap, 2003; Consortium et al., 2005), which aimed to map common shared sequences of extended LD across diverse populations.

An extension to the indirect genotyping approach is imputation, in which genotyped markers are used to predict missing genotypes using out-of-sample reference panels. This effectively increases the number of 'observed' genotypes which can be tested for associa-

[26]Common variants are defined as variants with a population minor allele frequency, MAF, of >1% or >5% depending on the source.

[27]In addition to the CD/CV hypothesis, capturing common variants as opposed to rare variants allowed a more straightforward and convenient design for high-throughput, indirect genotyping.

tion. Drawn from the same population, a larger number of variants are genotyped in an independent cohort for use as a reference panel. A missing genotype can then be imputed in a case-control GWAS based on the frequency with which it is detected together with the nearby markers in the reference panel. Imputation facilitates greater commensurability between independent genotyping platforms for meta-analysis by imputing variants present on other arrays. The confidence or quality of imputation decreases when there is less LD between a variant and the genotyped markers, but improves when reference panels capture more variation using denser genotypes and larger sample sizes.

Following genotyping, individual SNPs are tested for evidence of disease association, typically using the single-marker regression (SMR) to regress genotype against case-control status or a quantitative trait[28], with optional covariants including age, sex and BMI. The test of association is fitted to an inheritance model which can either be recessive, additive or dominant, with the genotypes coded as shown below. The most commonly used model is the co-dominant model, since this offers the most consistent power when the true inheritance is non-additive (Lettre et al., 2007).

	AA	Aa	aa
Recessive	0	0	1
Additive	0	1	2
Dominant	0	1	1

SMR was shown to increase power over the conventional Cochran-Armitage test (Dizier et al., 2017), which combines allele frequencies in a contingency table and calculates a trend-test statistic similar to a Pearson χ^2 statistic.

Based on the number of estimated SNPs and the correlated LD structure between them, Risch and Merikangas (1996) estimated that there would be approximately one million independent tests of association in a GWAS. The p-value from each test represents the probability of observing the resulting test statistic given that the null hypothesis (no association) is true. A conventional p-value threshold of 0.05 represents a 5% probability

[28]Alternative methods proposed included haplotype or multimarker tests of association and pooled analysis; these are reviewed elsewhere (Carlson et al., 2004a; Botstein and Risch, 2003)

of observing a significant test despite there being no actual association, hence a total of 50,000 tests would be expected to be false positives for 1 million independent tests. There are several methods used to correct for this 'multiple testing problem'. These include the Bonferroni-correction which controls the family-wise error rate (FWER) to achieve a probability of no more than one false positive. The significance level is divided by the number of independent tests, making the new p-value threshold 0.05/1,000,000 = 5×10^{-8} (Risch and Merikangas, 1996). Several less stringent thresholds have been suggested, such as 5×10^{-5} suggested by Colhoun et al. (2003) to improve replication rates hindered by publication bias, chance, and inadequate sample sizes. Control of type I error rates (false positives) can also be achieved using the false discovery rate (FDR) introduced by Benjamini and Hochberg (1995), which represents the expected proportion of false positives out of all the statistically significant results for a specified significance threshold. For example, significant SNPs may be identified using the q-value proposed by Storey (2002) which corrects based on FDR (Storey et al., 2003; Storey and Tibshirani, 2003). 5×10^{-8} is the widely accepted GWAS threshold[29].

2.4.3 GWAS of T2D

The published T2D GWAS are briefly reviewed below, including their findings and the evolving methodology. Sample sizes are highlighted in bold for easy comparison with the LDU-based study of Lau et al. (2017), in which 111 T2D loci were replicated at genome-wide significance in a sample of **5,800 T2D cases** and **9,691 controls**.

In the first T2D GWAS, Sladek et al. (2007) genotyped a French cohort of **694 T2D cases** and **669 healthy controls** (Sladek et al., 2007). The study replicated a known signal at the *TCF7L2* locus and reported two novel loci, which were themselves replicated by the Icelandic company deCODE (Steinthorsdottir et al., 2007). In the same year, coordinated publications from the UK Wellcome Trust Case Control Consortium (WTCCC), the Finland-United States Investigation of NIDDM Genetics (FUSION) and the Diabetes

[29]which assumes a high proportion (over 99%) of the total genomic tests will be consistent with the null hypothesis of no association. Where this is not true, for example with gene expression arrays (where the proportion of null tests may be nearer two thirds), the FDR can be used to introduce less stringent significance thresholds while still controlling the Type I error rate.

Genetics Initiative (DGI) together brought the number of independent, replicated GWAS signals to a total of five (Consortium et al., 2007b; Scott et al., 2007; Saxena et al., 2007). Of these, the WTCCC genotyped the largest number of cases at **1,924 T2D cases** (Consortium et al., 2007b; Zeggini et al., 2007). In 2008, the DIAGRAM Consortium (Diabetes Genetics Replication and Metaanalysis) combined data from WTCCC, FUSION and DGI enabling a larger sample size of **4,549 T2D cases and 5,579 controls** (Zeggini et al., 2007). This meta-analysis imputed up to ∼2.2 million common SNPs using the HapMap phase II reference panel, which was based on 60 European individuals genotyped for over 3 million SNPs (Consortium et al., 2007a)[30]. A total of six new loci were identified.

Despite a total of 19 loci robustly associated with T2D (McCarthy and Zeggini, 2009), Manolio et al. (2009) reported that only 6% of the estimated T2D heritability could be accounted for (T2D heritability is described in **Chapter 1, Section 1.3.1: T2D heritability** and the proposed explanations for this 'missing heritability' are discussed in detail in **Section 2.6: Missing heritability**). Several studies reported that additional heritability could be explained by including variants of small effect which failed to reach genome-wide significance, arguing in favour of larger GWAS to capture this 'hidden' heritability (Gibson, 2010). Denser genotyping arrays, denser imputation reference panels and consistent study designs[31] were also expected to capture more variation (De Bakker et al., 2008; McCarthy et al., 2008; Pe'er et al., 2008).

By 2011, a total of 44 loci had been associated with T2D (Wheeler and Barroso, 2011; Billings and Florez, 2010; Imamura and Maeda, 2011)[32], of which 12 loci were reported in an expanded European meta-analysis of **8,130 T2D cases and 38,987 controls** from the DIAGRAM+ Consortium, now comprising an additional five cohorts (Voight

[30]HapMap phase II genotyped approximately one SNP every kilobase and was estimated to contain between 25-35% of all common SNPs (MAF ≥0.05) (Consortium et al., 2007a).

[31]Best practices included consistent subject ascertainment, marker selection and array design, analysis methods, validation, quality control, replication and the use of 5×10^{-8} as a threshold for genome-wide significance, as well as the replication of novel signals in at least one independent cohort.

[32]Billings and Florez (2010); Imamura and Maeda (2011) also list 24 SNPs associated with continuous glycemic traits. GWAS of glycemic traits are not the focus of this section and are reviewed elsewhere (Marullo et al., 2014)

et al., 2010). Up to 2.4 million SNPs were imputed (MAF >1%)[33] and six of the 12 were reported to be detected 'wholly or predominantly' from imputed data. However, ∼90% of T2D heritability remained unaccounted for (Imamura and Maeda, 2011). Further improvements, including more trans-ethnic studies (see **Section 2.4.4: Trans-ethnic GWAS and defining replication**) and a focus on low frequency variants were called for (Imamura and Maeda, 2011; Wheeler and Barroso, 2011; Visscher et al., 2012a).

In 2010, the pilot phase of the 1000 Genomes Project provided the locations and frequencies of ∼15 million SNPs, following low-coverage sequencing of 179 genomes and high-coverage sequencing of 697 exomes (Consortium et al., 2010a). The 1000 Genomes data improved the accuracy of imputation compared to the HapMap reference panel. For example, Huang et al. (2012) detected one additional significant SNP absent from the original WTCCC phase 1 publication (Consortium et al., 2007b)[34]. Using 1000 Genomes data, it was estimated that previous arrays had captured <60% of common variants (Sanghera and Blackett, 2012). Denser genotyping arrays were developed, as well as custom arrays which included a range of allele frequencies designed for the replication and fine-mapping of previously identified loci.

Two subsequent GWAS by Saxena et al. (2012) and Morris et al. (2012) used custom arrays and both reported evidence of independent, secondary signals at several T2D loci, providing clear examples of allelic heterogeneity, in which variants arise independently at the same locus to potentially disrupt the same functional unit. The Morris et al. study ('DIAGRAMv3') meta-analysed **12,171 T2D cases and 56,862 controls** to discover ten new loci using the custom cardiometabolic chip (Metabochip), designed using 1000 Genomes Project data to genotype cardiometabolic trait-associated loci, of which 21,774 SNPs were at previously identified T2D loci. The Metabochip was estimated to capture ∼90% of common SNPs (MAF >5%) and 60% of lower-frequency variants (1%> MAF <5%) at the target loci, compared to ∼77% and ∼32% using HapMap data,

[33]One of the eight stage 1 cohorts genotyped ∼500,000 SNPs on an Illumina bead array, with the rest genotyping <400,000 SNPs.

[34]The total of four significant SNPs is in comparison to the 98 genome-wide significant and replicated T2D locations detected in the same data by Lau et al. (2017) using LDU-based gene mapping.

respectively[35].

In 2014, Mahajan et al. published seven novel loci from a trans-ethnic meta-analysis of **26,488 T2D cases and 83,964 controls**, combining DIAGRAM with East Asian, South Asian, Mexican and Mexican American meta-GWAS. Each imputed up to 2.5 million SNPs (MAF >1%) using HapMap reference panels (Consortium et al., 2007a, 2010b). By 2015, ∼153 variants had been associated with T2D, including through targeted association studies and exome-sequencing (Prasad and Groop, 2015). In 2017, Scott et al. carried out a European meta-analysis of **26,676 T2D cases and 132,532 controls**, imputing over 12 million SNVs[36] using new 1000 Genomes data[37]. 13 novel SNPs were identified and replicated. The authors reported no significant associations with low-frequency alleles, despite a reported study power of 80% to detect variants with MAF of 0.5%, 1% or 5% with odds ratios of 1.80, 1.48 and 1.16, respectively. In the same year, Zhao et al. (2017) reported 13 novel loci[38] following the analysis of **73,337 T2D cases and 192,341 controls** split between European and South Asian ancestry, imputing SNPs with >1% MAF using 1000 Genomes. Xue et al. (2018) later detected significant association at a total of 139 common and 4 rare variants in a total of **62,892 T2D cases and 596,424 controls**. Xue et al. (2018) imputed over 5 million SNPs (MAF >1%).

An improved European reference panel of >39 million SNPs (MAF ≥0.1%) was published by the Haplotype Reference Consortium (HRC), combining low-coverage WGS data from 32,488 individuals (McCarthy et al., 2016). McCarthy et al. demonstrated improved accuracy when imputing using the HRC vs 1000 Genomes and more recently, Belsare et al. (2019) reported that the low-coverage 1000 Genomes data failed to reliably impute

[35]Morris et al. suggest that their results favoured the CD/CV hypothesis. Notably, the Metabochip fine-mapped loci previously found by GWAS optimised for the detection of common variants. Therefore it should be considered that the target loci may predominantly be driven by common variants. Furthermore, Morris et al. excluded variants with a MAF <1% from their study.

[36]Single nucleotide polymorphisms (SNPs) is used to refer to polymorphic loci with common variants, whereas single nucleotide variants (SNVs) includes low-frequency and rare variants which are not observed to be polymorphic in the population.

[37]This included ∼38 million SNPs, 1.4 million indels and >14,000 larger deletions from the sequencing of 1,092 human genomes (Consortium et al., 2012a) of which 3.9 million had a MAF between 0.5 and 5%

[38]The authors report 16 novel loci, but as noted in a footnote of the Zhao et al. (2017) paper, Scott et al. (2017) published while the manuscript was under review and detected three of the 16 loci.

rare variants, advocating the denser HRC panel for imputation in European cohorts[39]. The HRC reference panel was used by Mahajan et al. (2018) to impute ∼27 million variants in a meta-analysis of **74,124 T2D cases and 824,006 controls**, of which ∼21 million variants had a MAF <5%. Mahajan et al. reported 243 genome-wide significance loci (231 in a BMI-unadjusted analysis and 152 in a BMI-adjusted analysis), of which 135 were reported to be novel. 403 distinct signals were detected across the total loci, demonstrating the expected allelic heterogeneity. 56 low-frequency and 24 rare T2D-associated variants were detected, including 14 with odds ratios >2.

Other recent GWAS include those by Suzuki et al. (2019) and Spracklen et al. (2020) who analysed individuals of Japanese and East Asian ancestry, respectively. Suzuki et al. reported 28 new loci from the analysis of **36,614 T2D cases** and **155,150 controls**, while Spracklen et al. meta-analysed **77,418 T2D cases and 356,122 controls** to identify 301 distinct signals at 183 loci, of which 61 loci were novel. The most recent T2D GWAS was published by Vujkovic et al. (2020), who reported the results from a multi-ethnic meta-analysis of **228,499 T2D cases and 1,178,783 controls**. The authors detected 568 significant variants, as well as 25 ancestry-specific variants, of which 286 were reported to be novel. The loci were estimated to explain 19% of T2D risk. Variants with MAF >0.1% in Europeans and >1% in other cohorts were included. The authors observed that the 286 novel SNPs had smaller average effect sizes than replicated SNPs, likely due to the increased sample size and power to detect small effects. Consistent with previously observed allelic heterogeneity, Vujkovic et al. reported 233 conditionally independent SNPs surrounding 49 novel and 108 replicated SNPs in Europeans.

2.4.4 Trans-ethnic GWAS and defining replication

The LDU-based study of Lau et al. (2017) included a trans-ethnic analysis of two European and one African American cohorts. This section will describe the advantages of trans-ethnic analysis and will provide a brief overview of the trans-ethnic GWAS and GWAS

[39]Notably, the HRC reference panel also largely used low-coverage WGS data, removing sites with a minor allele count <5 and calculating genotype likelihoods, which may also impact the accuracy of very rare variant calls.

for T2D which have been carried out using African American individuals. For direct comparison, Lau et al. (2017) used LDU-based gene mapping to replicate 93 T2D loci in **956 African American T2D cases** and **1,029 controls**. 57 loci were individually genome-wide signifiant in the African American cohort.

Trans-ethnic GWAS can aid locus discovery by exploiting population-specific LD patterns. LD is influenced by multiple population parameters including the number of generations of meiotic recombinations (Campbell and Tishkoff, 2008; Pengelly et al., 2015; Vergara-Lope et al., 2019b). As a result, younger populations with more extensive LD such as Europeans are useful for locus discovery, since a causal variant is more likely to be tagged and imputed correctly. Older populations with LD breakdown such as Africans are useful for fine-mapping, providing a narrower search space since markers will be in high LD with fewer variants. Improved power and resolution can be achieved by increasing sample size in a multi-ethnic analysis, while a smaller number of variants will be consistently associated with disease due to the different LD patterns (assuming the causal variant is shared across populations, i.e. is cosmopolitan) (Cooper et al., 2008; Zaitlen et al., 2010; Rosenberg et al., 2010; Li and Keating, 2014; Mahajan et al., 2014a). Novel loci can be detected due to differences in allele frequencies and effect sizes, including population-specific disease loci[40]. These may also implicate distinct aetiologies, since there is evidence that T2D in European, Finnish and American populations is characterised by insulin resistance (Elmansy and Koyutürk, 2019), while Arab populations associate risk with obesity (Abuyassin and Laher, 2015) and East Asian populations with β-cell function (Narayan and Kanaya, 2020). The incidence of T2D also differs between populations (Spanakis and Golden, 2013), being reported at 2.4-fold and 1.5-fold greater in African American women and men, who have an average of 80% African ancestry and 20% European ancestry (Brancati et al., 2000), compared to white individuals.

Some limitations of trans-ethnic GWAS include potential population substructure or admixture which can cause spurious associations, while less LD equates to more independent

[40]A classic example is the *KCNQ1* locus first identified in East Asian GWAS due to a MAF of 40% compared to 5% in Europeans causing significant difference in statistical power (McCarthy, 2008).

SNPs and a greater multiple testing burden (Pe'er et al., 2008). Early arrays were also designed using reference panels of predominantly European ancestry and captured significantly less variation in African populations compared to European populations, for example (Barrett and Cardon, 2006). Increasingly sophisticated methods for genotyping, imputation and meta-analysis have gradually improved gene mapping in diverse populations, including through denser and custom arrays (Charles et al., 2014; Wang et al., 2013; Harlemon et al., 2019). One particular complication of trans-ethnic GWAS is the replication of lead SNPs, which can differ due to population-specific LD even if the causal variant is shared[41]. Accurate imputation also requires population-specific reference panels (Teo et al., 2010). It has been widely acknowledged that a lack of replication does not guarantee that a signal is population-specific (Fu et al., 2011; Kato, 2012; McCarthy et al., 2008; Visscher et al., 2012a). Alternative approaches used to define replication include exact replication: nominal significance of the same lead SNP, often p-value <0.05, local replication: significance of the same lead SNP or any SNP in strong LD, and local transferability: the significance of any SNP within a pre-defined physical distance (Clarke et al., 2007; Charles et al., 2014; Ng, 2015). Local replication greatly improves replication rates over exact replication by accounting for differences in LD structure (Shriner et al., 2009)[42]. It should be noted that exact and local replication assume that a signal is driven either by the same causal variant or variants on the same haplotype, whereas local transferability may detect allelic heterogeneity of different causal variants disrupting the same functional element; a gene or enhancer for example.

LDU-based gene mapping and African-American GWAS

A limited number of T2D loci have been identified or replicated in African Americans to date. This may be attributed to the less extensive LD, allelic heterogeneity and different allele frequencies (Ng, 2015), as well as the accuracy of imputation and variant calling (Rosenberg et al., 2010; Huang et al., 2009). For example, Ng et al. (2014) mapped five

[41]The independent lead SNPs may be in high LD with the causal variant in each population, but may not be in LD with each other.

[42]The term 'transferability' may also refer to the same direction of effect rather than significant replication (Shriner et al., 2009; Mahajan et al., 2014a).

significant T2D loci in an African-American meta-analysis consisting of **8,284 cases and 15,543 controls**. In the recent T2D GWAS published by Vujkovic et al. (2020), which included a meta-analysis of **228,499 T2D cases and 1,178,783 controls** including 19.5% African Americans, a total of 21 significant T2D-associated SNPs were reported in African Americans including three which were population-specific.

By comparison, Lau et al. (2017) used LDU-based gene mapping to map **57 loci in 965 T2D cases and 1,029 controls** of African American ancestry at genome-wide significance. Lau et al. further reported a total of 93 T2D locations which replicated in African Americans and Europeans based on nominal significance (p-value $<10^{-3}$) in both cohorts, co-location within $\pm$100 kb and Bonferroni-corrected genome-wide significance in a combined meta-analysis. LDU-based gene mapping offers substantial advantages for trans-ethnic gene mapping, since the model of association incorporates population-specific LD structure rather than using it *post hoc* to investigate replication. There is also a reduced multiple testing burden since each test of association involves multiple markers (described in detail in **Section 2.7: Association mapping using LDU maps**).

In conclusion, the analysis of diverse populations has successfully increased the number of known risk loci and revealed novel disease-associated pathways. Trans-ethnic GWAS have shown that common T2D risk variants tend to replicate across populations (Waters et al., 2010), consistent with complex disease being largely driven by shared, common variants. However, many loci also show population-specific effects, which may be used to investigate phenotypic heterogeneity, including population-specific risk and prognosis (Spanakis and Golden, 2013).

2.5 Beyond GWAS: lessons learnt from sequencing

In contrast to the GWAS described in the previous sections, sequencing studies do not rely on indirect genotyping or imputation and instead directly genotype all variants. They have been used to successfully map novel risk variants including structural and rare variants which are not included on arrays, hence revealing novel insights into T2D genetic architecture and challenging previous conclusions that common variants drive complex

disease (in addition to the efficacy of the indirect, array-based GWAS design)[43] (Morris et al., 2012). Several sequencing studies for T2D are described below, with particular focus on the novel findings for which single-SNP tests are underpowered to detect, thus emphasising the need for alternative methods such as LDU-based gene mapping. Sequencing studies may be used in hypothesis-free locus discovery by targeting the whole-genome or whole-exome, while targeted sequence data can be used to fine-map candidate loci and obtain more accurate effect-size estimates (this is the approach used to fine-map a candidate T2D NEMG locus in Chapter 5) (Nasykhova et al., 2019).

Whole-exome sequencing in T2D

Whole-exome sequencing (WES) studies target only the protein-coding exome, thus reducing costs and allowing for larger sample sizes. WES studies frequently make use of aggregate tests, which analyse the combined impact of independent (often rare) variants on pre-defined functional units; most commonly protein-coding genes (Liu and Leal, 2010; Lee et al., 2014). Aggregate tests, also known as gene-based, variant-set and aggregate unit tests, are a powerful method with which to detect allelic heterogeneity and effectively reduce the multiple testing burden by testing multiple variants per test. Methods can be grouped into collapsing approaches, which compare the numbers of cases and controls with at least one variant and burden tests, which assess the combined effects of multiple, often rare variants (Lee et al., 2014; Nicolae, 2016; Povysil et al., 2019)[44].

Examples of WES in T2D include that of **20,791 T2D cases 24,440 controls** by Flannick et al. (2019). Gene-level tests reported significant enrichment of rare variants in cases for four genes, *SLC30A8, MC4R, PAM* and *UBE2NL*. The authors highlighted limitations with array-based GWAS, demonstrating that 95.3% and 74.6% of the SNPs contributing to the gene-level association could not be imputed using the 1000 Genomes

[43]While more recent GWAS arrays and reference panels have included increasing numbers of rare variants, the accurate calling of rare variants is still challenging, see Section 2.6: Missing heritability.

[44]Of interest is the recent report by Cirulli et al. (2020), in which a gene-based, collapsing rare variant (MAF <0.1%) approach was used to analyse thousands of traits in over 70,000 exomes from the UK Biobank and Healthy Nevada Project (HNP). The authors report significant associations driven by alleles below the MAF detectable by arrays, including substantial contributions from ultra-rare singletons (variants found in only one individual).

and Haplotype Reference Consortium (HRC) reference panels, respectively. Flannick et al. (2019) acknowledged that more genes would achieve exome-wide significance with a considerably larger sample size. Further examples include the WES of **574 T2D cases** and **290 controls** of Qatari ancestry in which six genes were significantly enriched for low-frequency, protein-altering mutations (*CTNNB1, DLL1, DTNB, DVL1, EPB41L3*, and *KIF12*), although none of the genes were replicated (perhaps being population-specific variants). WES has also been used to identify rare variants in genes which associate with the risk of proliferative diabetic retinopathy (Ung et al., 2017; Shtir et al., 2016).

Whole-genome sequencing in T2D

Compared to WES, whole-genome sequencing (WGS) is considerably more expensive. Studies with high-quality WGS are limited to small sample sizes, while larger studies may reduce costs by carrying out low-coverage sequencing supplemented using genotyping by array and imputation. However, the importance of whole-genome discovery is emphasised by the findings from GWAS indicating that the majority of T2D loci are non-coding (Tak and Farnham, 2015; Fuchsberger et al., 2016). For example, Cirillo et al. (2018) investigated published T2D GWAS SNPs, of which 98% were non-coding (this is discussed further in Chapter 3).

In 2014, Steinthorsdottir et al. combined WGS from **2,630 Icelandic individuals** with imputation in a further ∼11,000 cases and ∼276,000 controls to successfully identify several novel variants associated with T2D. These included two low-frequency missense mutations in *PAM*, a rare frameshift mutation in *PDX1* and a low-frequency, self-regulatory variant in *CCND2* intron 1. A total of 34.2 million variants were tested, representing a substantial increase in the number of variants investigated compared with genotyping arrays. More recently, the GoT2D and T2D-GENES Consortia published the results of low-coverage WGS in **1,326 T2D cases** and **1,331 controls**, combined with high-coverage WES in **6,504 T2D cases** and **6,436 controls** from five ancestry groups, plus genotyping and imputation in a further 111,548 subjects (Fuchsberger et al., 2016). In total, 26.7 million variants were investigated, including 1.5 million indels and 8,876 large deletions.

The authors reported 2.4 million low-frequency SNVs which were poorly tagged by arrays (r^2 <0.3). No genes were significantly enriched for rare or low-frequency variants, however rare variants were nominally enriched in genes implicated in monogenic diabetes and this was driven by SNVs with MAF <1% (p-value = 2.8×10^{-5}). A number of genome-wide significant T2D-associated SNPs were identified, including several novel signals. Fuchsberger et al. (2016) observed larger effect sizes for lower frequency alleles, but concluded that low-frequency variants explained limited heritability compared to common variants. However, it is important to note that the WGS data in this study consisted of low-coverage reads at an average of 5× coverage, which was supplemented by genotyping arrays and imputation. Belsare et al. (2019) recently demonstrated that this approach limits the accuracy with which rare variants are called. In independent work, Ros-Freixedes et al. (2018) demonstrated that low-coverage sequence data may cause biases in favour of the reference allele. WGS was also carried out in 20 large Mexican-American families with high prevalence of T2D, consisting of **600 individuals**. No significant associations were observed for individual rare variants or for gene-level tests and the authors concluded that large variations in T2D risk were unlikely to be explained by rare variants with large effects in these pedigrees.

It is worth noting that gene-based tests may also have limited power. Moutsianas et al. (2015) demonstrated that common gene-based methods had around ∼5-20% power to detect loci explaining ∼1% of phenotypic variance in 1,500 cases and 1,500 controls, depending on the genetic architecture (MAF, number of causal variants etc) which increased to a modest ∼60% in 10,000 individuals.

2.6 Missing heritability

Following over a decade of T2D GWAS, a substantial amount of T2D heritability remains unexplained. Two recent GWAS have estimated that the total SNPs genotyped in each respective study explained 18% and 19% of T2D heritability (Mahajan et al., 2018; Vujkovic et al., 2020). A recent study by Willemsen et al. (2015) estimated T2D heritability at 72% (discussed in **Chapter 1: Section 1.3.1: T2D heritability**). The popularly

termed 'missing heritability' (Maher, 2008) has fuelled debates regarding the genetic architecture of complex disease and remains the focus of various publications (Young, 2019; López-Cortegano and Caballero, 2019).

Several studies have reported that missing heritability can in fact be explained using all variants. Wainschtein et al. (2019) recently showed that the heritability of height and BMI could be entirely recovered using mixed linear modelling of WGS data, with the classic 'missing heritability' explained by rare variants, particularly those in regions of low LD. These results highlight the need for methods which can map causal variants with challenging genetic architectures, including low-frequency and rare variants in regions of low LD.

Commenting on the future use of GWAS, Visscher et al. (2017) noted that no complex trait had yet shown a plateau of locus discovery with increasing sample size and anticipated further discoveries with larger cohorts, as well as a future shift towards GWAS using WGS data (discussed in **Section 2.5: Beyond GWAS: lessons learnt from sequencing**). A prominent criticism of increasing sample size is that the collective contribution of variants with very small effects (which may require impractical sample sizes to detect at genome-wide significance) may eventually implicate almost all of the genome, challenging any useful biological insight (Goldstein et al., 2009). This may be because almost all genes can be related in one way or another to a subset of core disease-related genes due to the inter-connection of gene regulatory networks (Boyle et al., 2017). Under such a model, large-effect, rare variants may offer more insight into core disease genes (Povysil et al., 2019). With this in mind, an important aim of gene mapping is to identify core genes and pathways which underlie disease, disease heterogeneity and variable risks of micro- and macro-vascular complications (Ahlqvist et al., 2018; Udler et al., 2018; Udler, 2019). Alternative gene mapping methods which identify additional, large-effect disease loci therefore hold promise for providing important insights towards this goal. One of these methods is LDU-based gene mapping.

The missing heritability has been attributed to many causes, of which different meth-

ods have different levels of sensitivity to. These include: overestimated heritability, caused by gene-gene interaction effects (Zuk et al., 2012) or by estimating heritability for outbred populations using family or twin data (Groop and Pociot, 2014); gene-gene, gene-environment and non-additive effects, including the genetic nurture effect resulting from parental genotypes (Kong et al., 2018), sex-specific effects (Small et al., 2018) and parent-of-origin effects (Lyssenko et al., 2015)[45]; the epigenome (Ling and Rönn, 2019), the microbiome (Sandoval-Motta et al., 2017) and the *in utero* environment (Smith and Ryckman, 2015); as well as genetic and phenotypic heterogeneity (Groop and Pociot, 2014; Udler, 2019). Other causes include aspects of genetic architecture which single-SNP tests of association are underpowered to detect: LD breakdown and inaccurate imputation; structural variation such as insertions, deletions, copy number and structural variants, which may not be imputed or tested in array-based GWAS (Mahajan et al., 2018; Vujkovic et al., 2020); and rare variants and allelic heterogeneity. An additional limitation preventing the detection of significant risk loci is the stringent multiple testing burden. Several of these are directly addressed by the LDU-based gene mapping followed by targeted sequencing approach used by Lau et al. (2017) and in this study. These are discussed below.

2.6.1 LD breakdown, rare variants and impuation

Several previous studies have observed that low frequency variants and regions of low LD are enriched for complex trait heritability (Gazal et al., 2017; Zeng et al., 2018; Wainschtein et al., 2019). However, it is particularly difficult to map causal variants in regions of low LD using indirect, single-SNP tests of association, since accurate imputation relies on high LD with markers. Rare variants are typically in very low LD with common marker SNPs (Li et al., 2013) and thus indirect genotyping is very inefficient at capturing rare variant associations, often resulting in underestimated effect sizes and heritability

[45]Considering interaction effects may also improve the detection of risk loci (Woo and Reifman, 2018). Jyothi and Reddy (2015) reported two variants which became significantly associated with T2D when adjusted for various factors including sex and BMI. Keaton et al. (2018) identified a novel risk locus at the *CMIP* gene based on the analysis of genetic interactions with an intronic variant within *MTNR1B*, which was itself associated with insulin secretion.

(Yang et al., 2010; Zeng et al., 2018). By extension, rare variants are often imputed incorrectly (Huang et al., 2009; Marchini and Howie, 2010), including in low-coverage sequence data (Belsare et al., 2019) (such as in the T2D WGS analysis by Fuchsberger et al. (2016) or the 1000 Genomes data (Consortium et al., 2015a)). Rare variants may be surprisingly common[46] in regions of LD breakdown since mutation rate is correlated positively with recombination rate and negatively with LD (Smith et al., 2005; Hellmann et al., 2005; Spencer et al., 2006; Webster and Hurst, 2012; Arbeithuber et al., 2015; Halldorsson et al., 2019; Castellano et al., 2020). *De novo* mutations are also enriched near recombination events (Besenbacher et al., 2016)[47].

Interestingly, patterns of LD are influenced by genomic function. Selection against increased rates of mutations likely drives the observed depletion of recombination hotspots within genes (McVean et al., 2004; Myers et al., 2005; Mackiewicz et al., 2013; Liu et al., 2017), particularly highly conserved genes (Liu et al., 2017; Castellano et al., 2020). Genetic diversity is greater at regions with more recombination since selection can act on independent variants rather than on haplotype blocks[48]. Consistent with this, there are low levels of LD (high recombination rates) surrounding genes which are thought to benefit from increased allelic diversity; these include the immune response and sensory perception (Chuang and Li, 2004; Smith et al., 2005; Sun et al., 2011; Gibson et al., 2013). The LD surrounding genes relating to T2D will be an intriguing area of further study, since metabolic pathways involved in evolutionary adaptation may also have benefitted from higher levels of recombination and increased allelic diversity[49]. Rare and *de novo* mutations may also contribute significantly to phenotypes with a large mutational target, defined as the proportion of the genome which contributes to trait heritability (Lupski

[46]For example, 93.5% of the exonic variants found ∼45,000 T2D cases and controls had a MAF <5% (Flannick et al., 2019). This may be explained by relaxed selection pressures due to improved life expectancy, increasing mutation rates with increasing parental ages and accelerating population growth (Coventry et al., 2010; Lupski et al., 2011).

[47]Potentially due to additional rounds of DNA replication, as well as biased gene conversion in which one allele is more likely to be the donor for recombination (Duret and Arndt, 2008; Berglund et al., 2009).

[48]known as Hill-Robertson interference (Keightley and Otto, 2006; McVean and Charlesworth, 2000).

[49]It might also be further investigated whether the relationship between recombination and greater allelic diversity may differ for pathways involved in nutrition and insulin resistance compared to those regulating pancreatic β-cell function.

et al., 2011; Stanley and Kulathinal, 2016). Phenotypes associated with evolutionary adaptation are also associated with enhancers susceptible to deactivating mutations (Li et al., 2019), potentially implicating additional non-coding mutations.

Single-SNP tests of association are underpowered to detect rare variants. Purcell et al. (2003) estimated that a cohort of 1,500 cases and 1,500 controls offered ∼5% power to detect a variant with MAF = 0.5% and relative risk = 3 at genome-wide significance. Visscher et al. (2017) noted that a sample size of over one million would be required to achieve 80% power to detect an association for a variant with 0.001% MAF and odds ratio >4[50]. This is assuming accurate genotyping, since genotyping errors will cause further reductions in power (Gordon et al., 2002; Kang et al., 2004). Rare variants are particularly likely to be imputed incorrectly or called incorrectly from low-coverage sequence data, non-optimised bioinformatics methods or genotyping arrays (which require multiple data points to accurately separate genotype clusters) (Huang et al., 2009; Marchini and Howie, 2010; Belsare et al., 2019; Yang et al., 2010; Zeng et al., 2018; Ren et al., 2018; Anderson et al., 2010; Weedon et al., 2019; Wright et al., 2019). Furthermore, multi-ethnic GWAS are also unable to increase power for variants which arose after population divergence. As a consequence, rare variants have particularly underestimated effect sizes and rarely pass genome-wide significance (Stringer et al., 2011).

Multiple rare and low-frequency variants have been associated with T2D, partly through array-based GWAS with larger sample sizes and denser reference panels (Mahajan et al., 2018), but also through gene-based aggregate tests (Bonnefond and Froguel, 2015; Flannick, 2019). Many rare variants observed to contribute to gene-wide significance can not be accurately imputed (Flannick et al., 2019). Importantly, methods for aggregate tests are comparatively more advanced for coding compared to non-coding DNA[51], despite the majority of T2D signals being non-coding (Cirillo et al., 2018). This arises from easily

[50]Notably, the largest T2D GWAS to date included a sample size of ∼1.4 million, but truncated the analysis for SNPs with a MAF >1% in non-Europeans and >0.1% in Europeans (Vujkovic et al., 2020).

[51]For a brief review of potential methods used to aggregate and weight variants within non-coding elements, see Povysil et al. (2019).

defined gene boundaries, more apparent effects on protein-function[52] and the reduced expense of exome studies. The functional impact of non-coding variation is also difficult to predict. Enhancer elements, for example, may have different levels of susceptibility to deactivating mutations (Li et al., 2019). Allelic heterogeneity in regulatory elements may therefore be expected to explain additional T2D heritability[53].

A further problem with imputation arises when haplotype frequencies differ significantly between cases and controls, since imputation is based on general population reference panels. Maniatis et al. (personal communication) demonstrated that risk variants identified using targeted sequence data could not be imputed correctly at three candidate loci, due to significant differences in haplotype frequencies.

Multiple testing threshold

The conventional genome-wide significance threshold of 5×10^{-8} remains the subject of continued debate. Fadista et al. (2016) advocated more stringent thresholds for lower-frequency variants, since these are present at greater numbers in the genome and a greater number of tests must therefore be corrected for. Conversely, Irizarry (2017) argued that lowering the GWAS threshold would save millions of dollars and allow for more true positives to be identified. Several approaches have been suggested for reducing the multiple testing burden, such as restricting association analysis to pre-defined candidate regions or by aggregating SNPs in blocks of high LD (Huang et al., 2007; Wu et al., 2010; Li and Meyre, 2013; Yoo et al., 2017; Guinot et al., 2018). Eskin (2008) suggested that LD information may be used to set significance thresholds which are specific to each marker and the number of variants they correlate with, while Kaler and Purcell (2019) proposed an alternative p-value correction based on the estimated heritability of the phenotype.

[52]Coding mutations can be weighted based on predicted functional impact, for example gene sub-regional intolerance (Hayeck et al., 2019) or gene-based effect size (Lali et al., 2020).

[53]In their WGS study of T2D, Fuchsberger et al. (2016) tested the burden of rare variants in pre-defined pancreatic enhancers, reporting no evidence of enrichment. However, the WGS cohort included 2,657 combined cases and controls, offering only limited power (Moutsianas et al., 2015). Furthermore, the approach of supplementing low-coverage WGS data with imputed arrays may not achieve high levels of accuracy for calling rare variants (Belsare et al., 2019).

2.7 Association mapping using LDU maps

In single-SNP GWAS, LD is typically used to design genotyping arrays by placing marker SNPs to capture common variation, or to test the local replication of significant lead SNPs *post hoc*. Described in this section is a gene mapping method developed by Maniatis et al. (2004) in which LD is instead incorporated into the test for association. A likelihood framework is used to test multiple genotyped markers at once, both avoiding imputation and reducing the multiple testing burden. As opposed to methods which claim to incorporate LD by collapsing variants into 'blocks', the method of Maniatis et al. captures information from multiple genotyped SNPs, utilising the underlying LD structure and observed patterns of association to estimate the location of a causal variant. LD information is obtained from high-resolution population-specific genetic maps measured in linkage disequilibirum units (LDU). The construction, structure and applications of LDU maps are described in the next section. The use of LDU maps in testing for disease association is described in **Section 2.7.2: LDU-based gene mapping**, while the application of this method to T2D by Lau et al. (2017) is reviewed in **Section 2.7.3: LDU-based gene mapping in T2D**. Importantly, data from the Lau et al. study will be analysed in the following Chapters, which in turn will refer to methods detailed in this section.

2.7.1 Genetic and LDU maps

Genetic maps are an important tool in association-based gene mapping. Due to LD structure (described in **Section 2.3.2: linkage disequilibrium and association mapping**), nearby variants are inherited together with causal variants on shared haplotypes, causing all variants in high LD to be observed as associated with disease. This is crucial for indirect genotyping methods, since marker SNPs are only co-inherited with the phenotype if they are in high LD with the causal variant (assuming the marker is not itself the causal variant). LD depends on the frequency of recombination and this occurs non-randomly across the genome. Patterns of LD are captured in genetic maps as units of genetic distance, such that a large genetic distance corresponds to significant LD breakdown (two

variants are inherited together less often than two markers separated by a small genetic distance, even if they are separated by the same number of base pairs).

There are several different types of genetic maps. These include linkage maps, which can be constructed from current recombination events observed in family data and population-based maps built using coalescent theory (McVean et al., 2004). LDU maps are constructed by modelling the frequency at which alleles are found together across a population (the extent of LD) and therefore reflect historical recombinations and other population-specific demographic events which influence LD.

Genetic LDU maps: construction

In 2002, Maniatis et al. constructed the first genetic map measured in additive linkage disequilibrium units (LDU)[54]. Metric LDU maps are constructed using a Malecot model, first described by Malécot (1948) to model the decline in genetic relatedness of individuals with increasing geographical distance and later adapted by Collins and Morton (1998) to model the decline in LD with increasing genomic distance. The relationship between pairwise SNP association (ρ, see page 65) and physical distance (d) is plotted in Figure 4, to which an exponential decay curve is fitted using the Malecot equation:

$$\rho = (1 - L)Me^{-\epsilon d} + L$$

Where L is the residual association at large distances (the asymptote), M is the intercept, i.e. ρ at zero distance and ϵ is the exponential decline of association (how steep the curve is). Empirically observed pairwise SNP association, $\hat{\rho}$ is calculated below for two example loci A and B, assuming that b is the rarest allele and $D_{AB} = p_{AB} - p_A p_B$. The haplotype frequencies are denoted p_{AB}, p_{Ab}, p_{aB} and p_{ab} and the allele frequencies are p_A, p_B, p_a and p_b. ρ equals 1 when the SNPs are in complete LD.

$$\hat{\rho}_{AB} = \frac{D_{AB}}{p_A p_b}$$

		Locus B	
		B	b
Locus A	A	p_{AB}	p_{Ab}
	a	p_{aB}	p_{ab}

[54]The construction and uses of LDU maps are described according to Maniatis et al. (2002) and also Chapters 3 and 4 of Collins (2007).

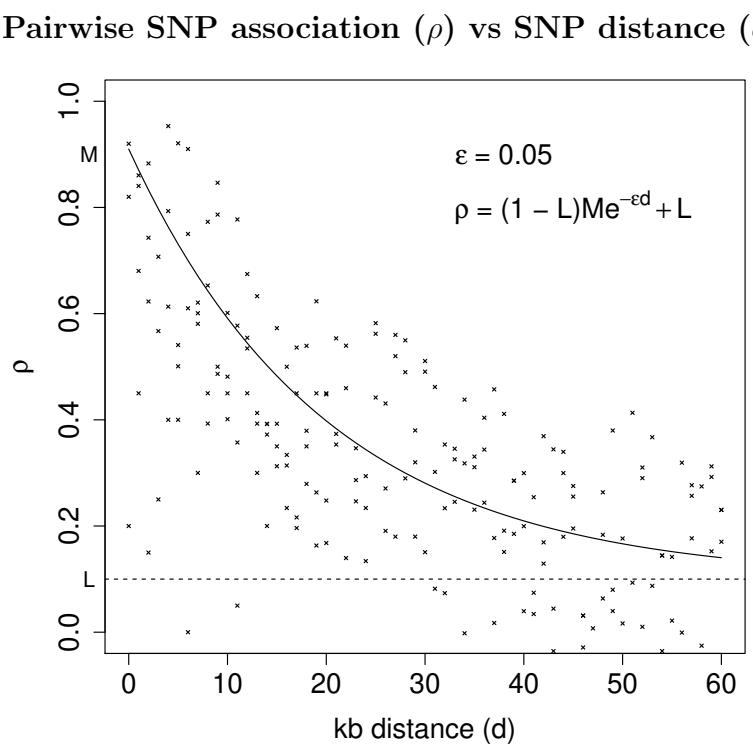
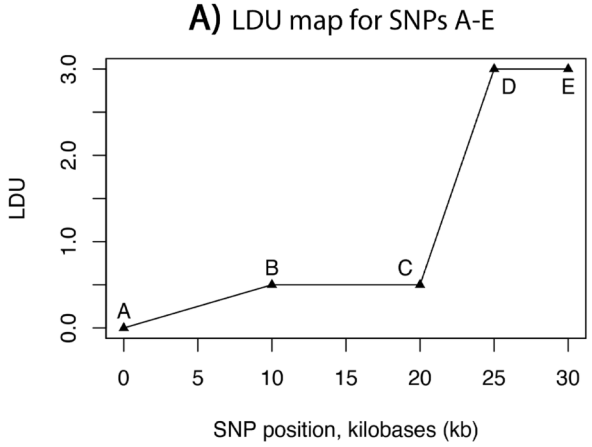


Figure 4: Theoretical LD between pairs of SNPs, measured as ρ, plotted against the distance between the SNP pairs (d). The fitted line is estimated using the Malecot equation as shown below.

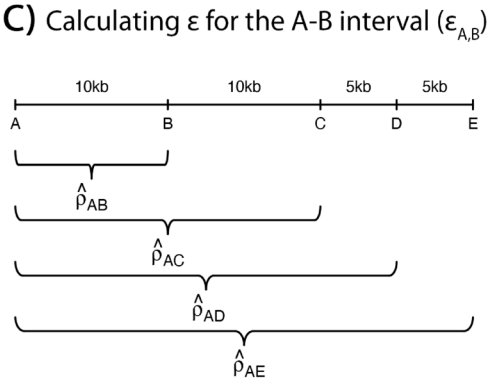
The metric ρ was shown by Morton et al. (2001) to be the most efficient metric for modelling the relationship between physical distance and LD compared to other measures such as D' or r^2, which are in turn more sensitive to the frequencies of the alleles being measured. LDU map construction requires ρ to be calculated for all SNP pairs using genotype data from multiple unrelated individuals. LDU are calculated using the exponential decline in association with distance, such that LDU $= \epsilon d$. To construct the LDU maps, the Malecot model is run iteratively to calculate the LDU between each pair of adjacent SNPs (e.g. A-B, B-C, C-D and D-E in the example below). LDU is additively measured as a sum of the LDU between the adjacent SNP pairs.

Figure 5 shows an example LDU map, adapted from the example shown by Collins (2007) and Tapper (2007). LDU for the ith interval is calculated as $\epsilon_i d_i$, where d_i is the distance between the two adjacent SNPs and ϵ_i is the Malecot parameter for that SNP interval. To calculate ϵ separately for each interval, the Malecot model is run on the pairwise SNP association, ρ, and distances, d, for all marker pairs which contain the interval, since these



B) Calculating LDU

SNPs	ϵ		Distance		LDU
A-B	0.05	x	10kb	=	0.5
B-C	0.00	x	10kb	=	0.0
C-D	0.50	x	5kb	=	2.5
D-E	0.00	x	5kb	=	0.0



$$\textbf{Total LDU} = \Sigma \epsilon_i d_i = \epsilon_{A,B} d_{A,B} + \epsilon_{B,C} d_{B,C} + \epsilon_{C,D} d_{C,D} + \epsilon_{D,E} d_{D,E}$$

$$\textbf{Total LDU} = 0.05{\times}10\text{kb} + 0.00{\times}10\text{kb} + 0.50{\times}5\text{kb} + 0.50{\times}0\text{kb} = 3\text{LDU}$$

Figure 5: LDU map construction, adapted from Collins et al. (2004) and Tapper (2007). A) LDU map for an example region containing SNPs A-E. B) The LDU for each adjacent pair is calculated as a product of the distance between the SNPs and the decline in association with distance (ϵd). C) ϵ is iteratively calculated for each adjacent SNP pair, starting with A-B (then B-C, C-D and D-E). The model is fitted to ρ and d for all informative SNP pairs which contain the adjacent SNP interval. In this example, $\epsilon_{A,B}$ is calculated by fitting the model to the intervals A-B, A-C, A-D and A-E.

contain information on the decline in association from the first marker (Figure 5C). The parameters L and M remain constant for each likelihood iteration. The pairs are weighted such that less weighting is given for pairs separated by larger distances, since LD declines with distance and larger intervals are less informative. The fitting of the Malecot model to the pairwise measures of ρ and d is achieved using a composite maximum likelihood approach, which estimates the optimal values of the parameters M, L and ϵ to minimise the differences between the observed and fitted values of ρ. The composite likelihood is calculated as $lnlk = -\Sigma K_\rho(\hat{\rho}_i - \rho_i)^2/2$ where $\hat{\rho}_i$ and ρ_i are the observed and fitted values

between the ith pair and $K_\rho = \chi^2/\hat{\rho}^2$.

Genetic LDU maps: properties

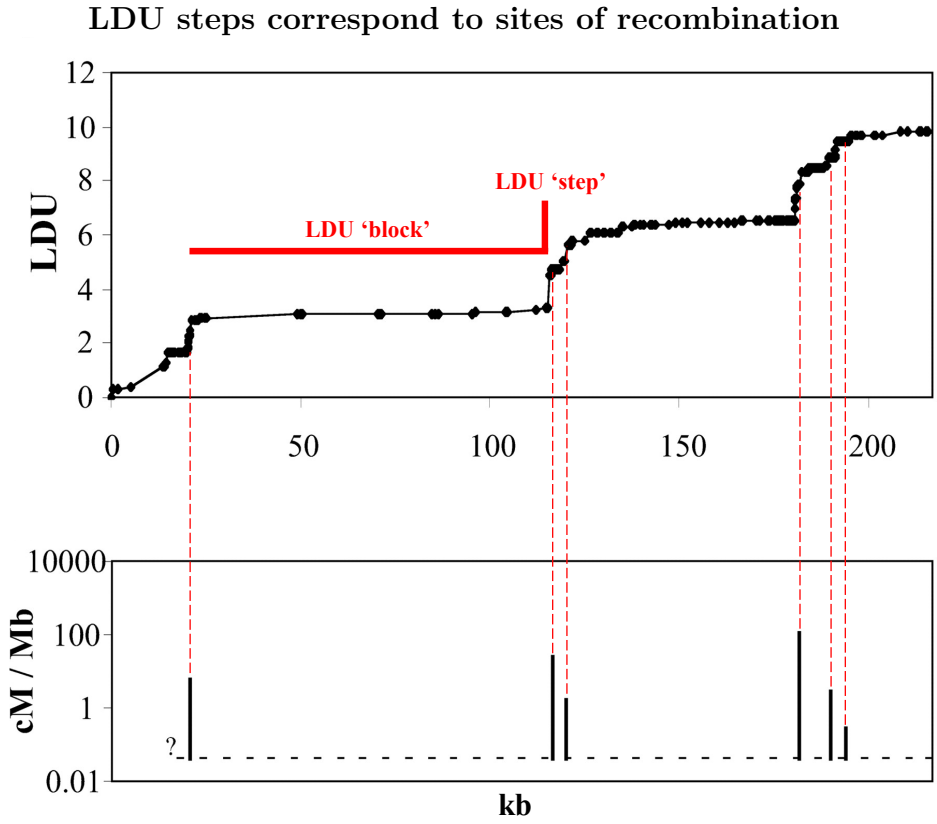


Figure 6: LDU and recombination maps for chromosome 6p21.3 region. This image is adapted from Zhang et al. (2002) (Figure 2), with recombination data from Jeffreys et al. (2001). Added notation is shown in red. This early version of the metric LDU map shows that LDU steps coincide with observed sites of recombination in sperm data.

LDU maps show a 'step' and 'block' structure which is illustrated in Figure 6, where steps have a large increase in LDU relative to the physical distance and blocks have no increase in LDU. Zhang et al. (2002) constructed LDU maps for two datasets capturing sites of recombination (Jeffreys et al., 2001) and low haplotype diversity (Daly et al., 2001), confirming that LDU steps have a striking overlap with recombination events while blocks align to regions of low haplotype diversity and extended LD. Figure 6 shows a comparison by Zhang et al. (2002) of LDU with recombination frequency available at high resolution from the analysis of sperm by Jeffreys et al. (2001). This study confirmed that LDU maps can accurately capture recombination hotspots and LD structure, despite the large

stochastic variation observed in population data.

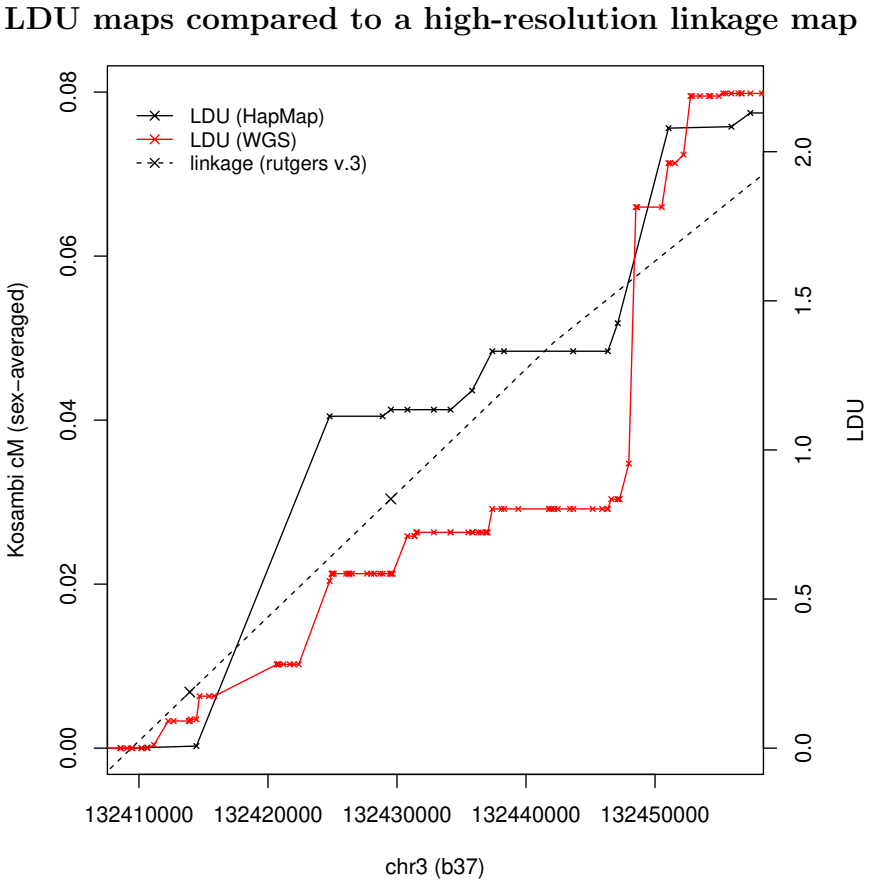


Figure 7: European LDU and linkage maps plotted for a 47 kb region at the chr3q22.1 T2D locus. LDU maps are based on genotype data from HapMap (Lau et al., 2017) and whole genome sequence (WGS) data (Jabalameli et al., 2019). The rutgers v.3 linkage map has been smoothed using local quadratic curves with interpolated SNP positions (there is no cM increase within this region on the non-smoothed map, demonstrating its limited resolution) (Kong et al., 2004; Matise et al., 2007). Genotyped markers are shown as crosses.

LDU maps have several important strengths. Firstly, high resolution is achieved since many historical recombinations are captured by assessing the frequency with which alleles are found together across a population. This difference in resolution is illustrated in Figure 7, which plots LDU and linkage maps for a 47 kb region at the chr3q22.1 T2D locus (this locus is fine-mapped in Chapter 5). Two separate LDU maps are shown, one constructed using HapMap release 28 genotypes for 56 individuals (∼2.2 million markers) (Lau et al., 2017) and a second constructed more recently using whole-genome sequence (WGS) data for 454 individuals (∼7.5 million markers) (Pengelly et al., 2015; Vergara-Lope et al.,

2019b). The HapMap LDU map is used for analysis in the following Chapters, although it is worth noting that the greater density of the WGS map allows additional resolution to be captured (96 markers are genotyped within the 47 kb region plotted in Figure 7 compared to 19 in the HapMap map). The high-resolution rutgers cM linkage map is also shown, built using ∼28,500 markers (Kong et al., 2004; Matise et al., 2007). Notably, the linkage map includes only two genotyped markers within the genomic segment region in Figure 7.

Secondly, LDU maps indirectly model all historical population demographic events which influence observed LD, including mutation, selection, effective population size, population age, bottlenecks and founder effects, allele frequencies, outbreeding, gene conversion and genetic drift (Zhang et al., 2004). These population-specific LD structures influence the tagging of causal variants in GWAS. Therefore while sperm recombination maps accurately capture recombination events, LDU maps capture allelic association at a population level which is required for association mapping using population data. On this note, population-specific LDU maps can be generated and used to study similarities between different populations such as recombination hotspot concordance (Gibson et al., 2005). Other uses of LDU maps include to investigate evidence of selection as well as population parameters such as divergence and estimated effective population size. LDU maps may identify novel recombination hotspots (Collins et al., 2001) (with the caveat that some recent hotspots may be missed and others may no longer be active (Jeffreys and Neumann, 2002; Jeffreys et al., 2005)). High-resolution LDU maps can be used to inform the selection of genotype markers, with high densities of markers in regions of LD breakdown. More recently, LDU maps were used to investigate how LD differs depending on the genomic context (Sved and Hill, 2018; Vergara-Lope et al., 2019a).

2.7.2 LDU-based gene mapping

A major application of genetic LDU maps is in disease gene mapping. In 2004, Maniatis et al. described positional cloning by linkage disequilibrium[55]; a gene mapping method

[55]This method is referred to as LDU-based gene mapping throughout this thesis.

which estimates the location of causal variants using LDU maps. The same Malecot model is adapted to model the association between each marker and a single phenotype rather than marker-by-marker association, as shown in Figure 8. The phenotype can be either dichotomous or continuous, such as disease status or gene expression levels. If a causal variant is present, the magnitude of trait association is expected to decline for SNPs which are increasingly further away, as a function of the declining LD. Distance is therefore measured in LDU. The Malecot equation is applied as:

$$y = (1 - L)Me^{-\epsilon|Si-S|} + L$$

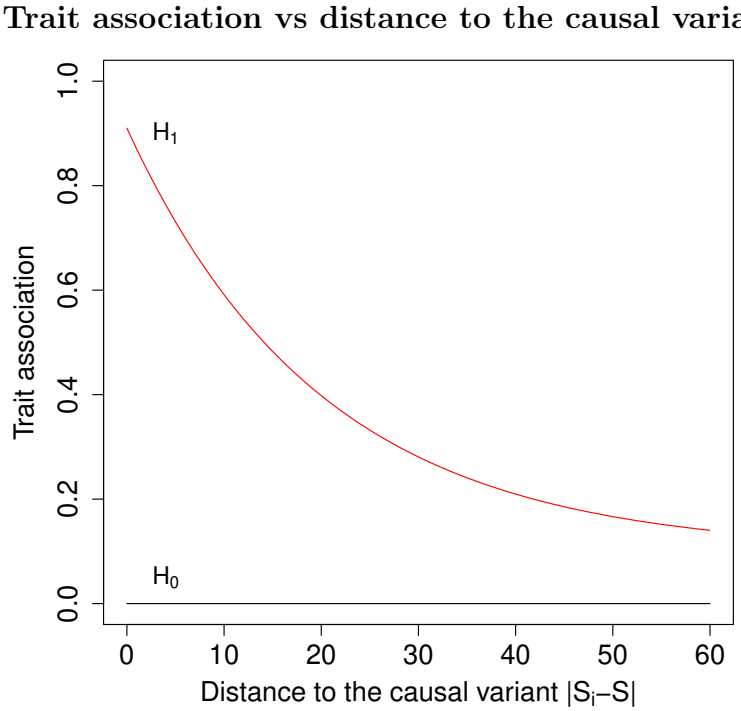


Figure 8: The relationship between trait association and the absolute distance between the causal variant of unknown location (S) and the ith SNP of known location (S_i), shown as $|S_i - S|$. There is no association under the Null hypothesis, H_0. Under the alternate hypothesis, H_1, the trait-SNP association decreases as the distance from the causal variant increases.

where y is a measure of trait association, such as a z-score of association with disease or a β-regression coefficient for a continuous trait. $|Si - S|$ represents the absolute distance between the ith SNP of known location (S_i) and the causal variant of unknown location (S). S is freely estimated along with the M, L and ϵ parameters.

The model effectively tests for a significant deviation from the Null hypothesis of no association in which the y-intercept, $M = 0$ (H_0), and the alternative hypothesis of association, $M > 0$ (H_1); these are shown in Figure 8. Both models are fitted and a χ^2 statistic[56] is calculated by comparing the composite likelihood of both models. For regions with significant evidence of trait association, the χ^2 will be large and optimal estimates of the model parameters will be returned for the best fit model, including the likely location of the causal variant (S). Therefore, the LDU-based test of association provides the estimated location of a causal variant (S), rather than a lead SNP as is the case with conventional single-SNP GWAS.

The LDU-based Malecot model has several important advantages compared to conventional single-SNP GWAS. Firstly, multiple genetic markers and underlying LD structure are used to inform gene mapping; this information is lost when analysing single-SNPs. Others include:

Improved power in regions of low LD: By simulating individual SNPs one at a time to act as the causal variant and using the remaining SNPs to estimate their location, Maniatis et al. (2004) showed that SNP positions in LDU rather than physical coordinate (kb) improved the model accuracy, since kb positions fail to consider the step and block structure of allelic association (Maniatis et al., 2002). The greatest improvement was seen in a region of LD breakdown. This is further demonstrated in original analysis presented in Figure 10 on page 76, showing that $\hat{S}_{T2D}$ from Lau et al. (2017) are surrounded by significantly greater LDU (LD breakdown) compared to GWAS lead SNPs mapped using dense imputation. LDU-based gene mapping also avoids the risk of incorrect imputation in regions of LD breakdown by using only directly genotyped SNPs.

Reduced multiple testing burden: The model is fitted once per analytical window, which substantially reduces the multiple testing burden compared to GWAS which test every SNP individually. For example, Lau et al. (2017) divided the genome into ∼4,800 analytical windows, making a stringent genome-wide significance threshold equal to ap-

[56]For small candidate regions, an F-test is more efficient than a χ^2 test.

proximately 1×10^{-5} (0.05/5,000). The significance threshold in single-SNP GWAS is set at 5×10^{-8} to correct for ∼1 million independent tests (Risch and Merikangas, 1996). A less stringent threshold, due to the reduced burden of multiple testing, makes it more likely that a larger number of true positives will achieve genome-wide significance.

Integrating independent and trans-ethnic data: Replicating lead SNPs in trans-ethnic GWAS is complicated by different LD structures, meaning that lead SNPs may differ between populations and may themselves not even be in LD, despite tagging the same causal variant. In comparison, LDU-based mapping incorporates population-specific LDU maps and outputs accurate estimates of causal variant locations. Furthermore, the increased power and lower multiple-testing burden makes it more likely that causal loci will achieve genome-wide significance in independent studies, allowing for confident replication.

The LDU-based methods of Maniatis et al. (2004) have since been applied to several complex diseases, including to map novel loci and refine estimates for Crohn's disease (Elding et al., 2013) and for T2D (Lau et al., 2017), described in detail below.

2.7.3 LDU-based gene mapping in T2D (Lau et al., 2017)

The Lau et al. (2017) study had three main aims, (1) to map the locations of T2D risk loci ($\hat{S}_{T2D}$) in Europeans and African Americans, (2) to map variants which regulate gene expression levels in adipose tissue, otherwise known as expression quantitative trait loci (eQTL)[57] ($\hat{S}_{eQTL}$) and (3) to assign target genes to T2D loci based on the assumption that co-locating $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ represent the same causal variant(s). The study included African Americans (AA) since few T2D loci had been identified and replicated for this informative population in which the prevalence of T2D is almost twice that of Europeans. African populations may also provide more accurate location estimates, since LD is generally less extensive in this older population (discussed in **Section 2.4.4: Trans-ethnic GWAS and defining replication**). The use of population-specific LDU maps

[57]eQTL are expression quantitative trait loci, defined as genetic loci where genotype associates with gene expression levels. These are described in more detail in Section 3.2.2.

also provided a powerful framework to test for shared risk loci.

(1) Mapping T2D loci ($\hat{S}_{T2D}$)

Dataset	# cases	# controls	Array (# of SNPs)
(1) European (WTC)	1,925	2,938	Affymetrix (∼500,000)
(2) European (MTC)	2,910	5,724	Metabochip[a] (∼200,000)
(3) African American (AA)	965	1,029	Affymetrix (∼1 million)

Table 1: European and African American T2D case-control datasets analysed by Lau et al. (2017). [a]The Metabochip is a custom array designed to capture genetic loci previously implicated in metabolic, cardiovascular and anthropometric traits (Voight et al., 2012).

Table 1 shows the T2D case-control datasets for two independent European cohorts and one African-American cohort. The European datasets (1) and (2) were obtained from the Wellcome Trust Case Control Consortium (WTCCC) phase I (Consortium et al., 2007b) and phase II (Voight et al., 2012), respectively. The African American dataset (3) was obtained from the National Institute of diabetes and Digestive and Kidney Diseases (NIDDK) (Palmer et al., 2012). The LDU-based Malecot model was applied to each dataset using population-specific LDU maps constructed from HapMap, on 4,800 analytical windows with a minimum length of 10 LDU comprising the autosomal genome. Replicated loci were defined where two or more datasets gave $\hat{S}_{T2D}$ within 100 kb of each other[58] and if the locus passed Bonferroni-corrected significance in meta-analysis (p-value $<10^{-5}$). Figure 9 plots the chr3q22.1 T2D locus and shows two significant $\hat{S}_{T2D}$ from the European WTC and MTC datasets; these are located 7 kb apart, or 0.086 LDU apart on the HapMap LDU map. This locus is European-specific since the AA dataset did not return a significant $\hat{S}_{T2D}$.

In total, Lau et al. (2017) mapped 111 novel T2D loci which were replicated in at least two datasets, of which 93 (84%) were shared between Europeans and African Americans. Including the replication of previously known loci, Lau et al. mapped 175 significant T2D loci using a combined sample size of **5,800 T2D cases and 9,691 controls**. In

[58]The use of genetic vs physical distance to define co-location are discussed further in Chapters 3/4.

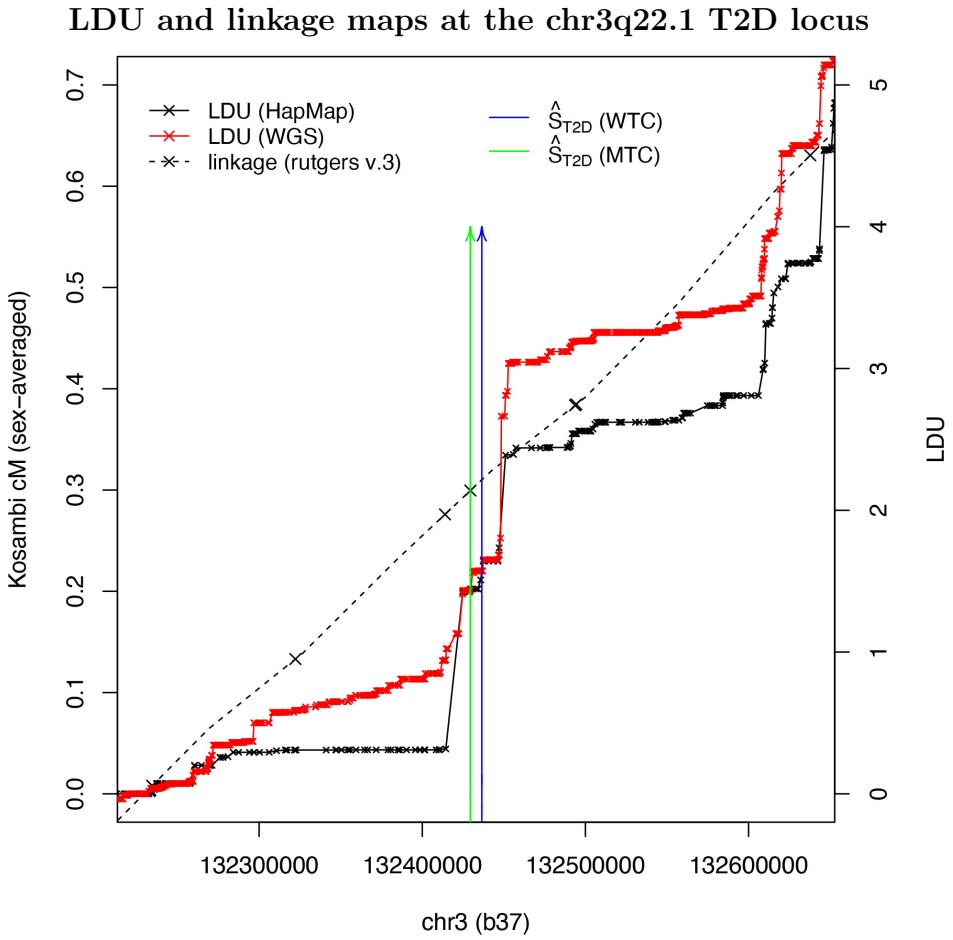


Figure 9: T2D location estimates ($\hat{S}_{T2D}$) plotted for ∼407 kb at the chr3q22.1 region with the HapMap LDU map (Lau et al., 2017), WGS LDU map (Jabalameli et al., 2019) and the smoothed rutgers linkage map v.3 (there is no cM increase within this region on the non-smoothed map) (Kong et al., 2004; Matise et al., 2007). The T2D locus is European-specific, with two independent $\hat{S}_{T2D}$ from the European WTC and MTC datasets located 7 kb and 0.086 LDU apart based on the physical and HapMap LDU maps, respectively.

comparison, the single-SNP GWAS by Mahajan et al. (2014a) included **26,488 T2D cases and 83,964 controls** and identified 76 significant loci. The greater number of loci can be attributed to the greater power of LDU-based gene mapping. This is also demonstrated by previous single-SNP analysis of the same datasets reporting substantially fewer results. For example, four significant SNPs were identified in the European WTC dataset using single SNP analysis (Consortium et al., 2007b; Huang et al., 2012), compared to 98 significant $\hat{S}_{T2D}$ (Lau et al., 2017).

There are several reasons why LDU-based analysis may identify more disease loci compared to single-SNP analysis, including increased power in regions of LD breakdown. To

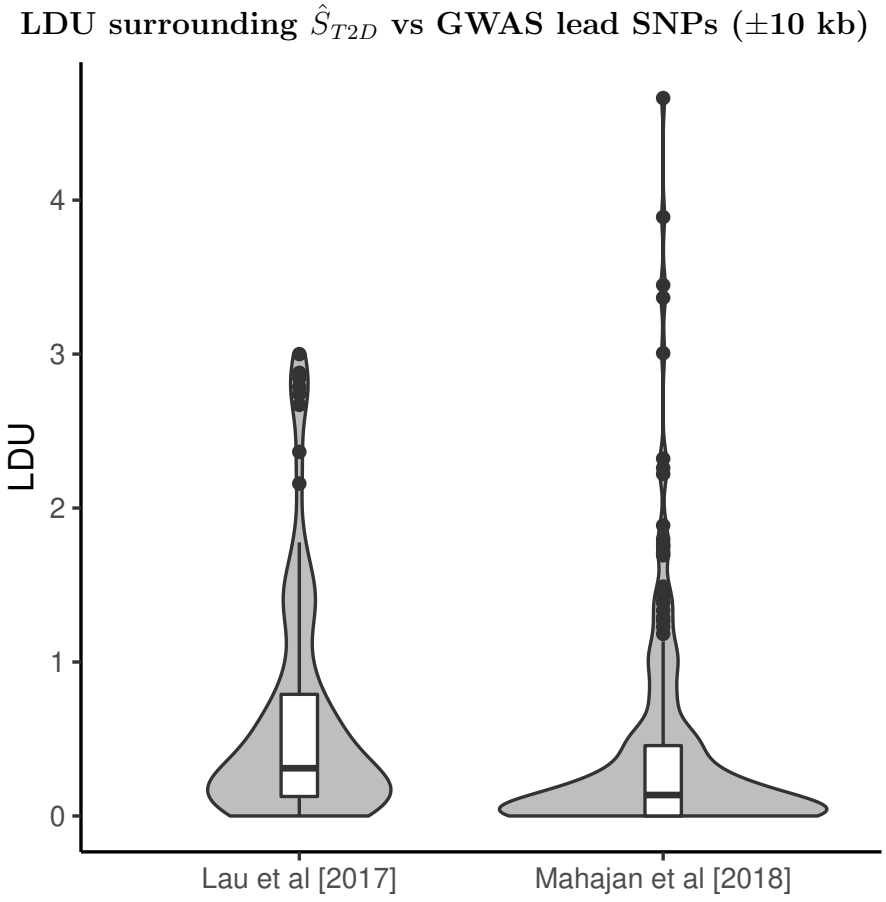


Figure 10: The LDU distance extending ±10 kb from Lau et al. (2017) $\hat{S}_{T2D}$ (n=111) and Mahajan et al. (2018) lead SNPs (n=243). Lower LDU corresponds to more extensive LD while larger LDU reflects a breakdown in surrounding LD.

investigate this, the surrounding LDU was compared for 111 $\hat{S}_{T2D}$ from Lau et al. (2017) and 243 lead SNPs from Mahajan et al. (2018). The physical coordinates of both $\hat{S}_{T2D}$ and lead SNPs were extended by ±10 kb and converted to genetic coordinates on the HapMap LDU map (European). The LDU distances for the surrounding 20 kb intervals are plotted in Figure 10. $\hat{S}_{T2D}$ are surrounded by significantly higher LDU (p-value = $2.87e$-06, Wilcoxon rank sum test) and therefore greater LD breakdown compared to lead SNPs. This is consistent with LDU-based gene mapping having increased power to map causal variants in regions of LD breakdown. Notably, 40 of the 111 novel T2D loci from Lau et al. (2017) have since been replicated in single-SNP GWAS, including 20 reported as novel sin the most recent T2D GWAS with 228,499 T2D cases using their replication criteria of ±500 kb (Vujkovic et al., 2020).

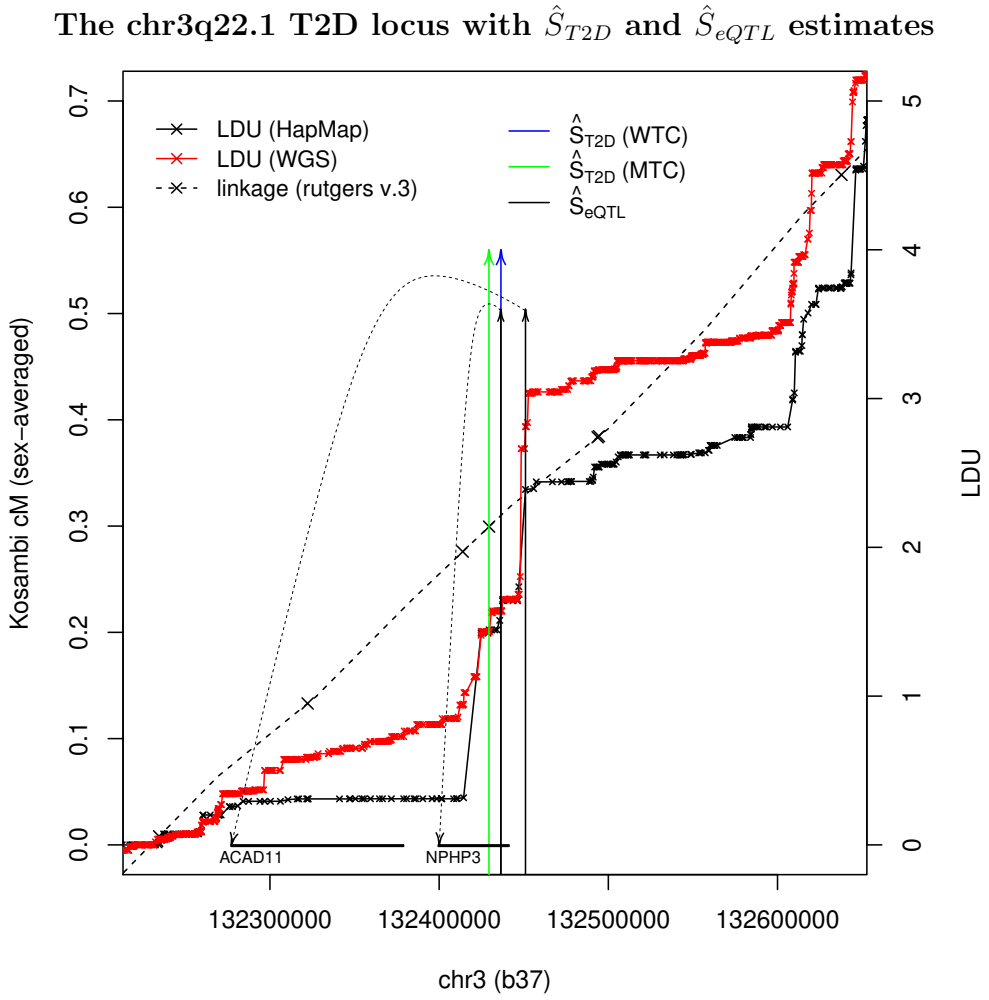


Figure 11: European T2D location estimates ($\hat{S}_{T2D}$) and eQTL location estimates ($\hat{S}_{eQTL}$) are plotted for the chr3q22.1 region with the HapMap LDU, WGS LDU and rutgers linkage maps. Two significant $\hat{S}_{eQTL}$ for the *cis*-genes *ACAD11* and *NPHP3* are shown as examples (other $\hat{S}_{eQTL}$ and genes are not shown for simplicity).

(2) MAPPING eQTL ($\hat{S}_{eQTL}$) and (3) ASSIGNING *cis*–GENES

Next, Lau et al. obtained adipose gene expression data for an ageing, population-based European sample from the MuTHER Consortium (TwinsUK) (Grundberg et al., 2012) and carried out association mapping using quantitative gene expression levels as the outcome. Genes within ±1.5Mb of replicated $\hat{S}_{T2D}$ were included in the analysis, which resulted in location estimates of causal variants associated with gene expression levels: adipose eQTL ($\hat{S}_{eQTL}$). Figure 11 shows two example $\hat{S}_{eQTL}$ at the chr3q22.1 T2D locus for the nearby genes *ACAD11* and *NPHP3*. Where $\hat{S}_{eQTL}$ co-located within 50 kb of previously replicated $\hat{S}_{T2D}$, Lau et al. defined shared signals (T2D-eQTL). This approach assumes that the same causal variant is responsible for both the association with T2D

($\hat{S}_{T2D}$) and gene expression ($\hat{S}_{eQTL}$). The genes associated with the co-locating $\hat{S}_{eQTL}$ were considered to be regulated in *cis* by the T2D-eQTL and are subsequently referred to as T2D *cis*-genes. In total, 104 T2D loci (including 33 previously known and 71 novel) co-located with $\hat{S}_{eQTL}$ for a total of 266 *cis*-genes.

The $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ mapped by Lau et al. will be analysed in Chapter 3 in order to (1) repeat T2D and eQTL co-location analyses using genetic, rather than physical distance and (2) identify functional target genes regulated in *cis* by T2D-eQTL loci (*cis*-genes), including those involved in mitochondrial function. Following this, Chapter 4 will aim to validate the identified T2D *cis*-genes by demonstrating differential expression in T2D cases compared to controls, using independent case-control gene expression datasets.

2.8 Discussion

The design of different gene mapping methods for complex disease is an important point of discussion, since any result may be influenced by the underlying assumptions or biases of the methods used. The most widely used method to date has been the single-SNP genome-wide association study (GWAS), with three recent GWAS by Mahajan et al. (2018); Vujkovic et al. (2020); Spracklen et al. (2020) reporting 403, 568 and 301 significant loci for T2D, respectively. These studies have aimed to improve power through larger sample sizes, higher resolution arrays and denser imputation panels. However, indirect genotyping and single-SNP analysis remain subject to limitations discussed in this Chapter, while large sample sizes require large expenditure. Furthermore, it has been suggested that increasing sample sizes to identify loci of smaller effects will eventually implicate the entire genome, placing a limit on how informative large studies can be (Goldstein et al., 2009; Boyle et al., 2017).

The variety of gene mapping methods available each have their own strengths and limitations. Sequencing studies address the limitations of indirect genotyping and avoid the inaccurate calling of very rare variants from genotyping arrays (Wright et al., 2019; Weedon et al., 2019). However, they come with high costs which may limit sample size and therefore power, as well as reduced quality across GC-rich regions and Alu repeats

(Freeman et al., 2020). Family-based designs offer more power to detect rare variants with very large effects (Terwilliger and Ott, 1994) and linkage studies are robust to allelic heterogeneity, despite their low resolution and need for multi-generational pedigrees (Flaquer and Strauch, 2012; Ott et al., 2015). Aggregate tests can detect independent rare variants, yet different methods are sensitive to different genetic architectures (e.g. MAF, selection pressure, number of causal variants) (Moutsianas et al., 2015) and they currently have limited use in non-coding regions (Povysil et al., 2019). Non-coding regions are also less well covered by sequencing studies, which often target the exome with high coverage and the non-coding genome with low coverage and imputation (Fuchsberger et al., 2016; Belsare et al., 2019).

This Chapter also discussed an alternative gene mapping method: LDU-based gene mapping, also called positional cloning by linkage disequilibrium. The application of this method to T2D by Lau et al. (2017) identified substantially more T2D loci compared to previous single-SNP analysis of the same data, for example detecting 98 significant, replicated $\hat{S}_{T2D}$ compared to only four significant SNPs in the WTCCC Phase I cohort (Consortium et al., 2007b; Huang et al., 2012). Crucially, the LDU-based method detected more results without the need for excessively large cohorts. Direct comparison with lead SNPs from single-SNP GWAS showed that T2D loci mapped using the LDU-based method were surrounded by significantly lower LD, suggesting that the two methods map loci with different genetic architectures. The results demonstrate that alternative methods which challenge the limitations of the conventional GWAS design can be used to successfully identify additional risk loci, including those in low LD (using LDU-based gene mapping) and rare variants (using aggregate tests). In other words, a diverse set of methods should be used to identify a diverse set of T2D risk loci.

As gene mapping methods have evolved it has become increasingly clear that T2D has a surprisingly diverse genetic architecture, to which its extensive phenotypic heterogeneity may be attributed (Udler, 2019). Common, intermediate-frequency, rare and *de novo* variants of various effect sizes all contribute to the heritability of T2D and other complex diseases, prompting the revision of exclusive hypotheses such as common disease, com-

mon variant (CD/CV) and common disease, rare variant (RD/RV) (Iyengar and Elston, 2007; Schork et al., 2009; Lupski et al., 2011; Visscher et al., 2012b). Further analyses may continue to study the unique genetic risk factors for T2D subtypes and associated complications (Ahlqvist et al., 2018; Udler, 2019). However, it is equally as important to establish the resulting molecular mechanisms which drive the risk of disease at both known and novel risk loci. As such, the following Chapters describe follow-up analysis to Lau et al. (2017), with Chapter 3 investigating the target genes of T2D risk loci in order to address a specific biological question: whether mitochondrial function is regulated by T2D genetic risk factors. The subsequent analyses aim to validate the *cis*–genes by testing their expression in independent cohorts of T2D cases and controls (Chapter 4) and to fine-map a candidate locus using targeted sequence data (Chapter 5).

2.9 Conclusions

This Chapter reviewed the current knowledge of T2D genetics and described a selection of popular gene mapping methods available to investigate the diverse genetic architecture of T2D. The potential biases and limitations of individual gene mapping methods motivates the conclusion that multiple methods should be used to gain a comprehensive view of T2D genetics. LDU-based gene mapping, the method used to map the T2D loci investigated in the following Chapters, addresses several limitations of indirect single-SNP tests of association and offers an effective tool to map novel risk loci, as well as to integrate independent datasets. In Chapter 3, genetic LDU maps will be used to integrate gene expression data and to identify the target genes of non-coding T2D loci which implicate mitochondrial function.

To conclude, there is evidence to suggest that neither GWAS nor sequencing studies may fully explain the familial heritability estimates of T2D (Fuchsberger et al., 2016; López-Cortegano and Caballero, 2019; Young, 2019). Although it has been suggested that larger studies will only achieve diminishing returns and the eventual potential implication of the entire genome (Goldstein et al., 2009; Boyle et al., 2017), others argue in favour of the continued mapping of loci which may have functional and clinical importance (Hirschhorn

et al., 2009). In addition, it is becoming increasingly clear that genetic heterogeneity contributes to the phenotypic heterogeneity seen in T2D, suggesting that informative cohorts and study designs may further identify phenotypically-relevant pathways (Udler, 2019). As such, addressing the current limitations of gene mapping methods will complement efforts to discover novel, clinically important risk factors and drivers of phenotypic heterogeneity.

3 Chapter 3: T2D loci regulate nuclear-encoded mitochondrial genes

3.1 Overview

The main aim of Chapter 3 is to investigate T2D-associated genetic risk loci (T2D loci) mapped by Lau et al. (2017) to assess their potential to alter mitochondrial function. To achieve this, T2D loci will be investigated for evidence of association with the expression levels of nuclear-encoded mitochondrial genes (NEMGs) using both published and unpublished data provided by Lau et al.

This analysis aims to identify testable genetic mechanisms through which a heritable predisposition to mitochondrial dysfunction may increase risk of T2D, through the identification of T2D risk loci which regulate the expression of NEMGs.

To introduce this Chapter, the next section will review the functions of regulatory, non-coding DNA, in which most T2D loci are found. **Section 3.2.2: eQTL analysis as a tool to interpret non-coding disease loci** will review eQTL analysis, the method of choice used by Lau et al. (2017) and in this Chapter to identify genes regulated in *cis* by nearby T2D loci (*cis*–genes). The LDU-based method of Lau et al. will be compared with other methods which are commonly used to integrate eQTL analysis with GWAS. The following sections will present three main aims:

1. Integrate eQTL and T2D loci to identify target *cis*-genes (Section 3.3)
2. Filter the total *cis*–genes for NEMGs (Section 3.4)
3. Test the *cis*–genes for enrichment of mitochondrial functions (Section 3.5).

3.2 Introduction

3.2.1 Complex traits and non-coding DNA

Crucial to the design of this study is that most genetic loci associated with complex diseases, including T2D, are found in intronic or intergenic non-coding regions (Hindorff

et al., 2009; Welter et al., 2014). For example, Cirillo et al. (2018) reported that 98% of T2D-associated GWAS SNPs were non-coding. Multiple studies have reported that GWAS variants are enriched in regulatory DNA, which can be categorised into several different types including enhancers, silencers, locus control regions, core and proximal promoters, non-coding RNAs and boundary elements or long non-coding RNAs which maintain the higher order structure of the 3D genome (discussed below) (Maston et al., 2006; Morris and Mattick, 2014; Engreitz et al., 2016; Moszyńska et al., 2017; Giral et al., 2018; Rowley and Corces, 2018). There is estimated to be over one million enhancers in the human genome (Consortium et al., 2012b). Regulatory elements such as enhancers have tissue-specific activity (Ong and Corces, 2011, 2012; Ko et al., 2017) and both GWAS variants and realised trait heritability estimates are particularly enriched in the regulatory elements of trait or disease-relevant cell types (Maurano et al., 2012; Gusev et al., 2014; Torres et al., 2014; Kundaje et al., 2015; Farh et al., 2015). For T2D in particular, GWAS variants are enriched in pancreas-specific regulatory DNA, including 3D enhancer 'hubs' (Parker et al., 2013; Pasquali et al., 2014; Varshney et al., 2017; Miguel-Escalada et al., 2019). This relationship has been observed particularly for variants associated within insulin secretion, while insulin action and lipid-associated variants show enrichment in adipocyte, pre-adipocyte, monocyte and hepatocyte enhancers (Scott et al., 2017; Udler et al., 2018; Torres et al., 2014).

While coding mutations alter the amino acid sequence of a protein and thus provide a clear mechanism of effect, mutations which disrupt regulatory elements are considerably less clear-cut. The regulation of gene expression is dependent on multiple factors including cell-type (Consortium et al., 2015b, 2017; Gamazon et al., 2018), temporal patterns such as the circadian clock (Zhang et al., 2014b; Mermet et al., 2017) and time of feeding (Vollmers et al., 2009), as well as nutritional intake (Pellatt et al., 2016) and stress (De Nadal et al., 2011), among others. Around half of human genes may show tissue-specific expression (Yang et al., 2018). Regulatory variants may therefore influence gene expression in several ways, including the level or timing of expression, splicing, transcript stability or translational efficiency. Further complicating the understanding of regulatory

mutations is the extensive folding of nuclear DNA into dynamic compartments, facilitating both short and long-range interactions between regulatory DNA and gene promoters (Bonev and Cavalli, 2016; Schoenfelder and Fraser, 2019). Hence, the target gene (*cis*-gene) of a regulatory element can be a neighbouring or distal gene, as opposed to the closest gene on a linear map of the chromosome. This 3D structure is also dynamic, for example Mermet et al. (2018) found that interactions between enhancer elements and gene promoters in mouse oscillated in response to the circadian rhythm.

The interpretation of non-coding disease loci hence requires several layers of information in order to identify (1) the causal variant(s), since the causal SNP is nearly always not the genotyped SNP but may be any SNP in reasonably high LD (discussed in detail in **Chapter 5: Fine-mapping**) and (2) the downstream effect, including the target *cis*-gene(s) and implicated cell-type. There are many different methods available to determine the *cis*–genes regulated by non-coding T2D loci (Kyono et al., 2019; Lin and Musunuru, 2018; Cebola, 2019). This chapter will focus on high-throughput eQTL analysis, although other methods include: 3D chromatin interaction maps, high-throughput reporter assays and CRISPR gene editing technologies to interrogate *in vivo* or *in vitro* gene perturbations.

3.2.2 eQTL analysis as a tool to interpret non-coding disease loci

eQTL analysis involves the mapping of variants which alter gene expression by testing genotype association with quantitative RNA levels. This concept is illustrated in Figure 12. Variants which associate with RNA levels are known as expression quantitative trait loci (eQTL), or eSNPs[59], and in 2010, Nicolae et al. systematically demonstrated that trait-associated SNPs were likely to be eQTL. Other evidence suggests that common GWAS SNPs are likely to be highly connected in tissue-specific gene expression networks, influencing the expression of multiple genes[60] (Fagny et al., 2017).

[59]Prior to fine-mapping, an associated SNP implicates every variant at the locus which is in high LD. Thus, while 'eSNP' might be used to refer to a lead SNP associated with gene expression levels, eQTL (expression quantitative trait *locus*) may more accurately reflect the true resolution until fine-mapping confirms the causal variant (the true eSNP).

[60]Although these results suggest that disease-related SNPs are more likely to perturb gene expression networks, they may also reflect the higher power of GWAS to detect variants of larger effect.

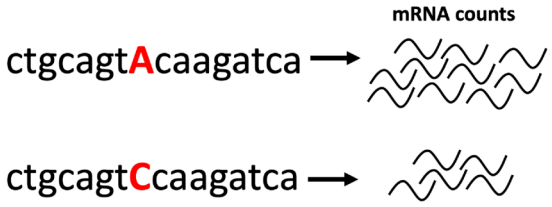


Figure 12: A simple diagram of an expression quantitative trait locus (eQTL). The A>C variant is associated with lower gene expression, measured as the number of transcribed mRNA molecules.

A disease-associated SNP which is also a significant eQTL may be expected to confer risk of disease via its effects on the *cis*-gene expression. Therefore, eQTL mapping can reveal the functional target gene(s) and therefore the likely biological mechanism which confers risk of disease. eQTL mapping is necessary because the regulated gene(s) are unlikely to be those closest to the disease locus. Previous studies have observed that between 60 and 90% of complex trait-associated variants associated with the expression of genes which were *not* the nearest gene (Nica et al., 2010; Maurano et al., 2012; Zhu et al., 2016; Mumbach et al., 2017; Gusev et al., 2018; Khamis et al., 2019). An example of a study which used eQTL analysis among other methods to map target gene(s) is seen in Claussnitzer et al. (2015) who discovered that, out of eight neighbouring genes, *IRX3* and *IRX5* expression levels were regulated by an obesity-associated variant in intron 1 of the *FTO* gene. This is despite a distance of 516 kb and 1,164 kb between gene and variant, respectively.

Integrating eQTL analysis with GWAS can be used to investigate differentially expressed genes for 'cause or consequence'. In other words, genes may show differential expression as a *consequence* of disease status, however it is the genes which *cause* disease when differentially expressed which are of the most interest, since disease prevention requires knowledge of the mechanisms which are causal to disease development and increase risk prior to disease onset (also known as aetiological). Mapping *cis*–genes regulated by disease loci can therefore reveal heritable mechanisms which can be used as potentially informative biomarkers prior to disease onset and can inform the design of preventative and therapeutic treatments. This approach will be used to investigate evidence that T2D

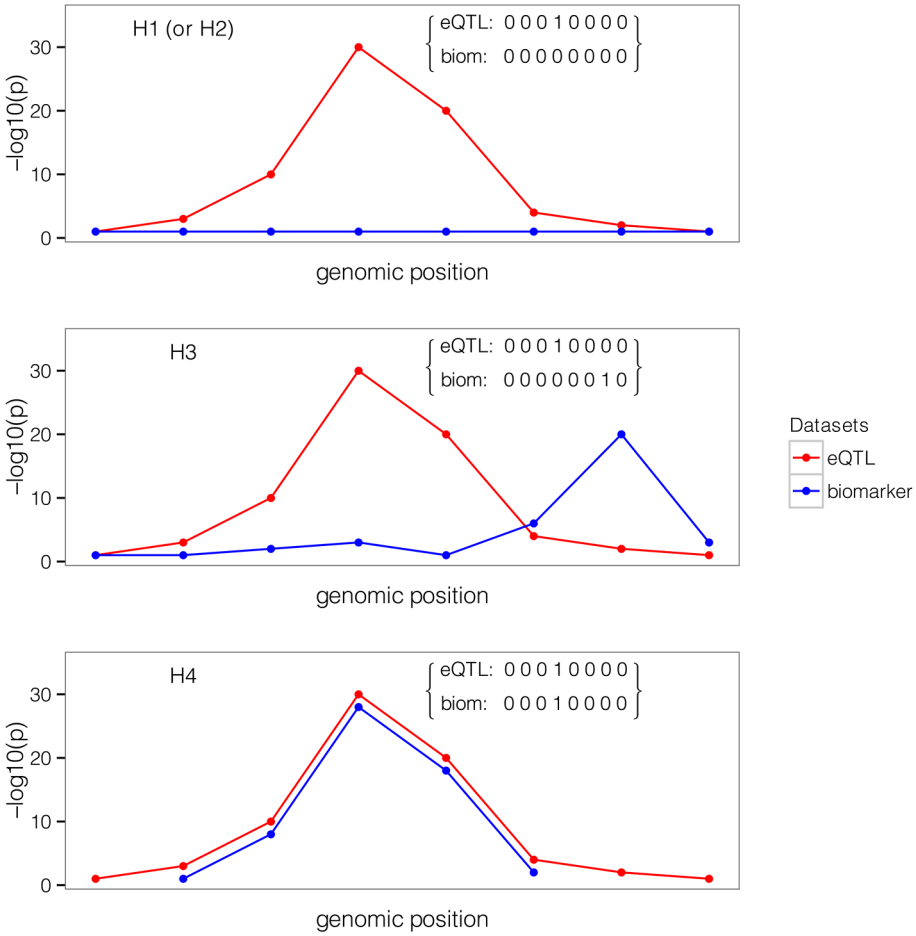


Figure 13: This is Figure 1 from Giambartolomei et al. (2014) and is used to describe the scenarios of co-localisation tested by the Bayesian method COLOC. Two independent tests are run: the red line shows SNP association (as the -$\log_{10}$ p-value) with gene expression and the blue line shows SNP association with disease status (for example). H1 (or H2): only one test shows evidence of association at this locus. H3: there are two independent signals of association. H4: the causal variant is shared and the surrounding patterns of association are the same.

loci regulate mitochondrial function, in order to provide evidence to the ongoing debate as to whether mitochondrial dysfunction is a 'cause or consequence' of T2D (see Chapter 1, **Section 1.5: Mitochondrial dysfunction in Type 2 diabetes**).

3.2.3 Methods for eQTL-GWAS integration

Integrating GWAS with gene expression data is now largely routine; this is made possible by publicly available databases of genome-wide regulatory elements such as the GTEx portal[61], which reports eQTL across 44 tissues or 'regions' (including different brain subre-

[61]The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and

gions, for example) (Consortium et al., 2017). As described in Chapter 2, **Section 2.4.4: Trans-ethnic GWAS and defining replication**, there are several different ways to integrate independent tests of association, in this case GWAS for disease with eQTL analysis. For example, Scott et al. (2017) and van de Bunt et al. (2015) classified shared GWAS/eQTL signals if lead SNPs for each were in high LD (r^2>0.8). Other groups have developed more sophistical methods including formal tests for shared causality. These can be broadly grouped into those which test for co-localisation and those which test intermediate gene expression (Barbeira et al., 2016). Co-localisation tests assume that, if a causal variant is shared, then the surrounding patterns of association will be the same. This is illustrated in Figure 13 as used by the popular tool COLOC (Giambartolomei et al., 2014). Other tools include eCAVIAR (Hormozdiari et al., 2016) and RTC (Nica et al., 2010; Wallace et al., 2012). Alternative methods assess shared causality by testing for the association of a SNP with a phenotype via intermediate gene expression. When gene expression data is unavailable for a GWAS cohort, it is imputed based on the variance in gene expression explained by individual or aggregate SNPs in an independent eQTL analysis. The imputed gene expression is then correlated with the phenotype. Example tools include SMR (summary data-based Mendelian Randomisation) (Zhu et al., 2016), TWAS (Gusev et al., 2016), and PrediXcan/MetaXcan (Gamazon et al., 2015; Barbeira et al., 2016).

The above tools including SMR, COLOC and MetaXcan have been used to identify novel *cis*-genes at T2D loci (Xue et al., 2018; Liang et al., 2017; Torres et al., 2017; Viñuela et al., 2019), while other studies have tested for high LD between lead T2D and eQTL SNPs (van de Bunt et al., 2015; Scott et al., 2017; Khamis et al., 2019; Raulerson et al., 2019)[62]. Two recent T2D GWAS which systematically applied these methods include Mahajan et al. (2018), in which eQTL from GTEx were tested for co-localisation with GWAS credible sets (sets of SNPs which have 99% confidence of containing the causal

NINDS. The data used for the analyses described in this Chapter were obtained from the GTEx Portal prior to August 2020.

[62] As a further example, Fernández-Tajes et al. (2019) integrated GWAS with eQTL using a regression-based approach with a 'positional candidacy score' and 'variant link score'.

variant) using COLOC and Vujkovic et al. (2020), who imputed gene expression using S-PrediXcan and tested for co-localisation using COLOC.

3.2.4 Methods for eQTL-GWAS integration: LDU maps

An alternative method to integrate GWAS with eQTL data is described by Lau et al. (2017), who estimated the locations of T2D risk variants and eQTL using the Malecot model described in Chapter 2, **Section 2.7: Gene-mapping with LDU maps**. The Lau et al. (2017) study is reviewed in detail in Chapter 2, **Section 2.7.3: LDU-based gene mapping in T2D**. The LDU-based approach provides greater commensurability between studies where lead SNPs differ (for example due to population-specific LD structures), but location estimates overlap. Location estimates ($\hat{S}$) represent the most likely location of a causal variant and are estimated using a powerful multi-marker model which incorporates population-specific LD.

Lau et al. (2017) described a simple test of co-localisation where $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ were required to co-locate within a physical distance of ±50 kb. Alternatively, the use of genetic LDU maps facilitates a test for co-location which incorporates genetic distance. As described in Chapter 2, LDU distance reflects the decline in allelic association, LD, with physical distance. A co-location threshold corresponding to a set genetic distance, for example 1 LDU (which reflects the physical distance over which ρ has declined to e^{-1} or 0.37 of its starting value) would therefore test whether the mapped loci are in high LD. This is similar to the local replication described on page 54, which requires lead SNPs to be in high LD, while a co-location threshold of ±50 kb would correspond to a local transferability approach. There are advantages to both approaches; local replication is considerably more effective at removing loci for which the causal variant is not shared, however local transferability is more robust to allelic heterogeneity, where the causal variants are independent but disrupt the same functional element. A formal test of co-localisation may also be developed to test the overlap between the $\hat{S}$ confidence intervals or by comparing the likelihood surfaces output from the Malecot model; this may be of interest for future work.

In the following analysis, T2D location estimates ($\hat{S}_{T2D}$) will be integrated with eQTL location estimates ($\hat{S}_{eQTL}$) by means of genetic co-location within 1 LDU, with the assumption that co-locating signals correspond to the same causal variant[63]. The *a priori* defined threshold of 1 LDU was assessed *post hoc* and compared to the threshold of 50 kb in **Section 3.3.3: Physical vs genetic distance: co-location using 1 LDU**.

3.2.5 Pathway analysis of *cis*-genes

The systematic, genome-wide integration of GWAS with eQTL offers the power to detect networks of perturbed *cis*–genes and thus biological pathways involved in disease (Wang et al., 2007; Chen et al., 2008; Cookson et al., 2009; Goh et al., 2007; Sieberts and Schadt, 2007; Wang et al., 2010a). This involves testing identified *cis*–genes for the enrichment of pre-defined sets of genes (gene set enrichment analysis, GSEA) (García-Campos et al., 2015; Kao et al., 2017). Prior to tools for eQTL mapping being made widely available, several studies carried out pathway analysis for T2D based on the closest gene(s) to GWAS lead SNPs (Torkamani et al., 2008; Yazdanpanah et al., 2013; Perry et al., 2009; Cirillo et al., 2018)[64]. However, as described in the previous section, GWAS variants are unlikely to target the closest gene and other studies have instead interrogated *cis*-genes implicated by eQTL analysis. This approach has shown T2D *cis*–genes to be enriched for mTOR pathways, FOXA2 targets and genes involved adipocyte differentiation, cholesterol biosynthesis and lipid metabolism (Scott et al., 2017; Liang et al., 2017; Small et al., 2018). Using this approach, the current study will test for evidence of the enrichment of mitochondrial pathways within the list of identified T2D *cis*–genes.

3.2.6 Aims

Following the availability of genetic location estimates associated with (1) T2D ($\hat{S}_{T2D}$) and (2) subcutaneous adipose gene expression levels ($\hat{S}_{eQTL}$) by Lau et al., this Chapter

[63]Independent location estimates which are within a small genetic distance may reflect either the same causal variant or independent causal variants in high LD. As described by Zhu et al. (2016), these reflect scenarios of shared 'causality' or 'linkage', respectively.

[64]Torkamani et al. (2008) assigned GWAS SNPs to genes within 5 kb; Yazdanpanah et al. (2013) assigned low-frequency variants to a gene if they were within the gene coordinates or ±50 kb; Perry et al. (2009) assigned genes to the most significant SNP within 200 kb; Cirillo et al. (2018) assign SNPs to genes if they are located within 1 kb upstream and 1 kb downstream of the gene coordinates.

will aim to address the following objectives:

1. Integrate $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ based on co-location within a pre-defined genetic distance of 1 LDU.

2. Identify *cis*-genes associated with co-locating $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ which are involved in mitochondrial function.

3. Test the total *cis*-genes for enrichment of mitochondrial pathways.

3.3 Aim one: filter and define T2D-eQTL

This study presents a follow-up to the analysis published by Lau et al. (2017), who mapped the most likely locations of variants associated with T2D ($\hat{S}_{T2D}$) and gene expression levels ($\hat{S}_{eQTL}$) using LDU maps and an adapted Malecot model. The first aim of this follow-up analysis was to define the putative *cis*–genes regulated by T2D loci by filtering published and unpublished $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ from Lau et al. (2017) which co-located within 1 LDU. Nominally significant loci were considered in order to include additional *cis*–genes and increase the power to detect subsequent pathway enrichment (with the caveat of potentially including more false positive results). T2D locations were mapped in one African American and two European cohorts of T2D case-controls, while eQTL were mapped using adipose gene expression data from an ageing, population-based European sample (TwinsUK). The number of T2D cases and controls for each dataset is shown in Table 1 on page 74, with a total of 5,800 T2D cases and 9,691 controls.

Type 2 diabetes case-control cohorts

- WTCCC phase 1. *Europeans* (WTC) (Consortium et al., 2007b)
- WTCCC phase 2. *Europeans* (MTC) (Voight et al., 2012)[65]
- NIDDK. *African Americans* (AA) (Palmer et al., 2012)

Subcutaneous adipose gene expression cohort

[65]The WTCCC phase 2 cohort was genotyped using the custom Metabochip array (Voight et al., 2012) with ∼200K SNPs strategically placed around known or candidate T2D loci.

- eQTL: MuTHeR. *Healthy Europeans* (Grundberg et al., 2012)

3.3.1 Methods

The analysis described in this chapter reproduces to some extent the test to define shared $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ by Lau et al. (2017), based on co-location within ±50 kb. However, the current study introduces several differences to the filtering stage; these are shown in Table 2 and are described in more detail below.

T2D-eQTL filtering criteria, Lau et al. (2017) vs the current study

Study	$\hat{S}_{T2D}$ p-value[a]	$\hat{S}_{eQTL}$ p-value	$\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance	$\hat{S}_{T2D}$-$\hat{S}_{T2D}$ distance
Lau et al. (2017)	$<10^{-5}$	Bonferroni-corrected[b]	<50 kb	<100 kb
Current	$<10^{-3}$	<0.05	<1 LDU	<1 LDU

Table 2: T2D-eQTL inclusion criteria for Lau et al. (2017) and the current study. LDU = linkage disequilibrium unit. [a]10^{-5} is Bonferroni-corrected, whereas 10^{-3} is nominally significant. [b]p-value thresholds were corrected for the total number of genes tested for each locus within ±1.5Mb of the replicated $\hat{S}_{T2D}$.

INCLUSION CRITERIA 1: T2D loci

As reported in Lau et al. (2017), $\hat{S}_{T2D}$ were mapped by running the adapted Malecot model on 4,800 genomic windows[66], equating to a stringent Bonferroni-corrected statistical threshold of 10^{-5} (a conservative estimate of 0.05/5,000). Lau et al. (2017) defined replicated T2D loci as windows containing two or more independent $\hat{S}_{T2D}$ within 100 kb of each other, each of which had nominal significance and passed Bonferroni-correction when meta-analysed using Fisher's method. The authors calculated that, for 111 $\hat{S}_{T2D}$ replicated using the criteria of 100 kb, the average D' for all HapMap SNP pairs found within the $\hat{S}_{T2D}$ intervals was 0.86 in Europeans and 0.78 for African Americans, confirming that an interval of 100 kb corresponded to generally high LD. In the current study, $\hat{S}_{T2D}$ were included with meta-analysed p-value of $<10^{-3}$ in order to facilitate pathway analyses with larger numbers of *cis*-genes.

INCLUSION CRITERIA 2: eQTL p-value

For each replicated disease locus (two or more independent $\hat{S}_{T2D}$ within ±100 kb), Lau et al. (2017) carried out eQTL mapping using the expression of neighbouring genes within ±1.5Mb. A strict Bonferroni-corrected threshold was applied to each locus in the Lau et al. study depending on the number of neighbouring genes tested. In this analysis, $\hat{S}_{eQTL}$ were required to have a nominally significant p-value of <0.05. An added requirement was that the standard error (stderr) of the $\hat{S}_{eQTL}$ estimate was less than 900, in order to exclude estimates where the likelihood model did not successfully converge.

INCLUSION CRITERIA 3: $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ co-location

As described previously, Lau et al. (2017) defined T2D-eQTL where $\hat{S}_{eQTL}$ co-located within ±50 kb of a replicated $\hat{S}_{T2D}$. In comparison, the current study required $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ to co-locate within a genetic distance of 1 LDU, the advantages of which include the removal of $\hat{S}$ which are physically close but are separated by a large genetic distance,

[66]the criteria of 10 LDU was defined using the European genetic map. The physical coordinates of each window were converted to the African American map, resulting in an average of 16 LDU per window due to the more extensive LD breakdown in this older population.

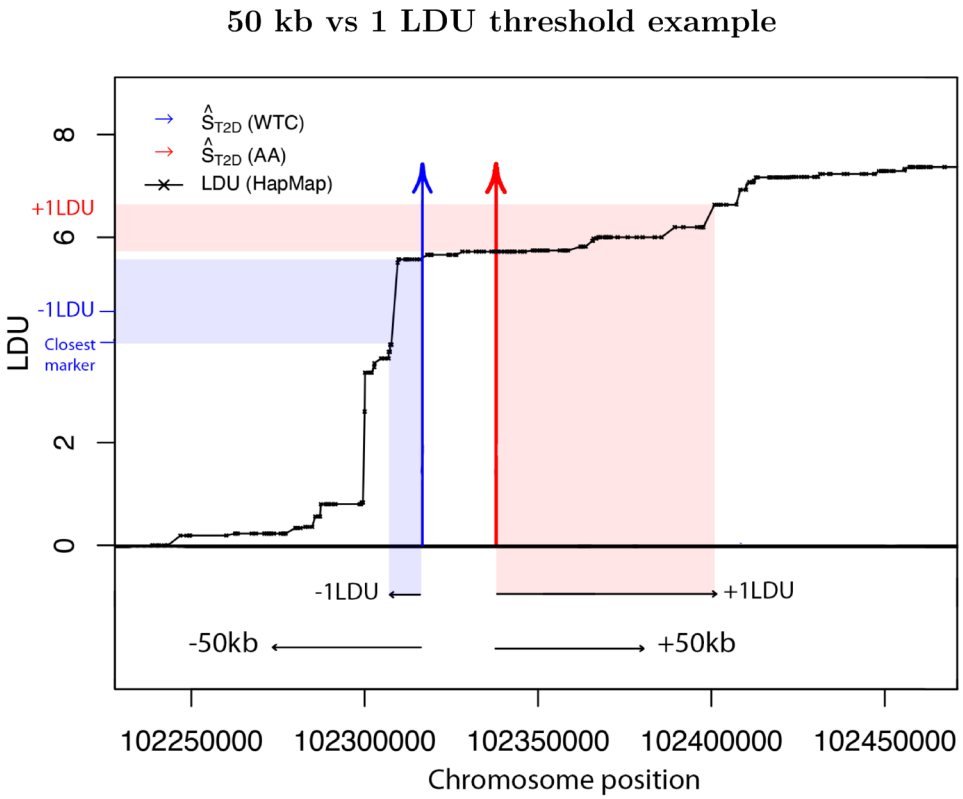


Figure 14: A theoretical T2D locus with two independent $\hat{S}_{T2D}$ estimates from the European WTCC phase 1 (WTC) and NIDDK African American (AA) cohorts. A co-location threshold of 1 LDU is used to define co-location of $\hat{S}_{eQTL}$. The threshold of 50 kb used by Lau et al. (2017) is also shown. The -1 LDU threshold is defined using the closest genetic marker on the LDU map.

making them unlikely to tag a shared causal variant. The difference between a physical and genetic co-location threshold is illustrated as a theoretical T2D locus in Figure 14. The example shows that the downstream distance corresponding to 1 LDU is significantly less than 50 kb due to a breakdown in LD (seen as a large LDU step). In comparison, the upstream distance of 1 LDU is greater than 50 kb due to more extensive LD. Filtering using a genetic co-location threshold therefore excludes $\hat{S}$ which are separated by LD breakdown, while including those which are in high LD. The co-location of 50 kb and 1 LDU are further compared in **Section 3.3.3: Physical vs genetic distance: co-location using 1 LDU.** To summarise, T2D-eQTL were defined where the following criteria were satisfied: two or three independent $\hat{S}_{T2D}$ and one or more $\hat{S}_{eQTL}$ within ± 1 LDU; $\hat{S}_{T2D}$ p-value $<10^{-3}$; $\hat{S}_{eQTL}$ p-value <0.05.

Assigning gene identifiers

eQTL location estimates ($\hat{S}_{eQTL}$) were each associated with gene expression as measured by a probe from the Illumina HumanHT-12 v3.0 array. Probes were initially assigned to a gene ID using the array annotation package in R (Dunning et al., 2015). However, following the observation that some of the probes did not overlap with the annotated gene, gene IDs were subsequently updated using gene coordinates from Ensembl GRCh37. Probe coordinates were converted from build 36 to build 37 and cross-referenced with Ensembl code coordinates. If the probe was located within multiple genes, the original gene ID assigned to the array was used. The Ensembl gene ID was used if the Ensembl ID differed from the array ID, or if there was no gene assigned by the array annotation.

3.3.2 Results

The filtering steps and results are shown in Figure 15 and are described below. Data was provided from Lau et al. for a total of 265 analytical windows for which the $\hat{S}_{T2D}$ achieved a p-value of $<10^{-3}$ in meta-analysis and the $\hat{S}_{T2D}$ locations were within 100 kb. For the current analysis, the windows were filtered to include those with $\hat{S}_{T2D}$ within 1 LDU, as shown in Figure 16. 91 results were removed and 174 were retained.

For the 174 T2D loci, eQTL mapping for genes within ±1.5Mb gave a total of 7,960 independent $\hat{S}_{eQTL}$ estimates associated with 3,530 annotated genes. Invalid $\hat{S}_{eQTL}$ estimates resulting from failed convergence of the Malecot model were assigned a high standard error and were excluded. Valid $\hat{S}_{eQTL}$ estimates were required to have a p-value <0.05 and to be within 1 LDU of one or more $\hat{S}_{T2D}$. After filtering for these criteria, a total of 1,066 $\hat{S}_{eQTL}$ co-located with 166 T2D loci and associated with 763 *cis*–genes.

3.3.3 Physical vs genetic distance: co-location using 1 LDU

In this study, $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ were required to co-locate within 1 LDU, since a small genetic distance reflects high LD. For example, $\hat{S}$ separated by a small genetic distance of 0.5 LDU and 150 kb represent variants in high LD, such that any variants within the 150 kb region may be driving the association and the $\hat{S}$ may tag the same causal variant. Conversely, $\hat{S}$ separated by 2 LDU and 50 kb are in low LD and the variants will

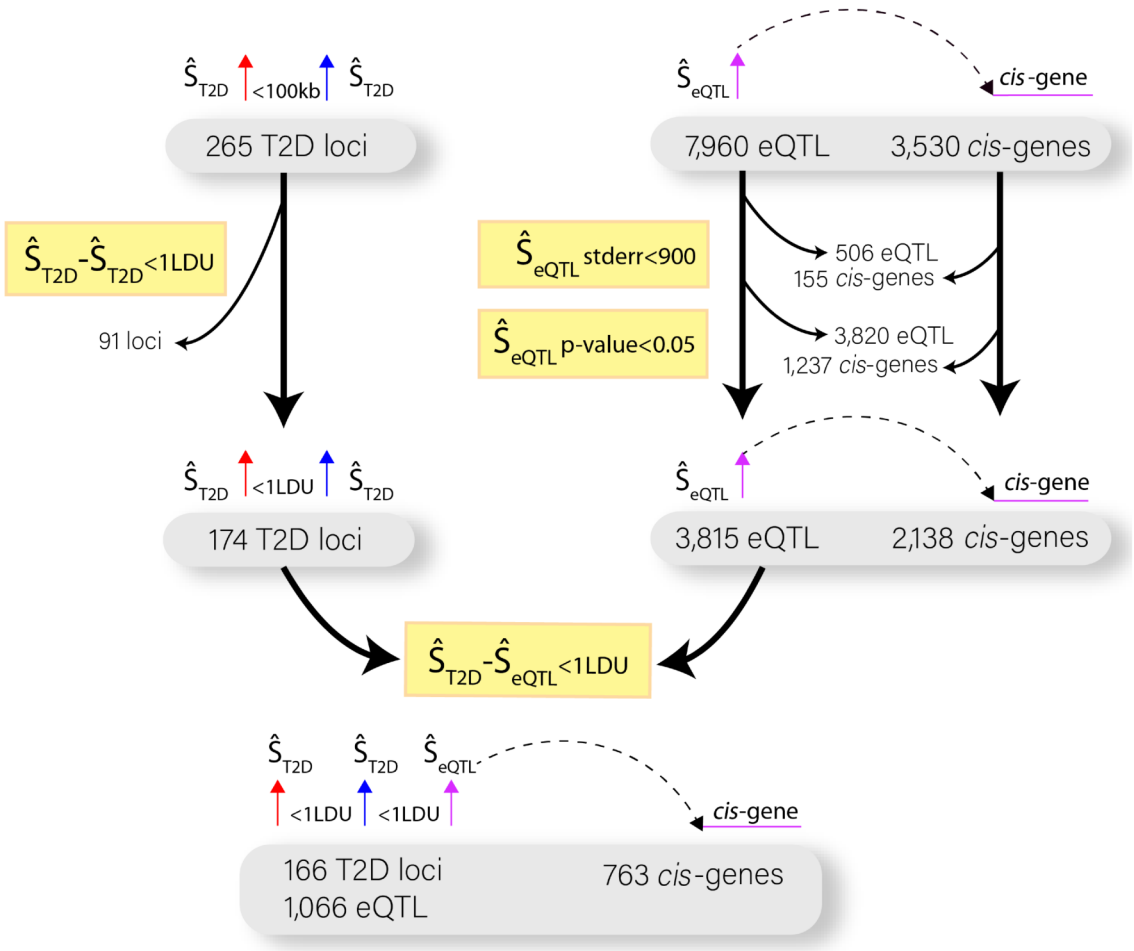


Figure 15: T2D-eQTL filtering. The filtering steps are shown in yellow and the total number of retained and excluded $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ are shown with each step. The 7,960 eQTL are for genes within ±1.5Mb of the 174 identified T2D loci, at which $\hat{S}_{T2D}$ co-located within 1 LDU. stderr = standard error.

be independently inherited. Therefore filtering by genetic distance, rather than physical distance, will remove physically close but independent estimates. This is shown in Figure 16, which plots the physical (kb) vs genetic (LDU) distance between the 265 $\hat{S}_{T2D}$ estimates provided by Lau et al. The co-location threshold of 1 LDU used in this study is highlighted in red, showing that $\hat{S}_{T2D}$ separated by a large genetic distance are excluded, despite being located within a small physical distance.

Figure 17 plots the physical (kb) vs genetic (LDU) distances between 3,815 nominal $\hat{S}_{eQTL}$ estimates (p-value <0.05 and stderr <900) and the nearest $\hat{S}_{T2D}$ made for neighbouring genes at the 174 T2D loci. As seen the plot, physical and genetic distance can vary

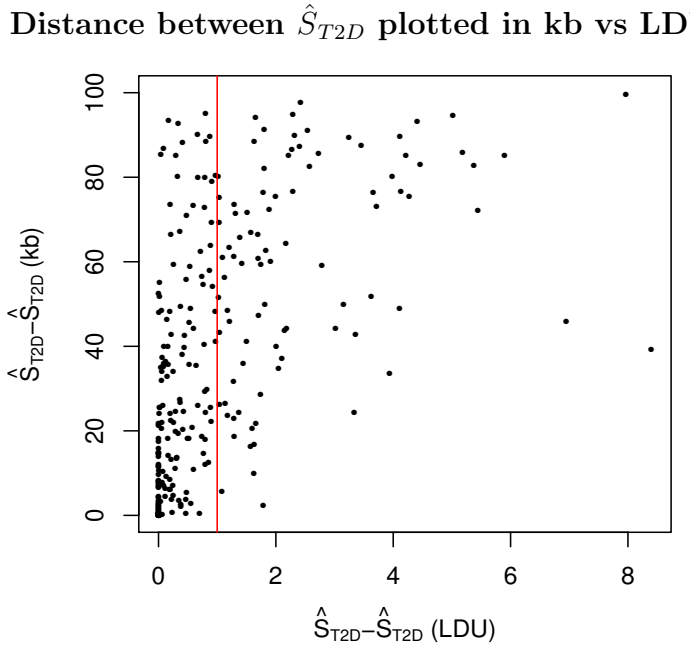


Figure 16: The distance between T2D locations ($\hat{S}_{T2D}$) measured in kb vs in LDU. The threshold of 1 LDU is shown in red.

widely; data points with a small kb and large LDU distance are located in regions of significant LD breakdown such as recombination hot spots, whereas data points with a large physical distance and small genetic distance are located in regions of extensive LD, such as recombination cold spots. These data are subsetted in Figure 18 to show a filtering criteria based on a physical distance of 50 kb and a genetic distance of 1 LDU. As shown in Figure 18, subsetting $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ based on a co-location of <50 kb also retains $\hat{S}$ estimates which are separated by large genetic distances, i.e. significant LD breakdown. Conversely, filtering by a genetic distance of <1 LDU retains $\hat{S}$ separated by a large physical distance, but which are in high LD. These filtering criteria are further compared in Chapter 4, **Section 4.4: Physical vs genetic distance: differential expression.**

Distance between $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ plotted in kb vs LDU

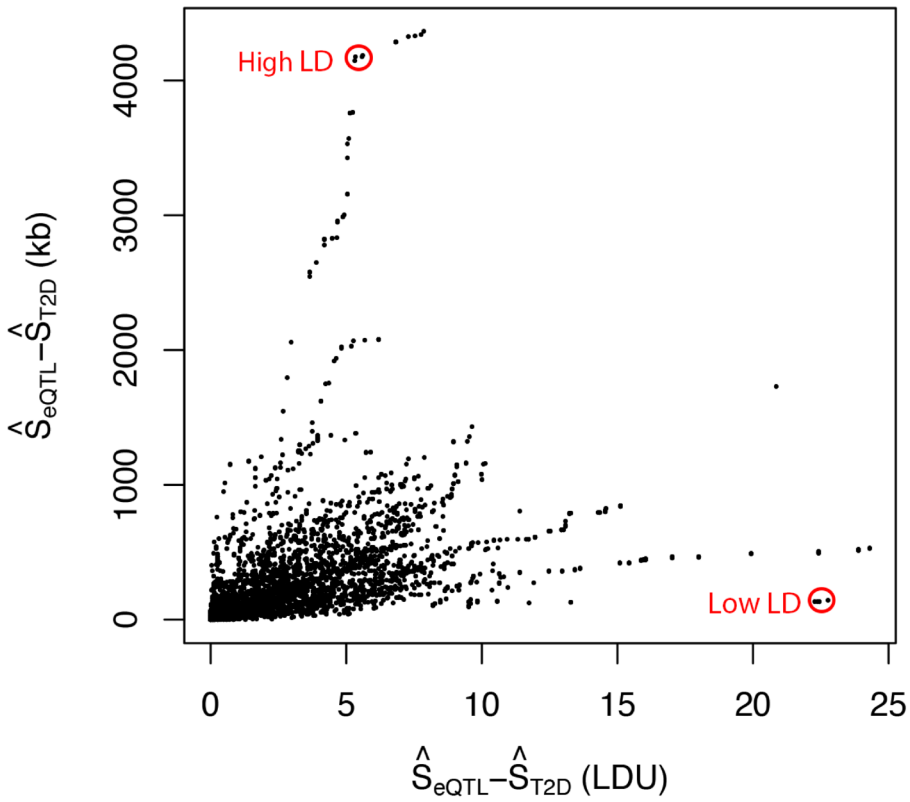


Figure 17: The distance between $\hat{S}_{eQTL}$, which are estimated for genes within ±1.5Mb of a replicated T2D locus and the nearest $\hat{S}_{T2D}$ measured in kb and LDU. High or extended LD reflects low recombination rates while low LD reflects high recombination rates.

$\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ co-location distances when filtered for <50 kb and <1 LDU

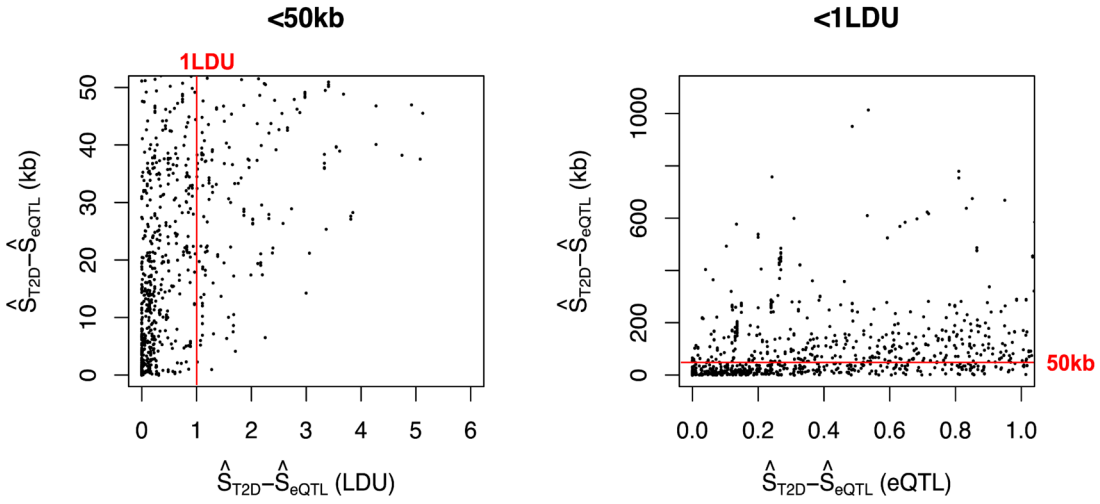


Figure 18: The distance between $\hat{S}_{eQTL}$ and $\hat{S}_{T2D}$ measured in kb and LDU.

3.4 Aim two: Identify *cis*-genes involved in mitochondrial function

The following section aims to identify nuclear-encoded mitochondrial genes (NEMGs) among the total 763 T2D *cis*-genes identified in Section 3.3.

3.4.1 Nuclear encoded mitochondrial genes (NEMGs)

Nuclear encoded mitochondrial genes (NEMGs) are genes located in the nuclear genome which encode proteins transported to the mitochondria. Early efforts to characterise the mitochondrial proteome hinted at up to 1,500 mitochondrial proteins (Rabilloud et al., 1998; Lopez et al., 2000). High-throughput techniques have more recently facilitated large databases of the mitochondrial proteome (Lemkin et al., 1996; Scharfe et al., 2000; Attimonelli et al., 2002; Gaucher et al., 2004; Forner et al., 2006; Ingman and Gyllensten, 2006; Godin and Eichler, 2017). The two databases used in this analysis result from the integration of multiple data sources which effectively detect low abundance mitochondrial proteins and reduce false positives. The first of these is MitoCarta2.0 (Pagliarini et al., 2008; Calvo et al., 2015a). Briefly, Calvo et al. (2015a) applied a Bayesian integration of known mitochondrial protein training sets with seven genome-scale datasets including: detection of the protein in purified mitochondria using highly sensitive, tandem mass spectrometry; homology with known yeast mitochondrial proteins; co-expression with known mitochondrial proteins; presence of a mitochondrial-specific protein domain; presence of an N-terminal mitochondrial targeting sequence; endosymbiont ancestry (i.e. homology with *Rickettsia prowazekii* proteins); and mRNA up-regulation in cellular models of mitochondrial proliferation. Mitocarta2.0 contains 1,158 NEMGs. This resource has since been revised by Floyd et al. (2016), yielding the second database used in this study: Mitocarta+. Floyd et al. (2016) combined the published MitoCarta2.0 with additional studies of the mitochondrial matrix proteome (Rhee et al., 2013) and inter-membrane space (Hung et al., 2014). The T2D *cis*–genes identified in this study were compared to 1,204 NEMGs present in both MitoCarta2.0 and Mitocarta+ to define *cis*–NEMGs.

3.4.2 Methods

The list of 763 T2D *cis*-genes defined in the previous section were filtered for NEMGs, referred to as T2D *cis*–NEMGs, by cross-referencing with the MitoCarta2.0 and MitoCarta+ databases. The MitoCarta2.0 database was downloaded from the Broad Institute webpage[67] and both HGNC (HUGO Gene Nomenclature Committee) and Ensembl gene IDs were extracted. The MitoCarta+ database was obtained from Floyd et al. (2016) and Entrez gene IDs were converted to Ensembl using the R bioconductor package (Carlson, 2019). Both Ensembl and HGNC gene IDs were used for cross-referencing.

3.4.3 Results

In total, 56 $\hat{S}_{eQTL}$ were associated with probes annotated to 50 unique NEMGs. These are shown in Table 3. The 50 NEMGs accounted for 6.55% of the total number of identified *cis*–genes. The 50 T2D *cis*–NEMGs are grouped in Figure 19 by common biological functions, following annotation according to KEGG pathways and KEGG ontology terms (Kanehisa et al., 2015), in addition to GeneCards summaries (Rebhan et al., 1997) where KEGG data was lacking. Summary information for the biological functions of each NEMG is shown in Appendix A.1. The potential relationship between candidate genes and pathways shown in Figure 19 and T2D are discussed further in **Section 4.5: Discussion**.

[67] https://www.broadinstitute.org/files/shared/metabolism/mitocarta/human.mitocarta2.0.html

Table 3: $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ for the 50 *cis*–NEMGs

Bonferroni-corrected T2D loci T2D locations p-value $< 10^{-5}$						
Chr	T2D location (WTC)	T2D location (AA)	T2D location (MTC)	*cis*–NEMG	eQTL location[d]	eQTL p-value
1	26006625	NA	26003695	MTFR1L	25900571	1.20e-04
1	26006625	NA	26003695	CLIC4	26002616	6.32e-05
					25900772	4.80e-02
1	234270220	234273988	NA	COA6	234338743	6.63e-04
1	234270220	234273988	NA	TOMM20	234309446	0.017
2	NA	204023399	203970861	NIF3L1	203798637	0.005
2	NA	204023399	203970861	C2orf69	204126463	0.005
2	NA	204023399	203970861	MARS2	203820504	0.002
2	NA	204023399	203970861	HSPD1	204204077	0.037
2	NA	204023399	203970861	CPS1	204126514	0.020
2	227080369	NA	227021099	MFF	227151692	0.037
3	67744088	67685265	NA	SLC25A26	67637196	0.049
3	120573472	120555505	NA	NDUFB4	120573615	6.21e-06
3	123048537	NA	123061689	CCDC58	122876198	0.002
3	132436519	NA	132429438	ACAD11	132451038	1.62e-09
3	183260285	183210822	NA	MCCC1	183260250	0.028
4	91942692	91950656	NA	PDHA2	91947875	1.21e-07
4	104004185	NA	103936988	CISD2	104141231	2.66e-04
6	127539286	NA	127502744	TRMT11	127479024	0.001
6	127539286	NA	127502744	HINT3	127357421	0.010
10	94499812	NA	94479016	MARCHF5	94488072	0.001
10	104786704	NA	104841790	SFXN2	104732175	2.18e-09
11	8551677	NA	8637191	CYB5R2	8667032	0.040
11	43879353	NA	43879882	ALKBH3	43607412	0.004
11	65575917	NA	65600493	MRPL11	65466334	0.002
12	56618300	NA	56622584	SUOX	56621500	0.005
12	56618300	NA	56622584	GLS2	56628724	0.018
12	106406604	106384733	NA	MTERF2	106410234	2.03e-04
12	121317223	NA	121243696	ACADS	121372772	0.003
12	121317223	NA	121243696	GATC	121186959	3.45e-06
12	123387213	123750895	123447928	DIABLO	123386489	1.10e-03
					123832253	1.09e-02
					123394850	3.53e-05
13	102408534	102452781	NA	PCCA	102428282	0.001
13	111049674	111004953	111035483	NAXD	111060677	0.003
15	NA	63345547	63425768	LACTB	63453484	2.34e-28
20	NA	25769672	25727136	ACSS1	25740998	0.013
22	33046025	33046036	NA	PISD	33058307	1.86e-04

Nominally significant T2D loci
T2D locations p-value $< 10^{-3}$

Chr	T2D location (WTC)	T2D location (AA)	T2D location (MTC)	*cis*–NEMG	eQTL location[d]	eQTL p-value
2	NA	194739106	194690969	COQ10B	194842026	3.76e-04
2	NA	194739106	194690969	HSPD1	194213959	0.010
2	200305909	NA	200330147	C2orf69	200457932	0.015
2	200305909	NA	200330147	MAIP1	200070891	3.48e-07
6	NA	76217291	76199095	COX7A2	76466553	4.19e-04
7	48732315	48812202	NA	ABCA13	48628852	2.71e-06
7	140349763	140367908	NA	MRPS33	140378682	0.010
10	112924900	112866891	NA	GPAM	112917454	0.004
11	61284211	61258729	NA	FADS2	61260634	5.89e-04
12	12633435	12621259	NA	HEBP1	12634131	0.004
12	123387253	NA	123470526	ABCB9	123260234	2.95e-02
					123244794	2.40e-02
					123386756	1.11e-06
12	123387213	123750895	123447928	ABCB9	123386769	7.63e-07
					123260234	2.95e-02
12	123387213	123750895	123447928	COXPD7	123439939	0.038
12	123387253	NA	123470526	DIABLO	123386662	0.003
					123797762	0.008
					123387709	6.55e-05
12	133105848	NA	133168320	PGAM5	133182753	0.018
15	77270791	NA	77310648	IDH3A	77085108	0.003
16	9759261	NA	9794698	ABAT	9799479	0.005
21	44327412	44352930	NA	NDUFV3	44210786	0.009

Table 3: T2D and eQTL location estimates associated with 50 T2D *cis*–NEMGs. All locations were estimated on population-specific genetic LDU maps and converted to physical coordinates (B37). All eQTL locations ($\hat{S}_{eQTL}$) are within ± 1 LDU of a T2D location estimate ($\hat{S}_{T2D}$). $\hat{S}_{T2D}$ are presented for the European WTCCC1 (WTC), African American NIDDK (AA) and European WTCCC2 MetaBoChip study cohorts (MTC) (signals with low SNP coverage indicated by N/A, were not analysed). $\hat{S}_{eQTL}$ were generated using subcutaneous adipose gene expression for a population based sample of European individuals from the MuTHER consortium (TwinsUK).

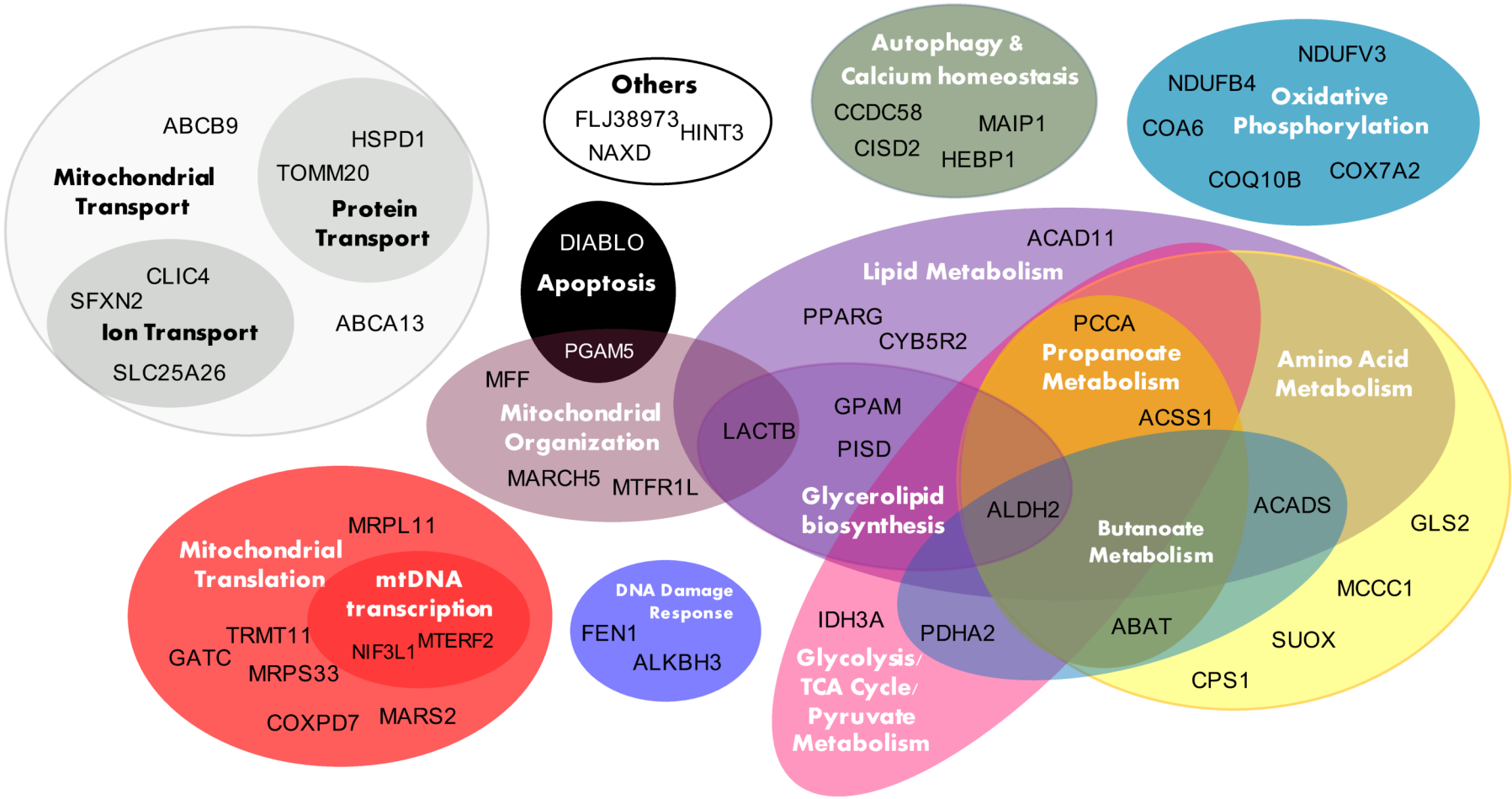


Figure 19: The 50 T2D *cis*–NEMGs are here grouped by general biological functions. The corresponding, more detailed descriptions of protein function are listed in Appendix A.1.

3.5 Aim three: test *cis*-genes for enrichment of mitochondrial pathways

3.5.1 Methods

The 763 T2D *cis*-genes were tested for evidence of mitochondrial pathway enrichment by cross-referencing with 41 curated gene sets downloaded from the Molecular Signatures Database (Subramanian et al., 2005; Liberzon et al., 2011), of which each had ≥25% genes in MitoCarta2.0/+ (see Appendix A.2). The 763 *cis*–genes were cross-referenced with the gene sets to generate observed counts (*obs*). Expected counts (*exp*) were generated using 10,000 permutations, in which genes were randomly selected from the same array used to measure the *cis*–genes using either:

- **Random approach:** 763 genes were randomly selected.
- **Structured approach:** the random gene selection was controlled for the local structure seen in the T2D *cis*-genes (selected within ±1.5Mb of a disease locus). There were ∼4.6 *cis*–genes per locus (763/166), therefore each permutation randomly selected 190 genes, then randomly selected three genes from within ±1.5Mb. Additional genes were randomly selected to bring the total to 763. The rationale for this approach was to emulate any potential correlation structure observed between local *cis*–genes.

For each gene set, an empirical *p*-value was calculated as $\Sigma(exp \geq obs)/10{,}000$, where $\Sigma(exp \geq obs)$ is the total number of permutations for which the number of random genes in the gene set (*exp*) was ≥ to the observed count of T2D *cis*–genes in the gene set (*obs*).

3.5.2 Results

The T2D *cis*–genes were not significantly enriched for NEMGs per se (gene set defined as the combined Mitocarta2.0/+, *p*-value = 0.14). However, the permutation analysis returned evidence of enrichment for four mitochondrial-related gene sets with a *p*-value threshold of 0.05. These are shown in Table 4.

Mitochondrial gene set	*cis*–genes	(a) *p*-value (random)	(b) *p*-value (structured)
Valine, leucine and isoleucine degradation (KEGG)	5	0.020	0.027
Biotin carboxylases (manually defined)	2	0.022	0.019
Propanoate metabolism (KEGG)	4	0.025	0.042
Butanoate metabolism (KEGG)	4	0.032	0.042

Table 4: Mitochondrial pathways with evidence of enrichment in the total T2D *cis*-genes (p-value <0.05). Gene sets corresponding to mitochondrial pathways are listed, along with the source, either KEGG (Kanehisa et al., 2015) or manually defined.

The enrichment of these gene sets are largely driven by a core set of overlapping genes. The T2D *cis*–NEMGs in each gene set are shown below:

- Valine, leucine and isoleucine degradation: *ABAT, ACADS, ALDH2, MCCC1* and *PCCA*
- Biotin carboxylases: *MCCC1* and *PCCA*
- Propanoate metabolism: *ABAT, ACSS1, ALDH2* and *PCCA*
- Butanoate metabolism: *ABAT, ACADS, ALDH2* and *PDHA2*

3.6 Discussion

This Chapter aimed to investigate the hypothesis that T2D loci regulate genes involved in mitochondrial function. As such, T2D *cis*–genes were assigned to T2D loci based on the co-location of $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ location estimates provided by Lau et al. This hypothesis is based on strong prior evidence that perturbed mitochondrial function may cause T2D (see Chapter 1, **Section 1.5.2: Mitochondrial function and T2D**), making an aim of this study to identify candidate genes which may constitute underlying genetic mechanisms. The number of potential candidate genes was increased by including $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ of nominal significance. A total of 50 nuclear-encoded mitochondrial genes (NEMGs) were identified as putative T2D *cis*–genes.

A test of enrichment did not provide evidence of NEMG enrichment in the total *cis*–genes, however more specific tests returned significant results for five mitochondrial-related pathways: branched chain amino acid (BCAA) degradation, biotin carboxylases, propanoate

metabolism and butanoate metabolism. These pathways largely overlapped, with a core set of *cis*–NEMGs driving the observed enrichment. These results may be consistent with a contributory, rather than principal role for mitochondrial dysfunction in T2D onset, although specific features of mitochondrial metabolism such as those highlighted here may be of particular importance. However, future work investigating eQTL identified in other tissues may reveal additional contributions of mitochondrial dysfunction. Specific candidate genes and pathways, including their potential relationship with T2D are discussed below. A summary of the NEMG functions is available in Appendix A.1.

Candidate T2D *cis*–NEMGs and T2D

All of the 50 identified T2D *cis*–NEMGs present mechanisms of interest which may potentially contribute to T2D onset. As seen in Figure 19 (page 102), the implicated pathways include mitochondrial transcription, translation and organisation, as well as more specific metabolic processes including lipid, amino acid, butanoate and propanoate metabolism, oxidative phosphorylation, pyruvate metabolism and the TCA cycle, in addition to mitochondrial protein and iron transport, autophagy & calcium homeostasis and apoptosis. Many of these pathways have been previously implicated in T2D (some are discussed in Chapter 1, **Section 1.5.2: Mitochondrial function and T2D** and lipid metabolism is discussed specifically in Chapter 5, **Section 5.1.2: *ACAD11*, fatty acid oxidation and diabetes**). Several examples of candidate genes and pathways which may be related to T2D aetiology are discussed below.

The first noteworthy example is the pathway of branched chain amino acid (BCAA) catabolism, which showed evidence of enrichment in the total T2D *cis*–genes by count. Five of the identified T2D *cis*–NEMGs were annotated to this pathway, demonstrating potential T2D-associated genetic regulation of multiple adjacent steps, as shown in Figure 20. The five *cis*–NEMGs include two biotin-dependent carboxylases, *MCCC1* and *PCCA*, as well as *ABAT* (encodes GABA transaminase), *ACADS* (encodes acyl-CoA dehydrogenase, short chain) and *ALDH2* (encodes aldehyde dehydrogenase 2 family member). An eQTL for *MCCC1* was previously shown to be highly associated with BMI, further sug-

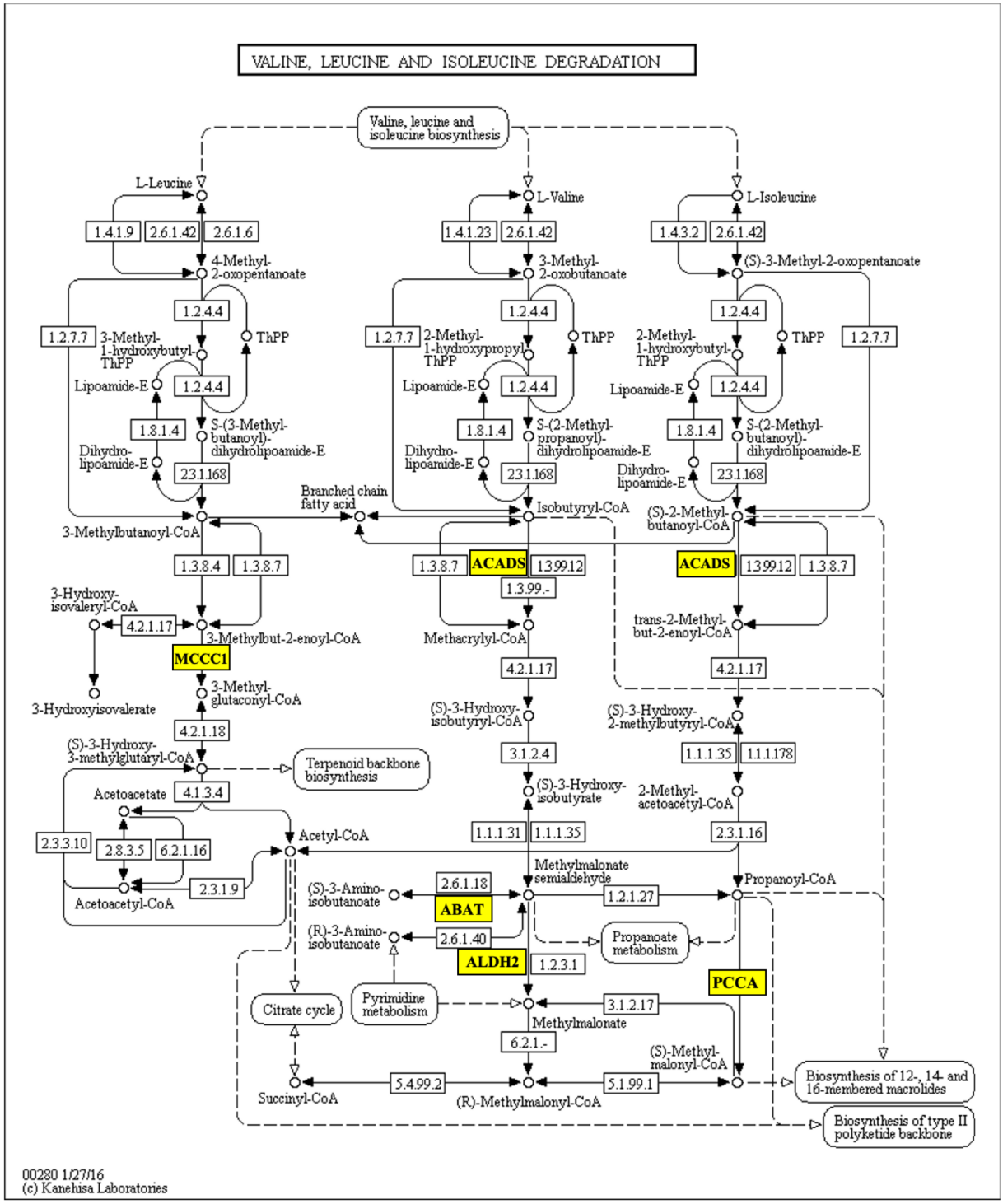


Figure 20: The branched chain amino acid catabolism pathway adapted from the KEGG database (Kanehisa et al., 2019). Five T2D *cis*–NEMGs were annotated to this pathway and are highlighted in yellow: *MCCC1, ACADS, ABAT, ALDH2* and *PCCA*.

gesting that this T2D locus could increase risk through an adiposity-related mechanism (Sajuthi et al., 2016).

There is a vast literature regarding the role of BCAA catabolism in T2D. For example,

circulating levels of the BCAAs, valine, leucine and isoleucine are increased in individuals who go on to develop T2D, demonstrating potential predictive power (Flores-Guerrero et al., 2018; Guasch-Ferré et al., 2016). The catabolism of the three BCAAs may influence mitochondrial function, metabolism and insulin sensitivity (reviewed in detail by Gannon et al. (2018)), as well as adipogenesis, lipogenesis (Green et al., 2016), inflammation (Nicastro et al., 2012; Zhenyukh et al., 2018) and appetite (Solon-Biet et al., 2019). Consistent with a previous study which reported a link between T2D genetic risk and impaired BCAA catabolism (Lotta et al., 2016), the results here present five candidate genes which may potentially drive this risk, providing testable genetic mechanisms for further analysis. For example, propionyl CoA-carboxylase, for which the α chain is encoded by the T2D *cis*-gene *PCCA*, catalyses a step directly upstream of the B_{12}-dependent methylmalonyl-CoA mutase (Mut) which induces impaired glucose tolerance in heterozygous knock-out mice (Lerin et al., 2016). Mut catalyzes the conversion of methylmalonyl-CoA, produced by the catabolism of BCAAs and odd-chain fatty acids, to the TCA cycle intermediate succinyl-CoA. It will also be of interest to further investigate BCAA *cis*–genes in other tissue types, such as omental adipose in which suppressed BCAA catabolism is associated with insulin resistance (Lackey et al., 2013).

As mentioned above, *MCCC1* and *PCCA* encode two biotin-dependent carboxylases, of which there are a total of five encoded in the human genome (Tong, 2013). Biotin levels are notably reduced in individuals with T2D (Valdés-Ramos et al., 2015) and biotin supplementation may improve glucose control (Sahin et al., 2013), making these T2D *cis*–NEMGs of potential therapeutic interest. An additional strong candidate is *CISD2*, in which mutations cause Wolfran Syndrome of which a known symptom is diabetes (Rouzier et al., 2017). *CISD2* encodes an integral membrane protein which facilitates the interaction between the mitochondria and endoplasmic reticulum (ER) membranes; a process which is itself associated with T2D and insulin resistance (Wang et al., 2014, 2015; Thivolet et al., 2017; Tubbs et al., 2018). The current study identified evidence that *CISD2* may be regulated by a genetic locus associated with risk of common T2D.

Three T2D *cis*–NEMGs regulate mitochondrial fission: *MARCH5*, *MFF* and *MTFRL1*.

Mitochondrial fission has been shown to cause insulin resistance when induced and improved insulin signalling when inhibited (Jheng et al., 2012; Rovira-Llopis et al., 2017; Lin et al., 2018a). Additional T2D *cis*-genes which were included in the list of NEMGs due to the proteins not being present in the mitochondria include *PPARG* and *SPATA18*. *PPARG* encodes the ligand-activated transcription factor PPAR-γ, while the protein encoded by *SPATA18* facilitates the removal of oxidised proteins in the mitochondria. These two genes regulate mitochondrial function and quality, respectively (Yeligar et al., 2018; Dan et al., 2020).

Strengths, limitations and future work

A common approach used in conventional GWAS is to assign significant SNPs from eQTL analysis (eSNPs) which are in high LD with GWAS lead SNPs (Scott et al., 2017; van de Bunt et al., 2015). Similarly, $\hat{S}_{eQTL}$ were here assigned to $\hat{S}_{T2D}$ if they were within a close genetic distance of 1 LDU, reflecting high LD. There are multiple advantages to using an LDU-based approach for mapping T2D loci and their *cis*–genes; these are discussed in detail in Chapter 2, **Section 2.7.2: LDU-based gene mapping**. Briefly, these include increased power to detect associations which makes it more likely that *cis*–genes will be identified, due to the integration of multiple genotyped markers, population-specific LD and a reduced multiple testing burden.

There are several assumptions which must be considered in the design of this study. Firstly, $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ within 1 LDU are assumed to represent a shared causal variant; this is an assumption of all approaches which test co-location based on high LD. As discussed on page 86, alternative methods can be used to formally assess shared causality. A more formal test for co-localisation could be developed for use with the $\hat{S}$ location estimates. For example, similarities between the likelihood surface outputs from the Malecot could be compared, similarly to the comparison of trait-associations carried out by COLOC (see Figure 13 on page 86). The Malecot model assumes that there is just one causal variant per window and outputs the highest peak of the likelihood surface. An additional step might be to run conditional analysis in a search for secondary signals, in

order to investigate allelic heterogeneity.

Future analysis may build on these results by further developing the methods used. Additional $\hat{S}_{eQTL}$ may be detected by incorporating RNA-seq datasets, which offer significant improvement over expression arrays. Array data has several complications including systematic biases introduced during sample preparation and potential SNPs within gene expression probe which may change probe affinity and appear as differential expression (Akey et al., 2007; Spielman et al., 2007; Alberts et al., 2007). Arrays are also limited to the genes included and importantly, gene annotation is continuously updated. With this in mind, the probes used in this analysis could be compared to gene coordinates according to Ensembl or GENCODE (GRCh38) to obtain the most up-to-date annotations. RNA-seq also captures genes which may not be functionally annotated such as long non-coding RNAs (lncRNAs). Follow-up analysis such as that by Small et al. (2018) may investigate genome-wide changes in gene expression in response to candidate risk variants to identify *trans*-regulatory networks. The study of *trans*–eQTL may also reveal additional tissue-specific mechanisms (Fagny et al., 2017; Consortium et al., 2017).

On the topic of tissue specificity, the $\hat{S}_{eQTL}$ for this study were mapped only for subcutaneous adipose, therefore the current *cis*–genes implicate mechanisms which are either specific to subcutaneous adipose or are shared across tissues. There is current ongoing work by Lau et al. to map $\hat{S}_{eQTL}$ using RNA-seq data available from the GTEx database for multiple tissues. These additional data will be important to (1) replicate the $\hat{S}_{eQTL}$ in this study since they were generated for only one dataset and (2) identify *cis*–NEMGs potentially regulated by other tissue-specific eQTL. This will facilitate the mapping of mechanisms which underlie tissue-specific mitochondrial functions (Fernández-Vizarra et al., 2011; Pacheu-Grau et al., 2018; Kappler et al., 2019). Beyond tissue specificity, regulatory elements can also be activated under stimulated conditions and are highly dynamic (Siersbæk et al., 2017; Freire-Pritchett et al., 2017; Rubin et al., 2017; Ramos-Rodríguez et al., 2019; Miguel-Escalada et al., 2019). For example, insulin stimulation was required to observe the increased expression of *ANK1* in the presence of the rs508419 T2D-risk allele (Yan et al., 2016). While most publicly available chromatin maps are representative

of a basal state, independent data may be useful for mapping eQTL active under stimulated conditions (Kyono et al., 2019). Additional NEMGs may also be implicated through the mapping of other QTL, such as loci associated with metabolites levels or measures of mitochondrial function. Expression data stratified for different heterogenous subtypes of T2D patients (see Udler (2019)) may also inform on NEMGs which are dysregulated with specific phenotypes.

An important future work will be to validate the 50 T2D *cis*–NEMGs identified in this analysis. The next Chapter 4 describes a step towards this, by investigating whether the *cis*–genes show differential expression in independent datasets of T2D case and control gene expression. Gene expression data may also be used to further investigate the choice of 1 LDU as a co-location threshold. One way to do this may be to compare the expression of *cis*–genes in an independent cohort of T2D cases and controls identified using different statistical thresholds. The enrichment of differential expression may be used to inform an optimal co-location threshold for identifying true positives. This approach is tested in the next Chapter 4, **Section 4.4: Physical vs genetic distance: differential expression**. Another important step to validation will be to fine-map the causal variants at each locus and provide evidence that they regulate the assigned *cis*–gene(s). This may be achieved through functional validation and by integrating chromatin interaction maps, which can provide evidence that a causal variant makes physical contact with a *cis*–gene promoter, for example. Chapter 5 discusses fine-mapping in detail and presents the fine-mapping of one candidate T2D NEMG locus.

Causation or correlation?

In this analysis, 763 *cis*–genes were assigned to 166 T2D loci, equivalent to an average of ∼4.6 *cis*–genes per locus. This is consistent with T2D risk being increased by mutations in regulatory elements which target multiple genes. For any one locus, it is an important question as to whether the risk of T2D is conferred by the altered expression of all the *cis*–genes, or by a subset. This question may be further investigated by integrating these results with targeted sequence data as well as follow-up functional studies which can

perturb individual genes in *in vitro* or *in vivo* to assess the impact on diabetes-related phenotypes. *Cis*-genes are also more likely to be causal if they are identified as *cis*–genes for multiple independent T2D loci.

3.6.1 Conclusions

A total of 50 *cis*–NEMGs were identified as potential target genes of T2D risk loci. Many of these genes implicate pathways which are well-established as related to the onset of T2D, while others implicate novel genetic pathways. The following Chapter 4 will aim to validate the role of these *cis*–genes in T2D by investigating their expression in T2D cases compared to controls in independent datasets. This will investigate the hypothesis that the identified *cis*–genes are regulated by T2D-risk variants present at different allele frequencies in T2D cases.

4 Chapter 4: T2D *cis*-gene expression in cases vs controls

4.1 Introduction

In Chapter 3, 763 T2D *cis*–genes were identified based on the co-location of the associated $\hat{S}_{eQTL}$ within 1 LDU of an independent, replicated $\hat{S}_{T2D}$. A total of 166 T2D loci co-located with $\hat{S}_{eQTL}$, defined as putative T2D-eQTL. Chapter 4 will aim to provide independent validation by investigating whether the same *cis*–genes are observed to be differentially expressed in independent case-control gene expression data. If the *cis*–genes are regulated by shared T2D risk variants then gene expression is expected to differ between cases and controls; this is illustrated in Table 5.

eQTL allele frequencies cause differential gene expression

	Gene expression	Allele frequency	Average gene expression level
Healthy Controls			
ACTGAG**T**ACGGAT	100%	80%	$(100 \times 0.8) + (50 \times 0.2) = 90\%$
ACTGAG**G**ACGGAT	50%	20%	
T2D Cases			
ACTGAG**T**ACGGAT	100%	60%	$(100 \times 0.6) + (50 \times 0.4) = 80\%$
ACTGAG**G**ACGGAT	50%	40%	

Table 5: In this illustrative example, a regulatory variant T>G causes a 50% reduction in gene expression levels. The higher frequency of the G allele in T2D cases causes an overall reduction in the average gene expression level.

Observing differential expression of the T2D *cis*–genes in T2D cases compared to controls will provide an important step towards validating $\hat{S}_{eQTL}$ which were originally mapped using adipose gene expression data for an independent population-based cohort. Gene expression will be analysed in publicly available gene expression datasets for skeletal muscle, liver and pancreas in addition to adipose tissue, in order to investigate potential ubiquitous effects in multiple tissues relevant to T2D.

4.1.1 Aims

The main aim of Chapter 4 is to provide independent evidence that the T2D *cis*-genes associated with T2D-eQTL mapped in Chapter 3 are also differentially expressed in T2D or insulin resistant (IR) cases compared to healthy controls. To do so, the following aims will be met:

1. Identify publicly available case-control gene expression datasets based on pre-defined inclusion and exclusion criteria (Section 4.2)

2. Test for differential gene expression and carry out meta-analysis of the T2D *cis*–genes, *cis*–NEMGs and mitochondrial pathways using gene set enrichment analysis (GSEA) (Section 4.3).

4.2 Aim one: identify gene expression datasets

4.2.1 Dataset search

Datasets measuring gene expression in T2D or IR cases were obtained from the public Gene Expression Omnibus (GEO) repository (Edgar et al., 2002; Barrett et al., 2012). Inclusion and exclusion criteria are listed in Table 6. Both T2D and IR phenotypes were included since the T2D *cis*-genes were mapped using subcutaneous adipose expression data and are likely to capture either adipose-specific or multi-tissue mechanisms involved in peripheral insulin resistance. Datasets were required to have the original, baseline gene expression measures (prior to any intervention), measured on Affymetrix arrays to allow for a consistent normalisation and meta-analysis pipeline. Cases with Type 1 diabetes or those taking insulin medication were excluded in order to obtain baseline gene expression measures. The search string used to identify datasets from GEO is shown below. Datasets with European samples were included in order to minimise heterogeneity.

Inclusion criteria	Exclusion criteria
T2D or insulin resistant case-control study	Ongoing medications at time of sampling (excluding pancreas samples)*
Baseline gene expression measures	Non-European cohort
Original data available as CEL format	Measurements from cultured cells
Data generated on an Affymetrix array	
Either skeletal muscle, adipose, liver or pancreas	

Table 6: Inclusion and exclusion criteria for GEO gene expression datasets. *Pancreas datasets were exempt from this exclusion criteria, since the donors are typically deceased and unable to stop medication prior to sampling.

GEO search string:

(((((T2D OR Type 2 Diab* OR IGT OR insulin resistan* OR IFG OR pre*diab* OR impaired fasting glucose OR impaired glucose tolerance)) AND (homo sapiens[Organism] OR human[Organism])) AND cel[Supplementary Files]) AND expression profiling by array[DataSet Type]) AND (muscle OR skeletal muscle OR adipose OR adipo* OR omental OR subcutaneous OR skeletal muscle OR vastus lateralis OR rectus abdominus OR liver OR hepat* OR pancreas OR islet OR beta cell* OR myotub*)

4.2.2 Results

132 datasets were returned from the initial GEO search. Filtering according to the inclusion and exclusion criteria led to the exclusion of 106 results (see Figure 21). The remaining 26 datasets were subject to a full text review, leading to the exclusion of an additional 13 results. Of the final 13 datasets, three were skeletal muscle, five were adipose, two were liver, three were pancreas and one dataset had both skeletal muscle and adipose samples. Summary information for these 13 datasets is provided in Table 7. Of the final 13 datasets, one skeletal muscle dataset included healthy controls with zero, one or two parents affects by T2D (GSE25462) (see Table 7).

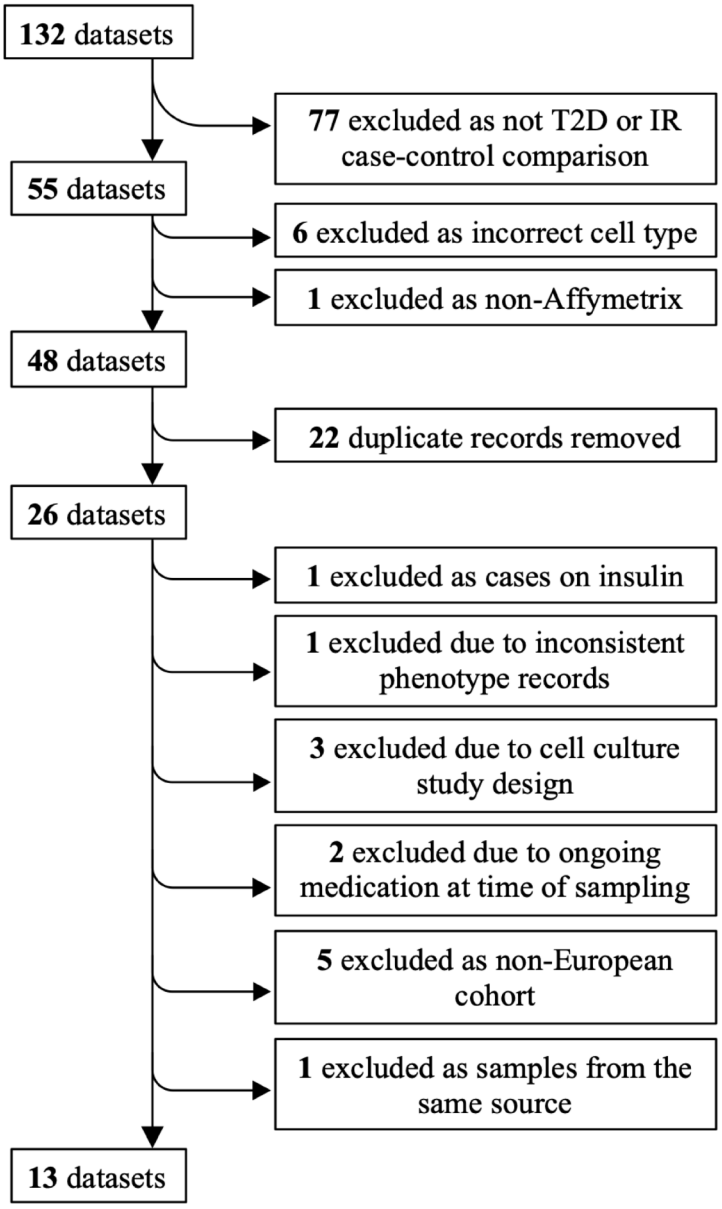


Figure 21: Gene expression datasets from GEO were investigated according to the *a priori* defined inclusion and exclusion criteria, shown in Table 6. 132 datasets were returned using the search string shown on page 114. Following a manual review of the inclusion and exclusion criteria, 119 results were excluded.

4.3 Aim two: differential gene expression

4.3.1 Differential gene expression

13 datasets passed the inclusion and exclusion criteria; these are listed in Table 7 with more detail available in Appendix A.3. For each of these datasets, raw expression data was downloaded from GEO, normalised and regressed on case control status to calculate differential expression. The steps used to test for differential gene expression (DGE)

Dataset ID	Tissue	Phenotype(s)	Cases / Controls	No. of genes
GSE13070[a]	Skeletal Muscle	IR	51/18	21,276
GSE25462[b]	Skeletal Muscle	T2D (& FH+)	10/15	21,276
GSE22435	Skeletal Muscle	T2D/IGT	7/10	21,276
GSE101492	Adipose	IR	40/40	24,254
GSE26637	Adipose	IR	5/5	20,212
GSE94752	Adipose	IR	18/21	21,276
GSE20950	Adipose	IR	9/10	21,276
GSE27949	Adipose	T2D (& IGT)	12/11	21,276
GSE13070[a]	Adipose	IR	28/6	21,276
GSE64998	Liver	T2D	7/8	13,931
GSE15653	Liver	T2D	9/4	20,212
GSE76894	Pancreas	T2D	19/83	21,276
GSE25724	Pancreas	T2D	6/7	13,931
GSE41762	Pancreas	T2D	20/57	20,212

Table 7: Summary information for the 13 gene expression datasets obtained from GEO using the inclusion and exclusion criteria listed in Table 6. [a]The dataset GSE13070 provided both skeletal muscle and adipose samples. [b]Controls included family history information as the number of parents with T2D. The No. of genes reports the number of genes present on the gene expression array used for each study. T2D = Type 2 diabetes, IR = insulin resistant, IGT = impaired glucose tolerance, FH+ = family history positive.

are shown in Figure 22. Briefly, raw gene expression data were downloaded in CEL format and normalised using robust multi-array averaging (RMA) (Irizarry et al., 2003), as implemented by the R *oligo* package (Carvalho and Irizarry, 2010). Phenotype data was extracted using the *GEOquery* package (Davis and Meltzer, 2007) and the corresponding array annotation package was downloaded (see Appendix A.3 for the array used by each study). Each gene was tested for evidence of differential expression between cases and controls by regressing the expression, as measured by one or more probes, on disease status. Age and BMI were included as co-variates where available[68]. If more than one probe was assigned to a gene then the probes were fitted as random effects in a linear mixed-effects model (Bates et al., 2015). This "gene-centric" model assumes that probes

[68]Of the three T2D cohorts analysed by Lau et al. (2017), the African American NIDDK cohort was matched for BMI (Palmer et al., 2012), in order to minimise confounding of loci associated with BMI. However, this has been suggested to bias against the discovery of loci which increase the risk of T2D via their effects on adiposity and BMI (Billings and Florez, 2010). An additional step might be to further study the expression of the T2D *cis*–genes and *cis*–NEMGs without correcting for BMI, however the current analysis included BMI as a covariate for consistency with the Palmer et al. (2012) and Lau et al. (2017) designs.

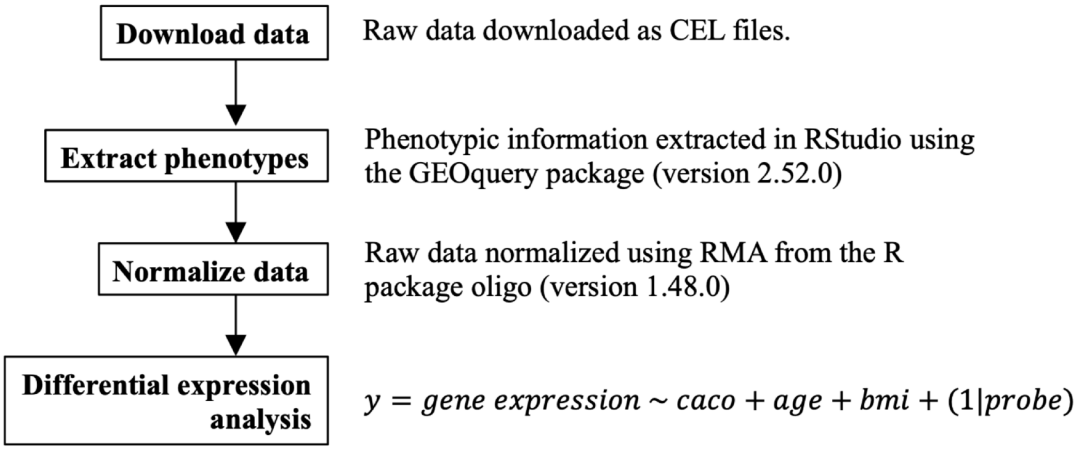


Figure 22: Analysis pipeline for differential gene expression analysis. *GEOquery* (Davis and Meltzer, 2007) and *oligo* (Carvalho and Irizarry, 2010) packages are used. Gene-centric scores of differential expression were calculated by combining raw expression measures from multiple probes as random effects in a linear mixed-effects model.

show the same direction of effect and may therefore be limited if transcripts for a single gene behave differently from one another. The regression model fitted for each gene is shown below, with *caco* representing case or control status (coded as 1 or 0):

$$gene\ expression \sim caco + age + BMI + (1|probe)$$

Where $(1|probe)$ indicates a random effect intercept term for the probes in each gene, effectively allowing a different intercept for each gene probe and a single, fixed-effect estimate for the caco term (with a non-zero β regression estimate providing evidence for a mean difference between cases and controls and hence differential gene expression). The total number of regressions for each dataset equaled the number of genes captured by each array, as shown in Table 7.

Since differential gene expression can also be induced by and thus confounded by disease status, an additional DGE analysis was carried out for one skeletal muscle dataset (GSE25462) which included healthy, normoglycemic but insulin resistant individuals with zero, one or two parents with T2D. This dataset represents an informative study design in which to investigate heritable changes in gene expression with reduced confounding caused by disease onset. Gene expression for each unaffected individual was regressed

against the number of affected parents (0, 1 or 2):

$$gene\ expression \sim \text{number of affected parents} + age + BMI + (1|probe)$$

A Z-score of differential expression was calculated for each gene by dividing the β regression coefficient by its standard error for the explanatory variable of interest (case control status or the number of affected parents).

4.3.2 Additional quality control

The Z-scores were correlated between datasets to test for dataset homogeneity as described by Väremo et al. (2015). A Pearson correlation coefficient was calculated for each pair of datasets using Z-scores for genes with Z <-1 or Z >1 in either dataset, in order to minimise uninformative variation. The results are shown in Figure 23. The correlations are shown by red (-1 to 0) and green (0 to +1) shading and the within-tissue correlations are highlighted in red boxes. The within-tissue datasets were generally positively correlated. One adipose dataset, GSE20950, had a weak negative correlation with the other adipose datasets, although it was strongly positively correlated with the three skeletal muscle datasets and so was retained. The two liver datasets had a weak negative correlation. As expected, the between-tissue correlations were generally weaker compared to the within-tissue comparisons.

4.3.3 Meta-analysis

Individual datasets were included in a tissue-specific meta-analysis by combining gene summary Z-scores and variance of differential expression using a random effects model (REM). As described by Choi et al. (2003), a REM may be used to meta-analyse gene summary scores across studies. As opposed to a fixed effects model which assumes that variation between datasets is due to sampling error alone, the REM considers heterogeneity between datasets caused by random sampling of the study cohorts from the population, by assuming study-specific means. This suitability of the REM for the data in question can be formally tested by estimating Cochran's Q statistic, which uses a non-parametric

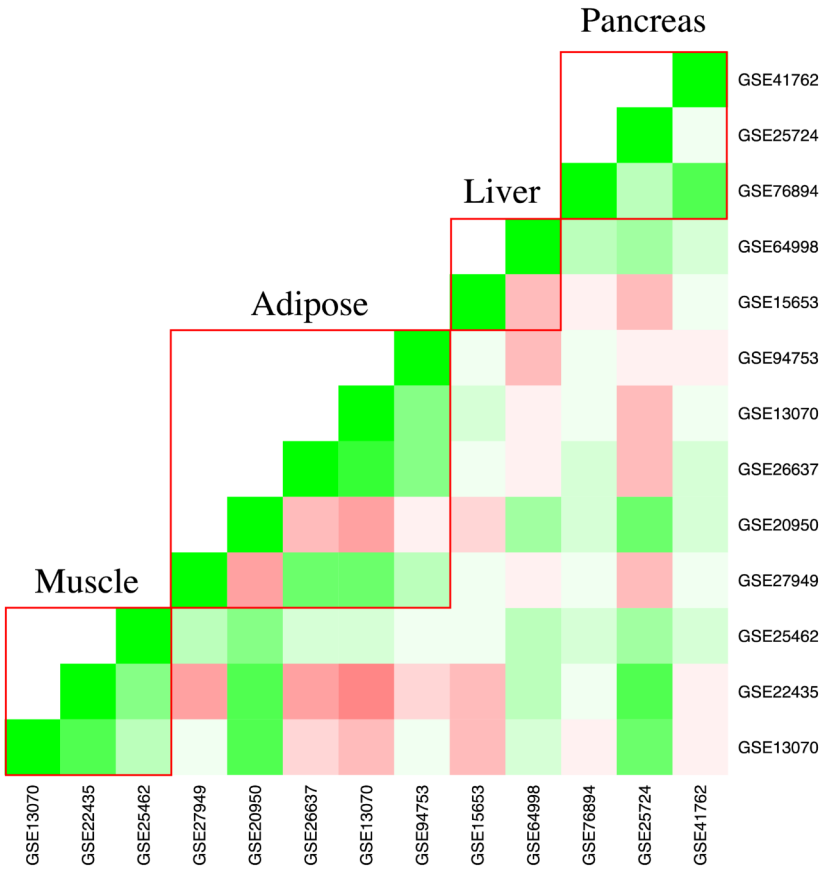


Figure 23: Global correlation of genome-wide Z-scores between datasets, included if the absolute Z-score was >1 in both datasets. The correlations range from -0.42 to 0.78, with positive values shown in green and negative values shown in red.

test with a null hypothesis that the mean gene expression is identical between studies. This test confirmed the need for REM for each tissue-specific meta-analysis. As a result, meta-analysis was carried out using an REM according to Choi et al. (2003), implemented using the R package *GeneMeta* (Lusa et al., 2019) (version 1.56). Each meta-analysis included the number of genes on the smallest array.

4.3.4 Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) refers to methods used to analyse the differential expression of sets of genes (Mootha et al., 2003; Subramanian et al., 2005), such as those representing a particular biological pathway. GSEA typically tests for enrichment of differential expression by comparing the gene-level summary statistics of differential expression of a gene set to the genomic background (e.g Z-scores or p-values). This can increase the power by analysing trends in the expression of multiple genes without requiring them all

to achieve statistical significance, while also providing biological insight.

GSEA was carried out in R using the package *piano* (Väremo et al., 2013), which incorporates different GSEA methods into a single, user-friendly interface[69]. *Piano* allows the user to test gene sets for enrichment which can be either distinct-directional, in which the gene set is enriched for decreased *or* increased expression or mixed-directional, in which the gene set is enriched for genes with both significantly increased *and* significantly decreased expression. For the following analyses, GSEA was carried out by calculating a Wilcoxon test statistic using *piano*, unless otherwise stated. The non-parametric Wilcoxon rank-sum test has been shown to have the highest reproducibility and sensitivity compared to other commonly used GSEA test statistics (Hung et al., 2011). Significance was calculated through gene sampling, in which sets of randomly selected genes were compared to the query gene set. The number of permutations was set at 10,000. GSEA were carried out to test the case vs control differential expression of the T2D *cis*-genes and *cis*-NEMGs identified in this study, as well as pre-defined mitochondrial pathways. These are described in more detail below.

(1) Differential expression of T2D *cis*-genes and *cis*-NEMGs

The first set of GSEA tested whether the T2D *cis*-genes and *cis*-NEMGs were enriched for differential expression in T2D/IR cases compared to controls. Z-scores that reflect the differential expression of each gene between cases and controls were used to make two initial comparisons:

1. T2D *cis*-genes (n=763) against the genomic background
2. T2D *cis*-NEMGs (n=50) against the genomic background

Three control gene sets were randomly selected for each GSEA. Three sets of 763 genes were selected as controls for the first GSEA from a list of adipose *cis*-genes obtained from the MuTHeR Consortium (Grundberg et al., 2012), from which summary data was

[69]Väremo et al. (2013) described 11 different competitive tests which compare a set of genes to the genomic background. These differ by permutations either randomising the genes or case-control status, or by the test statistic.

used to map the $\hat{S}_{eQTL}$ locations (Lau et al., 2017). The control genes were significant *cis*–genes associated with eQTL greater than 2Mb away from a T2D location, in order to test whether $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ co-location conferred a greater enrichment of differential expression. The *cis*–genes were associated with eQTL from the MuTHeR single-SNP analysis which achieved a p-value of <0.05, the same threshold used to include $\hat{S}_{eQTL}$. For the second GSEA, three sets of 50 genes were randomly selected from the list of total known NEMGs described in Chapter 3, **Section 3.4.1: Nuclear encoded mitochondrial genes (NEMGs)** (MitoCarta+ and MitoCarta2.0 combined), with the 50 observed T2D *cis*-NEMGs excluded.

Since mitochondrial dysfunction may also result as a consequence of developing T2D or IR (see Chapter 1, **Section 1.5: Mitochondrial dysfunction in Type 2 diabetes**), a prior hypothesis might be that all NEMGs show differential expression as a consequence of disease. In order to investigate whether the observed *cis*–NEMGs represent a subset of NEMGs regulated by T2D loci, the expression of the 50 *cis*–NEMGs were compared to the background of all known NEMGs. The third GSEA was therefore:

3. T2D *cis*-NEMGs (n=50) against all known NEMGs

The same three control sets of 50 NEMGs were included as controls as above.

(2) Differential expression of mitochondrial pathways

Previously in Chapter 3, **Section 3.5: Aim three: test *cis*-genes for enrichment of mitochondrial pathways**, four gene sets representing mitochondrial pathways showed evidence of enrichment in the total T2D *cis*–genes: valine, leucine and isoleucine degradation, biotin-dependent carboxylases, propanoate metabolism and butanoate metabolism. The same four gene sets, downloaded from the curated Molecular Signatures Database (MSigDB) (Liberzon et al., 2015), were tested for differential gene expression in these case-control datasets.

4.3.5 Results

DGE analysis was carried out on each dataset and the meta-analysed tissues. The resulting summary statistics for differential expression are plotted in Figure 24, ranked by absolute Z-score. Individual datasets are plotted in black with the meta-analysis in red. The steeper curve of the red, meta-analysed Z-scores reflect an increased discovery rate of significant differential expression following meta-analysis.

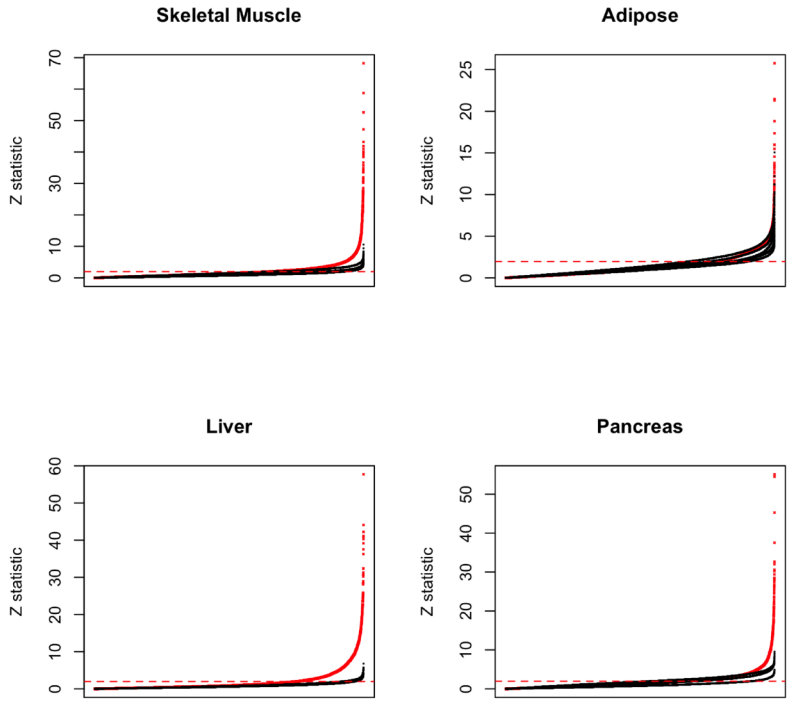


Figure 24: Absolute Z-scores for the case-control gene expression datasets. The absolute Z-scores are plotted in red for the meta-analysis and black for the individual datasets. A horizontal dotted red line is shown at Z = 1.96, corresponding to significant differential expression at a p-value threshold of 0.05.

(1) GSEA RESULTS: T2D *cis*-genes vs the genomic background

Table 8 shows the results of the first GSEA. The T2D *cis*–genes were enriched for differential expression in all four meta-analysed tissues, showing decreased expression in muscle, liver and pancreas and mixed differential expression[70] in adipose (FDR-adjusted p-value $\leq$0.05).

[70] A 'mixed' result is here used to refer to a significant test for differential expression including both increased and decreased expression (Väremo et al., 2013)

GSEA results: T2D *cis*-genes vs the genomic background

Dataset	Tissue	763 T2D *cis*–genes vs ALL	763 random *cis*–genes vs ALL (3 control gene sets)		
Meta-analysis	Muscle	↓ 0.001	↓ 0.016	n.s.	↓ 0.003
Meta-analysis	Adipose	0.006	n.s.	n.s.	n.s.
Meta-analysis	Liver	↓ 0.020	n.s.	n.s.	↑ 0.003
Meta-analysis	Pancreas	↓ 0.003	↑ 0.009	n.s.	n.s.
GSE13070	Muscle	↓ 0.050	n.s.	n.s.	↓ 0.003
GSE25462	Muscle	n.s.	n.s.	n.s.	0.012
GSE22435	Muscle	↓ 0.002	0.003	n.s.	↓ 0.036
GSE101492	Adipose	↓ 0.025	n.s.	n.s.	n.s.
GSE26637	Adipose	↑ 0.027	n.s.	n.s.	n.s.
GSE94752	Adipose	n.s.	n.s.	n.s.	n.s.
GSE20950	Adipose	↓ 2.5*e*-04	n.s.	n.s.	n.s.
GSE27949	Adipose	↑ 0.017	n.s.	↑ 0.009	↑ 0.034
GSE13070	Adipose	↑ 0.003	↑ 0.002	↑ 0.035	↑ 0.002
GSE64998	Liver	↓ 0.010	n.s.	n.s.	n.s.
GSE15653	Liver	n.s.	n.s.	n.s.	↑ 0.018
GSE76894	Pancreas	0.035	↑ 0.015	n.s.	0.048
GSE25724	Pancreas	n.s.	↑ 0.003	n.s.	n.s.
GSE41762	Pancreas	0.018	n.s.	n.s.	n.s.
Family history dataset					
GSE25462	Muscle	↓ 0.021	n.s.	n.s.	n.s.

Table 8: Gene set enrichment analysis results comparing the expression of the 763 T2D *cis*-genes to the genomic background (ALL). FDR-adjusted p-values ≤ 0.05 are shown for significant enrichment of increased (↑), decreased (↓) or mixed increased and decreased (no arrow). GSEA used a Wilcoxon rank-sum test and 10,000 permutations. n.s. = not significant. The family history dataset regressed gene expression in healthy individuals against the number of parents affected by T2D.

The T2D *cis*–genes were significantly enriched for differential expression in 10 out of the 14 individual datasets. Expression was decreased in two muscle, two adipose and one liver dataset, mixed in two pancreas datasets and increased in three adipose datasets. This is consistent with the T2D *cis*–genes being differentially expressed in individuals with T2D, with some genes showing increased expression and others showing decreased expression. Notably, all three datasets showing increased expression were adipose, possibly reflecting tissue-specific mechanisms. The T2D *cis*–genes were also enriched for decreased expression in the highly informative dataset of healthy, normoglycemic individuals with an increasing number of parents affected by T2D (family history dataset, Table 8), while

all three control datasets were non-significant. This dataset is expected to more accurately reflect heritable changes in gene expression with reduced confounding that may arise as a consequence of T2D onset, since the study subjects are healthy, first-degree relatives of T2D cases.

The three control gene sets showed evidence of differential expression in two, zero and two meta-analysed tissues, as well as four, two and seven of the individual datasets, respectively. When considering the proportion of individual datasets with evidence of *cis*–gene differential expression, the T2D *cis*–genes were enriched for differential expression in 11/15 or 73% of the observed datasets, which was significantly greater than the proportion of control gene sets (13/45 datasets or 29%) with a two-proportion, one-tailed z-test providing a p-value = 0.003. These results provide supporting evidence that the differential expression of the 763 *cis*-genes are plausibly driven, at least in part, by heritable mechanisms rather than as a consequence of disease onset.

(2) GSEA RESULTS: T2D *cis*-NEMGs vs the genomic background

The expression of the 50 T2D *cis*-NEMGs were significantly decreased compared to the genomic background in cases compared to controls in three of the four meta-analysed tissues, nine out of the 14 datasets and in the family history data (see Appendix A.4). However, the three control sets of 50 random *cis*–NEMGs were also consistently down-regulated across the datasets, including in the family history dataset. This is consistent with a general observation of mitochondrial dysfunction in T2D, which may result as a consequence of developing T2D or as driven by a subset of dysregulated genes. In this context, the expression of the T2D *cis*-NEMGs cannot be concluded to be different from that of randomly selected NEMGs. This further motivates the next analysis, in which the 50 T2D *cis*–NEMGs are instead compared to the background of all NEMGs.

(3) GSEA RESULTS: T2D *cis*-NEMGs vs all NEMGs

To further investigate if $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ co-location identified a subset of 50 NEMGs which are regulated by T2D risk variants (T2D-eQTL), the expression of the T2D *cis*-NEMGs were compared to the background of known NEMGs (n = 1,203 in the combined

MitoCarta2.0 and MitoCarta+). Due to the general reduced expression of NEMGs observed in T2D/IR patients compared to controls and the expected high correlation between NEMG expression, this comparison is likely to have limited power. As a result, Table 9 presents the raw p-values resulting from GSEA rather than the FDR-adjusted p-values as in Table 8.

Gene set enrichment analysis results: T2D *cis*-NEMGs vs all NEMGs

Dataset	**Tissue**	50 T2D *cis*–NEMGs vs all NEMGs	50 random *cis*–NEMGs vs all NEMGs (3 control gene sets)		
Meta-analysis	Muscle	n.s.	n.s.	n.s.	n.s.
Meta-analysis	Adipose	0.029	n.s.	0.029	n.s.
Meta-analysis	Liver	n.s.	n.s.	n.s.	n.s.
Meta-analysis	Pancreas	↓ 0.009	n.s.	n.s.	n.s.
GSE13070	Muscle	n.s.	n.s.	n.s.	n.s.
GSE25462	Muscle	↓ 0.035	n.s.	n.s.	n.s.
GSE22435	Muscle	n.s.	n.s.	n.s.	n.s.
GSE101492	Adipose	n.s.	n.s.	↓ 0.045	n.s.
GSE26637	Adipose	n.s.	n.s.	n.s.	n.s.
GSE94752	Adipose	n.s.	n.s.	↑ 0.003	n.s.
GSE20950	Adipose	n.s.	↓ 0.030	↓ 0.011	n.s.
GSE27949	Adipose	↑ 0.024	n.s.	n.s.	n.s.
GSE13070	Adipose	n.s.	n.s.	n.s.	n.s.
GSE64998	Liver	n.s.	n.s.	0.009	n.s.
GSE15653	Liver	n.s.	n.s.	n.s.	n.s.
GSE76894	Pancreas	n.s.	n.s.	n.s.	n.s.
GSE25724	Pancreas	↓ 0.016	n.s.	n.s.	n.s.
GSE41762	Pancreas	↓ 0.008	n.s.	n.s.	n.s.
Family history dataset					
GSE25462	Muscle	↓ 0.043	n.s.	n.s.	n.s.

Table 9: Gene set enrichment analysis comparing the expression of the 50 T2D *cis*-NEMGs to the background of all known NEMGs. Non-adjusted p-values $\leq$0.05 are shown, reflecting significant enrichment in the gene set for decreased (↓) or increased (↑) expression. GSEA used a Wilcoxon rank-sum test and 10,000 permutations. n.s. = non-significant. The family history dataset regressed gene expression in healthy individuals against the number of parents affected by T2D.

As shown in Table 9, the 50 T2D *cis*-NEMGs showed evidence of differential expression compared to the total NEMGs in two out of four meta-analysed tissues and four of the 14 datasets and the family history dataset. Most strikingly, the T2D *cis*-NEMGs showed evidence of significant decreased expression relative to all NEMGs in the family history

dataset which, as noted above, is expected to provide the most power to detect heritable changes in gene expression independent of disease status. None of the control gene sets were significant. One control NEMG set showed evidence of differential expression in one meta-analysis (control gene set 2 in adipose). In the individual datasets, the three control NEMG sets were enriched for differential expression in a total of 5/45 (11%) datasets, compared to 5/15 (33%) datasets for the T2D *cis*–NEMGs. A two-proportion, one-tailed z-test provides a nominal p-value = 0.055. These results tentatively support the hypothesis that $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ co-location offers the power to detect *cis*-genes involved in heritable disease risk, rather than disease onset.

GSEA RESULTS: enriched mitochondrial pathways

As described in Chapter 3, Section 3.5: **Aim three: test *cis*-genes for enrichment of mitochondrial pathways**, four mitochondrial pathways showed evidence of gene count enrichment in the total T2D *cis*–genes: valine, leucine and isoleucine degradation, biotin carboxylases, propanoate metabolism and butanoate metabolism. The differential expression of these four gene sets were also tested in the meta-analysed GEO datasets. The results are shown in Table 10. The four enriched mitochondrial gene sets showed consistent enrichment for decreased expression in T2D/IR cases compared to controls. Liver, being the smallest analysis with only 2 datasets, returned three non-significant results.

Gene Set (source)	Muscle	Adipose	Liver	Pancreas	Muscle (FH)
Valine leucine and isoleucine degradation	<2.50*e*-04	<1.67*e*-04	6.00*e*-03	2.50*e*-04	<2.50*e*-04
Biotin carboxylases	0.009	6.25*e*-04	n.s.	7.50*e*-04	0.054
Propanoate metabolism	<2.50*e*-04	<1.67*e*-04	n.s.	3.33*e*-04	<2.50*e*-04
Butanoate metabolism	0.009	<1.67*e*-04	n.s.	2.50*e*-04	1.00*e*-03

Table 10: False discovery rate (FDR)-corrected p-values for gene set decreased expression in meta-analysed T2D case-control gene expression datasets, calculated using 10,000 permutations and a Wilcoxon statistic. All gene sets were sourced from KEGG, excluding biotin carboxylases which was manually defined. The Muscle (FH) dataset contains family history information, with gene expression for healthy individuals regressed against the number of parents with T2D.

4.4 Physical vs genetic distance: differential expression

As discussed in Chapter 3, **Section 3.3.3: physical vs genetic distance: co-location using 1 LDU**, the co-location between $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ can be defined using a physical (e.g. 50 kb) or genetic distance (e.g. 1 LDU). A threshold of 1 LDU was used in this study. This approach assumes that independent $\hat{S}$ which co-locate within a tight genetic distance represent shared causal variant(s), which in this case both increase risk of T2D and regulate gene expression. The argument raised in Chapter 3 is that a physical co-location threshold ignores the 'step' and 'block' nature of LD, such that physically close $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ may still be separated by a large LDU distance (LD breakdown) and the variants will be independently inherited. Defining shared signals using a small genetic distance may therefore be expected to reduce the type I error rate (the number of false positive results). This section aims to investigate how the size of the co-location threshold influences *cis*–gene discovery, by using differential gene expression in independent T2D cases to as a proxy for the potential enrichment of true positive T2D *cis*–genes.

Adipose differential expression plotted against $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance

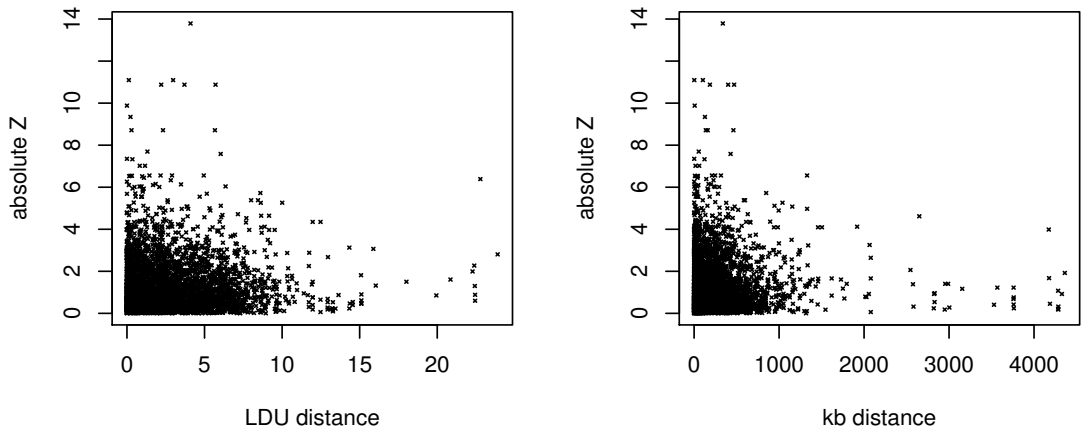


Figure 25: Adipose differential expression, shown as absolute Z-score from the adipose meta-analysis of T2D/IR cases vs control gene expression, plotted against $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance in either genetic LDU or physical kb.

The strength of the relationship between $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance and the *cis*–gene differential expression, as measured by summary Z-score, was investigated using regression. Differential gene expression was used from the adipose meta-analysis since the $\hat{S}_{eQTL}$ were mapped using subcutaneous adipose expression data. All nominally significant $\hat{S}_{eQTL}$

(p-value $\leq$0.05, stderr <900) estimated by Lau et al. for *cis*–genes within ±1.5Mb of a replicated $\hat{S}_{T2D}$ were considered. The total 265 T2D loci for which two or more $\hat{S}_{T2D}$ co-located within 100 kb, as provided by Lau et al. (2017), were included in order to maximise the data available. The distance between $\hat{S}_{eQTL}$ and the nearest $\hat{S}_{T2D}$ was measured in genetic LDU or physical kb (see Figure 25). Two models were run, either a linear regression or with distance included as a quadratic term to test for non-linearity:

$$\textbf{Model 1: } lm(|Z| \sim LDU)$$

$$\textbf{Model 2: } lm(|Z| \sim LDU + LDU^2)$$

$$\textbf{Model 3: } lm(|Z| \sim kb)$$

$$\textbf{Model 4: } lm(|Z| \sim kb + kb^2)$$

The results are shown in Tables 11 and 12. As shown in Table 11, $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance in LDU was negatively associated with the magnitude of *cis*–gene differential expression (Z-score), with Model 2 providing a significant fit (p-value = 0.016). These results provide independent validation that a smaller genetic distance between the $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ corresponds to increased *cis*–gene differential expression between T2D cases and controls. This might be interpreted as an increasing enrichment for true positive T2D *cis*–genes which are regulated by shared T2D-risk variants.

***Cis*-gene differential expression regressed on $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance (LDU)**

	Model 1 (linear)		Model 2 (quadratic)	
	β	p-value	β	p-value
T2D-eQTL distance (LDU)	-0.013	0.081	-0.044	0.004**
T2D-eQTL distance^2 (LDU^2)	-	-	0.003	0.022*
Adjusted R^2		5.82e-04		0.002
Omnibus p-value		0.081		0.016*

* p-value <0.05, ** p-value <0.01, *** p-value <0.001

Table 11: Regression of genetic distance between $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$, measured in LDU, against adipose case-control differential expression (absolute Z scores).

Regressing $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ physical distance (kb) against Z-score, shown in Table 12, showed

***Cis*-gene differential expression regressed on $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ distance (kb)**

	Model 3 (linear)		Model 4 (quadratic)	
	β	p-value	β	p-value
T2D-eQTL distance (kb)	-7.47e-05	0.166	-9.30e-05	0.385
T2D-eQTL distance^2 (kb^2)	-	-	6.93e-09	0.843
Adjusted R^2		2.60e-04		-1.29e-05
Omnibus p-value		0.166		0.376

* p-value <0.05, ** p-value <0.01

Table 12: Regression of physical distance between $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$, measured in kb, against adipose case-control differential expression (absolute Z scores).

no evidence of a significant relationship between physical co-location distance and *cis*–gene differential expression. This may result from a lower specificity with which T2D *cis*-genes are identified, since a co-location threshold based on physical distance does not utilise LD information to quantify shared association. Testing genetic co-location, however, can exclude physically close $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ which are separated by LD breakdown and therefore are likely to be independent signals.

4.5 Discussion

The aim of this chapter was to investigate if the T2D *cis*-genes and *cis*-NEMGs identified in Chapter 3 showed evidence of differential expression in T2D or IR cases compared to healthy controls. To recap, T2D *cis*-genes were identified based on association with subcutaneous adipose $\hat{S}_{eQTL}$ mapped in a population-based European cohort, which co-located within 1 LDU of independent $\hat{S}_{T2D}$. The associated *cis*-genes are hypothesised to be regulated by the same causal variant(s) which increase risk of T2D and as such, were expected to show differential expression in T2D cases compared to healthy controls.

Differential gene expression (DGE) analysis was carried out using gene expression data from 13 datasets of T2D or IR cases vs healthy controls from the GEO database (of which one contained both skeletal muscle and adipose samples and one contained family history data for the healthy controls, bringing the total to 15 separate analyses). Gene set enrichment analysis (GSEA) provided significant evidence that the 763 T2D *cis*-genes were

enriched for differential expression in T2D cases compared to controls, across multiple datasets and tissues. This is consistent with the identification of T2D *cis*–genes with multi-tissue effects, despite $\hat{S}_{eQTL}$ being mapped in subcutaneous adipose only. There was also evidence of tissue specificity, with some data sets showing an enrichment for decreased expression, while several adipose datasets showed enrichment for increased expression and two pancreas datasets showed mixed differential expression (both increased and decreased). The enrichment for differential expression was significant when compared to three control gene sets of adipose *cis*–genes with eQTL >2Mb away from a $\hat{S}_{T2D}$. This is consistent with $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ co-location being required to identify true T2D *cis*–genes which are regulated by T2D risk variants.

Both the 50 T2D *cis*–NEMGs and the NEMG control sets were enriched for decreased expression in cases across datasets, compared to the genomic background. This is consistent with the observation of general mitochondrial dysfunction in T2D, which may be either induced by disease onset or potentially driven by heritable changes in the expression of a subset of NEMGs. In order to address this confounding, a GSEA was carried out to test the expression of the 50 T2D *cis*–NEMGs compared to all known NEMGs. The T2D *cis*–NEMGs showed significant enrichment for differential expression compared to all NEMGs in both the adipose and pancreas meta-analyses and across several individual datasets. Most strikingly, the expression of the T2D *cis*–NEMGs was also decreased relative to all NEMGs in the skeletal muscle of healthy individuals with an increasing number of parents with T2D. The control sets for these data showed no evidence of differential expression. The analysis of healthy individuals with a positive family history is expected to reduce confounding due to disease onset and improve the power to detect heritable changes in gene expression. These results provide independent evidence for a primary effect of the 50 *cis*–NEMGs and supports their follow-up as strong candidate genes. The four mitochondrial gene sets which were enriched in the total T2D *cis*–genes also showed enrichment for decreased expression across datasets: valine, leucine and isoleucine degradation, biotin carboxylases, propanoate metabolism and butanoate metabolism.

The potential considerations in interpreting these results include the limited power con-

ferred from using gene expression arrays and small cohorts of potentially heterogenous individuals. The "gene-centric" analysis used here assumes that the probes measuring each gene have the same direction of effect. However, probes may different isoforms and potentially different patterns of expression. While probes could be measured individually, a combined score for each gene is required for the meta-analysis of studies which use independent expression arrays. In future analysis, RNA-seq case-control data may be used, in which individual transcripts can be annotated.

The inclusion of potentially heterogenous cohorts, for example T2D cases who may represent different T2D sub-phenotypes (Udler, 2019), may also reduce the power to detect differential gene expression. The meta-analysis of heterogeneous cohorts has previously been cautioned against, since the analysis may converge on generic signals of differential expression (Crow et al., 2019). Strict inclusion and exclusion criteria were used in this study to minimise potential heterogeneity. An additional quality control step might be to compare the genes identified in this analysis to be differentially expressed with a list of genes showing a high probability of generic differential expression, as reported by Crow et al. (2019). In order to minimise heterogeneity and enrich for individuals with a heritable predisposition to T2D, a family history dataset was included to measure gene expression in healthy individuals depending on the number of parents with T2D. Both the total T2D *cis*–genes and *cis*–NEMGs were enriched for decreased expression with an increasing number of affected parents, while all control datasets were non-significant. The observed differential expression of the control gene sets in some individual T2D/IR vs control datasets, but not in the family history dataset, is consistent with the family history dataset providing a cleaner analysis enriched for detecting heritable changes in gene expression. Furthermore, it shows that mapping T2D *cis*–genes based on $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ co-location has the power to identify changes in gene expression prior to disease onset.

A further question that remains is whether the T2D *cis*–NEMGs, for which evidence suggests are regulated by T2D risk variants, drive the observed mitochondrial dysfunction and differential expression of other NEMGs. To investigate this, networks of mitochondrial

genes could be used to determine whether the 50 *cis*–NEMGs represent core genes which are highly correlated with other NEMGs. Perturbing gene expression through *in vitro* functional studies will provide further evidence regarding the impact on overall NEMG expression.

4.5.1 Conclusions

To conclude, the analysis of independent gene expression datasets comparing T2D and insulin resistant cases and controls to healthy individuals validated the T2D *cis*–genes as being enriched for differential expression. The 50 T2D *cis*–NEMGs showed evidence of differential expression compared to the background of NEMGs, providing evidence of a primary effect for this subset of NEMGs. An informative family history dataset confirmed that $\hat{S}_{T2D}$-$\hat{S}_{eQTL}$ co-location has the power to detect T2D *cis*–genes which are differentially expressed prior to disease onset.

5 Fine-mapping a candidate locus

5.1 Introduction

5.1.1 Fine-mapping

Why fine-map?

Fine-mapping aims to identify the putative causal variant(s)[71] at a trait or disease-associated locus. Mapping the causal variant(s) is important for several reasons. Firstly, confirming the causal variant allows for the downstream biological mechanism to be studied. This is particularly important when concerning non-coding variants, since the mechanism can differ greatly depending on the tissue or context-specific activity of regulatory elements. Determining the downstream biological mechanism can potentially reveal novel therapeutic candidates and provide a greater understanding of disease aetiology, including causal mechanisms which precede disease onset. In addition, knowing the causal variant will allow direct genotyping to accurately assess the effect on disease risk, as well as its association with disease complications, prognosis and therapeutic success. Follow-up studies may investigate potential epistatic interactions with genetic background and environment and the impact of the variant across populations and subtypes of disease.

Multiple variants are associated with disease at a typical GWAS signal due to LD between the causal variant(s) and SNPs inherited on the same haplotype (see **Chapter 2, Section 2.4.2: Linkage disequilibrium and association mapping** for a more in-depth discussion on haplotype blocks and LD). Since these SNPs are inherited with the causal variant(s), they are themselves associated with disease. The resolution to determine the causal variant is therefore limited by the extent of LD around it, although there are methods discussed throughout this Chapter which can be used to address this.

[71] At any give trait or disease associated locus there may be just one causal variant. However there may also be allelic heterogeneity, which occurs where more than one allele associates independently with the trait or disease.

Is the lead SNP the causal SNP?

The lead SNP at a significant GWAS locus is defined as the SNP with the smallest p-value in a test of association. However, the lead SNP is not guaranteed or even likely to be the causal variant. Several simulation studies have confirmed that the lead SNP is unlikely to be causal and that even if directly genotyped, the causal variant may not have the smallest p-value (Zaykin and Zhivotovsky, 2005; Van de Bunt et al., 2015).

It is important to consider that a p-value can be influenced by multiple factors. As discussed in detail in Chapter 2, conventional GWAS use indirect genotyping to genotype select 'tag' or 'marker' SNPs and impute all other genotypes using reference panels. Inaccurate imputation of a causal variant may lead to an underestimated p-value. This is particularly likely to occur when haplotype frequencies differ significantly between cases and controls, since imputation is carried out for both cases and controls using generic population reference panels. Causal variants in low LD may also be more prone to inaccurate imputation since reliable inference depends on high pairwise LD with genotyped tag SNPs. Notably, this class of variants are enriched for disease heritability (Gazal et al., 2017; Wainschtein et al., 2019). Missing data, including failed genotyping, also reduces marker coverage.

GWAS arrays exclusively genotype single nucleotide polymorphisms (SNPs) and as such, the lead SNP is unlikely to be causal when the causal variant is a small insertion or deletion (indel), copy number variant (CNV), variable number of tandem repeats (VNTR) or large structural variant (SV). While genotyped SNPs may capture the association of non-SNP variants, the causal variant must be accurately imputed to achieve a highly significant p-value. Notably, the recent GWAS by Mahajan et al. (2018) carried out imputation using a reference panel from which indels are absent, while the more recent GWAS by Vujkovic et al. (2020) does not acknowledge any type of structural variants. Other scenarios in which the lead SNP is not the causal variant may include 'synthetic associations', where a common SNP achieves a significant p-value due to multiple rare causal variants arising on the same haplotype background (Dickson et al., 2010).

Methods for fine mapping

There is a plethora of approaches to fine-mapping, which are covered in detail in several excellent review papers (see Spain and Barrett (2015); Schaid et al. (2018); Lin and Musunuru (2018)) and are discussed briefly below.

Fine-mapping methods can be largely divided into computational *in silico* and functional *in vitro* or *in vivo* studies. Since any variant in LD may be driving the association of a lead SNP, it is impossible for a standard statistical test to improve the resolution beyond the extent of LD surrounding a causal variant. However, statistical methods have been developed to weight variants by systematically incorporating additional information, such as quantitative trait loci associated with gene expression (eQTL) (Zhu et al., 2016) and chromatin accessibility (caQTL) (Tehranchi et al., 2019), or functional and genomic annotation such as predicted measures of pathogenicity and chromatin annotation marks (see FGWAS, for example (Pickrell, 2014)). There are several categories of statistical fine-mapping methods (Schaid et al., 2018). Bayesian methods are most popular; these aim to identify the SNP with the highest probability of being causal. These include posterior probability and credible set[72] approaches which have been used in T2D GWAS. For example, Mahajan et al. (2018) fine-mapped T2D loci using credible sets, including those weighted for variants which overlapped pancreatic islet regulatory elements.

Another way to improve resolution is to carry out trans-ethnic fine-mapping (discussed in **Chapter 2, Section 2.4.4: Trans-ethnic GWAS and defining replication**). Assuming that the causal variant is shared, precision can potentially be improved since a smaller set of variants will be consistently associated across populations with different LD structures (Mahajan et al., 2014b; Van de Bunt et al., 2015; Asimit et al., 2016). Fine-mapping must also consider the possibility of allelic heterogeneity, where there is more than one causal variant at a disease locus. Conditional analysis is often used to identify independent signals, since including the putative causal variant as a variable in the test for association should remove the dependent association of SNPs in high LD.

[72]Credible sets are the minimum number of SNPs which explain the highest posterior probability of containing the causal SNP(s).

The number of causal variants is just one factor which influences the efficiency of different fine-mapping methods, others include the minor allele frequency and surrounding LD. For regions of LD breakdown where imputation has lower accuracy, sequence data can be used to obtain accurate genotypes. While this may be impractical for a genome-wide study, follow-up analysis can sequence replicated loci in cases and controls. A range of statistical methods can then be leveraged to help identify the causal variant(s). For non-coding regions, epigenomic annotation can help pinpoint which variants are more likely to be functional regulatory variants. These may include chromatin functional annotations, chromatin features (e.g. accessibility, histone modification enrichment, TF binding, 3D chromatin interactions, etc) as well as cross-species conservation (Cebola, 2019). Individual variants can be investigated for their potential to disrupt transcription factor binding sites and for independent evidence of association with quantitative traits, such as gene expression, methylation levels, metabolite levels or chromatin accessibility.

While *in silico* fine-mapping methods cannot determine causality, they can provide strong evidence for candidate causal variants. These can then be investigated using targeted functional studies such as enhancer reporter assays, or by altering the genotype of the putative causal variant or the expression of the predicted *cis*–gene transcript *in vitro* or *in vivo*. Several examples illustrating fine-mapping approaches in the context of T2D are described below.

Fine-mapping examples in T2D

Multiple T2D risk loci have been fine-mapped to a single causal variant. By way of example, Kycia et al. (2018) fine-mapped the intergenic rs7163757, nearby the *C2CD4A* gene, as the causal variant at the chr15q22.2 risk locus from a total of 16 strongly associated variants in high LD. Using *in silico* data, the authors demonstrated that these variants fell within a conserved pancreatic β-cell 'super-enhancer'. rs7163757 specifically fell within open chromatin and, using functional reporter assays, was shown to increase enhancer activity two-fold and cause differential binding of β-cell transcription factors. The authors identified the likely downstream mechanism by demonstrating the rs7163757

was associated with *C2CD4B* expression, itself shown to regulate inflammatory cytokines in islets.

Small et al. (2018) fine-mapped a female-specific T2D risk locus upstream of the *KLF14* gene, which encodes a transcriptional co-repressor associated with MODY. 29 risk variants in high LD were fine-mapped to identify five variants within an active adipose enhancer using ChromHMM[73] to annotate active regulatory elements. Gene expression data confirmed that the risk haplotype was an adipose-specific eQTL for *KLF14*. By investigating gene expression in response to altered *KLF14* expression, the risk haplotype was shown to reduce lipogenesis and adipocyte maturation, associating with fewer but larger adipocytes and favouring abdominal over gluteal adipose. Increased deposits of visceral, including abdominal adipose, are associated with insulin resistance and increased risk of T2D (Bjørndal et al., 2011; Direk et al., 2013).

Fine-mapping of the strongly associated *TCF7L2* region has involved targeted sequencing and conditional analysis to demonstrate that the association with disease status depended on the genotype at a single SNP: rs7903146 (Palmer et al., 2011; Maller et al., 2012). Gaulton et al. (2010) mapped rs7903146 to open chromatin in pancreatic islets, demonstrating that the SNP altered the chromatin state. rs7903146 is associated with *TCF7L2* expression in pancreatic islets (Viñuela et al., 2019) and deletion of the rs7903146 enhancer alters *TCF7L2* expression (Miguel-Escalada et al., 2019). Other examples include Gaulton et al. (2015) identifying rs10830963 as driving the T2D association at the *MTNR1B* gene, encoding the melatonin receptor 1B and Yan et al. (2016) reporting rs508419 as the causal variant at the *ANK1* locus.

Fine-mapping and genetic LDU maps

An example of improving fine-mapping resolution using LDU-based gene mapping can be seen in Direk et al. (2014). LDU-based gene mapping was used to obtain a precise location estimate for a causal variant at the *PARL/ABCC5* disease locus, which covers a

[73]ChromHMM tracks annotate likely chromatin states using histone modifications as input to a multivariate Hidden Markov Model (Ernst and Kellis, 2012).

~250 kb region of high LD. The locus had been previously implicated in T2D, but with contradictory evidence pointing to a strong biological candidate, *PARL*, as the potential candidate gene. *PARL* encodes the mitochondrial intramembrane cleaving protease, PARL. Using genetic LDU maps, Direk et al. (2014) reported an identical location estimate within the nearby *ABCC5* intron 26 associated with T2D, fasting insulin:glucose serum ratio levels, increased visceral fat deposits and *ABCC5* expression levels. Increased expression of *ABCC5*, which encodes the ATP binding cassette subfamily C member 5, was independently associated with an increased risk of T2D. Whereas the region had previously been implicated by linkage studies, but had failed to replicate in any subsequent single-SNP GWAS, the Direk et al. study achieved precise fine-mapping by integrating genetic maps with genotype data. Crucially, a follow-up knock-out of *ABCC5* in mice confirmed a diabetes-related phenotype of improved insulin sensitivity, decreased fat mass and increased levels of the incretin hormone GLP-1 (Cyranka et al., 2019).

5.1.2 *ACAD11*, fatty acid oxidation and diabetes

Chapter 5 aims to fine-map one candidate T2D locus from this study, for which targeted next generation sequencing data was available for an independent cohort of T2D cases and controls (described in the next section). The candidate locus was selected based on the co-location of $\hat{S}_{T2D}$ with a $\hat{S}_{eQTL}$ for the *cis*–NEMG *ACAD11*. ACAD11 or Acyl-CoA Dehydrogenase Family Member 11 is an acyl dehydrogenase involved in fatty acid β-oxidation, which catalyses the production of acetyl-CoA from long-chain fatty acids (LCFAs) and very long-chain fatty acids (VLCFAs) with carbon chains between 20 and 26 carbons long (He et al., 2011). β-oxidation is a pathway which has been repeatedly implicated in T2D, making *ACAD11* a particularly strong candidate gene. β-oxidation, including the link with T2D, is described below.

Fatty acid β-oxidation (FAO) and ACAD11

The breakdown of fats, glucose and amino acids comprise the three major energy sources, and mitochondrial fatty acid β-oxidation (FAO) is the primary pathway for the breakdown of fatty acids (FAs) (Houten and Wanders, 2010). FAO is of particular importance in

fasting conditions where glucose is limited. In the liver, FAs are converted to ketone bodies which are utilised by all tissues as an additional energy source, but particularly by the brain which cannot directly utilise fats for energy.

Free fatty acids (FFAs) formed by the break down of triglycerides (lipolysis) or the synthesis from acetyl-CoA (*de novo* lipogenesis) are degraded by mitochondrial or peroxisomal β-oxidation. Insulin stimulates lipogenesis and inhibits lipolysis. FFAs are esterified or 'activated' by an acyl-CoA synthetase (ACS) enzyme prior to their transport across the mitochondrial membrane via the carnitine shuttle, whereby acyl-CoAs are converted into acylcarnitines by carnitine palmitoyltransferase I (CPT1) on the mitochondrial outer membrane and then enter the mitochondria where they are converted back to acyl-CoAs by CPT2. β-oxidation is a cyclic pathway in which fatty acyl-CoAs are shortened by two carbons per cycle to form an acetyl-CoA molecule and two electron carriers. Acetyl-CoA enters the citric acid or 'TCA' cycle and the electrons are transported to the electron transport chain (ETC). Each β-oxidation cycle consists of four reactions, of which the first is catalysed by an acyl-coenzyme A dehydrogenase (ACAD) enzyme (Houten and Wanders, 2010), such as ACAD11. In the final cycle, even-chain fatty acids produce two acetyl-CoA molecules and odd-chain fatty acids produce one acetyl-CoA (2 carbons, 2C) and one propionyl-CoA (3C). Priopionyl-CoA is converted to methylmalonyl-CoA via the biotin-dependent propionyl-CoA carboxylase, which is then converted to succinyl-CoA and enters the TCA cycle. The α chain of propionyl-CoA carboxylase is encoded by the *PCCA* gene, which was also detected as a T2D *cis*–gene in this study.

Some fatty acyl-CoAs are too long to enter the mitochondria and are instead oxidised in the peroxisome; this may be the case for fatty acyl chains $\geq$22C. Peroxisomal β-oxidation (peroxidation) is oxidative and electrons are transferred to oxygen, forming hydrogen peroxide (H_2O_2). It has been debated whether ACAD11 is localised to the mitochondria or the peroxisome. ACAD11 was first identified in peroxisomes from rat liver (Kikuchi et al., 2004; Islinger et al., 2007; Wiese et al., 2007). Despite containing a peroxisomal targeting signal, He et al. (2011) found that the majority of ACAD11 localised to plasma membrane-associated vesicles and the mitochondria. ACAD11 is

present in the MitoCarta2.0 database of mitochondrial proteins (Calvo et al., 2016)[74]. Camões et al. (2015) recently presented evidence that ACAD11 is exclusively localised to the peroxisome.

Fatty acid β-oxidation and T2D

Fatty acid β-oxidation has been widely implicated in diabetes, insulin resistance and insulin deficiency (IS Sobczak et al., 2019). Perturbed lipid metabolism may be one of the earliest features of T2D (Bell et al., 2020). Observational studies have routinely reported the association of circulating and dietary saturated fatty acids with T2D. Circulating levels of medium odd-chain FAs are negatively associated with risk of T2D, while even-chain FAs are positively associated (Forouhi et al., 2014; Lu et al., 2016; Imamura et al., 2018; Huang et al., 2019) and that this relationship may be stronger for women (Imamura et al., 2018). Furthermore, higher levels of acylcarnitines[75] in T2D are indicative of incomplete β-oxidation, since fatty acyl-CoAs can be transported out of the mitochondria as acylcarnitines and then enter the blood stream (Koves et al., 2008). Increased levels of acylcarnitines have been observed across T2D and prediabetic states (Wang-Sattler et al., 2012; Mai et al., 2013; Ha et al., 2012; Adams et al., 2009; Mihalik et al., 2010; Floegel et al., 2013; Lin et al., 2018b; Möder et al., 2003), as well as in women with gestational diabetes (Batchuluun et al., 2018).

Circulating levels of short-chain FAs derived from gut microbes have also been associated with T2D (Sanna et al., 2019; Müller et al., 2019). At the other end of the spectrum, circulating levels of very-long chain fatty acids (VLCFAs), including those which are substrates for *ACAD11* (C20: arachidic acid, C22: behenic acid, C23: tricosanoic acid, C24: lignoceric acid) are inversely associated with risk of T2D, such that lower levels are associated with a higher risk of disease (Forouhi et al., 2014; Lemaitre et al., 2015; Fretts et al., 2019; Ardisson Korat et al., 2020). Changes in lipid levels and the adipocyte

[74]MitoCarta2.0 reported ACAD11 as having an exclusively mitochondrial protein domain, increased in expression in models of mitochondrial proliferation and co-expression with known nuclear-encoded mitochondrial genes (NEMGs).

[75]Odd-chain acylcarnitines (C3: propionylcarnitine and C5: isovalerylcarnitine) are produced during amino acid catabolism. Butyrylcarnitine (C4) results from the degragation of both FAs and amino acids as does acetylcarnitine (2C), which is produced from acetyl-CoA

fatty acid binding protein (A-FABP) have also been observed in the first-degree relatives of T2D patients (Jacob et al., 1999; Axelsen et al., 1999; Hu et al., 2016). Providing further evidence that altered fat levels may causally contribute to T2D are a multitude of functional studies. Intake of fatty acids in humans was shown to cause insulin resistance (Boden et al., 1991; Homko et al., 2003) and the chronic exposure of pancreatic β-cells to FAs caused impaired glucose-stimulated insulin release *in vitro* and in *in vivo* rodent models (Zhou and Grill, 1994, 1995; Roomp et al., 2017). High-fat diets cause insulin resistance in mice (Liu et al., 2015b; Kothari et al., 2017; Avtanski et al., 2019; Lang et al., 2019) and are often used to model T2D (Heydemann, 2016).

Although it was previously thought that increased β-oxidation may cause hyperglycemia by increasing the utilisation of fats relative to glucose (Randle, 1963), it is now more widely accepted that reduced or incomplete β-oxidation is characteristic of T2D. For example, the *in vitro* capacity of skeletal muscle myocytes for FAO was shown to reflect *in vivo* insulin sensitivity (Ukropcova et al., 2005). Higher expression of *ACC2*, which is known to inhibit β-oxidation, is observed in T2D patients (Debard et al., 2004). Directly inhibiting β-oxidation caused the T cells from lean subjects to become more similar to T cells from T2D patients, with increased use of 16C-fatty acylcarnitine in the production of Th17 inflammatory cytokines (Nicholas et al., 2019). Lee et al. (2017) reported that the loss of hepatic β-oxidation in mice improved glucose tolerance, however Lundsgaard et al. (2020) demonstrated that the long-term inhibition of β-oxidation caused increased glucose production, increased circulating fatty acids, hepatic steatosis, reduced insulin sensitivity and glucose intolerance. Alternatively, it has been suggested that an increased rate of β-oxidation may cause incomplete oxidation if the rate is greater than that of the TCA cycle (Koves et al., 2008). The β-oxidation end-product, acetyl-CoA, can be transported out of the mitochondria as acetylcarnitine (Schroeder et al., 2012) or converted to ketone bodies in the liver, consistent with the raised levels of acetylcarnitine and the ketone body β-hydroxybutyrate observed in T2D (Mahendran et al., 2013).

Several mechanisms have been proposed to impact insulin sensitivity and glucose-stimulated insulin release as a consequence of inefficient β-oxidation (reviewed by Park and Seo

(2020)). Increased intracellular levels of fatty acyl-CoAs can activate signalling cascades and inhibit insulin receptor signalling through the phosphorylation of insulin receptor substrate-1 (IRS-1) (Lowell and Shulman, 2005). Intracellular FAs and fatty acyl-CoAs can cause widespread alterations to signalling networks, epigenetic modifications and metabolic processes and may disrupt the translocation of the glucose receptor GLUT4 to the cellular membrane (Dittmann et al., 2019; Gonzalez-Becerra et al., 2019; Park and Seo, 2020). Raised levels of saturated FAs may also increase endogenous glucose production, reduce the potency of incretin hormones to stimulate insulin secretion (Astiarraga et al., 2018) and cause lipotoxicity and death of pancreatic β-cells (Boden and Shulman, 2002; Shimabukuro et al., 1998; Yang et al., 2016; Acosta-Montaño and García-González, 2018); lipotoxicity in β-cells is reviewed by Oh et al. (2018); Acosta-Montaño and García-González (2018); Ye et al. (2019); Lytrivi et al. (2020).

Increased levels of medium-chain acylcarnitines may impair insulin secretion and induce NF-κB mediated inflammation (Lee et al., 2001, 2003, 2004; Weatherill et al., 2005; Zhao et al., 2007; Adams et al., 2009; Batchuluun et al., 2018). Chronic NF-κB activation may cause lipotoxicity-mediated death of pancreatic β-cells and immune-mediated diabetes (Salem et al., 2014; Bagnati et al., 2016; Chen et al., 2018a). Acylcarnitines may also contribute to diabetes pathology by influencing membrane permeability and intracellular calcium (Ca^{2+}) levels (McCoin et al., 2015). Low levels of fat oxidation have also been linked to the conversion of LCFAs to the intermediates diacylgylcerol (DAG) and ceramides, which have independently been implicated in insulin resistance (Erion and Shulman, 2010; Jornayvaz and Shulman, 2012; Perry et al., 2014; Jornayvaz and Shulman, 2012; Sokolowska and Błachnio-Zabielska, 2019).

The peroxisome, in which ACAD11 may potentially be located, has itself been associated with T2D. Elsner et al. (2011) reported that H_2O_2 produced by the peroxisome, but not the mitochondria, was responsible for lipotoxicity induced by long-chain FAs in pancreatic β-cells. ROS from the peroxisome are important regulators of adipogenesis and adipocyte metabolism (Liu et al., 2019). For example, peroxisomes initiate lipogenesis and metabolise poly-unsaturated fatty acids, which are the natural ligands for PPARγ,

a well-known regulator of fat and adipose metabolism. Increased lipid peroxidation has been observed in T2D patients (Colas et al., 2011).

In addition to the above, multiple genetic studies have implicated genetic perturbations of fat metabolism and β-oxidation in T2D. Genetic variants in the gene encoding long-chain acyl-CoA synthetase 1 (*ACSL1*), which activates and channels fats towards β-oxidation, have been associated with T2D and fasting glucose (Manichaikul et al., 2016). Knock-down of *ACSL1* reduced β-oxidation and increased glucose utilisation in mice (Ellis et al., 2011); this is consistent with the short term-effects of inhibiting β-oxidation observed by Lundsgaard et al. (2020). Genetic variants in the fatty acid desaturases *FADS1/2* were observed to drive the association of seven fatty acids with T2D (Yuan and Larsson, 2020). T2D also associates with variants in the glucokinase regulatory protein *GCKR* (Ling et al., 2011; Stančáková et al., 2012; Shen et al., 2013; Simons et al., 2016) and with missense mutations in the adipokine adiponectin *ADIPOQ* which regulates both glucose and lipid metabolism (Stumvoll et al., 2002; Hivert et al., 2008; Gao et al., 2013; Tao et al., 2014). Common and rare variants in *PPARG*, which encodes the nuclear receptor PPARγ, have also been associated with risk of T2D (Vergotine et al., 2014; Majithia et al., 2014). PPARγ is highly expressed in adipose tissue and regulates whole-body lipid metabolism (Ahmadian et al., 2013). PPARγ agonists including thiazolidinedione (TZD), a known diabetes treatment, improve insulin sensitivity (Tontonoz and Spiegelman, 2008), while genetic variants which disrupt PPARγ binding may modulate response to TZD (Soccio et al., 2015).

Several of the T2D *cis*–genes identified in the current study are involved in fat metabolism. In addition to *ACAD11*, other *cis*–genes include *ACADS*, the short-chain acyl-CoA dehydrogenase (SCAD) gene and *ACSS1*, which encodes acetyl-CoA synthetase 1 and *PPARG*, which regulates the uptake of FAs, glucose homeostasis and inflammation in adipose and skeletal muscle. Also included are *PISD*, which encodes phosphatidylserine decarboxylase, a protein involved in phospholipid metabolism, and *LACTB*, which regulates *PISD* (Keckesova et al., 2017).

ACAD11 AND T2D

ACAD11 is likely to confer risk of T2D through its role as an acyl-CoA dehydrogenase involved in the β-oxidation of long and very-long chain FAs. ACAD11 has a particular affinity for saturated C22-CoA (He et al., 2011) and is highly expressed in the brain, particularly white matter, as well as the liver, heart and kidney (He et al., 2011). According to GTEx, ACAD11 has near ubiquitous expression with this being the highest in the liver, uterus and ovaries. ACAD11 has been reported to contribute to variation in serum metabolite levels (Hong et al., 2013), residual food intake in cattle (Karisa et al., 2013) and to be necessary for cell survival during glucose starvation as part of an evolutionary conserved targeting by the cell survival protein p53 (Jiang et al., 2015). p53 has itself been implicated in diabetes-related phenotypes (Minamino et al., 2009; Kung and Murphy, 2016; Strycharz et al., 2017; Itahana and Itahana, 2018). Both p53 and ACAD11 may play a role in adipogenesis, with ACAD11 implicated in the formation of beige adipocytes in mice (Liang et al., 2019).

Of interest are several studies reporting a link between *ACAD10*, a paralog of *ACAD11*, and T2D. Genetic variants in *ACAD10* have been previously associated with T2D in Pima Indians (Hanson et al., 2007; Bian et al., 2010), showing nominal associations with insulin resistance, lower lipid oxidation rate and larger subcutaneous abdominal adipocyte size. *ACAD10* knock out in mice caused impaired glucose tolerance and hyperinsulinemia (Bloom et al., 2018).

5.1.3 Aims and hypothesis

The following analysis aims to fine-map and prioritise putative causal variants at the *ACAD11/NPHP3* chr3q22.1 locus, at which T2D association was detected in two independent European cohorts by Lau et al. (2017) using LDU-based gene mapping. Previously generated targeted sequence data for a carefully selected cohort of 94 T2D cases with a positive family history of T2D and 94 healthy controls with no family history, will be used to carry out fine-mapping. The chr3q22.1 locus has previously been associated with T2D-end-stage renal disease (Guan et al., 2016), but has not been identified by any

T2D GWAS. As such, it might be hypothesised that the causal variant(s) at this locus represent a genetic architecture which is challenging to map using single-SNP GWAS, instead requiring the more powerful LDU-based gene mapping. This may include low levels of LD which prevent accurate imputation, low-frequency or rare variants, non-SNP variation such as indels or structural variants, or multiple causal variants on different haplotypes (alleleic heterogeneity). As such, the methods described in the following sections will test for general evidence of association, as well as low-frequency, rare and non-SNP variation.

5.2 Methods

5.2.1 Targeted sequencing: the data

Targeted sequencing of 104 T2D loci was previously carried out by Lau et al. for a French cohort of 94 T2D cases and 94 controls. As described in Lau et al. (2017), T2D cases and controls were 1:1 matched for age, BMI, and sex. Cases with a family history of T2D were selected from a cohort previously recruited for a T2D linkage analysis (Vionnet et al., 2000) and controls were selected from families without a family history of T2D previously recruited for an obesity study (Meyre et al., 2004). The case and control characteristics are shown below in Table 13, adapted from Supplementary Table S2 in Lau et al. (2017).

Targeted sequencing was carried out using the Agilent SureSelectXT2 capture kit for 100ng of DNA. The raw data, sequenced on an Illumina HiSeq 2500 as 150bp paired-end reads, was re-analysed as part of this project to call variants using current gold-standard tools according to the Genome Analysis Toolkit (gatk) (McKenna et al., 2010). Pre-processing and variant calling was carried out for the entire dataset prior to variant and sample filtering based on the subsetted chr3q22.1 sequenced region, covering ∼220 kb. The analysis steps are described below.

	Gender	#	Age	Bmi
Cases	Male	49	43.5±7.5	25.9±3.5
	Female	57	47.4±7.0	27.1±4.6
Controls	Male	49	40.7±7.1	27.0±3.6
	Female	57	47.8±7.3	27.7±4.0

Table 13: Case-control cohort used for targeted NGS sequencing. The number of samples (#) is presented along with the mean ± standard deviation of age and BMI. A total of 94 cases and 94 controls were included in the following analysis.

5.2.2 Pre-processing, variant calling and filtering

Pre-processing

FastQC quality reports were generated for all samples. Adapter sequences were trimmed using fastp (Chen et al., 2018b) and trimmed reads were aligned to the most recent reference genome (hg38) using bwa mem with default parameters (Li, 2013) (coordinates later converted to hg19 for consistency). Duplicate reads were marked using picard MarkDuplicates and base quality score recalibration (BQSR) from the gatk toolkit was carried out in order to correct for systematic technical errors in quality scores. For the structural variant (SV) calling, samblaster was used to exclude duplicate reads and extract splitter and discordant reads as recommended in the lumpy SV-calling pipeline (Layer et al., 2014). Briefly, splitter reads align across a SV breakpoint, resulting in separate parts of the read aligning to distinct locations in the reference genome. Discordant reads occur when paired reads do not align within the expected distance and orientation.

Variant calling

For SNPs and indels, the gatk HaplotypeCaller (Poplin et al., 2017) was used to call variants from pre-processed reads. HaplotypeCaller was run in GVCF mode to generate intermediate variant call files (gVCF) per individual. Intermediate files were consolidated using GenotypeGVCFs to generate one joint VCF file. Lumpyexpress was used to call structural variants (Layer et al., 2014).

Variant and sample filtering

The joint SNPs and indels vcf file was imported into R and assessed using the *vcfR*

package (version 1.10.0) (Knaus and Grünwald, 2016). The data was subsetted for the candidate locus (chr3q221). 12 samples, 10 cases and two controls, were removed due to failed sequencing (>50% missing data). Guidelines from Lin et al. (2014) regarding how to filter targeted sequencing data were adapted for the following steps. Samples for which >65% of the sites had a depth of <15× were removed, giving 82 cases and 89 controls. Multi-allelic variants were removed. Each genotype was required to have a read depth >10×, a minimum of two reads supporting the alternative allele and a genotype quality phred-score of >20. Following these steps, variants with >20% missing data were removed as per Lin et al. (2014), resulting in the removal of 209 variants with 549 remaining. The number of singletons (variants observed in only one individual) per sample was assessed across the 220 kb region. The mean was 1.48 singletons per sample, with a standard deviation of 3.59. One outlier (a case sample) was removed for having 45 singletons. A total of 482 variants over 170 samples remained after excluding monomorphic variants. Variants were tested for Hardy-Weinberg equilibrium, although all p-values were $>10^{-5}$. The mean coverage across the samples was 50×, with a standard deviation of 11.8 and a minimum of 33×.

For indels, individual genotypes with a depth of <10, genotype quality <20, or less than two alleles supporting the alternative allele at a heterozygous genotype were removed. Variants with an overall >20% missingness were removed. Multi-allelic indels were corrected to 0 for the reference allele and 1 for a non-reference allele. Two variants not in HWE were removed. 75 total indels across the samples were included in the following test of association. For structural variants, quality scores were assigned to the SVs called by lumpyexpress using SVtyper (Chiang et al., 2015). SVs were required to have a minimum phred quality score of 20.

5.2.3 Variant annotation and association

Annotation and predicting pathogenicity

SNPs and indels were annotated using ANNOVAR (Wang et al., 2010b) with gene-based annotation from refGene (hg38) and minor allele frequencies from ExAC (ExAC 65000

exome allele frequency data) and the 1000 Genome Project. Pathogenicity scores were added for coding variants from SIFT and PolyPhen-2. Briefly, SIFT assesses conservation across highly related sequences (Vaser et al., 2016) and PolyPhen-2 HDIV uses a naive Bayes classifier to incorporate sequence conservation in closely related mammalian species (Adzhubei et al., 2010). To estimate pathogenicity for all variants including non-coding variants, CADD (Combined Annotation Dependent Depletion) scores were calculated (Rentzsch et al., 2019). CADD models multiple genomic features, including genetic context, conservation, epigenetic modifications and functional predictions, with scores calculated using a machine-learning algorithm trained on *proxy-neutral* fixed variants which have become fixed following the human-ape split and are presumed to be benign (Rentzsch et al., 2019). CADD scores are scaled for the entire genome, such that a CADD score of 20 indicates the variant is in the top 1% of estimated deleterious variants in the genome.

Case-control associations

Variants were tested for disease association using the R package *SNPassoc* (version 1.9-2) (Gonzlez et al., 2014). *SNPassoc* fitted a logistic regression model for each SNP and case-control status, coded as case = 1 and control = 0, based on a co-dominant (additive) model of inheritance.

Annotate regulatory elements

The Roadmap Epigenomics Project was used as the primary source of regulatory annotations. A total of six datasets were investigated, including five primary tissues (adipose nuclei: E063, skeletal muscle, female: E108, adult liver: E066, pancreatic islets: E087, pancreas: E098) and the liver HepG2 carcinoma cell line (E118). ChiP-seq data was downloaded for histone modifications characteristic of enhancers and promoters (H3K27ac and H3K9ac) and promoters (H3M4me1). Detailed information regarding the six datasets is available via ROADMAP.

To define regulatory element boundaries and test for overlapping variants, ChromHMM tracks were downloaded from ROADMAP. ChromHMM tracks annotate likely chromatin

states using histone modifications as input to a multivariate Hidden Markov Model (Ernst and Kellis, 2012). The ROADMAP ChromHMM tracks annotate 15 states based on the integration of five chromatin marks: H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3. The 15 states include: (1) Active Transcription Start Site (TSS), (2) Flanking Active TSS, (3) Transcription at gene 5' and 3', (4) Strong transcription, (5) Weak transcription, (6) Genic enhancers, (7) Enhancers (Enh), (8) ZNF genes & repeats, (9) Heterochromatin, (10) Bivalent/Poised TSS, (11) Flanking Bivalent TSS/Enh, (12) Bivalent Enhancer, (13) Repressed PolyComb, (14) Weak Repressed PolyComb and (15) Quiescent/Low.

ChromHMM tracks were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/ChromHMMSegmentations/ChmmModels/coreMarks/jointModel/final/.

Transcription factor binding sites

The R package *motifbreakR* was used to test select variants for evidence of disrupting transcription factor binding sites (TFBS). TFBS were obtained from the Homo Sapiens Comprehensive Model Collection (HOCOMOCO) which contains manually curated sequence motifs represented by position weight matrices for 401 human transcription factors (Kulakovskiy et al., 2013). *MotifbreakR* was used with the recommended settings unless otherwise stated, with a threshold p-value of $1e$-4.

Linkage disequilibrium

The LD between candidate variants were queried from the 1000 Genomes Project using the LDpop Tool from the National Cancer Institute: https://ldlink.nci.nih.gov.

5.3 Results

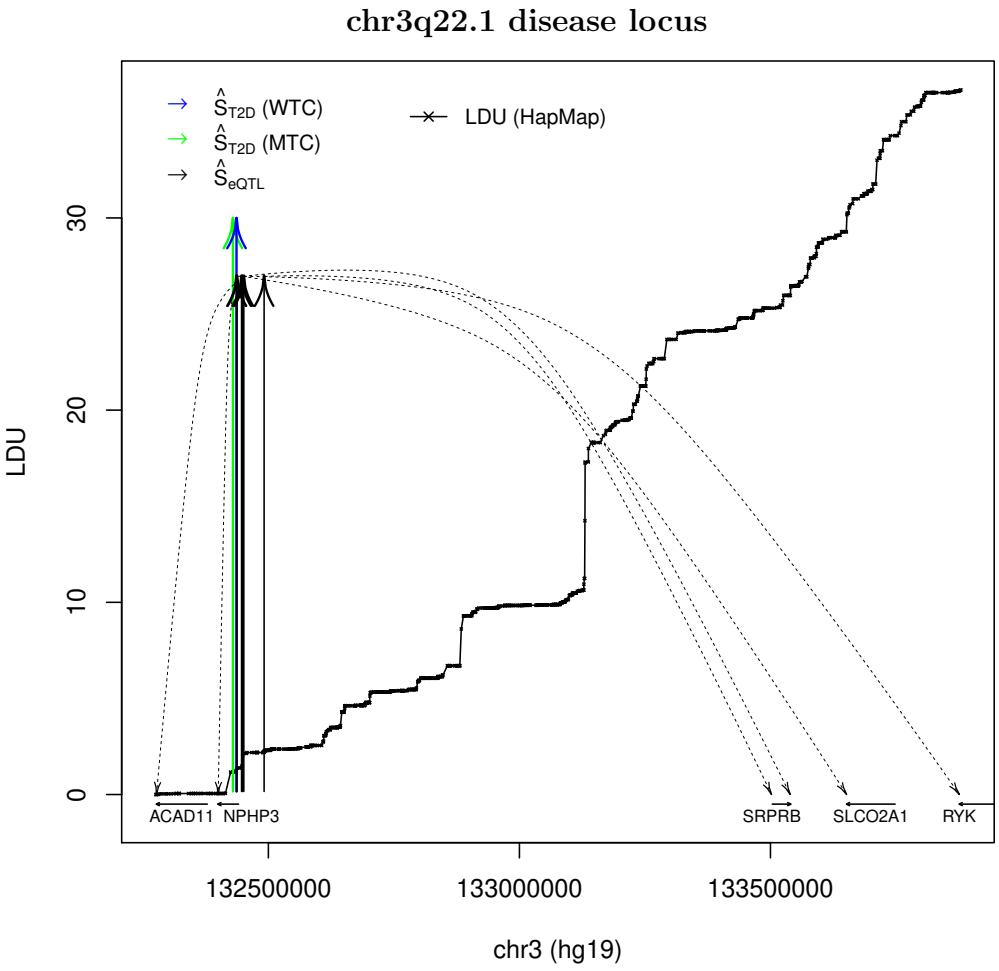


Figure 26: Plot showing the chr3q22.1 disease locus. $\hat{S}_{T2D}$ from the WTC and MTC European datasets are shown, along with $\hat{S}_{eQTL}$ within 1 LDU for the *cis*–genes *ACAD11*, *NPHP3*, *SRPRB*, *SLCO2A1* and *RYK*, with dotted lines connecting the $\hat{S}_{eQTL}$ estimates with the *cis*–genes. Other genes are not shown.

The chr3q22.1 T2D locus (shown in Chapter 2, see Figures 7, 9 and 11) was initially selected for follow-up fine-mapping due to the *cis*–NEMG *ACAD11* (Acyl-CoA Dehydrogenase Family Member 11) which is involved in the highly relevant pathway of fatty acid β-oxidation. This locus has been previously associated with T2D-associated end-stage kidney disease in African Americans, for which the lead SNPs are shown in Figure 27 (Guan et al., 2016).

$\hat{S}_{T2D}$ location estimates were generated by running the adapted Malecot model as described in Chapter 2, **Section 2.7: Association mapping using LDU maps** on the genomic coordinates chr3:132065563-132884215 (hg19). This analytical window ex-

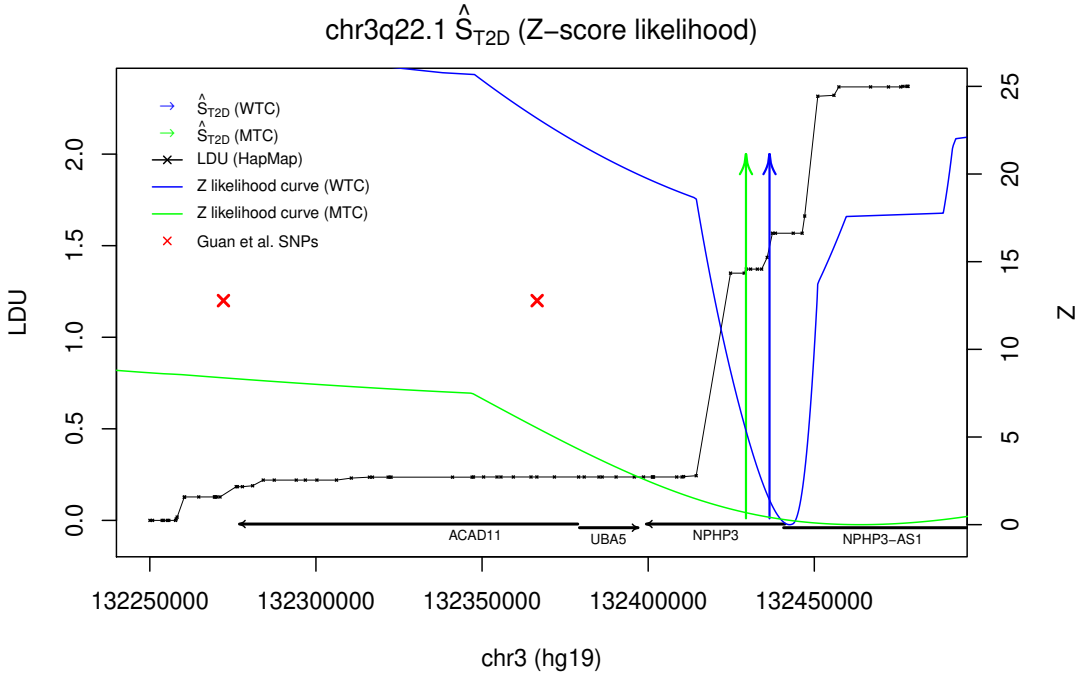


Figure 27: The chr3q22.1 disease locus with European T2D estimates ($\hat{S}_{T2D}$) from the WTC and MTC cohorts, showing Z-score likelihood curves and the HapMap LDU map (European). The minimum Z values reflect the estimated location of the causal variant, with $\hat{S}_{T2D}$ point estimates (shown as arrows) interpolated onto the high resolution LDU map. The locations of two SNPs previously reported to be associated with end-stage kidney disease in T2D are shown in red (Guan et al., 2016).

tends a total of 819 kb (10.5 LDU). The two European datasets gave significant $\hat{S}_{T2D}$ at chr3:132436519 (WTC), p-value = 4.83e-03 and chr3:132429438 (MTC), p-value = 1.03e-09. The output of the model included Z-score likelihood curves, with the most likely estimate of the causal variant located at the minimum Z. These curves are plotted in Figure 27. The $\hat{S}_{T2D}$ estimates were interpolated from the Z-curves, which depend on the resolution of the array used for the case-control genotyping[76], onto the higher resolution LDU map. The WTC and MTC $\hat{S}_{T2D}$ estimates are located 0.086 LDU apart or 7.08 kb when converted to physical coordinates. It is worth noting that many GWAS papers define replication where lead SNPs are within ±500 kb (Vujkovic et al., 2020); the distances used here are comparatively very small. Notably, if the $\hat{S}_{T2D}$ occurs in an LDU block, the $\hat{S}_{T2D}$ effectively corresponds to the entire block since the LDU distance is zero. A point estimate of the $\hat{S}_{T2D}$ is typically shown for simplicity. 1 LDU extends

[76]The metabochip included only 21 SNPs across this 819 kb analytical window.

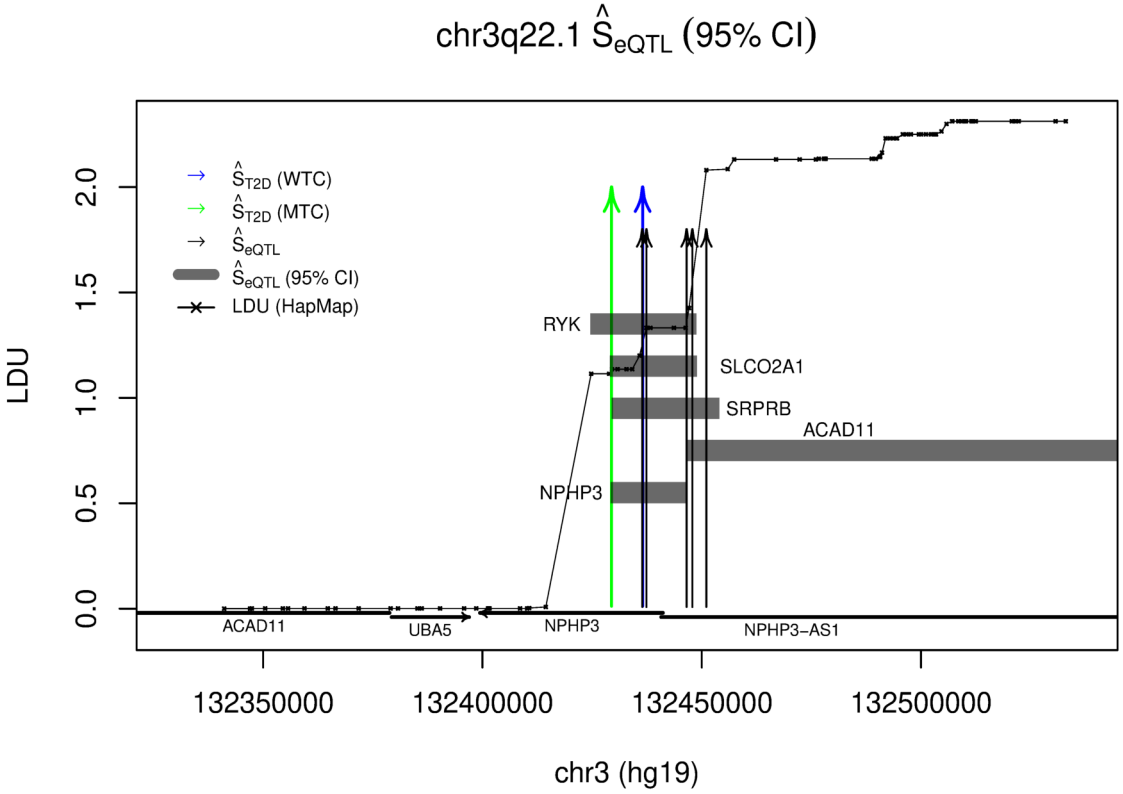


Figure 28: The chr3q22.1 disease locus showing the WTC (blue) and MTC (green) $\hat{S}_{T2D}$ estimates ±100 kb. The $\hat{S}_{eQTL}$ estimates for the five significant *cis*-genes are shown with grey blocks extending the 95% confidence intervals (CI).

21.38 kb downstream and 58.70 kb upstream of the WTC $\hat{S}_{T2D}$ and 19.70 kb downstream and 59.22 kb upstream of the MTC $\hat{S}_{T2D}$.

Five *cis*–genes were identified at this locus with an $\hat{S}_{eQTL}$ estimate within 1 LDU of the WTC or MTC $\hat{S}_{T2D}$. These are shown in Figure 28, with the 95% confidence intervals for each $\hat{S}_{eQTL}$ shown as grey blocks. The locations of all five *cis*–genes can be seen in Figure 26, of which *RYK* is the furthest at ∼1.4Mb upstream of the $\hat{S}_{T2D}$. The distances of each $\hat{S}_{eQTL}$ from the $\hat{S}_{T2D}$ are shown in Table 14 along with the individual $\hat{S}_{eQTL}$ p-values.

As seen in Table 14, the $\hat{S}_{eQTL}$ are all located at or less than 0.23 LDU (14.52 kb) from the nearest $\hat{S}_{T2D}$. Of these, *NPHP3* has the closet $\hat{S}_{eQTL}$ with a distance of 0 LDU (0.09 kb) from the WTC $\hat{S}_{T2D}$. The 95% confidence interval for the *ACAD11* $\hat{S}_{eQTL}$ does not overlap the $\hat{S}_{T2D}$ point estimate (note the confidence intervals for the $\hat{S}_{T2D}$ are not provided so these may still overlap). However, GTEx reported significant eQTL for

chr3q22.1 *cis*-genes

Cis-gene	$\hat{S}_{eQTL}$ (hg19)	eQTL p-value	LDU distance	kb distance
ACAD11	chr3:132451038	1.62e-09	0.23 LDU	14.52 kb
NPHP3	chr3:132436437	5.66e-25	0.00 LDU	0.09 kb
SRPRB	chr3:132447854	1.98e-03	0.23 LDU	11.34 kb
SLCO2A1	chr3:132446548	1.13e-06	0.13 LDU	10.03 kb
RYK	chr3:132437389	5.56e-04	0.13 LDU	0.87 kb

Table 14: *Cis*-genes of the chr3q22.1 disease locus. LDU and kb distance from the $\hat{S}_{eQTL}$ to the nearest $\hat{S}_{T2D}$.

ACAD11 across the region, with the closest eSNP located between the two $\hat{S}_{T2D}$, 225bp downstream of the WTC $\hat{S}_{T2D}$. A limitation of the LDU-based Malecot model is the assumption that there is one causal variant within the analytical window. Hence only one location estimate is given, even if there are multiple eQTL. The benefit of fine-mapping is that candidate SNPs can be identified and further studied to confirm if they associate with both disease status and gene expression levels.

Chromatin interactions (pancreatic islets)

Human islet chromatin interactions in the locus were investigated using the Capture HiC Plotter (https://www.chicp.org/) (Schofield et al., 2016). Capture HiC data from Miguel-Escalada et al. (2019) was available for pancreatic islets and is plotted for the chr3q22.1 locus in Figure 29. Significant interactions are shown with coloured lines and the $\hat{S}_{T2D}$ region can be seen to have extensive interactions, with most interactions for this locus contacting the $\hat{S}_{T2D}$ region. While the $\hat{S}_{eQTL}$ data for this project is for adipose, the interaction plot highlights potential other candidate genes for which promoters interact with the $\hat{S}_{T2D}$ locus in islets, including *TMEM108*, *BFSP2*, *CDV3*, *TOPBP1* and *TFP1*. Interestingly, *TMEM108* was detected as a T2D *cis*–gene in this study for a different $\hat{S}_{T2D}$ locus ∼830 kb upstream. To conclude, the HiC plot confirms that the $\hat{S}_{T2D}$ locus is highly interactive and may be implicated in the dysregulation of multiple genes. To further investigate tissue-specific effects on gene expression, $\hat{S}_{eQTL}$ for other tissues should be generated.

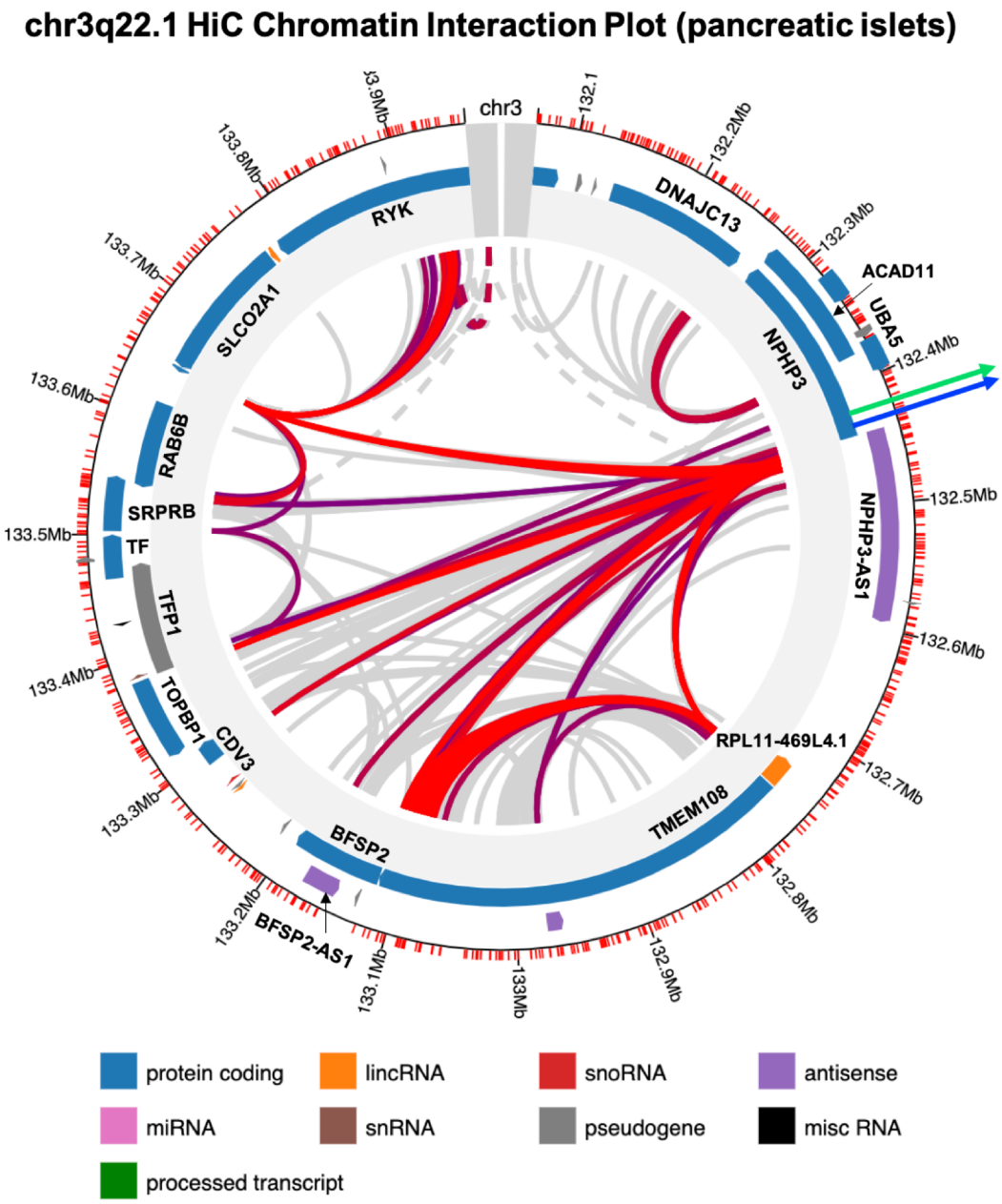


Figure 29: Chromatin interaction plot from the Capture HiC Plotter for pancreatic islet data (Schofield et al., 2016). The T2D location estimates ($\hat{S}_{T2D}$) are indicated with a green and blue arrow. Interactions with a score >10 are highlighted, with the redder colour indicating a more significant score. Grey lines represent interactions with a score <10.

5.3.1 chr3q22.1 *cis*-genes

While *ACAD11* is the only NEMG at this locus, the other four *cis*–genes present interesting candidates. These four *cis*–genes and their potential links with diabetes are discussed below.

NPHP3

NPHP3, or Nephrocystin 3, is implicated in the development of several organs through ciliogenesis (Zhou et al., 2010) and localises to the Inv compartment of the cilia shaft (Shiba et al., 2012). *NPHP3* mutations cause Nephronophthisis (NPHP), an autosomal recessive kidney disorder with co-morbitities of tapeto-retinal degeneration, hepatic fibrosis and end-stage kidney disease (Olbrich et al., 2003) and Senior-Løken Syndrome, which manifests with juvenile nephronophthisis and retinal degeneration (Ronquillo et al., 2012). Notably, retinopathy is the most common complication of diabetes (Simó-Servat et al., 2019), while kidney disease (nephropathy) occurs in up to 20-40% of diabetes patients and between 25-45% of patients with end-stage kidney disease have diabetes (Persson and Rossing, 2018). *NPHP3* mutations have also been linked with multi-organ polycystic disease, affecting both the kidneys and pancreas (Leeman et al., 2014) and are a documented cause for Renal-Hepatic-Pancreatic dysplasia (RHPD) (Neuhaus et al., 1996; Bergmann et al., 2008). Interestingly, renal cysts are a known co-morbidity of Maturity-onset diabetes of the young (MODY) caused by HNF-1β mutations (Bingham and Hattersley, 2004; Clissold et al., 2015). Patients with autosomal dominant polycystic kidney disease have dyslipidemia and are at a higher risk of developing diabetes (Fliszkiewicz et al., 2019) and mouse models of kidney disease have altered lipid metabolism (Menezes et al., 2016). This is consistent with the previous association of this locus with T2D-associated end-stage kidney disease (Guan et al., 2016).

A link between cilia function and diabetes has been discussed in the literature, with evidence suggesting a role for primary cilia in leptin and insulin signalling pathways, as well as satiety signalling in the hypothalamus (Lee et al., 2015a; Volta and Gerdes, 2017). Other ciliopathies, notably Alström syndrome and Bardet-Biedl syndrome have been observed to manifest with obsesity, early-onset diabetes and retinopathy (Kim et al., 2015a) and Gerdes et al. (2014) provided evidence that cilia influence insulin secretion and insulin signalling in rat pancreatic β-cells. Furthermore, both diabetes and ciliopathies are associated with male infertility (Inaba and Mizuno, 2016; Condorelli et al., 2018) and polycystic ovary syndrome (PCOS) (Gambineri et al., 2012; Tsang et al., 2018), presenting

further phenotypic overlap. Volta et al. (2019) discuss how defective cilia in pancreatic β-cells causes defective insulin secretion.

SRPRB

The Signal Recognition Particle (SRP) receptor (SR) is localised to the endoplasmic reticulum (ER) in eukaryotic cells and facilitates the import of secretory and membrane peptides into the ER. The ubiquitous RNA-protein SRP complex recognises translating peptides at ribosomes and transports them to the ER where it interacts with the SR (Lee et al., 2018). *SRPRB* encodes the β subunit of the SR (SR-β); a transmembrane GTPase that localises the α subunit, which interacts with the signal recognition particle, to the membrane (Ogg et al., 1998; Jadhav et al., 2015; Lee et al., 2018). SR-β is essential for protein translocation across the ER membrane (Fulga et al., 2001).

Although there is no direct evidence linking *SRPRB* to diabetes, the related gene *SEC61A1* caused diabetes in mice when mutated by triggering ER stress and β-cell apoptosis (Lloyd et al., 2010). *SEC61A1* encodes a subunit of the ER protein-translocation pore which transports peptides into the ER following the targeting of translating ribosomes to the ER via the SRP/SR (Lang et al., 2017). *SEC61A1* was shown to negatively regulate *SRPRB* which in turn may activate apoptosis via the NF-κB apoptosis pathway (Ma et al., 2017). The role of ER stress in β-cell apoptosis, hepatic insulin resistance and adipose insulin resistance is further reviewed in Meyerovich et al. (2018), Kim et al. (2015b) and Khan and Wang (2014), respectively.

SLCO2A1

SLCO2A1, or Solute Carrier Organic Anion Transporter Family Member 2A1, encodes the prostaglandin transporter (PGT) which mediates the degradation of prostaglandins (Liu et al., 2015a). Prostaglandins are lipids produced at sites of injury and infection which are involved in the regulation of inflammation and other biological processes. Their levels are reduced in diabetes presumably through enhanced degradation (Liu et al., 2015a). Nonsense mutations in *SLCO2A1* may cause familial digit clubbing, hypertrophic osteoarthropathy and colon cancer (Seifert et al., 2012; Guda et al., 2014), whereas missense

and less severe mutations may cause Hereditary Enteropathy (intestinal disease) (Umeno et al., 2015). Enteropathy is a less well-known complication of diabetes (Meldgaard et al., 2018). Conversely, inhibiting PGT has been shown to improve the reduced wound-healing seen in diabetes by increasing prostaglandin-induced angiogenesis (Syeda et al., 2012; Baltzis et al., 2014; Liu et al., 2015a).

PGT transports prostaglandins including PGD_2, PGE_1, PGE_2 and PGF_{2A}, which have been individually linked with diabetes-related phenotypes. For example, reducing levels of PGD_2 in mice via the knock out of lipocalin-type prostaglandin D2 synthase (L-PGDS) accelerated glucose intolerance and insulin-resistance (Ragolia et al., 2005). The PGE_2 receptor (EP3) was shown to blunt glucose-stimulated insulin secretion and antagonise GLP-1 signalling upon stimulation (Ragolia et al., 2005; Kimple et al., 2013) and PGE_2 levels are also significantly raised in patients with Diabetic retinopathy complications (Schoenberger et al., 2012) and with reduced glycemic control (Fenske et al., 2017).

RYK

RYK, or Receptor-Like Tyrosine Kinase, is a growth factor receptor of the non-canonical Wnt signalling pathway. RYK is one of two co-receptors of the Frz receptor which binds Wnt ligands and triggers β-catenin-independent Wnt signalling, activating either the Wnt/Ca^{2+} or Wnt/planar cell polarity (PCP) signalling pathways (Foulquier et al., 2018). Non-canonical Wnt signalling may mediate adipogenesis resulting from high levels of glucose (Keats et al., 2014) and a switch from canonical to non-canonical Wnt signalling has been linked to fatty liver phenotypes (Ackers and Malgor, 2018).

5.3.2 chr3q22.1 *cis*-gene expression

The case-control expression of the five *cis*–genes was tested using the gene expression datasets meta-analysed in Chapter 4. The results are shown in Table 15. *ACAD11* was not present on any of the arrays used in the expression studies. The other four *cis*–genes showed significant differential expression in several tissues. *NPHP3* was significantly down-regulated in T2D cases in skeletal muscle. Consistent with the literature, *SLCO2A1* expression was increased with T2D (Liu et al., 2015a). *SRPRB* did not show

***Cis*-gene expression in T2D cases vs. controls**

Cis-gene	Muscle	Adipose	Liver	Pancreas	Family history (muscle)
ACAD11	NA	NA	NA	NA	NA
NPHP3	**-3.18**	-0.30[a]	NA	-1.55[b]	-0.90
SRPRB	-0.91	1.83	1.12	-0.87	0.33
SLCO2A1	**5.91**	-0.27	**2.32**	-0.41	**2.06**
RYK	**-6.41**	-0.41	**4.34**	0.01	**-2.70**

Table 15: Case-control expression of the *cis*–genes. Meta-analysed Z scores are presented, with positive numbers representing higher expression in cases compared to controls and vice versa. NA values indicate missing data. [a]Meta-analysis for 5/6 adipose datasets (excluding GSE94752), [b]Z-score for 1/3 pancreas datasets (excluding GSE25724 and GSE41762).

independent evidence for differential expression in these datasets. *RYK* was significantly down-regulated in muscle, however showed increased expression in liver. The potential tissue-specific regulation of *RYK* may be an interesting area of further study. This analysis provides a convenient means to confirm that the putative *cis*–genes are differentially expressed in T2D. However, further fine-mapping and functional studies will be needed to confirm that these genes are regulated by the causal T2D-associated variant(s) at this locus. The following sections describe the fine-mapping results.

5.3.3 Targeted sequencing results

Targeted sequencing was carried out for French T2D cases and controls (n = 82 cases and n = 89 controls after quality control). The sequenced region is highlighted in Figure 30. The following sections describe the analysis of four classes of variants, (1) nominally signifiant variants, (2) coding variants, (3) rare variants and (4) promoter variants.

(1) Nominally significant variants

SNPs and indels were included in a test of association with disease status, as described in **Section 5.2.3: Variant annotation and association**. Five SNPs and one indel were nominally significant (p-value $\leq$0.05); these are listed in Table 16 and are plotted in Figure 31. According to the 1000 Genomes European cohort (CEU), the rs16839460 and rs3860501 SNPs are in high LD ($R^2 = 0.82$). rs114923567 and rs75185415 are located 477bp apart and are in complete LD ($R^2 = 1$), with the minor alleles found in the same four

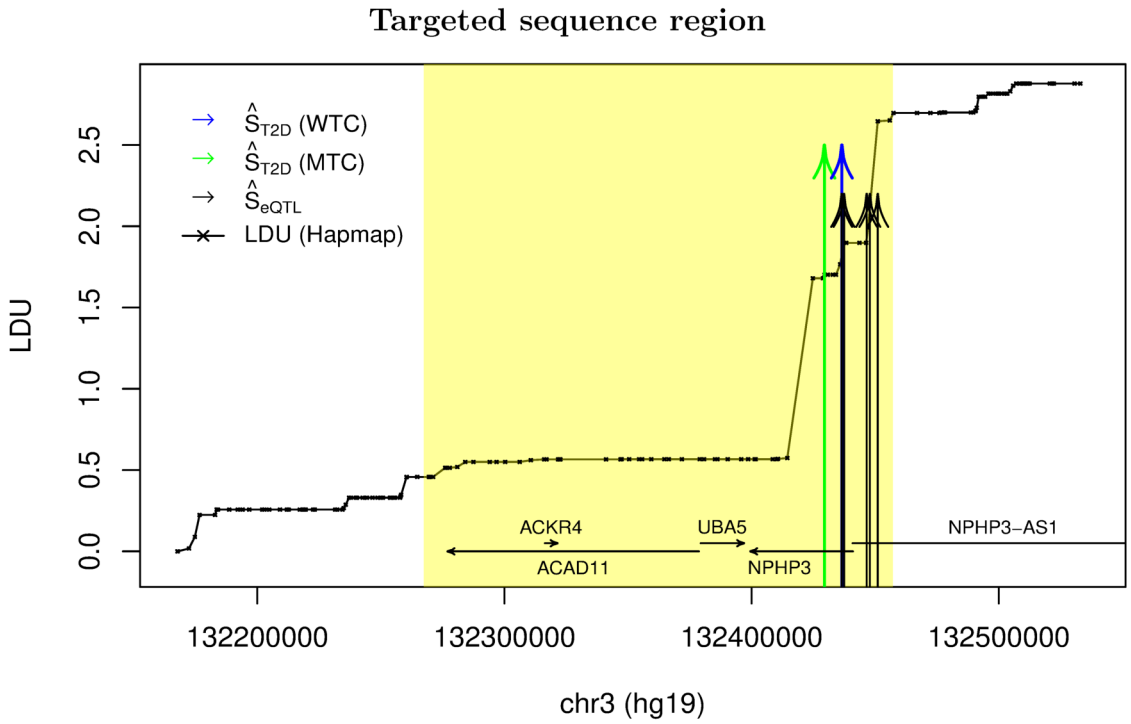


Figure 30: Plot showing the targeted sequence region for the chr3q22.1 T2D locus. Sequence data is available for the region highlighted in yellow. The European LDU map is shown, along with $\hat{S}_{T2D}$ and $\hat{S}_{eQTL}$ point estimates and nearby genes.

T2D cases. All other SNPs are in low LD. To test if the variants represent independent signals, a conditional test of association was carried out. As expected, neither rs16839460 or rs3860501 was significant when the analysis was conditioned on the other, which would

Nominally significant variants associated with T2D status

Position (hg19)	rsID	1000G MAF	Cases MAF (count)	Controls MAF (count)	*p*-value	CADD scores
SNPs						
chr3:132275482[b]	rs114923567	0.010	0.025 (4)	.	0.050	0.63
chr3:132275959[b]	rs75185415	0.010	0.025 (4)	.	0.050	1.58
chr3:132284255	rs2085316	0.004	.	0.034 (6)	0.030	10.70
chr3:132347414[a]	rs16839460	0.095	0.105 (17)	0.045 (8)	0.026	3.20
chr3:132424780[a]	rs3860501	0.112	0.148 (21)	0.062 (11)	0.023	5.28
Indels						
chr3:132428532[c]	rs138040526	-	0.037 (6)	0.006 (1)	0.049	0.79

Table 16: Significant variants. The population minor allele frequency (MAF) is shown according to the 1000 Genomes Project (1000G) along with the cohort MAFs and minor allele counts for the T2D cases and controls. [a]rs16839460 and rs3860501 are in high LD ($R^2 = 0.82$). [b]rs114923567 and rs75185415 are in complete LD.[c]The GT/- genotype was observed in five cases and one control, and GGT>TGT was observed in one case.

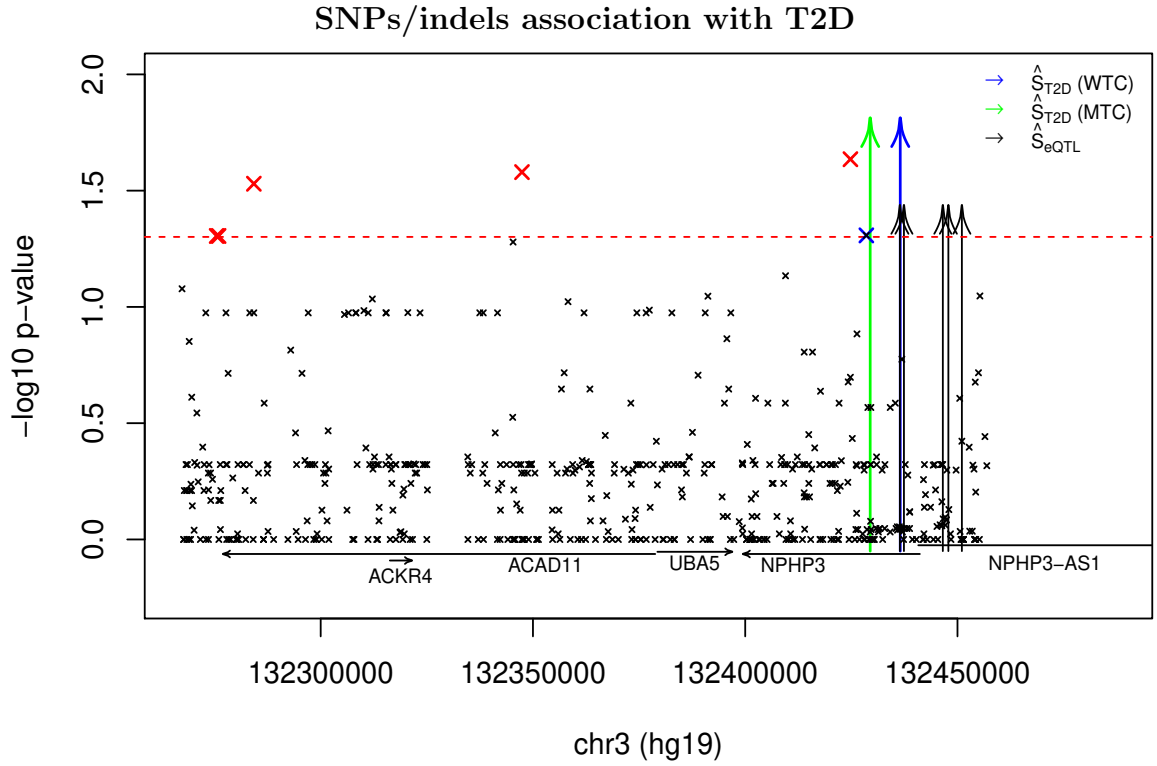


Figure 31: The association results for all variants across the chr3q22.1 T2D locus. The -log10 p-value is plotted, with SNPs giving a p<0.05 highlighted in red and indels in blue.

be consistent with just one causal SNP driving the association. When conditioned on these two SNPs, the p-values for rs114923567, rs75185415 and rs2085316 decreased to <0.01, demonstrating that these are independent signals.

When entered into the Genotype-Tissue Expression (GTEx) database, rs16839460 and rs3860501 were reported to be significant eQTL for *NPHP3*; rs16839460 in 18 different tissues, including in subcutaneous adipose (p = 4.1e-05), omental adipose (p = 8.8e-05) and skeletal muscle (p = 3.0e-06) and rs3860501 in 27 different tissues, including skeletal muscle (p = 2.4e-12), subcutaneous adipose (p = 2.8e-09), omental adipose (p = 6.0e-07) and pancreas (p = 3.3e-05). The minor alleles of both SNPs associated with higher expression of *NPHP3*; this can be seen in Figure 32 for rs16839460. This is notably inconsistent with the decreased expression of *NPHP3* observed in the T2D gene expression datasets (see Table 15). Further functional studies may be required to determine whether the rs16839460 minor allele actively alters *NPHP3* expression and increases T2D risk. The other four variants were not reported as significant eQTL, however these are all low-frequency variants and may be underpowered to detect association with gene

expression.

GTEx *NPHP3* expression and rs16839460

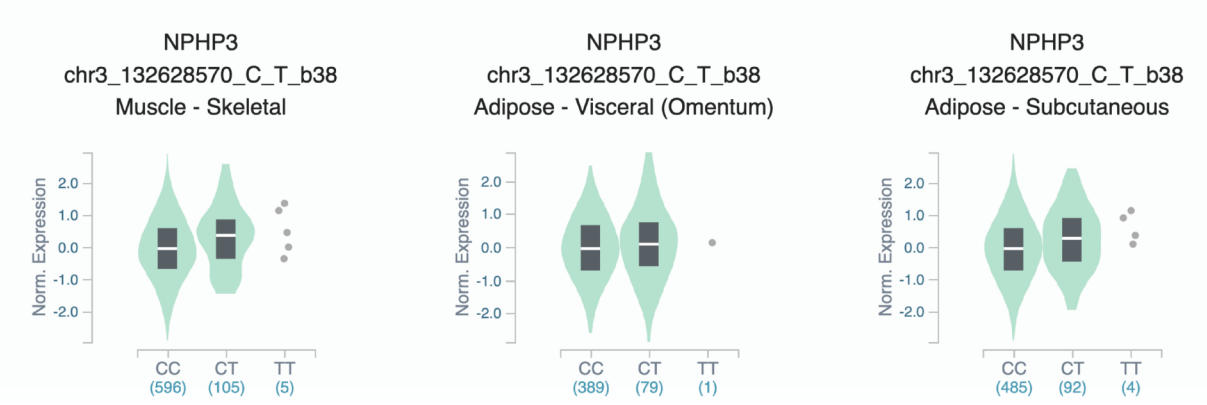


Figure 32: *NPHP3* expression with the rs16839460 genotype in skeletal muscle, omental and subcutaneous adipose according to GTEx.

The SNPs were investigated for evidence of overlapping annotated regulatory elements using ChromHMM tracks from the ROADMAP Epigenomics Project. Six datasets were downloaded: Adipose Nuclei (E063), Skeletal Muscle (Female) (E108), Adult Liver (E066), the liver HepG2 Hepatocellular Carcinoma Cell Line (E118), Pancreatic islets (E087) and Pancreas (E098). Chip-seq data for the same Roadmap datasets can be seen plotted in Figure 33. Three of the five variants overlapped with annotated enhancers: rs75185415 and rs16839460 in adipose and rs114923567 in adipose and pancreas. rs3860501 did not overlap any annotated regulatory element, suggesting that rs16839460 may be the stronger candidate of these two SNPs which are in high LD. These three SNPs were tested for evidence of disrupting transcription factor binding sites (TFBSs) using the R package *motifbreakR* (detailed method in **Section 5.2.3: Variant annotation and association**). The indel rs138040526 did not overlap with any annotated regulatory elements.

Running *motifbreakR* with the recommended p-value threshold of 1e-04 gave a significant match for rs16839460 to a FOXP3 binding site. The minor allele, which was found in 10.5% of cases compared to 4.5% of controls (see Table 16) increased the match suggesting stronger binding of FOXP3. Interestingly, one control had a 162bp deletion at the same site, shown in Figure 33. The observed deletion in one control is consistent with increased FOXP3 binding associating positively with T2D risk; presumably the deletion would be protective if FOXP3 binding is disrupted. The FOXP3 binding motif is shown below.

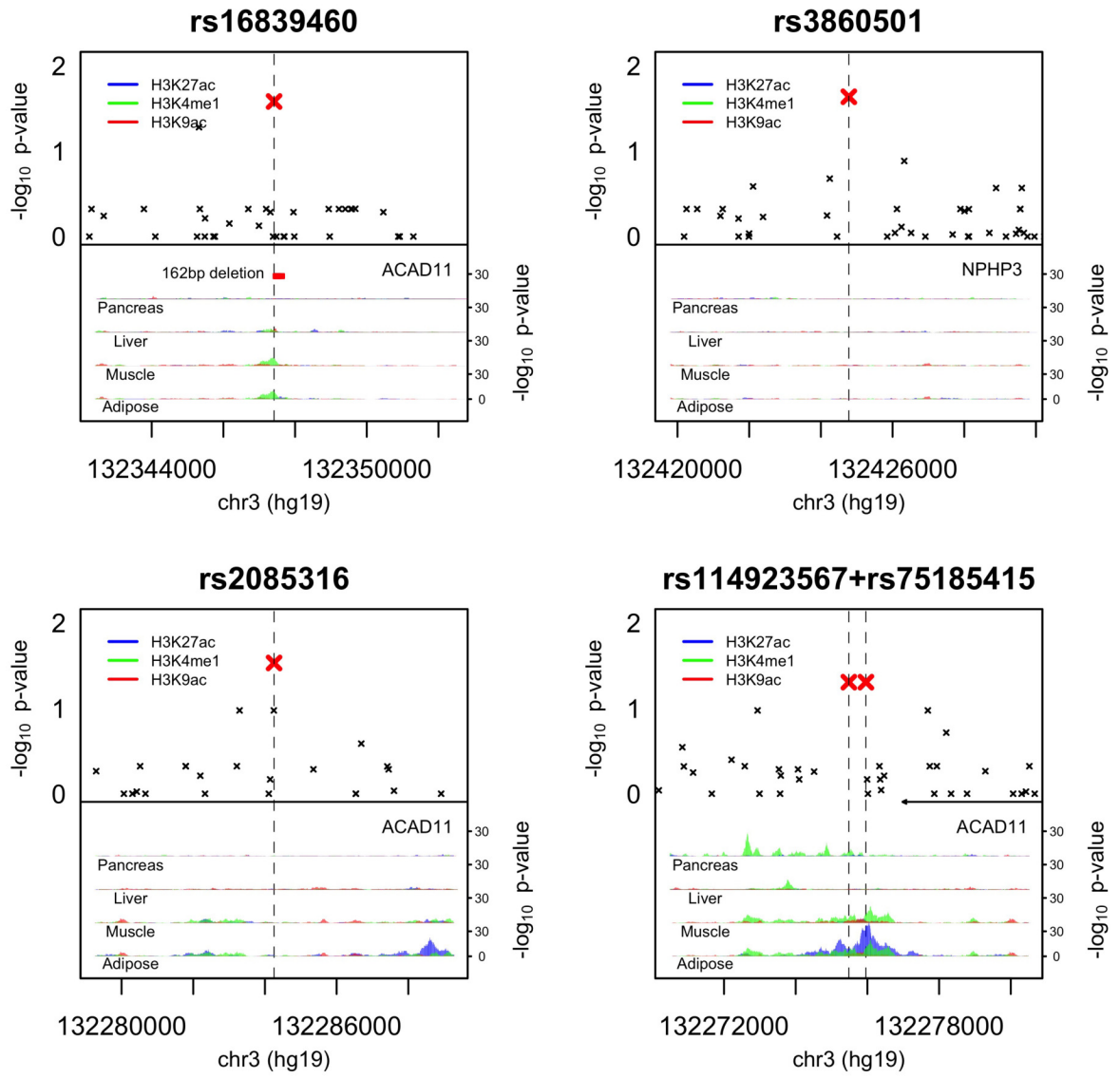


Figure 33: ROADMAP ChiP-seq tracks for the chromatin enhancer marks (H3K27ac, H3K4me1 and H3K9ac) surrounding the five nominally significant SNPs at the chr3q22.1 locus. The T2D-relevant tissues adipose, skeletal muscle, liver (liver and HepG2 are shown combined) and pancreas (pancreas and islet are combined) are shown. A 162bp deletion found in one control is shown at the rs16839460 locus.

FOXP3 expression is characteristic of T regulatory (Treg) cells (Onodera et al., 2015), which are decreased in patients with T2D and obesity and potentially contribute to high glucose levels, inflammation and glucose intolerance (Wagner et al., 2013; Zhang et al., 2014a). Treg are highly dependent on fatty acid oxidation (FAO) for proliferation, differentiation and protection from fatty acid-induced cell death (Howie et al., 2017; Chen et al., 2019b) and this is dependent on FOXP3 expression (Howie et al., 2017). This

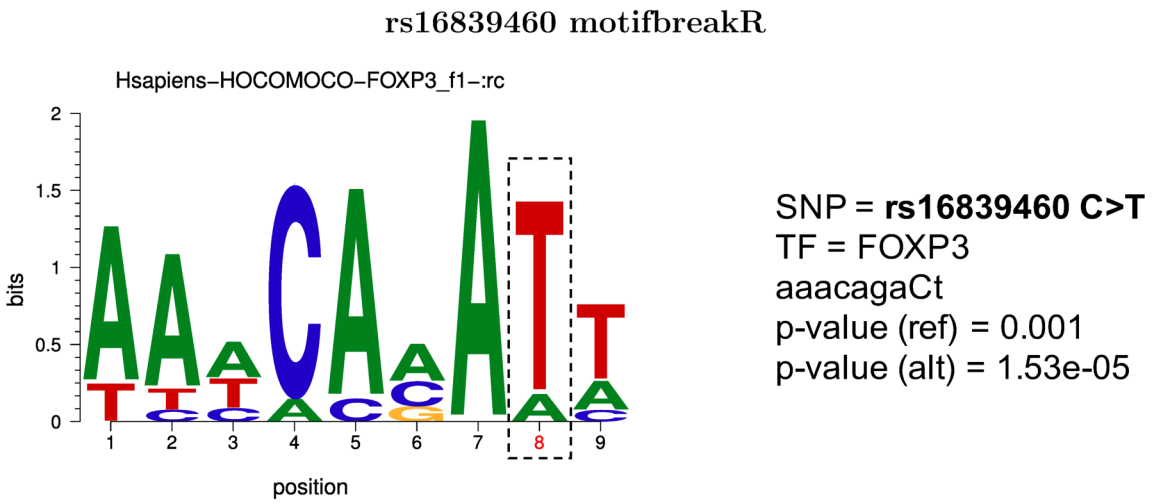


Figure 34: rs16839460 occurs at position 8 of a motif matching a FOXP3 transcription factor binding motif. The DNA sequence surrounding rs16839460 is shown on the right. The alternative allele (T) increases the significance of the match.

is particularly true for tissue-resident Treg such as visceral adipose tissue (VAT) Treg which act to suppress local inflammation (Kempkes et al., 2019). T cell differentiation is also influenced by the lengths of FA chains, with long-chain fatty acids (LCFAs) reducing Treg differentiation (instead encouraging T helper cell formation), while short-chain FAs enhance Treg differentiation (Haghikia et al., 2015). Since *ACAD11* oxidises LCFAs and VLCFAs (He et al., 2011), the relationship between *ACAD11*, rs16839460, FOXP3 and Treg cell differentiation may be an interesting area of further study. FOXP3 can act as both a repressor and activator of gene expression, so further investigation will be required to confirm if FOXP3 binds at rs16839460 and how this might affect the expression of *ACAD11*, *NPHP3* and other genes.

rs114923567 and rs75185415 also overlapped ChromHMM enhancers and can be seen in Figure 33 to overlap with significant ChiP-seq peaks for chromatin modifications characteristic of enhancer elements, particularly in adipose. Neither SNPs returned significant results from *motifbreakR* using the default statistical threshold. *motifbreakR* was repeated with a lower p-value threshold of $1e$-3, resulting in rs75185415 matching eight motifs and rs114923567 matching four; these are listed in Table 17. Of the TFs matched to rs114923567, SMAD3 and SMAD4 play a significant role in adipogenesis and have been repeatedly implicated in T2D aetiology (Tan et al., 2012). SMAD transcription

rs114923567 and rs75185415 motifbreakR

Transcription factor	*p*-value (ref)	*p*-value (alt)	Transcription factor	*p*-value (ref)	*p*-value (alt)
rs114923567			**rs75185415**		
MYB	0.022	7.32*e*-4	ALX1	0.006	3.17*e*-4
REST	9.74*e*-4	3.54*e*-3	E4F1	7.21*e*-4	0.013
SMAD3	4.96*e*-4	0.016	HOXD9	0.013	3.82*e*-4
SMAD4	4.77*e*-4	0.013	NR2E3	0.012	8.43*e*-4
			ONECUT2	0.010	5.27*e*-4
			POU1F1	0.014	7.74*e*-4
			POU3F2	0.011	4.28*e*-4
			POU5F1	0.005	9.02*e*-4

Table 17: *motibreakR* results for the SNPs rs114923567 and rs75185415. Results defined as having a 'strong' effect are shown, with the *p*-values showing the significance of reference (ref) allele and alternative (alt) allele match to the transcription factor (TF) motif.

factors may positively regulate FAO (Seong et al., 2018). The rs114923567C>T variant occurs at a fully conserved base in the SMAD3/4 binding motifs, shown in Figure 35. rs114923567C>T may be hypothesised to disrupt *ACAD11* expression and FAO by reducing SMAD3/4 binding.

rs114923567 motifbreakR

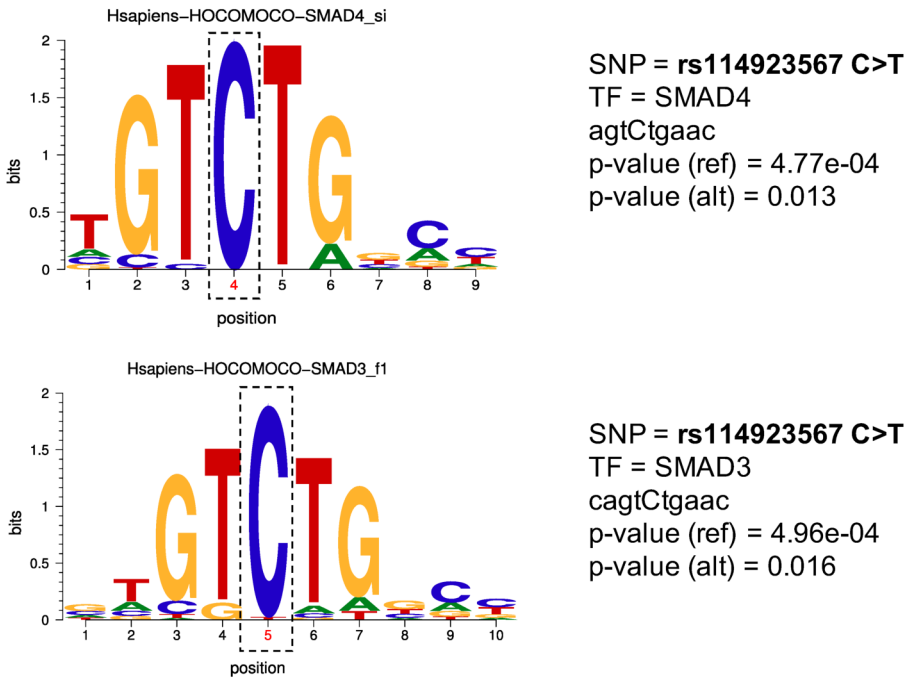


Figure 35: **rs114923567** occurs at position 4/5 of the SMAD4/SMAD3 binding motifs.

Of the TFs matched to rs75185415, E4F1 and NR2E3 offer particularly interesting candidates. E4F1 has been reported to regulate aspects of mitochondrial function (Hatchi

et al., 2011; Lacroix et al., 2016) and to modulate p53 transcriptional activity (Le Cam et al., 2006). As described on page 144, *ACAD11* responds to p53 activity to mediate cell survival. NR2E3 is expressed in the retina and has a potential link with retinopathy, as does *NPHP3* (Li et al., 2020). The other two nominally significant SNPs did not appear to overlap active regulatory elements in the four tissues tested (adipose, muscle liver and pancreas), however both were significantly matched to TF binding motifs when investigated using *motibreakR*; rs3860501A>G to a POU3F2 binding motif and rs2085316C>T to a HAND1 motif. The possibility that these variants may have other tissue-specific effects or act through other mechanisms, such as affecting DNA structure, should be further investigated to either include or exclude them as candidates.

(2) Coding variants

The genes *ACAD11*, *NPHP3*, *UBA5* and *ACKR4* were included in the targeted sequencing. *UBA5* encodes the E1-like activating enzyme UBA5, which is required for ufmylation (a post-translational protein modification in which the ubiquitin-like protein UFM1 is conjugated to target proteins) and *ACKR4* encodes the atypical chemokine receptor 4. A total of 28 coding mutations were detected within *ACAD11*, *NPHP3* and *ACKR4*. These included one splicing and 14 non-synonymous mutations shown in Table 18 with predicted pathogenicity scores from SIFT, PolyPhen-2 (HDIV) and CADD (see **Section 5.2.3: Variant annotation and association** for methods). Of these, six non-synonymous SNVs and one splicing mutation were detected in *ACAD11*. The splicing mutation was not observed in any controls, however was present in three cases and was predicted by both SIFT and PolyPhen-2 to be deleterious. It also obtained the highest CADD score of all the variants across the sequenced region (CADD score = 33). For the remaining variants, there was no evidence of a significant enrichment for potentially deleterious variants in T2D cases. However, the frequency of the *ACAD11* splicing mutation may be an interesting candidate for further study.

Gene	Type	rsID	position (hg19)	MAF (NFE)	MAF cases (#)	MAF controls (#)	SIFT	PolyPhen-2 (HDIV)	CADD
ACAD11	missense SNV	rs41272321[a]	chr3:132338346	0.113	0.130 (21)	0.135 (24)	**0.01** (D)	0.29 (B)	21.0
ACAD11	missense SNV	rs747458128	chr3:132278780	3.70*e*-05	.	0.006 (1)	**0.00** (D)	**1.00** (D)	23.3
ACAD11	missense SNV	.	chr3:132298370	.	.	0.006 (1)	0.08 (T)	0.94 (P)	18.9
ACAD11	missense SNV	.	chr3:132347194	.	0.006 (1)	.	0.21 (T)	0.17 (B)	22.6
ACAD11	missense SNV	rs1439156821	chr3:132349400	.	0.006 (1)	.	0.10 (T)	0.04 (B)	16.0
ACAD11	missense SNV	rs144771431	chr3:132360973	0.001	0.006 (1)	.	**0.04** (D)	**1.00** (D)	28.9
ACAD11	splicing SNV	rs41272317	chr3:132337477	0.014	0.020 (3)	.	**0.00** (D)	**1.00** (D)	33.0
NPHP3	missense SNV	rs746500844	chr3:132403514	1.50*e*-05	.	0.006 (1)	0.26 (T)	0.57 (P)	23.4
NPHP3	missense SNV	rs146890274[b]	chr3:132405152	7.00*e*-04	.	0.006 (1)	0.53 (T)	0.84 (P)	23.5
NPHP3	missense SNV	rs145166784[c]	chr3:132418279	3.1010*e*-05	.	0.006 (1)	0.20 (T)	**0.98** (D)	26.3
NPHP3	missense SNV	rs756330976	chr3:132401637	1.00*e*-04	.	0.006 (1)	**0.01** (D)	**1.00** (D)	28.4
NPHP3	missense SNV	rs373870292	chr3:132402331	3.01*e*-05	0.006 (1)	.	0.15 (T)	**0.96** (D)	31.0
ACKR4	missense SNV	rs138269944	chr3:132319690	0.002	0.006 (1)	0.006 (1)	**0.04** (D)	0.00 (B)	8.4
ACKR4	missense SNV	rs141386637	chr3:132319810	4.51*e*-05	0.006 (1)	.	0.57 (T)	0.00 (B)	12.3
ACKR4	missense SNV	rs145975679	chr3:132320035	0.003	0.006 (1)	.	0.19 (T)	0.35 (B)	23.9

Table 18: Exonic mutations across the chr3q22.1 sequenced region. The minor allele frequencies (MAFs) for non-Finnish Europeans (NFE) are shown according to ExAc (Lek et al., 2016). The MAF for the T2D cases (n=82) and controls (n=89) are shown, along with the minor allele counts (#). Predicted pathogenicity phred scores are shown according to SIFT (Vaser et al., 2016) and PolyPhen-2 (Adzhubei et al., 2010), which predict whether variants are likely to be tolerated (T), deleterious (D), benign (B) or possibly damaging (P). [a]Associated with neonatal cytokine/chemokine levels (Traglia et al., 2018). [b]Reported in ClinVar as unknown significance (observed with Nephronophthisis, accession: VCV000195784.2), [c]Reported in ClinVar as unknown significance (observed in a case of juvenile nephronophthisis, VCV000578257.1).

(3) Rare variants

Case-only and control-only variants were compared using CADD scores of estimated pathogenicity (see methods **Section 5.2.1: NGS methods**). Of a total 482 variants, 215 were singletons (present in only one individual). 107 occurred in the 82 cases and 108 occurred in the 89 controls. There was no significant difference between the CADD scores of these variants. All case-only and control-only variants not limited to singletons included 131 case-only variants and 118 control-only variants. There was no significant difference between the distribution of CADD scores for the case-only and control-only variants. Interestingly, the top three scoring variant were case-only, these included the *ACAD11* splicing variant observed in 3 cases (rs41272317, CADD = 33.0) and two variants observed in one case each: rs373870292, a *NPHP3* missense variant and rs144771431, an *ACAD11* missense variant. Case and control-only variants were investigated for intersections with enhancer or promoter elements using ChromHMM tracks, however there was no clear enrichment of case-only variants overlapping annotated elements compared to control-only variants. The results are shown in Table 19.

Case-only and control-only variants overlapping ChromHMM elements

	Case		**Control**	
	Enhancer	Tss	Enhancer	Tss
Adipose (E063)	15	6	14	8
Muscle (E108)	2	2	1	6
Liver (E066)	0	5	0	7
HepG2 (E118)	8	5	7	5
Islet (E087)	7	4	8	6
Pancreas (E098)	9	1	7	4

Table 19: The number of case-only and control-only variants which overlap ROADMAP ChromHMM annotated enhancers or transcription start sites (Tss).

(4) Promoter variants: *ACAD11/UBA5*

Variants intersecting the ChromHMM Active TSS or Flanking Active TSS states at the *ACAD11/UBA5* back-to-back promoters were further investigated. Two case-only, two control-only and three common SNPs were located in the annotated TSS. The ChromHMM tracks are shown in Figure 36 and ChiP-seq tracks are plotted in Figure 37 for adipose

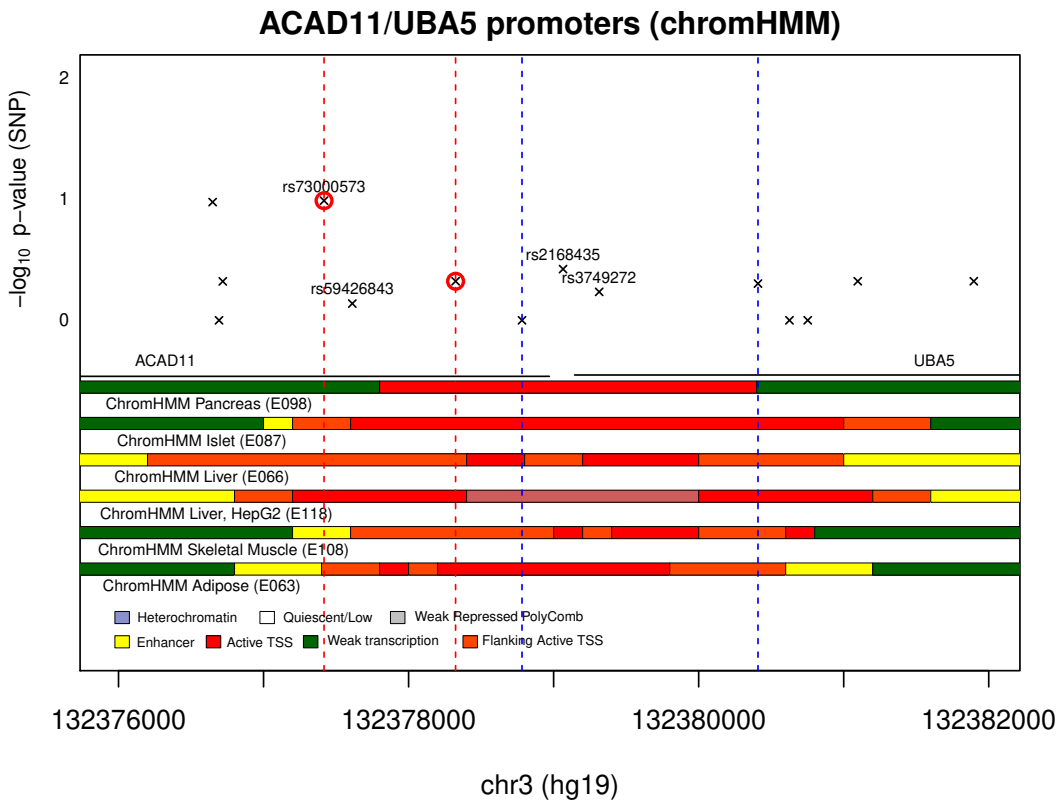


Figure 36: ROADMAP ChromHMM data for the *ACAD11 / UBA5* promoters. Case-only and control-only variants are shown with red and blue dotted lines, respectively. Two variants disrupting transcription factor binding sites are circled in red.

nuclei. In both figures, the case-only and control-only variants are highlighted with red and blue dotted lines, respectively. All seven variants were tested for evidence of altering TF binding sites using *motifbreakR* with the recommended p-value $<1e$-4 and a filter for 'strong' effects. The three common SNPs (rs59426843, rs2168435 and rs3749272) as well as the rare control-only variant did not return significant results. However the two case-only variants, circled in red in Figures 36 and 37, were both predicted to disrupt TF binding sites. rs73000573 was observed in three T2D cases and chr3:132378324C>T was observed in one case. In Figure 36, rs73000573 can also be seen to overlap an active enhancer annotated in muscle.

rs73000573 was predicted to have a strong effect on TF binding for four TFs, these are shown in Figure 38: NR1H2 (LXR-β), NR2F6 (COUP-TFγ), NR2F2 (COT2/COUP-TFII, or COUP-TFβ) and NR2A1 (HNF4α). All of these TFs have been implicated in the regulation of FAO or related processes, making them strong candidates for the

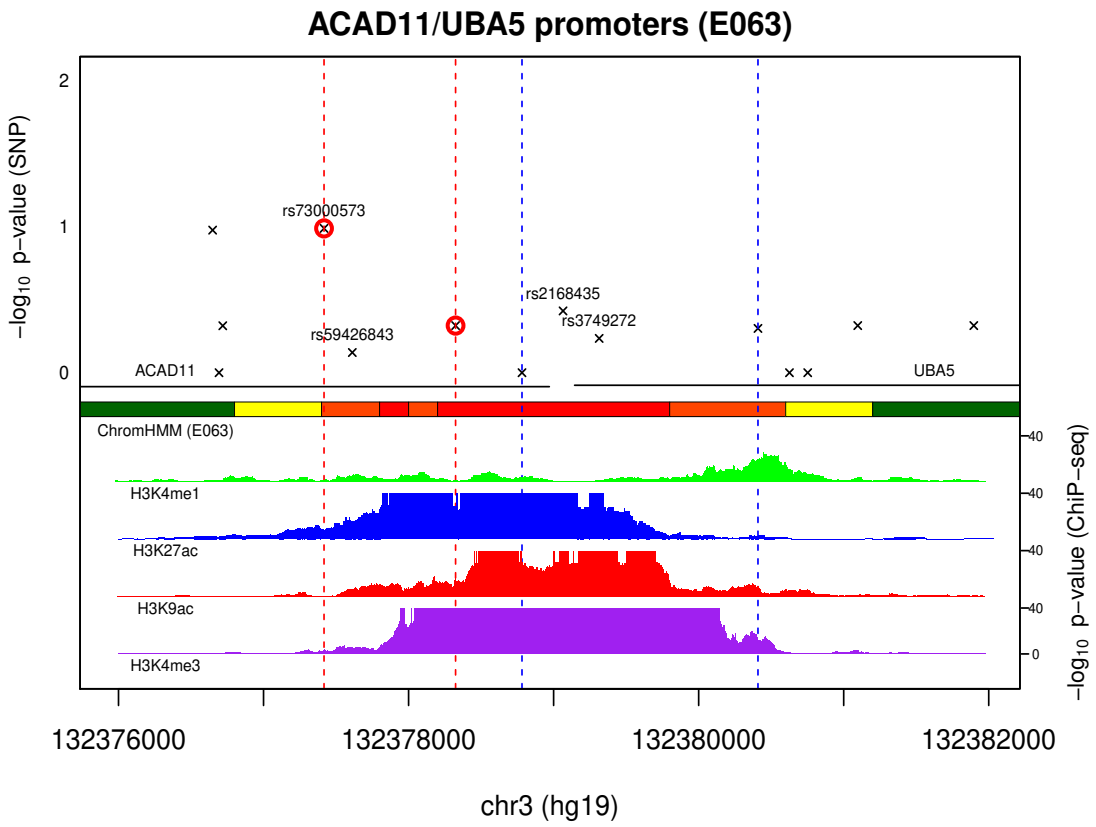


Figure 37: ROADMAP ChiP-seq and ChromHMM data for adipose nuclei (E063) around the *ACAD11/UBA5* back-to-back promoters. ChiP-seq tracks are truncated at a -$\log_{10}$ p-value of 40. Chromatin modifications typical of enhancers (H3K4me1), enhancers and promoters (H3K27ac, H3K9ac) and promoters (H3K4me3) are shown.

regulation of *ACAD11*. All four TFs may bind DNA as heterodimers with the retinoid X receptor (RXR). COUP-TFII, COUP-TFγ and HNF4α also bind fatty acids, while LXR-β is regulated by a variety of lipophilic molecules including thyroid hormone, fatty acids, bile acids, and sterols (Weikum et al., 2018). COUP-TFII positively regulates β-fatty acid oxidation in the liver (Ashraf et al., 2019) and has been quoted as a master regulator of metabolism, influencing adipogenesis, gluconeogenesis, insulin sensitivity and even insulin secretion (Perilhou et al., 2008; Ashraf et al., 2019; Polvani et al., 2020). COUP-TFII may act to repress gene expression by acting competitively with HNF4α to dimerise with RXR (Stroup and Chiang, 2000; McMullen et al., 2014; Ashraf et al., 2019). HNF4α is itself involved in the formation and maintenance of the liver and pancreas (Lau et al., 2018) and mutations in HNF4α cause MODY1. LXR-β has been implicated in the growth of adipocytes, glucose homeostasis and pancreatic β-cell function (Gerin et al.,

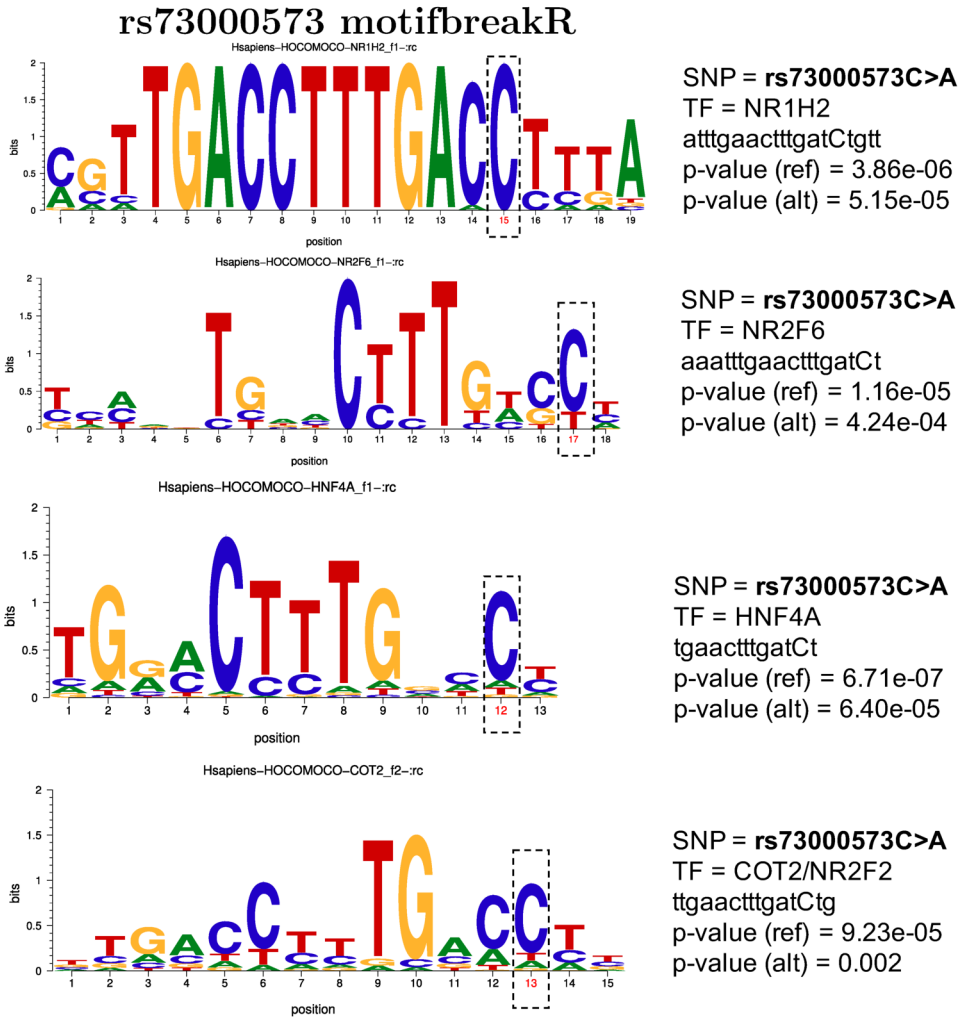


Figure 38: *motifbreakR* results for rs73000573. The C>A variant occurs at positions 15, 17, 12 and 13 of motifs matching the transcription factors NR1H2 (LXR-β), NR2F6 (COUP-TFγ), HNF4A and COT2/NR2F2 (COT2/COUP-TFII, or COUP-TFβ), respectively.

2005). LXR-β deficient mice develop insulin reistance on a high-fat diet (Gerin et al., 2005; Korach-André et al., 2010) and LXR-β agonists improve insulin sensitivity (Cao et al., 2003; Commerford et al., 2007; Schulman, 2017). Interestingly, LXR, RXR and COUP-TFγ (*NR2F6*) all enhance the differentiation of Foxp3+ Treg cells (Hermann-Kleiter et al., 2008; Hermann-Kleiter and Baier, 2014; Takeuchi et al., 2013; Herold et al., 2017).

To further investigate whether these TFs bind at the rs73000573 position, ChiP-seq data available for COUP-TFII and HNF4α was downloaded from ENCODE and plotted in Figure 39. Data was available for two liver samples, one 32 year old male (ENCODE

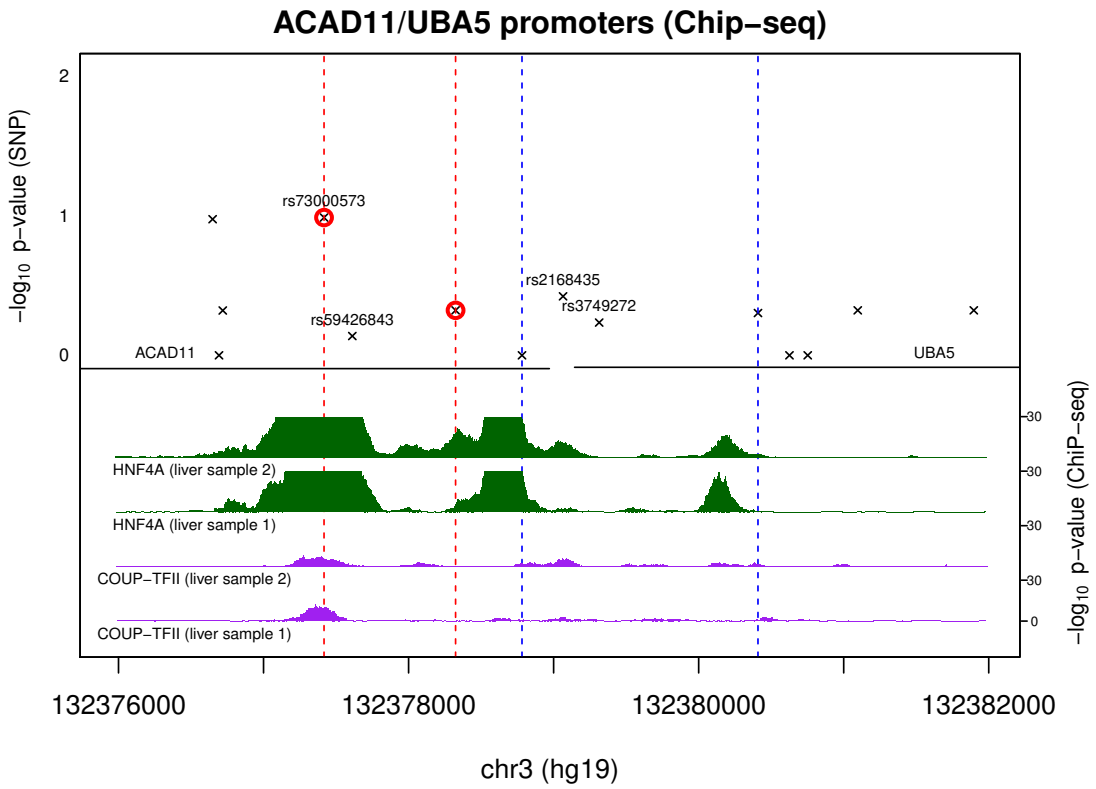


Figure 39: ChiP-seq for COUP-TFII and HNF4A plotted for two liver samples. ChiP-seq tracks are truncated at -$\log_{10}$ p-value of 30.

accessions: COUP-TFII = ENCSR338MMB and HNF4α = ENCSR601OGE) and one 4 year old female (COUP-TFII = ENCSR168SMX and HNF4α = ENCSR445QRF). As seen in Figure 39, rs73000573 is located in the middle of a peak for these two TFs, confirming the binding of COUP-TFII and HNF4α at this position. In GTEx, rs73000573 was reported as a significant eQTL associated with lower expression of both *ACAD11* and *NPHP3* across multiple tissues. The results from GTEx are shown in Figure 40.

The LD between rs73000573 and the previously reported candidate SNPs was investigated. rs73000573 was found to be in complete LD with the *ACAD11* splicing mutation and the minor alleles were found in the same three T2D cases. This low-frequency haplotype therefore has both an *ACAD11* splicing mutation and promoter/enhancer mutation disrupting HNF4α and COUP-TFII binding. The splicing mutation, rs41272317, was also associated with lower expression of *NPHP3* and *ACAD11* in GTEx, as expected. Since both variants are in complete LD, further study will be needed to determine if one or both SNPs alter gene expression.

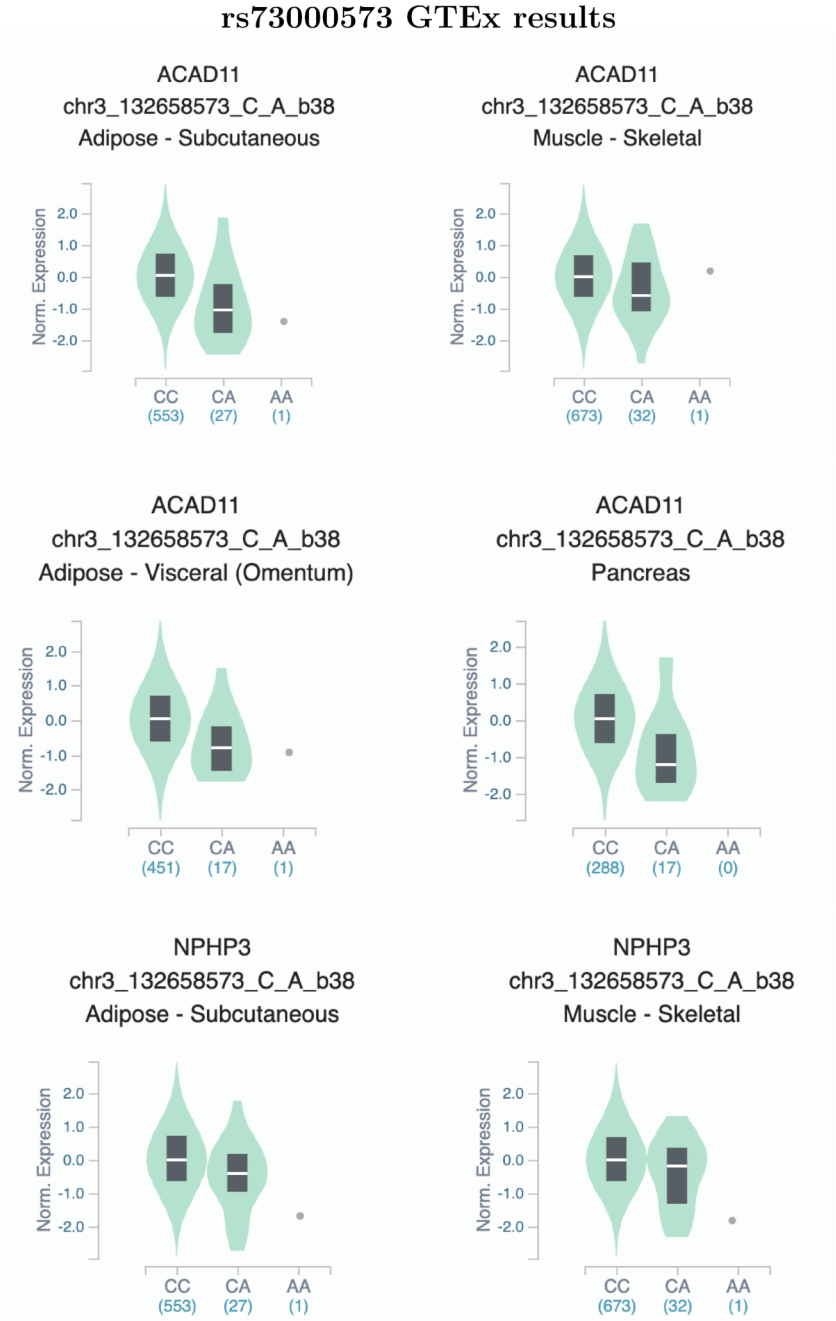


Figure 40: GTEx association of rs73000573 with *ACAD11* and *NPHP3* expression levels.

A second case-only mutation, chr3:132378324C>T, was reported to overlap the ChromHMM Tss state and was reported by by *motifbreakR* to match a KLF8 binding motif, the results of which are shown in Figure 41. Despite a significant *p*-value, both C and T are observed at position 8 of the binding motif, making it unlikely that the mutations would disrupt KLF8 binding.

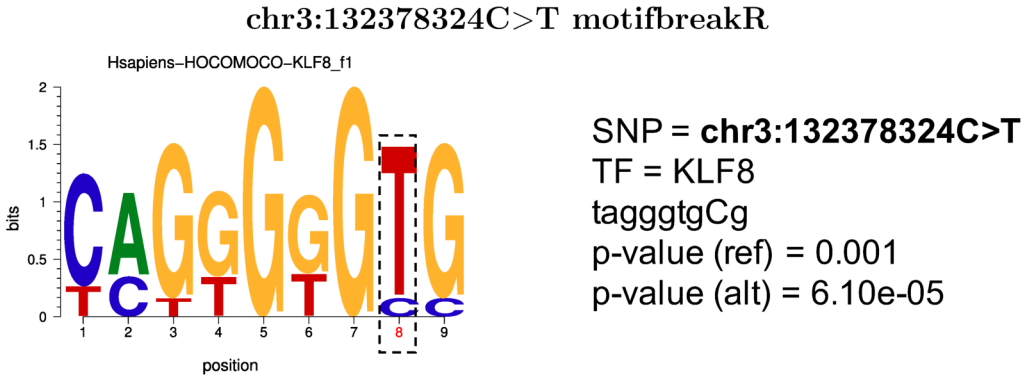


Figure 41: The chr3:132378324C>T variant observed in one T2D case was reported by *motifbreakR* to match a motif for the transcription factor KLF8.

(4) Promoter variants: *NPHP3/NPHP3-AS1*

Variants surrounding the $\hat{S}_{T2D}$, *NPHP3* and *NPHP3-AS1* promoters were investigated. Figure 42 shows ROADMAP Chip-seq data for this region. Based on this data, the regulatory element closest to the $\hat{S}_{T2D}$ location estimates was the *NPHP3/NPHP3-AS1* back-to-back promoters, with no other peaks seen in the region. Figure 43 shows the same image for the smaller promoter region. Four variants overlapping ChiP-seq peaks at the *NPHP3/NPHP3-AS1* promoters were predicted by *motifbreakR* to disrupt TF-BSs, these are circled in red in Figure 43. These include rs560793264G>T observed in one control, a common SNP rs66770376G>A and two previously unreported variants; chr3:132441927A>C observed in one control and chr3:132442215G>A observed in one T2D case.

Of interest is the rare variant chr3:132442215G>A observed in one T2D case, which disrupts a binding motif for NKX2-2. NKX2-2 is a TF involved in the development, function and maintenance of pancreatic β-cell identity (Doyle and Sussel, 2007; Bastidas-Ponce et al., 2017) and NKX2-2 mutations cause neonatal diabetes (Flanagan et al., 2014).

Through its role in organ development and cilia function, *NPHP3* mutations also cause pancreatic dysplasia (Fiskerstrand et al., 2010), suggesting an interesting link between NKX2-2, *NPHP3* expression and β-cell development which may warrant further investigation. However, the control-only mutations were also predicted to disrupt TF binding, suggesting that there is no significant differences between case and control mutations at this location.

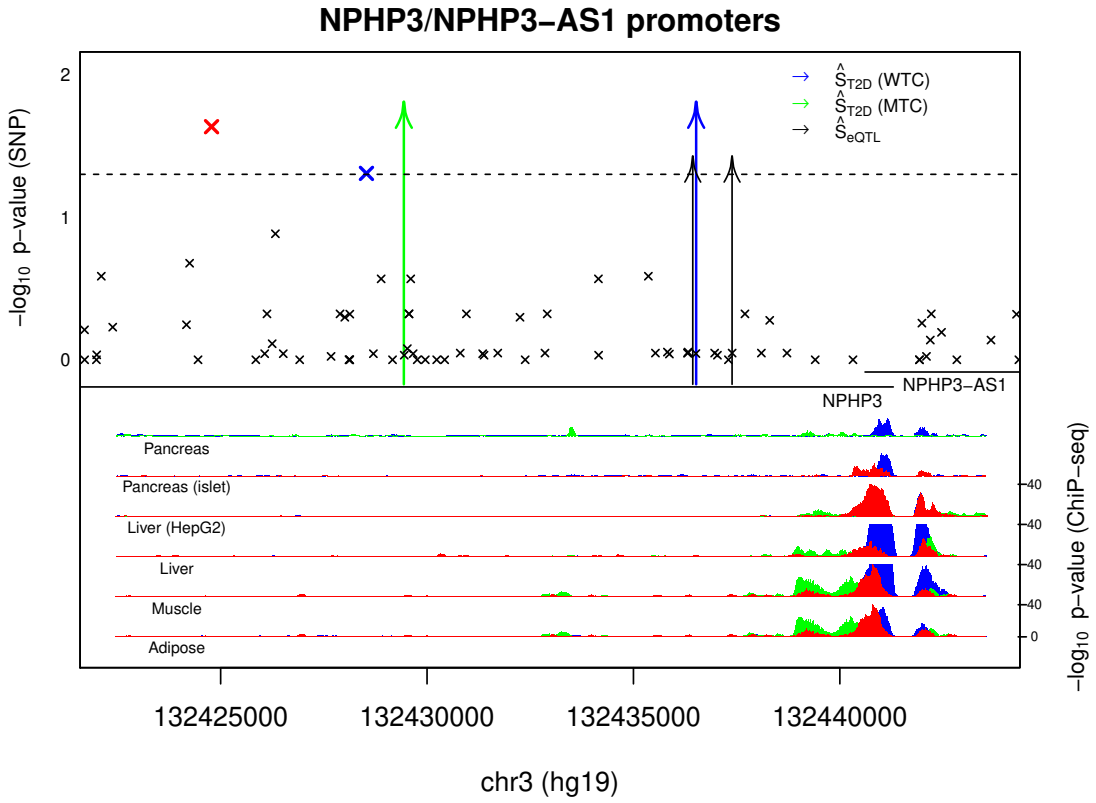


Figure 42: ROADMAP ChiP-seq data for a ∼21 kb region surrounding the $\hat{S}_{T2D}$ and *NPHP3/NPHP3-AS1* promoters. The nominally significant SNP (rs3860501) and indel (rs138040526) are highlighted as a red and blue cross, respectively. ChiP-seq data for H3K27ac (blue), H3K4me1 (green) and H3K9ac (red) are shown for the ROADMAP datasets: adipose nuclei (E063), skeletal muscle (E108), liver (E066), the liver HepG2 cell line (E118), pancreatic islet (E087) and pancreas (E098).

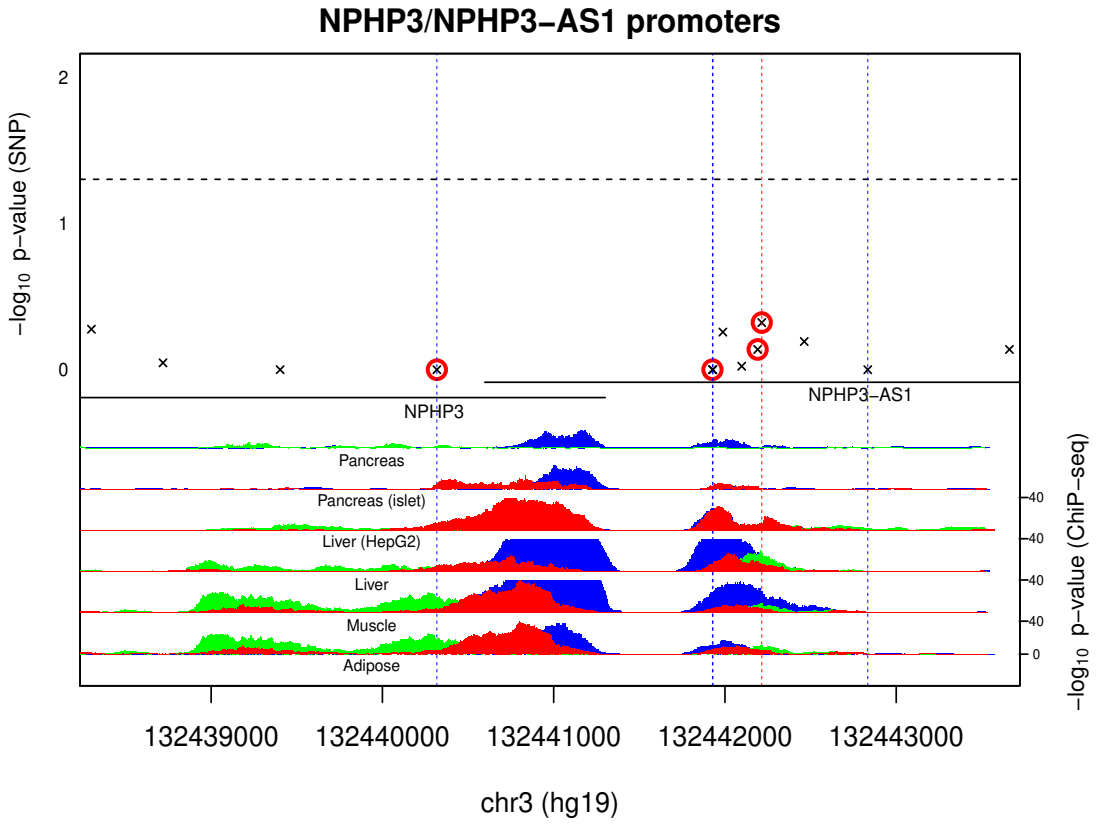


Figure 43: ROADMAP ChiP-seq data for the *NPHP3/NPHP3-AS1* promoter region (∼7 kb). Variants which were predicted to have a 'strong' effect on transcription factor binding by *motifbreakR* are circled in red. Case-only and control-only variants are indicated by red and blue dotted red lines, respectively.

5.4 Discussion

This Chapter aimed to fine-map the chr3q22.1 T2D disease locus using targeted sequence data for a cohort of 82 T2D cases and 89 controls and publicly available functional annotation data. The chr3q22.1 locus was observed to be associated with T2D in two independent European cohorts, with genetic LDU maps used to obtain precise location estimates of T2D causal variants ($\hat{S}_{T2D}$) at chr3:132436519 (WTC) and chr3:132429438 (MTC). These are separated by only 7 kb. eQTL mapped for adipose tissue in a cohort of ageing, population-based Europeans returned eQTL estimates ($\hat{S}_{eQTL}$) for five *cis*–genes all located within less than 15 kb: *ACAD11*, *NPHP3, SRPRB, SLCO2A1* and *RYK*. The closest of these was *NPHP3*, with the $\hat{S}_{eQTL}$ reported at the exact same coordinate as the WTC $\hat{S}_{T2D}$.

The causal variant(s) at this locus were hypothesised to either have low-frequency or be

in low LD with marker SNPs, due to the absence of this locus from previous T2D single-SNP GWAS. However, this locus was previously associated with T2D-associated end-stage kidney disease (Guan et al., 2016), suggesting that its detection required the increase in power achieved by the stratification of extreme phenotypes. The two lead SNPs from this study, rs74504809 and rs78174962, which were both low-frequency (MAFs between 2 to 3% in the 1000G Project), were not associated with gene expression in GTEx and did not overlap any active enhancers in the ChromHMM tracks used in this analysis. The association of this region with end-stage kidney disease in T2D is likely due to the altered expression of *NPHP3*, which is reported to cause renal cysts and kidney dysplasia when mutated. To further fine-map the locus, targeted sequence data for a 189.57 kb region covering the $\hat{S}_{T2D}$ estimates and the downstream *NPHP3* and *ACAD11* genes were used to investigate nominally significant, coding, rare and promoter variants. This analysis identified several candidates. The key findings are summarised below.

Two low-frequency haplotypes were identified as potential candidates. The first included two SNPs found in four T2D cases and no controls: rs114923567 and rs75185415. These variants, which were separated by 477bp, gave p-values of 0.05 when tested for association with T2D status, or <0.01 when conditioned on the rs16839460 genotype. Both SNPs overlapped an annotated enhancer, for which significant ChiP-seq peaks for chromatin marks characteristic of enhancers can be seen in Figure 33, particularly for adipose. *motifbreakR* returned nominal evidence of matches for both SNPs to several different TFBSs. These include several candidates of particular biological interest, including those which may be particularly active in adipose (SMAD3/SMAD4). Additional ChiP-seq data should be used to confirm whether the suggested TFs bind at this locus. Further analysis may investigate whether the strength of the ChiP-seq peaks depend on the genotype of these two SNPs, or whether the minor alleles disrupt TF binding or enhancer activity in enhancer reporter assays. The most significant associations of these two SNPs reported in the T2D Knowledge Portal are with BMI (p-value = 0.024 from the GIANT-UK Biobank GWAS) and neither were reported to be eQTL in GTEx.

The second low-frequency haplotype was found in three T2D cases and no controls

and included two candidate mutations: rs41272317, an *ACAD11* splicing mutation and rs73000573, a mutation within an enhancer at the *ACAD11* promoter region predicted to disrupt the binding of COUP-TFII/HNF4α. The binding of COUP-TFII and HNF4α was confirmed using ChiP-seq data, with rs73000573 located in the centre of a peak for both transcription factors (see Figure 38). Functional experiments such as enhancer reporter assays should be carried out to confirm that rs73000573 disrupts HNF4α binding. Alternatively, ChiP-seq may be stratified for individuals with the major and minor alleles to test for differential binding in individuals with the minor allele. Both SNPs are significant eQTL in GTEx and were associated with lower expression of *ACAD11* and *NPHP3* across multiple tissues. This is consistent with the lower expression of *NPHP3* observed in T2D cases. *ACAD11* was not included on any of the genotyping arrays used in the T2D case-control gene expression analysis (see Chapter 4) and may require further investigation using RNA-seq datasets. According to the T2D Knowledge Portal, rs73000573 and rs41272317 were associated with LDL cholesterol levels in the Japanese Biobank GWAS with modest p-values of 2.52e-03 and 1.96e-03, respectively. rs41272317 was also associated with BMI by the GIANT Consortium (p-value = 5.30e-03).

These interesting candidates were both low-frequency haplotypes and therefore were observed in only a few T2D cases; rs114923567 and rs75185415 had a MAF of 1% in the 1000 Genomes cohort, while rs41272317 had a MAF of 1.4% in the ExAc non-Finnish European cohort. An important next step will be to replicate the association of these haplotypes in an independent cohort. Investigating the quality with which these variants can be imputed will also address the prior hypothesis that these SNPs may not have been reported in previous GWAS due to their low frequency inaccurate imputation. Further studies may benefit from directly genotyping these variants in order to obtain accurate estimates of their frequency[77]. Functional studies may provide alternative validation that these variants affect *cis*–gene expression and T2D-related outcomes[78].

[77]However, there are few large case-control cohorts with high-quality whole-genome sequence data available. One approach may be to stratify individuals with the risk haplotypes in large Biobank projects with whole-genome sequence data and investigate any differences between phenotypic measures, such as T2D prevalence, BMI and cholesterol.

[78]If there is evidence to suggest that both haplotypes affect gene expression in a similar way, then their

The common SNP rs16839460 was associated with T2D status with a p-value of 0.026 and a MAF in cases of 10.5% compared to 4.5% in controls. rs16839460 overlapped an annotated enhancer and was predicted by *motifbreakR* to increase the binding of FOXP3. rs16839460 is in high LD with another nominally significant SNP, rs3860501, and neither SNP was significant when conditioned on the other. rs3860501 did not appear to overlap any annotated regulatory elements in the tissues tested, suggesting that rs16839460 may drive the association. Both minor alleles were associated with higher expression of *NPHP3* in GTEx, despite *NPHP3* expression being lower in T2D cases (see Table 15). Interestingly, FOXP3 expression is characteristic of T regulatory (Treg) cells, which are not included in the GTEx database. Treg cells are highly reliant on fatty acid oxidation and are also significantly decreased in T2D. The potential link between rs16839460, *ACAD11* expression and Treg cells may be an interesting area of further study. This may involve gene expression and ChiP-seq data for Treg cells, although it should be considered that Treg cells which reside in different tissues display markedly different gene expression profiles (Niedzielska et al., 2018). FOXP3 was observed to bind just downstream of *ACAD11* (chr3:132261091-132262533) (Sadlon et al., 2010), suggesting that it may be involved in regulating *cis*–genes at this locus. Further work may aim to replicated the association of this variant and to demonstrate that the genotype alters enhancer activity. Cohorts stratified for the reference and risk alleles may be compared to gain measures of inflammation and adiposity, for example.

It should be considered that these candidate variants were prioritised based on genomic annotation obtained from ROADMAP for four tissue types: adipose, skeletal muscle, liver and pancreas/pancreatic islets. Further analysis may consider ChiP-seq and eQTL data from other databases, such as ENCODE and GTEx, to capture enhancers which may differ in activity across samples. On this note, $\hat{S}_{eQTL}$ estimates for the entire GTEx dataset will soon be made available from ongoing work following the Lau et al. (2017) publication. In the current study, the limited selection of tissues types may have prevented the mapping of

combined frequency may be considered. In this dataset the combined haplotypes are present in seven T2D cases compared to zero controls (fisher test p-value = 0.005)

relevant enhancers in other tissues. ChromHMM tracks for multiple tissues could be used to identify the relevant tissue type based on active enhancers overlapping the significant SNPs, for example as carried out by Claussnitzer et al. (2015) to identify the relevant cell-type at the obesity-associated *FTO* region. Importantly, regulatory elements can be tissue-specific and may also be active at different times of development or in response to different stimuli. Therefore is is important to investigate a variety of datasets. This analysis also prioritised genomic annotation characteristic of enhancer elements, however mutations may also prevent the formation of heterochromatin or disrupt Scaffold/matrix attachment regions (S/MARs) to disrupt higher-order DNA structure (Narwade et al., 2019). These alternative mechanisms may be considered in additional analysis.

After fine-mapping, several questions remain:

1. **If there is allelic heterogeneity, do the independent risk alleles have the same functional impact, or might there be distinct molecular mechanisms?** For example, different variants may disrupt different tissue-specific regulatory elements.

2. **If the causal variant(s) regulate multiple *cis*–genes, do they all contribute to T2D, or can they be divided into *cis*–genes which drive the phenotype and co-regulated *cis*–genes which do not directly impact T2D risk?** This scenario may occur where a risk variant disrupts an enhancer which makes contact with multiple genes, such as an enhancer hub, or a chromatin insulator which alters the local DNA structure, for example. In a theoretical scenario where *ACAD11* was the causal gene, the co-regulated *cis*–genes including would not drive T2D risk directly. However, their dysregulation along with *ACAD11* may lead to other co-morbities such as renal-failure (*NPHP3*, causes the Mendelian disease pancreatic-renal hepatic dysplasia) and delayed wound healing (*SLCO2A1*, transports prostaglandins during wound healing).

If both scenarios are considered, independent risk variants at a locus would all be expected to regulate the causal gene, however some variants may also regulate other *cis*–genes. In

this scenario, allelic heterogeneity should be carefully considered since only some variants may increase the risk of co-morbidities. Functional studies may be used to further investigate the causal gene(s), for example by knocking-down individual *cis*–genes *in vitro* or *in vivo*. Alternatively, if a candidate gene is associated with more than one independent $\hat{S}_{T2D}$, then this provides additional evidence that it it is implicated in disease risk.

An additional consideration is the *NPHP3-AS1* transcript. The function of this long non-coding RNA is unclear. There are also several enhancers within the *NPHP3-AS1* region which were not captured in this analysis and two SNPs within *NPHP3-AS1* have been associated with response to antidepressants in Major Depressive Disorder, a known co-morbidity of T2D (Garriock et al., 2010; Bădescu et al., 2016). Further studies should aim to characterise the function of the *NPHP3-AS1* long non-coding RNA and also the reported *NPHP3-ACAD11* read-through transcript.

5.5 Conclusions

LDU-based gene mapping can be used to identify risk loci which are challenging to map using single-SNP association methods. Subsequent fine-mapping with targeted sequence data identified several candidate functional variants at the chr3q22.1 locus, including two low-frequency haplotypes (MAF ∼1%) and one candidate SNP (MAF ∼9.5%). Further study will be needed to replicate these results and functionally validate whether these variants may fail to be correctly imputed, hence providing an explanation to why this locus may have been missed by previous single-SNP GWAS.

References

Abuyassin, B. and Laher, I. (2015). Obesity-linked diabetes in the arab world: a review. *East Mediterr Health J*, 21(6):420–39.

Achari, A. E. and Jain, S. K. (2017). Adiponectin, a therapeutic target for obesity, diabetes, and endothelial dysfunction. *International journal of molecular sciences*, 18(6):1321.

Ackers, I. and Malgor, R. (2018). Interrelationship of canonical and non-canonical wnt signalling pathways in chronic metabolic diseases. *Diabetes and Vascular Disease Research*, 15(1):3–13.

Acosta-Montaño, P. and García-González, V. (2018). Effects of dietary fatty acids in pancreatic beta cell metabolism, implications in homeostasis. *Nutrients*, 10(4):393.

Adams, F. et al. (1856). *The extant works of Aretaeus, the Cappadocian*, volume 27. Sydenham Society.

Adams, S. H., Hoppel, C. L., Lok, K. H., Zhao, L., Wong, S. W., Minkler, P. E., Hwang, D. H., Newman, J. W., and Garvey, W. T. (2009). Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid β-oxidation and altered tricarboxylic acid cycle activity in type 2 diabetic african-american women. *The Journal of nutrition*, 139(6):1073–1081.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249.

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology*, 6(5):361–369.

Ahmad, S., Bannister, C., van der Lee, S. J., Vojinovic, D., Adams, H. H., Ramirez, A., Escott-Price, V., Sims, R., Baker, E., Williams, J., et al. (2018). Disentangling the biological pathways involved in early features of alzheimer's disease in the rotterdam study. *Alzheimer's & Dementia*, 14(7):848–857.

Ahmadian, M., Suh, J. M., Hah, N., Liddle, C., Atkins, A. R., Downes, M., and Evans, R. M. (2013). Pparγ signaling and metabolism: the good, the bad and the future. *Nature medicine*, 19(5):557–566.

Aires, V., Labbé, J., Deckert, V., de Barros, J.-P. P., Boidot, R., Haumont, M., Maquart, G., Le Guern, N., Masson, D., Prost-Camus, E., et al. (2019). Healthy adiposity and extended lifespan in obese mice fed a diet supplemented with a polyphenol-rich plant extract. *Scientific Reports*, 9(1):1–16.

Akey, J. M., Biswas, S., Leek, J. T., and Storey, J. D. (2007). On the design and analysis of gene expression studies in human populations. *Nature genetics*, 39(7):807–808.

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.-P., and Jansen, R. C. (2007). Sequence polymorphisms cause many false cis eqtls. *PloS one*, 2(7):e622.

Ali, A. T., Boehme, L., Carbajosa, G., Seitan, V. C., Small, K. S., and Hodgkinson, A. (2019). Nuclear genetic regulation of the human mitochondrial transcriptome. *eLife*, 8:e41927.

Ali, M. K., Siegel, K. R., Chandrasekar, E., Tandon, N., Montoya, P. A., Mbanya, J.-C., Chan, J., Zhang, P., and Narayan, K. (2017). Diabetes: An update on the pandemic and potential solutions. *Disease Control Priorities*, 5.

Allan, F. N. (1953). The writings of thomas willis, md: diabetes three hundred years ago. *Diabetes*, 2(1):74–78.

Allen, J. F. (2015). Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. *Proceedings of the National Academy of Sciences*, 112(33):10231–10238.

Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M.-R., Lyssenko, V., Tuomi, T., Groop, L., Group, B. S., et al. (2011). Heritability and familiality of type 2 diabetes and related quantitative traits in the botnia study. *Diabetologia*, 54(11):2811.

Almind, K., Bjørbaek, C., Vestergaard, H., Hansen, T., Echwald, S., and Pedersen, O. (1993). Aminoacid polymorphisms of insulin receptor substrate-1 in non-insulin-dependent diabetes mellitus. *The Lancet*, 342(8875):828–832.

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *The American Journal of Human Genetics*, 69(5):936–950.

Altshuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M.-C., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., et al. (2000). The common pparγ pro12ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature genetics*, 26(1):76.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564.

Anderson, E. J., Lustig, M. E., Boyle, K. E., Woodlief, T. L., Kane, D. A., Lin, C.-T., Price, J. W., Kang, L., Rabinovitch, P. S., Szeto, H. H., et al. (2009). Mitochondrial h 2 o 2 emission and cellular redox state link excess fat intake to insulin resistance in both rodents and humans. *The Journal of clinical investigation*, 119(3):573–581.

Applegarth, A. P. and Koneff, A. A. (1946). The effect of alloxan diabetes on the golgi apparatus and mitochondria of the thyroid gland in the rat. *The Anatomical Record*, 96(1):13–21.

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, 112(7):2109–2114.

Ardisson Korat, A. V., Malik, V. S., Furtado, J. D., Sacks, F., Rosner, B., Rexrode, K. M., Willett, W. C., Mozaffarian, D., Hu, F. B., and Sun, Q. (2020). Circulating very-long-chain sfa concentrations are inversely associated with incident type 2 diabetes in us men and women. *The Journal of Nutrition*, 150(2):340–349.

Ashraf, U. M., Sanchez, E. R., and Kumarasamy, S. (2019). Coup-tfii revisited: Its role in metabolic gene regulation. *Steroids*, 141:63–69.

Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P., and Zeggini, E. (2016). Trans-ethnic study design approaches for fine-mapping. *European journal of human genetics*, 24(9):1330–1336.

Association, A. D. et al. (2014). Diagnosis and classification of diabetes mellitus. *Diabetes care*, 37(Supplement 1):S81–S90.

Association, A. D. et al. (2015). 2. classification and diagnosis of diabetes. *Diabetes care*, 38(Supplement 1):S8–S16.

Astiarraga, B., Chueire, V. B., Souza, A. L., Pereira-Moreira, R., Alegre, S. M., Natali, A., Tura, A., Mari, A., Ferrannini, E., and Muscelli, E. (2018). Effects of acute nefa manipulation on incretin-induced insulin secretion in participants with and without type 2 diabetes. *Diabetologia*, 61(8):1829–1837.

Atlas, D. (2015). International diabetes federation. idf diabetes atlas. *Brussels: International Diabetes Federation*.

Attimonelli, M., Catalano, D., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., Santamaria, M., Pesole, G., and Saccone, C. (2002). Mitonuc: a database of nuclear genes coding for mitochondrial proteins. update 2002. *Nucleic Acids Research*, 30(1):172–173.

Avtanski, D., Pavlov, V. A., Tracey, K. J., and Poretsky, L. (2019). Characterization of inflammation and insulin resistance in high-fat diet-induced male c57bl/6j mouse model of obesity. *Animal models and experimental medicine*, 2(4):252–258.

Axelsen, M., Smith, U., Eriksson, J. W., Taskinen, M.-R., and Jansson, P.-A. (1999). Postprandial hypertriglyceridemia and insulin resistance in normoglycemic first-degree relatives of patients with type 2 diabetes. *Annals of internal medicine*, 131(1):27–31.

Bădescu, S., Tătaru, C., Kobylinska, L., Georgescu, E., Zahiu, D., Zăgrean, A., and Zăgrean, L. (2016). The association between diabetes mellitus and depression. *Journal of medicine and life*, 9(2):120.

Bagnati, M., Ogunkolade, B. W., Marshall, C., Tucci, C., Hanna, K., Jones, T. A., Bugliani, M., Nedjai, B., Caton, P. W., Kieswich, J., et al. (2016). Glucolipotoxicity initiates pancreatic β-cell death through tnfr5/cd40-mediated stat1 and nf-κ b activation. *Cell death & disease*, 7(8):e2329–e2329.

Baltzis, D., Eleftheriadou, I., and Veves, A. (2014). Pathogenesis and treatment of impaired wound healing in diabetes mellitus: new insights. *Advances in therapy*, 31(8):817–836.

Barbeira, A., Shah, K. P., Torres, J. M., Wheeler, H. E., Torstenson, E. S., Edwards, T., Garcia, T., Bell, G. I., Nicolae, D., Cox, N. J., et al. (2016). Metaxcan: summary statistics based gene-level association method infers accurate predixcan results. *BioRxiv*, page 045260.

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature genetics*, 38(6):659–662.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi geo: archive for functional genomics data sets-update. *Nucleic acids research*, 41(D1):D991–D995.

Bassot, A., Chauvin, M.-A., Bendridi, N., Ji-Cao, J., Vial, G., Monnier, L., Bartosch, B., Alves, A., Cottet-Rousselle, C., Gouriou, Y., et al. (2019). Regulation of mitochondria-associated membranes (mams) by no/sgc/pkg participates in the control of hepatic insulin response. *Cells*, 8(11):1319.

Bastidas-Ponce, A., Scheibner, K., Lickert, H., and Bakhti, M. (2017). Cellular and molecular mechanisms coordinating pancreas development. *Development*, 144(16):2873–2888.

Batchuluun, B., Al Rijjal, D., Prentice, K. J., Eversley, J. A., Burdett, E., Mohan, H., Bhattacharjee, A., Gunderson, E. P., Liu, Y., and Wheeler, M. B. (2018). Elevated medium-chain acylcarnitines are associated with gestational diabetes mellitus and early progression to type 2 diabetes and induce pancreatic β-cell dysfunction. *Diabetes*, 67(5):885–897.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bell, J. A., Bull, C. J., Gunter, M. J., Carslake, D., Mahajan, A., Smith, G. D., Timpson, N. J., and Vincent, E. E. (2020). Early metabolic features of genetic liability to type 2 diabetes: cohort study with repeated metabolomics across early life. *Diabetes Care*.

Belsare, P. V., Watve, M. G., Ghaskadbi, S. S., Bhat, D. S., Yajnik, C. S., and Jog, M. (2010). Metabolic syndrome: aggression control mechanisms gone out of control. *Medical hypotheses*, 74(3):578–589.

Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A. S., Kwok, P.-Y., Seshagiri, S., and Wall, J. D. (2019). Evaluating the quality of the 1000 genomes project data. *BMC genomics*, 20(1):1–14.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Berglund, J., Pollard, K. S., and Webster, M. T. (2009). Hotspots of biased nucleotide substitutions in human genes. *PLoS biology*, 7(1).

Bergmann, C., Fliegauf, M., Brüchle, N. O., Frank, V., Olbrich, H., Kirschner, J., Schermer, B., Schmedding, I., Kispert, A., Kränzlin, B., et al. (2008). Loss of nephrocystin-3 function can cause embryonic lethality, meckel-gruber-like syndrome, situs inversus, and renal-hepatic-pancreatic dysplasia. *The American Journal of Human Genetics*, 82(4):959–970.

Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., Jonasdottir, A., Magnusson, O. T., Thorsteinsdottir, U., Masson, G., et al. (2016). Multi-nucleotide de novo mutations in humans. *PLoS genetics*, 12(11).

Bian, L., Hanson, R., Muller, Y., Ma, L., Kobes, S., Knowler, W., Bogardus, C., Baier, L., Investigators, M., et al. (2010). Variants in acad10 are associated with type 2 diabetes, insulin resistance and lipid oxidation in pima indians. *Diabetologia*, 53(7):1349–1353.

Billings, L. K. and Florez, J. C. (2010). The genetics of type 2 diabetes: what have we learned from gwas? *Annals of the New York Academy of Sciences*, 1212:59.

Bingham, C. and Hattersley, A. T. (2004). Renal cysts and diabetes syndrome resulting from mutations in hepatocyte nuclear factor-1β. *Nephrology Dialysis Transplantation*, 19(11):2703–2708.

Bjørndal, B., Burri, L., Staalesen, V., Skorve, J., and Berge, R. K. (2011). Different adipose depots: their role in the development of metabolic syndrome and mitochondrial response to hypolipidemic agents. *Journal of obesity*, 2011.

Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K., and Gilad, Y. (2008). Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS genetics*, 4(11).

Bloom, K., Mohsen, A.-W., Karunanidhi, A., El Demellawy, D., Reyes-Múgica, M., Wang, Y., Ghaloul-Gonzalez, L., Otsubo, C., Tobita, K., Muzumdar, R., et al. (2018). Investigating the link of acad10 deficiency to type 2 diabetes mellitus. *Journal of inherited metabolic disease*, 41(1):49–57.

Boden, G., Jadali, F., White, J., Liang, Y., Mozzoli, M., Chen, X., Coleman, E., and Smith, C. (1991). Effects of fat on insulin-stimulated carbohydrate metabolism in normal men. *The Journal of clinical investigation*, 88(3):960–966.

Boden, G. and Shulman, G. (2002). Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and β-cell dysfunction. *European journal of clinical investigation*, 32:14–23.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661.

Bonnefond, A. and Froguel, P. (2015). Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell metabolism*, 21(3):357–368.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*, 33(3):228–237.

Boudina, S. and Graham, T. E. (2014). Mitochondrial function/dysfunction in white adipose tissue. *Experimental physiology*, 99(9):1168–1178.

Boushel, R., Gnaiger, E., Schjerling, P., Skovbro, M., Kraunsøe, R., and Dela, F. (2007). Patients with type 2 diabetes have normal mitochondrial function in skeletal muscle. *Diabetologia*, 50(4):790–796.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

Brancati, F. L., Kao, W. L., Folsom, A. R., Watson, R. L., and Szklo, M. (2000). Incident type 2 diabetes mellitus in african american and white adults: the atherosclerosis risk in communities study. *Jama*, 283(17):2253–2259.

Burgner, D. and Hull, J. (2000). In defence of genetic association studies. *The Lancet*, 356(9229):599.

Busfield, F., Duffy, D. L., Kesting, J. B., Walker, S. M., Lovelock, P. K., Good, D., Tate, H., Watego, D., Marczak, M., Hayman, N., et al. (2002). A genomewide search for type 2 diabetes–susceptibility genes in indigenous australians. *The American Journal of Human Genetics*, 70(2):349–357.

Calvo, S., Clauser, K., and Mootha, V. (2015a). Mitocarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research*, 44(D1):D1251–D1257.

Calvo, S. E., Clauser, K. R., and Mootha, V. K. (2015b). Mitocarta2. 0: an updated inventory of mammalian mitochondrial proteins. *Nucleic acids research*, 44(D1):D1251–D1257.

Calvo, S. E., Clauser, K. R., and Mootha, V. K. (2016). Mitocarta2. 0: an updated inventory of mammalian mitochondrial proteins. *Nucleic acids research*, 44(D1):D1251–D1257.

Camões, F., Islinger, M., Guimarães, S. C., Kilaru, S., Schuster, M., Godinho, L. F., Steinberg, G., and Schrader, M. (2015). New insights into the peroxisomal protein inventory: Acyl-coa oxidases anddehydrogenases are an ancient feature of peroxisomes. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1853(1):111–125.

Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.*, 9:403–433.

Cao, G., Liang, Y., Broderick, C. L., Oldham, B. A., Beyer, T. P., Schmidt, R. J., Zhang, Y., Stayrook, K. R., Suen, C., Otto, K. A., et al. (2003). Antidiabetic action of a liver x receptor agonist mediated by inhibition of hepatic gluconeogenesis. *Journal of Biological Chemistry*, 278(2):1131–1136.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics*, 22(3):231.

Carlson, C. S., Eberle, M. A., Kruglyak, L., and Nickerson, D. A. (2004a). Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004b). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics*, 74(1):106–120.

Carlson, M. (2019). *org.Hs.eg.db: Genome wide annotation for Human.* R package version 3.8.2.

Carvalho, B. S. and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–7.

Caspers, S., Röckner, M. E., Jockwitz, C., Bittner, N., Teumer, A., Herms, S., Hoffmann, P., Nöthen, M. M., Moebus, S., Amunts, K., et al. (2020). Pathway-specific genetic risk for alzheimer's disease differentiates regional patterns of cortical atrophy in older adults. *Cerebral Cortex*, 30(2):801–811.

Castellano, D., Eyre-Walker, A., and Munch, K. (2020). Impact of mutation rate and selection at linked sites on dna variation across the genomes of humans and other homininae. *Genome biology and evolution*, 12(1):3550–3561.

Cauchi, S., El Achhab, Y., Choquet, H., Dina, C., Krempler, F., Weitgasser, R., Nejjari, C., Patsch, W., Chikri, M., Meyre, D., et al. (2007). Tcf7l2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *Journal of molecular medicine*, 85(7):777–782.

Cebola, I. (2019). Pancreatic islet transcriptional enhancers and diabetes. *Current diabetes reports*, 19(12):145.

Cedikova, M., Kripnerová, M., Dvorakova, J., Pitule, P., Grundmanova, M., Babuska, V., Mullerova, D., and Kuncova, J. (2016). Mitochondria in white, brown, and beige adipocytes. *Stem Cells International*, 2016.

Chae, S., Kim, S.-J., Do Koo, Y., Lee, J. H., Kim, H., Ahn, B. Y., Ha, Y.-C., Kim, Y.-H., Jang, M. G., Koo, K.-H., et al. (2018). A mitochondrial proteome profile indicative of type 2 diabetes mellitus in skeletal muscles. *Experimental & molecular medicine*, 50(9):129.

Chakravarti, A. (1999). Population genetics—making sense out of sequence. *Nature genetics*, 21(1s):56.

Chakravarti, A., Buetow, K. H., Antonarakis, S., Waber, P., Boehm, C., and Kazazian, H. (1984). Nonuniform recombination within the human beta-globin gene cluster. *American journal of human genetics*, 36(6):1239.

Charles, B. A., Shriner, D., and Rotimi, C. N. (2014). Accounting for linkage disequilibrium in association analysis of diverse populations. *Genetic epidemiology*, 38(3):265–273.

Chasman, D. I., Giulianini, F., Demler, O. V., and Udler, M. S. (2020). Pleiotropy-based decomposition of genetic risk scores: Association and interaction analysis for type 2 diabetes and cad. *The American Journal of Human Genetics*.

Chen, H., Zhou, W., Ruan, Y., Yang, L., Xu, N., Chen, R., Yang, R., Sun, J., and Zhang, Z. (2018a). Reversal of angiotensin ll-induced β-cell dedifferentiation via inhibition of nf-κb signaling. *Molecular Medicine*, 24(1):43.

Chen, J., Sun, M., Adeyemo, A., Pirie, F., Carstensen, T., Pomilla, C., Doumatey, A. P., Chen, G., Young, E. H., Sandhu, M., et al. (2019a). Genome-wide association study of type 2 diabetes in africa. *Diabetologia*, pages 1–8.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018b). fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890.

Chen, Y., Colello, J., Jarjour, W., and Zheng, S. G. (2019b). Cellular metabolic regulation in the differentiation and function of regulatory t cells. *Cells*, 8(2):188.

Chen, Y., Zhu, J., Lum, P. Y., Yang, X., Pinto, S., MacNeil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., et al. (2008). Variations in dna elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435.

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., and Hall, I. M. (2015). Speedseq: ultra-fast personal genome analysis and interpretation. *Nature methods*, 12(10):966–968.

Choi, J., Choi, J.-Y., Lee, S.-A., Lee, K.-M., Shin, A., Oh, J., Park, J., Song, M., Yang, J. J., Lee, J.-k., et al. (2019). Association between family history of diabetes and clusters of adherence to healthy behaviors: cross-sectional results from the health examinees-gem (hexa-g) study. *BMJ open*, 9(6):e025477.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl_1):i84–i90.

Chondronikola, M., Volpi, E., Børsheim, E., Porter, C., Annamalai, P., Enerbäck, S., Lidell, M. E., Saraf, M. K., Labbe, S. M., Hurren, N. M., et al. (2014). Brown adipose tissue improves whole-body glucose homeostasis and insulin sensitivity in humans. *Diabetes*, 63(12):4089–4099.

Chuang, J. H. and Li, H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS biology*, 2(2).

Cirillo, E., Kutmon, M., Hernandez, M. G., Hooimeijer, T., Adriaens, M. E., Eijssen, L. M., Parnell, L. D., Coort, S. L., and Evelo, C. T. (2018). From snps to pathways: Biological interpretation of type 2 diabetes (t2dm) genome wide association study (gwas) results. *PloS one*, 13(4):e0193515.

Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., Fath, D. M., Sandoval, E., Isaksson, M., Schlauch, K. A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nature communications*, 11(1):1–10.

Clarke, G. M., Carter, K. W., Palmer, L. J., Morris, A. P., and Cardon, L. R. (2007). Fine mapping versus replication in whole-genome association studies. *The American Journal of Human Genetics*, 81(5):995–1005.

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., et al. (2015). Fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907.

Clissold, R. L., Hamilton, A. J., Hattersley, A. T., Ellard, S., and Bingham, C. (2015). Hnf1b-associated renal and extra-renal disease-an expanding clinical spectrum. *Nature Reviews Nephrology*, 11(2):102.

Colas, R., Sassolas, A., Guichardant, M., Cugnet-Anceau, C., Moret, M., Moulin, P., Lagarde, M., and Calzada, C. (2011). Ldl from obese patients with the metabolic syndrome show increased lipid peroxidation and activate platelets. *Diabetologia*, 54(11):2931.

Cole, L. W. (2016). The evolution of per-cell organelle number. *Frontiers in cell and developmental biology*, 4:85.

Colhoun, H. M., McKeigue, P. M., and Smith, G. D. (2003). Problems of reporting genetic associations with complex outcomes. *The Lancet*, 361(9360):865–872.

Collins, A., Ennis, S., Taillon-Miller, P., Kwok, P.-Y., and Morton, N. (2001). Allelic association with snps: metrics, populations, and the linkage disequilibrium map. *Human Mutation*, 17(4):255–262.

Collins, A., Lau, W., and Francisco, M. (2004). Mapping genes for common diseases: the case for genetic (ld) maps. *Human heredity*, 58(1):2–9.

Collins, A. and Morton, N. (1998). Mapping a disease locus by allelic association. *Proceedings of the National Academy of Sciences*, 95(4):1741–1745.

Collins, A. R. (2007). Linkage disequilibrium and association mapping. In *Linkage Disequilibrium and Association Mapping*, pages 1–15. Springer.

Collins, F. S. (1992). Positional cloning: let's not call it reverse anymore. *Nature genetics*, 1(1):3–6.

Collins, F. S., Guyer, M. S., and Chakravarti, A. (1997). Variations on a theme: cataloging human dna sequence variation. *Science*, 278(5343):1580–1581.

Commerford, S. R., Vargas, L., Dorfman, S. E., Mitro, N., Rocheford, E. C., Mak, P. A., Li, X., Kennedy, P., Mullarkey, T. L., and Saez, E. (2007). Dissection of the insulin-sensitizing effect of liver x receptor ligands. *Molecular Endocrinology*, 21(12):3002–3012.

Condorelli, R. A., La Vignera, S., Mongioì, L. M., Alamo, A., and Calogero, A. E. (2018). Diabetes mellitus and infertility: different pathophysiological effects in type 1 and type 2 on sperm function. *Frontiers in endocrinology*, 9:268.

Consortium, . G. P. et al. (2010a). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061.

Consortium, . G. P. et al. (2012a). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56.

Consortium, . G. P. et al. (2015a). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Consortium, E. P. et al. (2012b). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.

Consortium, G. et al. (2015b). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204.

Consortium, I. H. . et al. (2010b). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52.

Consortium, I. H. et al. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299.

Consortium, I. H. et al. (2007a). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851.

Consortium, W. T. C. C. et al. (2007b). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184.

Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1):23–34.

Cooper, R. S., Tayo, B., and Zhu, X. (2008). Genome-wide association studies: implications for multi-ethnic samples. *Human molecular genetics*, 17(R2):R151–R155.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications*, 1:131.

Crow, M., Lim, N., Ballouz, S., Pavlidis, P., and Gillis, J. (2019). Predictability of human differential gene expression. *Proceedings of the National Academy of Sciences*, 116(13):6491–6500.

Cypess, A. M., Lehman, S., Williams, G., Tal, I., Rodman, D., Goldfine, A. B., Kuo, F. C., Palmer, E. L., Tseng, Y.-H., Doria, A., et al. (2009). Identification and importance of brown adipose tissue in adult humans. *New England Journal of Medicine*, 360(15):1509–1517.

Cyranka, M., Veprik, A., McKay, E. J., van Loon, N., Thijsse, A., Cotter, L., Hare, N., Saibudeen, A., Lingam, S., Pires, E., et al. (2019). Abcc5 knockout mice have lower fat mass and increased levels of circulating glp-1. *Obesity*, 27(8):1292–1304.

Czech, M. P. (2020). Mechanisms of insulin resistance related to white, beige, and brown adipocytes. *Molecular Metabolism*, 195:79–150.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–232.

Damkondwar, D. R., Raman, R., Suganeswari, G., Kulothungan, V., and Sharma, T. (2012). Assessing framingham cardiovascular risk scores in subjects with diabetes and their correlation with diabetic retinopathy. *Indian journal of ophthalmology*, 60(1):45.

Dan, X., Babbar, M., Moore, A., Wechter, N., Tian, J., Mohanty, J. G., Croteau, D. L., and Bohr, V. A. (2020). Dna damage invokes mitophagy through a pathway involving spata18. *Nucleic Acids Research*.

Davis, S. and Meltzer, P. S. (2007). Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847.

De Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, 17(R2):R122–R128.

De Knijff, P., Dekker, J., Stolk, R., Nijpels, G., van der Does, F., Ruige, J., Grobbee, D., Heine, R., Maassen, J., et al. (1999). Variants in the sulphonylurea receptor gene: association of the exon 16–3t variant with type ii diabetes mellitus in dutch caucasians. *Diabetologia*, 42(5):617–620.

De Nadal, E., Ammerer, G., and Posas, F. (2011). Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12):833–845.

De Pauw, A., Tejerina, S., Raes, M., Keijer, J., and Arnould, T. (2009). Mitochondrial (dys) function in adipocyte (de) differentiation and systemic metabolic alterations. *The American journal of pathology*, 175(3):927–939.

Debard, C., Laville, M., Berbe, V., Loizon, E., Guillet, C., Morio-Liondore, B., Boirie, Y., and Vidal, H. (2004). Expression of key genes of fatty acid oxidation, including adiponectin receptors, in skeletal muscle of type 2 diabetic patients. *Diabetologia*, 47(5):917–925.

Dennis, J., Shields, B., Henley, W., Jones, A., and Hattersley, A. (2019). Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared to models based on simple clinical features: an evaluation using clinical trial data. *The Lancet*.

Devarshi, P. P., McNabney, S. M., and Henagan, T. M. (2017). Skeletal muscle nucleo-mitochondrial crosstalk in obesity and type 2 diabetes. *International journal of molecular sciences*, 18(4):831.

Deveaud, C., Beauvoit, B., Salin, B., Schaeffer, J., and Rigoulet, M. (2004). Regional differences in oxidative capacity of rat white adipose tissue are linked to the mitochondrial content of mature adipocytes. *Molecular and cellular biochemistry*, 267(1-2):157–166.

Diabetes, U. (2014). The cost of diabetes report. *London: Diabetes UK*.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS biology*, 8(1).

Direk, K., Cecelja, M., Astle, W., Chowienczyk, P., Spector, T. D., Falchi, M., and Andrew, T. (2013). The relationship between dxa-based and anthropometric measures of visceral fat and morbidity in women. *BMC cardiovascular disorders*, 13(1):25.

Direk, K., Lau, W., Small, K. S., Maniatis, N., and Andrew, T. (2014). Abcc5 transporter is a novel type 2 diabetes susceptibility gene in european and african american populations. *Annals of Human Genetics*, 78(5):333–344.

Dittmann, A., Kennedy, N. J., Soltero, N. L., Morshed, N., Mana, M. D., Yilmaz, Ö. H., Davis, R. J., and White, F. M. (2019). High-fat diet in a mouse insulin-resistant model induces widespread rewiring of the phosphotyrosine signaling network. *Molecular Systems Biology*, 15(8):e8849.

Dizier, M.-H., Demenais, F., and Mathieu, F. (2017). Gain of power of the general regression model compared to cochran-armitage trend tests: simulation study and application to bipolar disorder. *BMC genetics*, 18(1):24.

Dominguez, V., Raimondi, C., Somanath, S., Bugliani, M., Loder, M. K., Edling, C. E., Divecha, N., da Silva-Xavier, G., Marselli, L., Persaud, S. J., et al. (2011). Class ii phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic β cells. *Journal of Biological Chemistry*, 286(6):4216–4225.

Doyle, M. J. and Sussel, L. (2007). Nkx2. 2 regulates β-cell function in the mature islet. *Diabetes*, 56(8):1999–2007.

Drew, B. G., Ribas, V., Le, J. A., Henstridge, D. C., Phun, J., Zhou, Z., Soleymani, T., Daraei, P., Sitz, D., Vergnes, L., et al. (2014). Hsp72 is a mitochondrial stress sensor critical for parkin action, oxidative metabolism, and insulin sensitivity in skeletal muscle. *Diabetes*, 63(5):1488–1505.

Duggirala, R., Blangero, J., Almasy, L., Dyer, T. D., Williams, K. L., Leach, R. J., O'Connell, P., and Stern, M. P. (1999). Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in mexican americans. *The American Journal of Human Genetics*, 64(4):1127–1140.

Dumas, J.-F., Simard, G., Flamment, M., Ducluzeau, P.-H., and Ritz, P. (2009). Is skeletal muscle mitochondrial dysfunction a cause or an indirect consequence of insulin resistance in humans? *Diabetes & metabolism*, 35(3):159–167.

Dunning, M., Lynch, A., and Eldridge, M. (2015). *illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3)*. R package version 1.26.0.

Duret, L. and Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS genetics*, 4(5).

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

Edwards, C. M. and Cusi, K. (2016). Prediabetes: a worldwide epidemic. *Endocrinology and Metabolism Clinics*, 45(4):751–764.

Elding, H., Lau, W., Swallow, D. M., and Maniatis, N. (2013). Refinement in localization and identification of gene regions associated with crohn disease. *The American Journal of Human Genetics*, 92(1):107–113.

Ellis, J. M., Mentock, S. M., DePetrillo, M. A., Koves, T. R., Sen, S., Watkins, S. M., Muoio, D. M., Cline, G. W., Taegtmeyer, H., Shulman, G. I., et al. (2011). Mouse cardiac acyl coenzyme a synthetase 1 deficiency impairs fatty acid oxidation and induces cardiac hypertrophy. *Molecular and cellular biology*, 31(6):1252–1262.

Elmansy, D. and Koyutürk, M. (2019). Cross-population analysis for functional characterization of type ii diabetes variants. *BMC bioinformatics*, 20(12):320.

Elsner, M., Gehrmann, W., and Lenzen, S. (2011). Peroxisome-generated hydrogen peroxide as important mediator of lipotoxicity in insulin-producing cells. *Diabetes*, 60(1):200–208.

Engreitz, J. M., Ollikainen, N., and Guttman, M. (2016). Long non-coding rnas: spatial amplifiers that control nuclear structure and gene expression. *Nature Reviews Molecular Cell Biology*, 17(12):756.

Erion, D. M. and Shulman, G. I. (2010). Diacylglycerol-mediated insulin resistance. *Nature medicine*, 16(4):400–402.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216.

Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome research*, 18(4):653–660.

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in) famous gwas p-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8):1202–1205.

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C.-Y., Lopes-Ramos, C. M., Glass, K., Quackenbush, J., and Platig, J. (2017). Exploring regulation in tissues with eqtl networks. *Proceedings of the National Academy of Sciences*, 114(37):E7841–E7850.

Falconer, D. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics*, 31(1):1–20.

Falconer, D. and Mackay, T. (1996). Introduction to quantitative genetics. 1996. *Harlow, Essex, UK: Longmans Green*, 3.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.

Fazakerley, D. J., Minard, A. Y., Krycer, J. R., Thomas, K. C., Stöckli, J., Harney, D. J., Burchfield, J. G., Maghzal, G. J., Caldwell, S. T., Hartley, R. C., et al. (2018). Mitochondrial oxidative stress causes insulin resistance without disrupting oxidative phosphorylation. *Journal of Biological Chemistry*, 293(19):7315–7328.

Fenske, R., Weeks, A., Brill, A., Nall, R., Pabitch, S., Punt, M., Daniels, M., Blaha, S., Davis, D. B., and Kimple, M. (2017). Prostaglandin e2 (pge2) levels as a predictor of type 2 diabetes control in human subjects: A cross-sectional view of initial cohort study data. *The FASEB Journal*, 31(1_supplement):675–6.

Fernández-Tajes, J., Gaulton, K. J., van de Bunt, M., Torres, J., Thurner, M., Mahajan, A., Gloyn, A. L., Lage, K., and McCarthy, M. I. (2019). Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data. *Genome medicine*, 11(1):19.

Fernández-Vizarra, E., Enríquez, J. A., Pérez-Martos, A., Montoya, J., and Fernández-Silva, P. (2011). Tissue-specific differences in mitochondrial activity and biogenesis. *Mitochondrion*, 11(1):207–213.

Fex, M., Nicholas, L. M., Vishnu, N., Medina, A., Sharoyko, V. V., Nicholls, D. G., Spégel, P., and Mulder, H. (2018). The pathogenetic role of β-cell mitochondria in type 2 diabetes. *Journal of Endocrinology*, 236(3):R145–R159.

Fiskerstrand, T., Houge, G., Sund, S., Scheie, D., Leh, S., Boman, H., and Knappskog, P. M. (2010). Identification of a gene for renal-hepatic-pancreatic dysplasia by microarray-based homozygosity mapping. *The Journal of Molecular Diagnostics*, 12(1):125–131.

Flanagan, S. E., De Franco, E., Allen, H. L., Zerah, M., Abdul-Rasoul, M. M., Edge, J. A., Stewart, H., Alamiri, E., Hussain, K., Wallis, S., et al. (2014). Analysis of transcription factors key for mouse pancreatic development establishes nkx2-2 and mnx1 mutations as causes of neonatal diabetes in man. *Cell metabolism*, 19(1):146–154.

Flannick, J. (2019). The contribution of low-frequency and rare coding variation to susceptibility to type 2 diabetes. *Current diabetes reports*, 19(5):25.

Flannick, J., Mercader, J. M., Fuchsberger, C., Udler, M. S., Mahajan, A., Wessel, J., Teslovich, T. M., Caulkins, L., Koesterer, R., Barajas-Olmos, F., et al. (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*, 570(7759):71–76.

Flaquer, A. and Strauch, K. (2012). A comparison of different linkage statistics in small to moderate sized pedigrees with complex diseases. *BMC research notes*, 5(1):411.

Fletcher, B., Gulanick, M., and Lamendola, C. (2002). Risk factors for type 2 diabetes mellitus. *Journal of Cardiovascular Nursing*, 16(2):17–23.

Fliszkiewicz, M., Niemczyk, M., Kulesza, A., Łabuś, A., and Paczek, L. (2019). Glucose and lipid metabolism abnormalities among patients with autosomal dominant polycystic kidney disease. *Kidney and Blood Pressure Research*, 44(6):1416–1422.

Floegel, A., Stefan, N., Yu, Z., Mühlenbruch, K., Drogan, D., Joost, H.-G., Fritsche, A., Häring, H.-U., de Angelis, M. H., Peters, A., et al. (2013). Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes*, 62(2):639–648.

Flores-Guerrero, J. L., Osté, M. C., Kieneker, L. M., Gruppen, E. G., Wolak-Dinsmore, J., Otvos, J. D., Connelly, M. A., Bakker, S. J., and Dullaart, R. P. (2018). Plasma branched-chain amino acids and risk of incident type 2 diabetes: results from the prevend prospective cohort study. *Journal of clinical medicine*, 7(12):513.

Floyd, B. J., Wilkerson, E. M., Veling, M. T., Minogue, C. E., Xia, C., Beebe, E. T., Wrobel, R. L., Cho, H., Kremer, L. S., Alston, C. L., et al. (2016). Mitochondrial protein interaction mapping identifies regulators of respiratory chain function. *Molecular cell*, 63(4):621–632.

Forner, F., Foster, L. J., Campanaro, S., Valle, G., and Mann, M. (2006). Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Molecular & Cellular Proteomics*, 5(4):608–619.

Forouhi, N. G., Koulman, A., Sharp, S. J., Imamura, F., Kröger, J., Schulze, M. B., Crowe, F. L., Huerta, J. M., Guevara, M., Beulens, J. W., et al. (2014). Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the epic-interact case-cohort study. *The lancet Diabetes & endocrinology*, 2(10):810–818.

Foulquier, S., Daskalopoulos, E. P., Lluri, G., Hermans, K. C., Deb, A., and Blankesteijn, W. M. (2018). Wnt signaling in cardiac and vascular disease. *Pharmacological reviews*, 70(1):68–141.

Freeman, T. M., Wang, D., Harris, J., Ambrose, J. C., Arumugam, P., Baple, E. L., Bleda, M., Boardman-Pretty, F., Boissiere, J. M., Boustred, C. R., et al. (2020). Genomic loci susceptible to systematic sequencing bias in clinical whole genomes. *Genome Research*, 30(3):415–426.

Freese, J., Klement, R. J., Ruiz-Núñez, B., Schwarz, S., and Lötzerich, H. (2017). The sedentary (r) evolution: Have we lost our metabolic flexibility? *F1000Research*, 6.

Freimer, N. and Sabatti, C. (2004). The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nature genetics*, 36(10):1045–1051.

Freire-Pritchett, P., Schoenfelder, S., Varnai, C., Wingett, S. W., Cairns, J., Collier, A. J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C. S., Fraser, P., et al. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife*, 6:e21926.

Fretts, A. M., Imamura, F., Marklund, M., Micha, R., Wu, J. H., Murphy, R. A., Chien, K.-L., McKnight, B., Tintle, N., Forouhi, N. G., et al. (2019). Associations of circulating very-long-chain saturated fatty acids and incident type 2 diabetes: a pooled analysis of prospective cohort studies. *The American journal of clinical nutrition*, 109(4):1216–1223.

Fridlyand, L. E. and Philipson, L. H. (2010). Glucose sensing in the pancreatic beta cell: a computational systems analysis. *Theoretical Biology and Medical Modelling*, 7(1):15.

Fu, J., Festen, E. A., and Wijmenga, C. (2011). Multi-ethnic studies in complex traits. *Human molecular genetics*, 20(R2):R206–R213.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., et al. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41.

Fujimaki, S. and Kuwabara, T. (2017). Diabetes-induced dysfunction of mitochondria and stem cells in skeletal muscle and the nervous system. *International Journal of Molecular Sciences*, 18(10):2147.

Fulga, T. A., Sinning, I., Dobberstein, B., and Pool, M. R. (2001). Srβ coordinates signal sequence release from srp with ribosome binding to the translocon. *The EMBO journal*, 20(9):2338–2347.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956–967.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091.

Gambano, G., Anglani, F., and D'Angelo, A. (2000). Association studies of genetic polymorphisms and complex disease. *The Lancet*, 355(9200):308–311.

Gambineri, A., Patton, L., Altieri, P., Pagotto, U., Pizzi, C., Manzoli, L., and Pasquali, R. (2012). Polycystic ovary syndrome is a risk factor for type 2 diabetes: results from a long-term prospective study. *Diabetes*, 61(9):2369–2374.

Gannon, N. P., Schnuck, J. K., and Vaughan, R. A. (2018). Bcaa metabolism and insulin sensitivity–dysregulated by metabolic status? *Molecular nutrition & food research*, 62(6):1700756.

Gao, H., Kerr, A., Jiao, H., Hon, C.-C., Rydén, M., Dahlman, I., and Arner, P. (2018). Long non-coding rnas associated with metabolic traits in human white adipose tissue. *EBioMedicine*, 30:248–260.

Gao, M., Ding, D., Huang, J., Qu, Y., Wang, Y., and Huang, Q. (2013). Association of genetic variants in the adiponectin gene with metabolic syndrome: a case-control study and a systematic meta-analysis in the chinese population. *PloS one*, 8(4):e58412.

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. *Frontiers in physiology*, 6:383.

Garriock, H. A., Kraft, J. B., Shyn, S. I., Peters, E. J., Yokoyama, J. S., Jenkins, G. D., Reinalda, M. S., Slager, S. L., McGrath, P. J., and Hamilton, S. P. (2010). A genomewide association study of citalopram response in major depressive disorder. *Biological psychiatry*, 67(2):133–138.

Gaucher, S. P., Taylor, S. W., Fahy, E., Zhang, B., Warnock, D. E., Ghosh, S. S., and Gibson, B. W. (2004). Expanded coverage of the human heart mitochondrial proteome using multidimensional liquid chromatography coupled with tandem mass spectrometry. *Journal of proteome research*, 3(3):495–505.

Gaulton, K. J., Ferreira, T., Lee, Y., Raimondo, A., Mägi, R., Reschen, M. E., Mahajan, A., Locke, A., Rayner, N. W., Robertson, N., et al. (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics*, 47(12):1415.

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. *Nature genetics*, 42(3):255.

Gauthier, B. and Wollheim, C. (2006). The nuclear factor tfam restores mitochondrial dna and glucose-stimulated insulin secretion in islets deficient pdxl function: 1579-p. *Diabetes*, 55.

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421.

Gemmill, C. L. (1972). The greek concept of diabetes. *Bulletin of the New York Academy of Medicine*, 48(8):1033.

Gerdes, J. M., Christou-Savina, S., Xiong, Y., Moede, T., Moruzzi, N., Karlsson-Edlund, P., Leibiger, B., Leibiger, I. B., Östenson, C.-G., Beales, P. L., et al. (2014). Ciliary dysfunction impairs beta-cell

insulin secretion and promotes development of type 2 diabetes in rodents. *Nature communications*, 5(1):1–13.

Gerin, I., Dolinsky, V. W., Shackman, J. G., Kennedy, R. T., Chiang, S.-H., Burant, C. F., Steffensen, K. R., Gustafsson, J.-Å., and MacDougald, O. A. (2005). Lxrβ is required for adipocyte growth, glucose homeostasis, and β cell function. *Journal of Biological Chemistry*, 280(24):23024–23031.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5).

Gibson, G. (2010). Hints of hidden heritability in gwas. *Nature genetics*, 42(7):558–560.

Gibson, J., Tapper, W., Ennis, S., and Collins, A. (2013). Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. *Human genetics*, 132(2):233–243.

Gibson, J., Tapper, W., Zhang, W., Morton, N., and Collins, A. (2005). Cosmopolitan linkage disequilibrium maps. *Human Genomics*, 2(1):20.

Giral, H., Landmesser, U., and Kratzer, A. (2018). Into the wild: Gwas exploration of non-coding rnas. *Frontiers in cardiovascular medicine*, 5.

Godin, N. and Eichler, J. (2017). The mitochondrial protein atlas: A database of experimentally verified information on the human mitochondrial proteome. *Journal of Computational Biology*, 24(9):906–916.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Goldstein, D. B. et al. (2009). Common genetic variation and human traits. *New England journal of medicine*, 360(17):1696.

Gonzalez-Becerra, K., Ramos-Lopez, O., Barron-Cabrera, E., Riezu-Boj, J., Milagro, F., Martinez-Lopez, E., and Martínez, J. (2019). Fatty acids, epigenetic mechanisms and chronic diseases: a systematic review. *Lipids in health and disease*, 18(1):178.

Gonzalez-Franquesa, A. and Patti, M.-E. (2017). Insulin resistance and mitochondrial dysfunction. In *Mitochondrial Dynamics in Cardiovascular Medicine*, pages 465–520. Springer.

Gonzalez-Hurtado, E., Lee, J., Choi, J., and Wolfgang, M. J. (2018). Fatty acid oxidation is required for active and quiescent brown adipose tissue maintenance and thermogenic programing. *Molecular metabolism*, 7:45–56.

Gonzlez, J. R., Armengol, L., Guin, E., Sol, X., and Moreno, V. (2014). *SNPassoc: SNPs-based whole genome association studies*. R package version 1.9-2.

Gordon, D., Finch, S. J., Nothnagel, M., and Ott, J. (2002). Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Human heredity*, 54(1):22–33.

Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. (2006). Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. *Nature genetics*, 38(3):320.

Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor perspectives in biology*, 4(9):a011403.

Green, C. R., Wallace, M., Divakaruni, A. S., Phillips, S. A., Murphy, A. N., Ciaraldi, T. P., and Metallo, C. M. (2016). Branched-chain amino acid catabolism fuels adipocyte differentiation and lipogenesis. *Nature chemical biology*, 12(1):15–21.

Groop, L. and Pociot, F. (2014). Genetics of diabetes–are we missing the genes or the disease? *Molecular and cellular endocrinology*, 382(1):726–739.

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089.

Guan, M., Ma, J., Keaton, J. M., Dimitrov, L., Mudgal, P., Stromberg, M., Bonomo, J. A., Hicks, P. J., Freedman, B. I., Bowden, D. W., et al. (2016). Association of kidney structure-related gene variants with type 2 diabetes-attributed end-stage kidney disease in african americans. *Human genetics*, 135(11):1251–1262.

Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C. B., Martínez-González, M. A., Salas-Salvadó, J., and Hu, F. B. (2016). Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. *Diabetes care*, 39(5):833–846.

Guda, K., Fink, S. P., Milne, G. L., Molyneaux, N., Ravi, L., Lewis, S. M., Dannenberg, A. J., Montgomery, C. G., Zhang, S., Willis, J., et al. (2014). Inactivating mutation in the prostaglandin transporter gene, slco2a1, associated with familial digital clubbing, colon neoplasia, and nsaid resistance. *Cancer prevention research*, 7(8):805–812.

Guilherme, A., Virbasius, J. V., Puri, V., and Czech, M. P. (2008). Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nature reviews Molecular cell biology*, 9(5):367–377.

Guinot, F., Szafranski, M., Ambroise, C., and Samson, F. (2018). Learning the optimal scale for gwas through hierarchical snp aggregation. *BMC bioinformatics*, 19(1):1–14.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B. M., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, 50(4):538.

Ha, C. Y., Kim, J. Y., Paik, J. K., Kim, O. Y., Paik, Y.-H., Lee, E. J., and Lee, J. H. (2012). The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. *Clinical endocrinology*, 76(5):674–682.

Hackinger, S. and Zeggini, E. (2017). Statistical methods to detect pleiotropy in human complex traits. *Open biology*, 7(11):170125.

Haghikia, A., Jörg, S., Duscha, A., Berg, J., Manzel, A., Waschbisch, A., Hammer, A., Lee, D.-H., May, C., Wilck, N., et al. (2015). Dietary fatty acids directly impact central nervous system autoimmunity via the small intestine. *Immunity*, 43(4):817–829.

Hales, C. N. and Barker, D. J. (1992). Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia*, 35(7):595–601.

Hales, C. N. and Barker, D. J. (2001). The thrifty phenotype hypothesis: Type 2 diabetes. *British medical bulletin*, 60(1):5–20.

Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425):eaau1043.

Hanson, R. L., Bogardus, C., Duggan, D., Kobes, S., Knowlton, M., Infante, A. M., Marovich, L., Benitez, D., Baier, L. J., and Knowler, W. C. (2007). A search for variants associated with young-onset type 2 diabetes in american indians in a 100k genotyping array. *Diabetes*, 56(12):3045–3052.

Hapmap, C. (2003). The international hapmap project: The international hapmap consortium. *Nature*, 426:789–796.

Hardy, O. T., Perugini, R. A., Nicoloro, S. M., Gallagher-Dorval, K., Puri, V., Straubhaar, J., and Czech, M. P. (2011). Body mass index-independent inflammation in omental adipose tissue associated with insulin resistance in morbid obesity. *Surgery for Obesity and Related Diseases*, 7(1):60–67.

Harlan, L. C., Harlan, W. R., Landis, J. R., and Goldstein, N. G. (1987). Factors associated with glucose tolerance in adults in the united states. *American journal of epidemiology*, 126(4):674–684.

Harlemon, M., Ajayi, O., Kachambwa, P., Kim, M. S., Simonti, C. N., Quiver, M. H., Petersen, D. C., Mittal, A., Fernandez, P., Hsing, A. W., et al. (2019). A custom genotyping array reveals population-level heterogeneity for the genetic risks of prostate cancer and other cancers in africa. *bioRxiv*, page 702910.

Hatchi, E., Rodier, G., Lacroix, M., Caramel, J., Kirsh, O., Jacquet, C., Schrepfer, E., Lagarrigue, S., Linares, L. K., Lledo, G., et al. (2011). E4f1 deficiency results in oxidative stress–mediated cell death of leukemic cells. *Journal of Experimental Medicine*, 208(7):1403–1417.

Hattersley, A. T. and McCarthy, M. I. (2005). What makes a good genetic association study? *The Lancet*, 366(9493):1315–1323.

Hattersley, A. T. and Patel, K. A. (2017). Precision diabetes: learning from monogenic diabetes. *Diabetologia*, 60(5):769–777.

Hayeck, T. J., Stong, N., Wolock, C. J., Copeland, B., Kamalakaran, S., Goldstein, D. B., and Allen, A. S. (2019). Improved pathogenic variant localization via a hierarchical model of sub-regional intolerance. *The American Journal of Human Genetics*, 104(2):299–309.

Haythorne, E., Rohm, M., van de Bunt, M., Brereton, M. F., Tarasov, A. I., Blacker, T. S., Sachse, G., dos Santos, M. S., Exposito, R. T., Davis, S., et al. (2019). Diabetes causes marked inhibition of mitochondrial metabolism in pancreatic β-cells. *Nature communications*, 10(1):1–17.

He, M., Pei, Z., Mohsen, A.-W., Watkins, P., Murdoch, G., Van Veldhoven, P. P., Ensenauer, R., and Vockley, J. (2011). Identification and characterization of new long chain acyl-coa dehydrogenases. *Molecular genetics and metabolism*, 102(4):418–429.

Heaton, G. M., Wagenvoord, R. J., Kemp Jr, A., and Nicholls, D. G. (1978). brown-adipose-tissue mitochondria: photoaffinity labelling of the regulatory site of energy dissipation. *European Journal of Biochemistry*, 82(2):515–521.

Heilbronn, L. K., Gan, S. K., Turner, N., Campbell, L. V., and Chisholm, D. J. (2007). Markers of mitochondrial biogenesis and metabolism are lower in overweight and obese insulin-resistant subjects. *The Journal of Clinical Endocrinology & Metabolism*, 92(4):1467–1473.

Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S., and Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale? *Genome research*, 15(9):1222–1231.

Hemminki, K., Li, X., Sundquist, K., and Sundquist, J. (2010). Familial risks for type 2 diabetes in sweden. *Diabetes care*, 33(2):293–297.

Henstridge, D. C., Bruce, C. R., Drew, B. G., Tory, K., Kolonics, A., Estevez, E., Chung, J., Watson, N., Gardner, T., Lee-Young, R. S., et al. (2014). Activating hsp72 in rodent skeletal muscle increases mitochondrial number and oxidative capacity and decreases insulin resistance. *Diabetes*, 63(6):1881–1894.

Hermann-Kleiter, N. and Baier, G. (2014). Orphan nuclear receptor nr2f6 acts as an essential gatekeeper of th17 cd4+ t cell effector functions. *Cell Communication and Signaling*, 12(1):38.

Hermann-Kleiter, N., Gruber, T., Lutz-Nicoladoni, C., Thuille, N., Fresser, F., Labi, V., Schiefermeier, N., Warnecke, M., Huber, L., Villunger, A., et al. (2008). The nuclear orphan receptor nr2f6 suppresses lymphocyte activation and t helper 17-dependent autoimmunity. *Immunity*, 29(2):205–216.

Herold, M., Breuer, J., Hucke, S., Knolle, P., Schwab, N., Wiendl, H., and Klotz, L. (2017). Liver x receptor activation promotes differentiation of regulatory t cells. *PLoS One*, 12(9).

Hesselink, M. K., Schrauwen-Hinderling, V., and Schrauwen, P. (2016). Skeletal muscle mitochondria as a target to prevent or treat type 2 diabetes mellitus. *Nature reviews endocrinology*, 12(11):633.

Hex, N., Bartlett, C., Wright, D., Taylor, M., and Varley, D. (2012). Estimating the current and future costs of type 1 and type 2 diabetes in the uk, including direct health costs and indirect societal and productivity costs. *Diabetic medicine*, 29(7):855–862.

Heydemann, A. (2016). An overview of murine high fat diet as a model for type 2 diabetes mellitus. *Journal of diabetes research*, 2016.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. (2005). Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079.

Hirschhorn, J. N. and Altshuler, D. (2002). Once and again—issues surrounding replication in genetic association studies.

Hirschhorn, J. N. et al. (2009). Genomewide association studies–illuminating biologic pathways. *New England Journal of Medicine*, 360(17):1699.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, 4(2):45–61.

Hivert, M.-F., Manning, A. K., McAteer, J. B., Florez, J. C., Dupuis, J., Fox, C. S., O'Donnell, C. J., Cupples, L. A., and Meigs, J. B. (2008). Common variants in the adiponectin gene (adipoq) associated with plasma adiponectin levels, type 2 diabetes, and diabetes-related quantitative traits: the framingham offspring study. *Diabetes*, 57(12):3353–3359.

Hoeks, J., van Herpen, N. A., Mensink, M., Moonen-Kornips, E., van Beurden, D., Hesselink, M. K., and Schrauwen, P. (2010). Prolonged fasting identifies skeletal muscle mitochondrial dysfunction as consequence rather than cause of human insulin resistance. *Diabetes*, 59(9):2117–2125.

Holloway, G. P., Thrush, A. B., Heigenhauser, G. J., Tandon, N. N., Dyck, D. J., Bonen, A., and Spriet, L. L. (2007). Skeletal muscle mitochondrial fat/cd36 content and palmitate oxidation are not decreased in obese women. *American Journal of Physiology-Endocrinology and Metabolism*, 292(6):E1782–E1789.

Homko, C. J., Cheung, P., and Boden, G. (2003). Effects of free fatty acids on glucose uptake and utilization in healthy women. *Diabetes*, 52(2):487–491.

Hong, M.-G., Karlsson, R., Magnusson, P. K., Lewis, M. R., Isaacs, W., Zheng, L. S., Xu, J., Grönberg, H., Ingelsson, E., Pawitan, Y., et al. (2013). A genome-wide assessment of variability in human serum metabolism. *Human mutation*, 34(3):515–524.

Hormozdiari, F., Van De Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260.

Houten, S. M. and Wanders, R. J. (2010). A general introduction to the biochemistry of mitochondrial fatty acid β-oxidation. *Journal of inherited metabolic disease*, 33(5):469–477.

Howie, D., Cobbold, S. P., Adams, E., Ten Bokum, A., Necula, A. S., Zhang, W., Huang, H., Roberts, D. J., Thomas, B., Hester, S. S., et al. (2017). Foxp3 drives oxidative phosphorylation and protection from lipotoxicity. *JCI insight*, 2(3).

Hu, X., Pan, X., Ma, X., Luo, Y., Xu, Y., Xiong, Q., Xiao, Y., Bao, Y., and Jia, W. (2016). Contribution of a first-degree family history of diabetes to increased serum adipocyte fatty acid binding protein levels independent of body fat content and distribution. *International Journal of Obesity*, 40(11):1649–1654.

Huang, B., Amos, C., and Lin, D. (2007). Detecting haplotype effects in genomewide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(8):803–812.

Huang, J., Ellinghaus, D., Franke, A., Howie, B., and Li, Y. (2012). 1000 genomes-based imputation identifies novel and refined associations for the wellcome trust case control consortium phase 1 data. *European Journal of Human Genetics*, 20(7):801–805.

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250.

Huang, L., Lin, J.-s., Aris, I. M., Yang, G., Chen, W.-Q., and Li, L.-J. (2019). Circulating saturated fatty acids and incident type 2 diabetes: A systematic review and meta-analysis. *Nutrients*, 11(5):998.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2011). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291.

Hung, V., Zou, P., Rhee, H.-W., Udeshi, N. D., Cracan, V., Svinkina, T., Carr, S. A., Mootha, V. K., and Ting, A. Y. (2014). Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric apex tagging. *Molecular cell*, 55(2):332–341.

Hurrle, S. and Hsu, W. H. (2017). The etiology of oxidative stress in insulin resistance. *biomedical journal*, 40(5):257–262.

Huxtable, S. J., Saker, P. J., Haddad, L., Walker, M., Frayling, T. M., Levy, J. C., Hitman, G. A., O'Rahilly, S., Hattersley, A. T., and McCarthy, M. I. (2000). Analysis of parent-offspring trios provides evidence for linkage and association between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class iii variable number tandem repeat alleles. *Diabetes*, 49(1):126–130.

Imamura, F., Fretts, A., Marklund, M., Ardisson Korat, A. V., Yang, W.-S., Lankinen, M., Qureshi, W., Helmer, C., Chen, T.-A., Wong, K., et al. (2018). Fatty acid biomarkers of dairy fat consumption and incidence of type 2 diabetes: a pooled analysis of prospective cohort studies. *PLoS medicine*, 15(10):e1002670.

Imamura, M. and Maeda, S. (2011). Genetics of type 2 diabetes: the gwas era and future perspectives. *Endocrine journal*, pages 1107190592–1107190592.

Inaba, K. and Mizuno, K. (2016). Sperm dysfunction and ciliopathy. *Reproductive medicine and biology*, 15(2):77–94.

Ingman, M. and Gyllensten, U. (2006). mtdb: Human mitochondrial genome database, a resource for population genetics and medical sciences. *Nucleic acids research*, 34(suppl_1):D749–D751.

Inoue, H., Ferrer, J., Welling, C. M., Elbein, S. C., Hoffman, M., Mayorga, R., Warren-Perry, M., Zhang, Y., Millns, H., Turner, R., et al. (1996). Sequence variants in the sulfonylurea receptor (sur) gene are associated with niddm in caucasians. *Diabetes*, 45(6):825–831.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature genetics*, 29(3):306–309.

Irizarry, R. (2017). Lowering the gwas threshold would save millions of dollars.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

IS Sobczak, A., A Blindauer, C., and J Stewart, A. (2019). Changes in plasma free fatty acids associated with type-2 diabetes. *Nutrients*, 11(9):2022.

Islinger, M., Lüers, G. H., Li, K. W., Loos, M., and Völkl, A. (2007). Rat liver peroxisomes after fibrate treatment a survey using quantitative mass spectrometry. *Journal of Biological Chemistry*, 282(32):23055–23069.

Itahana, Y. and Itahana, K. (2018). Emerging roles of p53 family members in glucose metabolism. *International journal of molecular sciences*, 19(3):776.

Iyengar, S. K. and Elston, R. C. (2007). The genetic basis of complex traits. In *Linkage Disequilibrium and Association Mapping*, pages 71–84. Springer.

Jabalameli, M., Vergara Lope Gracia, N., Horscroft, C., Ennis, S., Collins, A., Pengelly, R., et al. (2019). Whole-genome linkage disequilibrium maps for european and african populations. *Scientific Data*, 6(208).

Jacob, S., Machann, J., Rett, K., Brechtel, K., Volk, A., Renn, W., Maerker, E., Matthaei, S., Schick, F., Claussen, C.-D., et al. (1999). Association of increased intramyocellular lipid content with insulin resistance in lean nondiabetic offspring of type 2 diabetic subjects. *Diabetes*, 48(5):1113–1119.

Jadhav, B., Wild, K., Pool, M. R., and Sinning, I. (2015). Structure and switch cycle of srβ as ancestral eukaryotic gtpase associated with secretory membranes. *Structure*, 23(10):1838–1847.

James, W., Johnson, R., Speakman, J., Wallace, D., Frühbeck, G., Iversen, P., and Stover, P. (2019). Nutrition and its role in human evolution. *Journal of internal medicine*, 285(5):533–549.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature genetics*, 29(2):217–222.

Jeffreys, A. J. and Neumann, R. (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature genetics*, 31(3):267–271.

Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S., and Donnelly, P. (2005). Human recombination hot spots hidden in regions of strong marker association. *Nature genetics*, 37(6):601–606.

Jheng, H.-F., Tsai, P.-J., Guo, S.-M., Kuo, L.-H., Chang, C.-S., Su, I.-J., Chang, C.-R., and Tsai, Y.-S. (2012). Mitochondrial fission contributes to mitochondrial dysfunction and insulin resistance in skeletal muscle. *Molecular and cellular biology*, 32(2):309–319.

Jiang, D., LaGory, E. L., Brož, D. K., Bieging, K. T., Brady, C. A., Link, N., Abrams, J. M., Giaccia, A. J., and Attardi, L. D. (2015). Analysis of p53 transactivation domain mutants reveals acad11 as a metabolic target important for p53 pro-survival function. *Cell reports*, 10(7):1096–1109.

Jin, W., Goldfine, A. B., Boes, T., Henry, R. R., Ciaraldi, T. P., Kim, E.-Y., Emecan, M., Fitzpatrick, C., Sen, A., Shah, A., et al. (2011). Increased srf transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance. *The Journal of clinical investigation*, 121(3):918–929.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nature genetics*, 29(2):233–237.

Jornayvaz, F. R. and Shulman, G. I. (2012). Diacylglycerol activation of protein kinase cε and hepatic insulin resistance. *Cell metabolism*, 15(5):574–584.

Joseph, J. W., Jensen, M. V., Ilkayeva, O., Palmieri, F., Alárcon, C., Rhodes, C. J., and Newgard, C. B. (2006). The mitochondrial citrate/isocitrate carrier plays a regulatory role in glucose-stimulated insulin secretion. *Journal of Biological Chemistry*, 281(47):35624–35632.

Jyothi, K. U. and Reddy, B. M. (2015). Gene–gene and gene–environment interactions in the etiology of type 2 diabetes mellitus in the population of hyderabad, india. *Meta gene*, 5:9–20.

Kaler, A. S. and Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC genomics*, 20(1):618.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in kegg. *Nucleic acids research*, 47(D1):D590–D595.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462.

Kaneto, H., Katakami, N., Matsuhisa, M., and Matsuoka, T.-a. (2010). Role of reactive oxygen species in the progression of type 2 diabetes and atherosclerosis. *Mediators of inflammation*, 2010.

Kang, H. M., Ahn, S. H., Choi, P., Ko, Y.-A., Han, S. H., Chinga, F., Park, A. S. D., Tao, J., Sharma, K., Pullman, J., et al. (2015). Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. *Nature medicine*, 21(1):37–46.

Kang, J. T. and Rosenberg, N. A. (2019). Mathematical properties of linkage disequilibrium statistics defined by normalization of the coefficient d= pab–papb. *Human Heredity*, 84(3):127–143.

Kang, S. J., Gordon, D., and Finch, S. J. (2004). What snp genotyping errors are most costly for genetic association studies? *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 26(2):132–141.

Kao, P. Y., Leung, K. H., Chan, L. W., Yip, S. P., and Yap, M. K. (2017). Pathway analysis of complex diseases for gwas, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(2):335–353.

Kappler, L., Hoene, M., Hu, C., von Toerne, C., Li, J., Bleher, D., Hoffmann, C., Böhm, A., Kollipara, L., Zischka, H., et al. (2019). Linking bioenergetic function of mitochondria to tissue-specific molecular fingerprints. *American Journal of Physiology-Endocrinology and Metabolism*, 317(2):E374–E387.

Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J., Stengård, J., and Kesäniemi, Y. (1992). Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in finland. *Diabetologia*, 35(11):1060–1067.

Karisa, B., Thomson, J., Wang, Z., Stothard, P., Moore, S., and Plastow, G. (2013). Candidate genes and single nucleotide polymorphisms associated with variation in residual feed intake in beef cattle. *Journal of animal science*, 91(8):3502–3513.

Katic, M., Kennedy, A. R., Leykin, I., Norris, A., McGettrick, A., Gesta, S., Russell, S. J., Bluher, M., Maratos-Flier, E., and Kahn, C. R. (2007). Mitochondrial gene expression and increased oxidative metabolism: role in increased lifespan of fat-specific insulin receptor knock-out mice. *Aging cell*, 6(6):827–839.

Kato, N. (2012). Ethnic diversity in type 2 diabetes genetics between east asians and europeans. *Journal of diabetes investigation*, 3(4):349.

Keaton, J. M., Gao, C., Guan, M., Hellwege, J. N., Palmer, N. D., Pankow, J. S., Fornage, M., Wilson, J. G., Correa, A., Rasmussen-Torvik, L. J., et al. (2018). Genome-wide interaction with the insulin secretion locus mtnr1b reveals cmip as a novel type 2 diabetes susceptibility gene in african americans. *Genetic epidemiology*, 42(6):559–570.

Keats, E. C., Dominguez, J. M., Grant, M. B., and Khan, Z. A. (2014). Switch from canonical to noncanonical wnt signaling mediates high glucose-induced adipogenesis. *Stem Cells*, 32(6):1649–1660.

Keckesova, Z., Donaher, J. L., De Cock, J., Freinkman, E., Lingrell, S., Bachovchin, D. A., Bierie, B., Tischler, V., Noske, A., Okondo, M. C., et al. (2017). Lactb is a tumour suppressor that modulates lipid metabolism and cell state. *Nature*, 543(7647):681–686.

Keightley, P. D. and Otto, S. P. (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92.

Keller, P., Gburcik, V., Petrovic, N., Gallagher, I. J., Nedergaard, J., Cannon, B., and Timmons, J. A. (2011). Gene-chip studies of adipogenesis-regulated microrna s in mouse primary adipocytes and human obesity. *BMC endocrine disorders*, 11(1):7.

Kelley, D. E., He, J., Menshikova, E. V., and Ritov, V. B. (2002). Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes*, 51(10):2944–2950.

Kempkes, R. W., Joosten, I., Koenen, H. J., and He, X. (2019). Metabolic pathways involved in regulatory t cell functionality. *Frontiers in Immunology*, 10:2839.

Kennedy, E. D., Maechler, P., and Wollheim, C. B. (1998). Effects of depletion of mitochondrial dna in metabolism secretion coupling in ins-1 cells. *Diabetes*, 47(3):374–380.

Khamis, A., Canouil, M., Siddiq, A., Crouch, H., Falchi, M., von Bulow, M., Ehehalt, F., Marselli, L., Distler, M., Richter, D., et al. (2019). Laser capture microdissection of human pancreatic islets reveals novel eqtls associated with type 2 diabetes. *Molecular metabolism*.

Khan, S. and Wang, C. H. (2014). Er stress in adipocytes and insulin resistance: mechanisms and significance. *Molecular medicine reports*, 10(5):2234–2240.

Khil, P. and Camerini-Otero, R. (2009). Variation in patterns of human meiotic recombination. In *Meiosis*, volume 5, pages 117–127. Karger Publishers.

Kikuchi, M., Hatano, N., Yokota, S., Shimozawa, N., Imanaka, T., and Taniguchi, H. (2004). Proteomic analysis of rat liver peroxisome presence of peroxisome-specific isozyme of lon protease. *Journal of Biological Chemistry*, 279(1):421–428.

Kim, J.-a., Wei, Y., and Sowers, J. R. (2008). Role of mitochondrial dysfunction in insulin resistance. *Circulation research*, 102(4):401–414.

Kim, M. K., Kwak, S. H., Kang, S., Jung, H. S., Cho, Y. M., Kim, S. Y., and Park, K. S. (2015a). Identification of two cases of ciliopathy-associated diabetes and their mutation analysis using whole exome sequencing. *Diabetes & metabolism journal*, 39(5):439–443.

Kim, O.-K., Jun, W., and Lee, J. (2015b). Mechanism of er stress and inflammation for hepatic insulin resistance in obesity. *Annals of Nutrition and Metabolism*, 67(4):218–227.

Kimple, M. E., Keller, M. P., Rabaglia, M. R., Pasker, R. L., Neuman, J. C., Truchan, N. A., Brar, H. K., and Attie, A. D. (2013). Prostaglandin e2 receptor, ep3, is induced in diabetic islets and negatively regulates glucose-and hormone-stimulated insulin secretion. *Diabetes*, 62(6):1904–1912.

Kirchner, H., Sinha, I., Gao, H., Ruby, M. A., Schönke, M., Lindvall, J. M., Barrès, R., Krook, A., Näslund, E., Dahlman-Wright, K., et al. (2016). Altered dna methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. *Molecular metabolism*, 5(3):171–183.

Knaus, B. J. and Grünwald, N. J. (2016). VcfR: an r package to manipulate and visualize VCF format data. *BioRxiv*.

Ko, J. Y., Oh, S., and Yoo, K. H. (2017). Functional enhancers as master regulators of tissue-specific gene regulation and cancer development. *Molecules and cells*, 40(3):169.

Köbberling, J. and Tattersall, R. (1982). *The genetics of diabetes mellitus*, volume 47. Academic Pr.

Koh, E. H., Park, J.-Y., Park, H.-S., Jeon, M. J., Ryu, J. W., Kim, M., Kim, S. Y., Kim, M.-S., Kim, S.-W., Park, I. S., et al. (2007). Essential role of mitochondrial function in adiponectin synthesis in adipocytes. *Diabetes*, 56(12):2973–2981.

Kolb, H. and Martin, S. (2017). Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC medicine*, 15(1):131.

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., et al. (2018). The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428.

Kong, X., Murphy, K., Raj, T., He, C., White, P., and Matise, T. (2004). A combined linkage-physical map of the human genome. *The American Journal of Human Genetics*, 75(6):1143–1148.

Korach-André, M., Parini, P., Larsson, L., Arner, A., Steffensen, K. R., and Gustafsson, J.-Å. (2010). Separate and overlapping metabolic functions of lxrα and lxrβ in c57bl/6 female mice. *American Journal of Physiology-Endocrinology and Metabolism*, 298(2):E167–E178.

Kothari, V., Luo, Y., Tornabene, T., O'Neill, A. M., Greene, M. W., Geetha, T., and Babu, J. R. (2017). High fat diet induces brain insulin resistance and cognitive impairment in mice. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1863(2):499–508.

Koves, T. R., Ussher, J. R., Noland, R. C., Slentz, D., Mosedale, M., Ilkayeva, O., Bain, J., Stevens, R., Dyck, J. R., Newgard, C. B., et al. (2008). Mitochondrial overload and incomplete fatty acid oxidation contribute to skeletal muscle insulin resistance. *Cell metabolism*, 7(1):45–56.

Kraja, A. T., Liu, C., Fetterman, J. L., Graff, M., Have, C. T., Gu, C., Yanek, L. R., Feitosa, M. F., Arking, D. E., Chasman, D. I., et al. (2019). Associations of mitochondrial and nuclear mitochondrial variants and genes with seven metabolic traits. *The American Journal of Human Genetics*, 104(1):112–138.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics*, 22(2):139–144.

Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nature genetics*, 27(3):234–236.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J. (2013). Hocomoco: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(D1):D195–D202.

Kulyte, A., Ehrlund, A., Arner, P., and Dahlman, I. (2017). Global transcriptome profiling identifies klf15 and slc25a10 as modifiers of adipocytes insulin sensitivity in obese women. *PLoS One*, 12(6):e0178485.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317.

Kung, C.-P. and Murphy, M. E. (2016). The role of the p53 tumor suppressor in metabolism and diabetes. *The Journal of endocrinology*, 231(2):R61.

Kycia, I., Wolford, B. N., Huyghe, J. R., Fuchsberger, C., Vadlamudi, S., Kursawe, R., Welch, R. P., Albanus, R. d., Uyar, A., Khetan, S., et al. (2018). A common type 2 diabetes risk variant potentiates activity of an evolutionarily conserved islet stretch enhancer and increases c2cd4a and c2cd4b expression. *The American Journal of Human Genetics*, 102(4):620–635.

Kyono, Y., Kitzman, J. O., and Parker, S. C. (2019). Genomic annotation of disease-associated variants reveals shared functional contexts. *Diabetologia*, 62(5):735–743.

Lackey, D. E., Lynch, C. J., Olson, K. C., Mostaedi, R., Ali, M., Smith, W. H., Karpe, F., Humphreys, S., Bedinger, D. H., Dunn, T. N., et al. (2013). Regulation of adipose branched-chain amino acid catabolism enzyme expression and cross-adipose amino acid flux in human obesity. *American Journal of Physiology-Endocrinology and Metabolism*, 304(11):E1175–E1187.

Lacroix, M., Rodier, G., Kirsh, O., Houles, T., Delpech, H., Seyran, B., Gayte, L., Casas, F., Pessemesse, L., Heuillet, M., et al. (2016). E4f1 controls a transcriptional program essential for pyruvate dehydrogenase activity. *Proceedings of the National Academy of Sciences*, 113(39):10998–11003.

Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., and Pare, G. (2020). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *bioRxiv*.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

Lane, N. and Martin, W. (2010). The energetics of genome complexity. *Nature*, 467(7318):929–934.

Lang, P., Hasselwander, S., Li, H., and Xia, N. (2019). Effects of different diets used in diet-induced obesity models on insulin resistance and vascular dysfunction in c57bl/6 mice. *Scientific Reports*, 9(1):1–14.

Lang, S., Pfeffer, S., Lee, P.-H., Cavalié, A., Helms, V., Förster, F., and Zimmermann, R. (2017). An update on sec61 channel functions, mechanisms, and related diseases. *Frontiers in physiology*, 8:887.

Las, G., Oliveira, M. F., and Shirihai, O. S. (2020). Emerging roles of β-cell mitochondria in type-2-diabetes. *Molecular Aspects of Medicine*, 71:100843.

Lau, H. H., Ng, N. H. J., Loo, L. S. W., Jasmen, J. B., and Teo, A. K. K. (2018). The molecular functions of hepatocyte nuclear factors–in and beyond the liver. *Journal of hepatology*, 68(5):1033–1048.

Lau, W., Andrew, T., and Maniatis, N. (2017). High-resolution genetic maps identify multiple type 2 diabetes loci at regulatory hotspots in african americans and europeans. *The American Journal of Human Genetics*, 100(5):803–816.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). Lumpy: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6):R84.

Lê, K.-A., Mahurkar, S., Alderete, T. L., Hasson, R. E., Adam, T. C., Kim, J. S., Beale, E., Xie, C., Greenberg, A. S., Allayee, H., et al. (2011). Subcutaneous adipose tissue macrophage infiltration is associated with hepatic and visceral fat deposition, hyperinsulinemia, and stimulation of nf-κb stress pathway. *Diabetes*, 60(11):2802–2809.

Le Cam, L., Linares, L. K., Paul, C., Julien, E., Lacroix, M., Hatchi, E., Triboulet, R., Bossis, G., Shmueli, A., Rodriguez, M. S., et al. (2006). E4f1 is an atypical ubiquitin ligase that modulates p53 effector functions independently of degradation. *Cell*, 127(4):775–788.

Lee, H., Song, J., Jung, J. H., and Ko, H. W. (2015a). Primary cilia in energy balance signaling and metabolic disorder. *BMB reports*, 48(12):647.

Lee, H.-Y., Choi, C. S., Birkenfeld, A. L., Alves, T. C., Jornayvaz, F. R., Jurczak, M. J., Zhang, D., Woo, D. K., Shadel, G. S., Ladiges, W., et al. (2010). Targeted expression of catalase to mitochondria prevents age-associated reductions in mitochondrial function and insulin resistance. *Cell metabolism*, 12(6):668–674.

Lee, J., Choi, J., Alpergin, E. S. S., Zhao, L., Hartung, T., Scafidi, S., Riddle, R. C., and Wolfgang, M. J. (2017). Loss of hepatic mitochondrial long-chain fatty acid oxidation confers resistance to diet-induced obesity and glucose intolerance. *Cell reports*, 20(3):655–667.

Lee, J., Ellis, J. M., and Wolfgang, M. J. (2015b). Adipose fatty acid oxidation is required for thermogenesis and potentiates oxidative stress-induced inflammation. *Cell reports*, 10(2):266–279.

Lee, J. H., Chandrasekar, S., Chung, S., Fu, Y.-H. H., Liu, D., Weiss, S., and Shan, S.-o. (2018). Sequential activation of human signal recognition particle by the ribosome and signal sequence drives efficient protein targeting. *Proceedings of the National Academy of Sciences*, 115(24):E5487–E5496.

Lee, J. Y., Sohn, K. H., Rhee, S. H., and Hwang, D. (2001). Saturated fatty acids, but not unsaturated fatty acids, induce the expression of cyclooxygenase-2 mediated through toll-like receptor 4. *Journal of Biological Chemistry*, 276(20):16683–16689.

Lee, J. Y., Ye, J., Gao, Z., Youn, H. S., Lee, W. H., Zhao, L., Sizemore, N., and Hwang, D. H. (2003). Reciprocal modulation of toll-like receptor-4 signaling pathways involving myd88 and phosphatidylinositol 3-kinase/akt by saturated and polyunsaturated fatty acids. *Journal of Biological Chemistry*, 278(39):37041–37051.

Lee, J. Y., Zhao, L., Youn, H. S., Weatherill, A. R., Tapping, R., Feng, L., Lee, W. H., Fitzgerald, K. A., and Hwang, D. H. (2004). Saturated fatty acid activates but polyunsaturated fatty acid inhibits toll-like receptor 2 dimerized with toll-like receptor 6 or 1. *Journal of Biological Chemistry*, 279(17):16971–16979.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Leeman, K., Dobson, L., Towne, M., Dukhovny, D., Joshi, M., Stoler, J., and Agrawal, P. (2014). Nphp3 mutations are associated with neonatal onset multiorgan polycystic disease in two siblings. *Journal of Perinatology*, 34(5):410–411.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

Lemaitre, R. N., Fretts, A. M., Sitlani, C. M., Biggs, M. L., Mukamal, K., King, I. B., Song, X., Djoussé, L., Siscovick, D. S., McKnight, B., et al. (2015). Plasma phospholipid very-long-chain saturated fatty acids and incident diabetes in older adults: the cardiovascular health study. *The American journal of clinical nutrition*, 101(5):1047–1054.

Lemkin, P., Chipperfield, M., Merril, C., and Zullo, S. (1996). A world wide web (www) server database engine for an organelle database, mitodat. *Electrophoresis*, 17(3):566–72.

Lerin, C., Goldfine, A. B., Boes, T., Liu, M., Kasif, S., Dreyfuss, J. M., De Sousa-Coelho, A. L., Daher, G., Manoli, I., Sysol, J. R., et al. (2016). Defects in muscle branched-chain amino acid oxidation contribute to impaired lipid metabolism. *Molecular metabolism*, 5(10):926–936.

Lettre, G., Lange, C., and Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(4):358–362.

Lewis, C. M. and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12:1–11.

Li, A. and Meyre, D. (2013). Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *International journal of obesity*, 37(4):559–567.

Li, B., Liu, D. J., and Leal, S. M. (2013). Identifying rare variants associated with complex traits via sequencing. *Current protocols in human genetics*, 78(1):1–26.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, S., Datta, S., Brabbit, E., Love, Z., Woytowicz, V., Flattery, K., Capri, J., Yao, K., Wu, S., Imboden, M., et al. (2020). Nr2e3 is a genetic modifier that rescues retinal degeneration and promotes homeostasis in multiple models of retinitis pigmentosa. *Gene Therapy*, pages 1–19.

Li, S., Kvon, E. Z., Visel, A., Pennacchio, L. A., and Ovcharenko, I. (2019). Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation. *Genome biology*, 20(1):140.

Li, Y. R. and Keating, B. J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome medicine*, 6(10):91.

Liang, X., He, A., Wang, W., Liu, L., Du, Y., Fan, Q., Li, P., Wen, Y., Hao, J., Guo, X., et al. (2017). Integrating genome-wide association and eqtls studies identifies the genes and gene sets associated with diabetes. *BioMed research international*, 2017.

Liang, X., Pan, J., Cao, C., Zhang, L., Zhao, Y., Fan, Y., Li, K., Tao, C., and Wang, Y. (2019). Transcriptional response of subcutaneous white adipose tissue to acute cold exposure in mice. *International journal of molecular sciences*, 20(16):3968.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.

Lim, J., Lee, J., Suh, Y. H., Kim, W., Song, J., and Jung, M. (2006). Mitochondrial dysfunction induces aberrant insulin signalling and glucose utilisation in murine c2c12 myotube cells. *Diabetologia*, 49(8):1924–1936.

Lin, H., Wang, M., Brody, J. A., Bis, J. C., Dupuis, J., Lumley, T., McKnight, B., Rice, K. M., Sitlani, C. M., Reid, J. G., et al. (2014). Strategies to design and analyze targeted sequencing data: cohorts for heart and aging research in genomic epidemiology (charge) consortium targeted sequencing study. *Circulation: Cardiovascular Genetics*, 7(3):335–343.

Lin, H.-Y., Weng, S.-W., Chang, Y.-H., Su, Y.-J., Chang, C.-M., Tsai, C.-J., Shen, F.-C., Chuang, J.-H., Lin, T.-K., Liou, C.-W., et al. (2018a). The causal role of mitochondrial dynamics in regulating insulin resistance in diabetes: link through mitochondrial reactive oxygen species. *Oxidative Medicine and Cellular Longevity*, 2018.

Lin, J. and Musunuru, K. (2018). From genotype to phenotype: a primer on the functional follow-up of genome-wide association studies in cardiovascular disease. *Circulation: Genomic and Precision Medicine*, 11(2):e001946.

Lin, J.-s., Dong, H.-l., Chen, G.-d., Chen, Z.-y., Dong, X.-w., Zheng, J.-s., and Chen, Y.-m. (2018b). Erythrocyte saturated fatty acids and incident type 2 diabetes in chinese men and women: A prospective cohort study. *Nutrients*, 10(10):1393.

Ling, C. and Rönn, T. (2019). Epigenetics in human obesity and type 2 diabetes. *Cell metabolism*, 29(5):1028–1044.

Ling, Y., Li, X., Gu, Q., Chen, H., Lu, D., and Gao, X. (2011). Associations of common polymorphisms in gckr with type 2 diabetes and related traits in a han chinese population: a case-control study. *BMC medical genetics*, 12(1):66.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS genetics*, 6(10).

Liu, J., Lu, W., Shi, B., Klein, S., and Su, X. (2019). Peroxisomal regulation of redox homeostasis and adipocyte metabolism. *Redox biology*, 24:101167.

Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J., and Kellis, M. (2017). Evidence of reduced recombination rate in human regulatory domains. *Genome biology*, 18(1):193.

Liu, Z., Benard, O., Syeda, M. M., Schuster, V. L., and Chi, Y. (2015a). Inhibition of prostaglandin transporter (pgt) promotes perfusion and vascularization and accelerates wound healing in non-diabetic and diabetic rats. *PloS one*, 10(7).

Liu, Z., Patil, I. Y., Jiang, T., Sancheti, H., Walsh, J. P., Stiles, B. L., Yin, F., and Cadenas, E. (2015b). High-fat diet induces hepatic insulin resistance and impairment of synaptic plasticity. *PloS one*, 10(5):e0128274.

Lloyd, D. J., Wheeler, M. C., and Gekakis, N. (2010). A point mutation in sec61α1 leads to diabetes and hepatosteatosis in mice. *Diabetes*, 59(2):460–470.

Loh, K., Deng, H., Fukushima, A., Cai, X., Boivin, B., Galic, S., Bruce, C., Shields, B. J., Skiba, B., Ooms, L. M., et al. (2009). Reactive oxygen species enhance insulin sensitivity. *Cell metabolism*, 10(4):260–272.

Lopez, M. F., Kristal, B. S., Chernokalskaya, E., Lazarev, A., Shestopalov, A. I., Bogdanova, A., and Robinson, M. (2000). High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *ELECTROPHORESIS: An International Journal*, 21(16):3427–3440.

López-Cortegano, E. and Caballero, A. (2019). Inferring the nature of missing heritability in human traits using data from the gwas catalog. *Genetics*, 212(3):891–904.

Lotta, L. A., Scott, R. A., Sharp, S. J., Burgess, S., Luan, J., Tillin, T., Schmidt, A. F., Imamura, F., Stewart, I. D., Perry, J. R., et al. (2016). Genetic predisposition to an impaired metabolism of the branched-chain amino acids and risk of type 2 diabetes: a mendelian randomisation analysis. *PLoS medicine*, 13(11):e1002179.

Love-Gregory, L. D., Wasson, J., Ma, J., Jin, C. H., Glaser, B., Suarez, B. K., and Permutt, M. A. (2004). A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4α gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an ashkenazi jewish population. *Diabetes*, 53(4):1134–1140.

Lowell, B. B. and Shulman, G. I. (2005). Mitochondrial dysfunction and type 2 diabetes. *Science*, 307(5708):384–387.

Lu, Y., Wang, Y., Ong, C.-N., Subramaniam, T., Choi, H. W., Yuan, J.-M., Koh, W.-P., and Pan, A. (2016). Metabolic signatures and risk of type 2 diabetes in a chinese population: an untargeted metabolomics study using both lc-ms and gc-ms. *Diabetologia*, 59(11):2349–2359.

Luca, F., Perry, G. H., and Di Rienzo, A. (2010). Evolutionary adaptations to dietary changes. *Annual review of nutrition*, 30:291–314.

Lundsgaard, A.-M., Fritzen, A. M., Nicolaisen, T. S., Carl, C. S., Sjøberg, K. A., Raun, S. H., Klein, A. B., Sanchez-Quant, E., Langer, J., Ørskov, C., et al. (2020). Glucometabolic consequences of acute and prolonged inhibition of fatty acid oxidation. *Journal of lipid research*, 61(1):10–19.

Lupski, J. R., Belmont, J. W., Boerwinkle, E., and Gibbs, R. A. (2011). Clan genomics and the complex architecture of human disease. *Cell*, 147(1):32–43.

Lusa, L., Gentleman, R., and Ruschhaupt, M. (2019). *GeneMeta: MetaAnalysis for High Throughput Experiments*. R package version 1.56.0.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Lyssenko, V., Groop, L., and Prasad, R. B. (2015). Genetics of type 2 diabetes: it matters from which parent we inherit the risk. *The Review of Diabetic Studies: RDS*, 12(3-4):233.

Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., Berglund, G., Altshuler, D., Nilsson, P., and Groop, L. (2008). Clinical risk factors, dna variants, and the development of type 2 diabetes. *New England Journal of Medicine*, 359(21):2220–2232.

Lytrivi, M., Castell, A.-L., Poitout, V., and Cnop, M. (2020). Recent insights into mechanisms of β-cell lipo-and glucolipotoxicity in type 2 diabetes. *Journal of Molecular Biology*, 432(5):1514–1534.

Ma, Q., Wu, X., Wu, J., Liang, Z., and Liu, T. (2017). Serp1 is a novel marker of poor prognosis in pancreatic ductal adenocarcinoma patients via anti-apoptosis and regulating srprb/nf-κb axis. *International journal of oncology*, 51(4):1104–1114.

Ma, Z. A., Zhao, Z., and Turk, J. (2012). Mitochondrial dysfunction and β-cell failure in type 2 diabetes mellitus. *Experimental diabetes research*, 2012.

Maassen, J. A., M't Hart, L., van Essen, E., Heine, R. J., Nijpels, G., Tafrechi, R. S. J., Raap, A. K., Janssen, G. M., and Lemkes, H. H. (2004). Mitochondrial diabetes: molecular mechanisms and clinical presentation. *Diabetes*, 53(suppl 1):S103–S109.

Macfarlane, W. M., Frayling, T. M., Ellard, S., Evans, J. C., Allen, L. I., Bulman, M. P., Ayres, S., Shepherd, M., Clark, P., Millward, A., et al. (2000). Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *The Journal of clinical investigation*, 106(5):717–717.

Mackiewicz, D., de Oliveira, P. M. C., de Oliveira, S. M., and Cebrat, S. (2013). Distribution of recombination hotspots in the human genome–a comparison of computer simulations with real data. *PloS one*, 8(6).

Mahajan, A., Go, M., Zhang, W., Below, J., Gaulton, K., Ferreira, T., Horikoshi, M., Johnson, A., Ng, M., Prokopenko, I., et al. (2014a). Diabetes genetics replication and meta-analysis (diagram) consortium; asian genetic epidemiology network type 2 diabetes (agen-t2d) consortium; south asian type 2 diabetes (sat2d) consortium; mexican american type 2 diabetes (mat2d) consortium; type 2 diabetes genetic exploration by nex-generation sequencing in muylti-ethnic samples (t2d-genes) consortium. genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*, 46(3):234–244.

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C., Prokopenko, I., et al. (2014b). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics*, 46(3):234–244.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics*, 50(11):1505.

Mahdi, T., Hänzelmann, S., Salehi, A., Muhammed, S. J., Reinbothe, T. M., Tang, Y., Axelsson, A. S., Zhou, Y., Jing, X., Almgren, P., et al. (2012). Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. *Cell metabolism*, 16(5):625–633.

Mahendran, Y., Vangipurapu, J., Cederberg, H., Stančáková, A., Pihlajamäki, J., Soininen, P., Kangas, A. J., Paananen, J., Civelek, M., Saleem, N. K., et al. (2013). Association of ketone body levels with hyperglycemia and type 2 diabetes in 9,398 finnish men. *Diabetes*, 62(10):3618–3626.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21.

Mai, M., Tönjes, A., Kovacs, P., Stumvoll, M., Fiedler, G. M., and Leichtle, A. B. (2013). Serum levels of acylcarnitines are altered in prediabetic conditions. *PloS one*, 8(12):e82459.

Majithia, A. R., Flannick, J., Shahinian, P., Guo, M., Bray, M.-A., Fontanillas, P., Gabriel, S. B., JHS, N., Rosen, E. D., Altshuler, D., et al. (2014). Rare variants in pparg with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proceedings of the National Academy of Sciences*, 111(36):13127–13132.

Malécot, G. (1948). Mathématiques de l'hérédité.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294.

Mallet, J., Le Strat, Y., Dubertret, C., and Gorwood, P. (2020). Polygenic risk scores shed light on the relationship between schizophrenia and cognitive functioning: Review and meta-analysis. *Journal of clinical medicine*, 9(2):341.

Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N. E. (2004). Positional cloning by linkage disequilibrium. *The American Journal of Human Genetics*, 74(5):846–855.

Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L., Hewett, D., Tapper, W., Ennis, S., Ke, X., and Morton, N. (2002). The first linkage disequilibrium (ld) maps: delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences*, 99(4):2228–2233.

Manichaikul, A., Wang, X.-Q., Zhao, W., Wojczynski, M. K., Siebenthall, K., Stamatoyannopoulos, J. A., Saleheen, D., Borecki, I. B., Reilly, M. P., Rich, S. S., et al. (2016). Genetic association of long-chain acyl-coa synthetase 1 variants with fasting glucose, diabetes, and subclinical atherosclerosis. *Journal of lipid research*, 57(3):433–442.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

Martin, W. (2003). Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proceedings of the National Academy of Sciences*, 100(15):8612–8614.

Marullo, L., Moustafa, J. S. E.-S., and Prokopenko, I. (2014). Insights into the genetic susceptibility to type 2 diabetes from genome-wide association studies of glycaemic traits. *Current diabetes reports*, 14(11):551.

Marwood, S. (1973). Diabetes mellitus–some reflections. *The Journal of the Royal College of General Practitioners*, 23(126):38.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59.

Matise, T. C., Chen, F., Chen, W., Francisco, M., Hansen, M., He, C., Hyland, F. C., Kennedy, G. C., Kong, X., Murray, S. S., et al. (2007). A second-generation combined linkage–physical map of the human genome. *Genome research*, 17(12):1783–1786.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195.

McBride, H. M., Neuspiel, M., and Wasiak, S. (2006). Mitochondria: more than just a powerhouse. *Current biology*, 16(14):R551–R560.

McCarthy, M. and Menzel, S. (2001). The genetics of type 2 diabetes. *British journal of clinical pharmacology*, 51(3):195.

McCarthy, M. I. (2003). Growing evidence for diabetes susceptibility genes from genome scan data. *Current diabetes reports*, 3(2):159–167.

McCarthy, M. I. (2004). Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. *Human molecular genetics*, 13(suppl_1):R33–R41.

McCarthy, M. I. (2008). Casting a wider net for diabetes susceptibility genes. *Nature genetics*, 40(9):1039–1040.

McCarthy, M. I. (2017). Painting a new picture of personalised medicine for diabetes. *Diabetologia*, 60(5):793–799.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369.

McCarthy, M. I. and Mahajan, A. (2018). The value of genetic risk scores in precision medicine for diabetes.

McCarthy, M. I., Smedley, D., and Hide, W. (2003). New methods for finding disease-susceptibility genes: impact and potential. *Genome biology*, 4(10):119.

McCarthy, M. I. and Zeggini, E. (2009). Genome-wide association studies in type 2 diabetes. *Current diabetes reports*, 9(2):164–171.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279–1283.

McCoin, C. S., Knotts, T. A., and Adams, S. H. (2015). Acylcarnitines-old actors auditioning for new roles in metabolic physiology. *Nature Reviews Endocrinology*, 11(10):617.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

McMullen, P. D., Bhattacharya, S., Woods, C. G., Sun, B., Yarborough, K., Ross, S. M., Miller, M. E., McBride, M. T., LeCluyse, E. L., Clewell, R. A., et al. (2014). A map of the pparα transcription regulatory network for primary human hepatocytes. *Chemico-biological interactions*, 209:14–24.

McVean, G. A. and Charlesworth, B. (2000). The effects of hill-robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, 155(2):929–944.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584.

Medici, F., Hawa, M., Ianari, A., Pyke, D., and Leslie, R. G. (1999). Concordance rate for type ii diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetologia*, 42(2):146–150.

Medina-Gómez, G. (2012). Mitochondria and endocrine function of adipose tissue. *Best Practice & Research Clinical Endocrinology & Metabolism*, 26(6):791–804.

Meigs, J. B., Cupples, L. A., and Wilson, P. (2000). Parental transmission of type 2 diabetes: the framingham offspring study. *Diabetes*, 49(12):2201–2207.

Meldgaard, T., Olesen, S. S., Farmer, A. D., Krogh, K., Wendel, A. A., Brock, B., Drewes, A. M., and Brock, C. (2018). Diabetic enteropathy: from molecule to mechanism-based treatment. *Journal of diabetes research*, 2018.

Menezes, L. F., Lin, C.-C., Zhou, F., and Germino, G. G. (2016). Fatty acid oxidation is impaired in an orthologous mouse model of autosomal dominant polycystic kidney disease. *EBioMedicine*, 5:183–192.

Mermet, J., Yeung, J., Hurni, C., Mauvoisin, D., Gustafson, K., Jouffe, C., Nicolas, D., Emmenegger, Y., Gobet, C., Franken, P., et al. (2018). Clock-dependent chromatin topology modulates circadian transcription and behavior. *Genes & development*, 32(5-6):347–358.

Mermet, J., Yeung, J., and Naef, F. (2017). Systems chronobiology: global analysis of gene regulation in a 24-hour periodic world. *Cold Spring Harbor perspectives in biology*, 9(3):a028720.

Meyerovich, K., Ortis, F., and Cardozo, A. K. (2018). The non-canonical nf-κb pathway and its contribution to β-cell failure in diabetes. *Journal of molecular endocrinology*, 61(2):F1–F6.

Meyre, D., Lecoeur, C., Delplanque, J., Francke, S., Vatin, V., Durand, E., Weill, J., Dina, C., and Froguel, P. (2004). A genome-wide scan for childhood obesity–associated traits in french families shows significant linkage on chromosome 6q22. 31-q23. 2. *Diabetes*, 53(3):803–811.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B. M., Rolando, D. M., Farabella, I., Morgan, C. C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature genetics*, page 1.

Mihalik, S. J., Goodpaster, B. H., Kelley, D. E., Chace, D. H., Vockley, J., Toledo, F. G., and DeLany, J. P. (2010). Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity. *Obesity*, 18(9):1695–1700.

Miller, J. B. and Colagiuri, S. (1994). The carnivore connection: dietary carbohydrate in the evolution of niddm. *Diabetologia*, 37(12):1280–1286.

Minamino, T., Orimo, M., Shimizu, I., Kunieda, T., Yokoyama, M., Ito, T., Nojima, A., Nabetani, A., Oike, Y., Matsubara, H., et al. (2009). A crucial role for adipose tissue p53 in the regulation of insulin resistance. *Nature medicine*, 15(9):1082.

Mishra, P. and Chan, D. C. (2016). Metabolic regulation of mitochondrial dynamics. *Journal of Cell Biology*, 212(4):379–387.

Möder, M., Kiessling, A., Löster, H., and Brüggemann, L. (2003). The pattern of urinary acylcarnitines determined by electrospray mass spectrometry: a new tool in the diagnosis of diabetes mellitus. *Analytical and bioanalytical chemistry*, 375(2):200–210.

Molina, A. J., Wikstrom, J. D., Stiles, L., Las, G., Mohamed, H., Elorza, A., Walzer, G., Twig, G., Katz, S., Corkey, B. E., et al. (2009). Mitochondrial networking protects β-cells from nutrient-induced apoptosis. *Diabetes*, 58(10):2303–2315.

Monnereau, C., Vogelezang, S., Kruithof, C. J., Jaddoe, V. W., and Felix, J. F. (2016). Associations of genetic risk scores based on adult adiposity pathways with childhood growth and adiposity measures. *BMC genetics*, 17(1):120.

Montgomery, M. K. and Turner, N. (2015). Mitochondrial dysfunction and insulin resistance: an update. *Endocrine connections*, 4(1):R1–R15.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273.

Morgan, T. H. (1911). Random segregation versus coupling in mendelian inheritance. *Science*, 34(873):384–384.

Mori, Y., Otabe, S., Dina, C., Yasuda, K., Populaire, C., Lecoeur, C., Vatin, V., Durand, E., Hara, K., Okada, T., et al. (2002). Genome-wide search for type 2 diabetes in japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate loci on 7p and 11p. *Diabetes*, 51(4):1247–1255.

Morino, K., Petersen, K. F., Dufour, S., Befroy, D., Frattini, J., Shatzkes, N., Neschen, S., White, M. F., Bilz, S., Sono, S., et al. (2005). Reduced mitochondrial density and increased irs-1 serine phosphorylation in muscle of insulin-resistant offspring of type 2 diabetic parents. *The Journal of clinical investigation*, 115(12):3587–3593.

Morino, K., Petersen, K. F., and Shulman, G. I. (2006). Molecular mechanisms of insulin resistance in humans and their potential links with mitochondrial dysfunction.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory rna. *Nature Reviews Genetics*, 15(6):423.

Morris, R. D., Rimm, D. L., Hartz, A. J., Kalkhoff, R. K., and Rimm, A. A. (1989). Obesity and heredity in the etiology of non-insulin-dependent diabetes mellitus in 32,662 adult white women. *American journal of epidemiology*, 130(1):112–121.

Morton, N., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y., and Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences*, 98(9):5217–5221.

Moszyńska, A., Gebert, M., Collawn, J. F., and Bartoszewski, R. (2017). Snps in microrna target sites and their potential role in human disease. *Open biology*, 7(4):170019.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., McVean, G., Boehnke, M., Altshuler, D., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS genetics*, 11(4).

Müller, M., Hernández, M. A. G., Goossens, G. H., Reijnders, D., Holst, J. J., Jocken, J. W., van Eijk, H., Canfora, E. E., and Blaak, E. E. (2019). Circulating but not faecal short-chain fatty acids are related to insulin sensitivity, lipolysis and glp-1 concentrations in humans. *Scientific reports*, 9(1):1–9.

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated dna elements. *Nature genetics*, 49(11):1602.

Muoio, D. M. and Neufer, P. D. (2012). Lipid-induced mitochondrial stress and insulin action in muscle. *Cell metabolism*, 15(5):595–605.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.

Narayan, K. M. V. and Kanaya, A. M. (2020). Why are south asians prone to type 2 diabetes? a hypothesis based on underexplored pathways. *Diabetologia*.

Narwade, N., Patel, S., Alam, A., Chattopadhyay, S., Mittal, S., and Kulkarni, A. (2019). Mapping of scaffold/matrix attachment regions in human genome: a data mining exercise. *Nucleic acids research*, 47(14):7247–7261.

Nasykhova, Y. A., Barbitoff, Y. A., Serebryakova, E. A., Katserov, D. S., and Glotov, A. S. (2019). Recent advances and perspectives in next generation sequencing application to the genetic research of type 2 diabetes. *World journal of diabetes*, 10(7):376.

Neale, B., Ferreira, M., Medland, S., and Posthuma, D. (2007). *Statistical genetics: gene mapping through linkage and association*. Taylor & Francis.

Nedergaard, J., Bengtsson, T., and Cannon, B. (2007). Unexpected evidence for active brown adipose tissue in adult humans. *American Journal of Physiology-Endocrinology and Metabolism*.

Neel, J. V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics*, 14(4):353.

Neuhaus, T., Sennhauser, F., Briner, J., Van Damme, B., and Leumann, E. (1996). Renal-hepatic-pancreatic dysplasia: an autosomal recessive disorder with renal and hepatic failure. *European journal of pediatrics*, 155(9):791–795.

Newman, B., Selby, J., King, M.-C., Slemenda, C., Fabsitz, R., and Friedman, G. (1987). Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*, 30(10):763–768.

Newsholme, P., Gaudel, C., and Krause, M. (2012). Mitochondria and diabetes. an intriguing pathogenetic role. In *Advances in mitochondrial medicine*, pages 235–247. Springer.

Ng, M. C. (2015). Genetics of type 2 diabetes in african americans. *Current diabetes reports*, 15(10):74.

Ng, M. C., Shriner, D., Chen, B. H., Li, J., Chen, W.-M., Guo, X., Liu, J., Bielinski, S. J., Yanek, L. R., Nalls, M. A., et al. (2014). Meta-analysis of genome-wide association studies in african americans provides insights into the genetic architecture of type 2 diabetes. *PLoS genetics*, 10(8).

Ngo, D. T., Sverdlov, A. L., Karki, S., Macartney-Coxson, D., Stubbs, R. S., Farb, M. G., Carmine, B., Hess, D. T., Colucci, W. S., and Gokce, N. (2019). Oxidative modifications of mitochondrial complex ii are associated with insulin resistance of visceral fat in obesity. *American Journal of Physiology-Endocrinology and Metabolism*, 316(2):E168–E177.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., and Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS genetics*, 6(4):e1000895.

Nicastro, H., Da Luz, C. R., Chaves, D. F. S., Bechara, L. R. G., Voltarelli, V. A., Rogero, M. M., and Lancha, A. H. (2012). Does branched-chain amino acids supplementation modulate skeletal muscle remodeling through inflammation modulation? possible mechanisms of action. *Journal of nutrition and metabolism*, 2012.

Nicholas, D. A., Proctor, E. A., Agrawal, M., Belkina, A. C., Van Nostrand, S. C., Panneerseelan-Bharath, L., Jones IV, A. R., Raval, F., Ip, B. C., Zhu, M., et al. (2019). Fatty acid metabolites combine with reduced β oxidation to activate th17 inflammation in human type 2 diabetes. *Cell metabolism*, 30(3):447–461.

Nicolae, D. L. (2016). Association tests for rare variants. *Annual review of genomics and human genetics*, 17:117–130.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4).

Niedzielska, M., Israelsson, E., Angermann, B., Sidders, B. S., Clausen, M., Catley, M., Malhotra, R., and Dumont, C. (2018). Differential gene expression in human tissue resident regulatory t cells from lung, colon, and blood. *Oncotarget*, 9(90):36166.

Noda, M., Yamashita, S., Takahashi, N., Eto, K., Shen, L.-M., Izumi, K., Daniel, S., Tsubamoto, Y., Nemoto, T., Iino, M., et al. (2002). Switch to anaerobic glucose metabolism with nadh accumulation in the β-cell model of mitochondrial diabetes characteristics of βhc9 cells deficient in mitochondrial dna transcription. *Journal of Biological Chemistry*, 277(44):41817–41826.

Ogg, S. C., Barz, W. P., and Walter, P. (1998). A functional gtpase domain, but not its transmembrane domain, is required for function of the srp receptor β-subunit. *The Journal of cell biology*, 142(2):341–354.

Oh, Y. S., Bae, G. D., Baek, D. J., Park, E.-Y., and Jun, H.-S. (2018). Fatty acid-induced lipotoxicity in pancreatic beta-cells during development of type 2 diabetes. *Frontiers in endocrinology*, 9:384.

Olbrich, H., Fliegauf, M., Hoefele, J., Kispert, A., Otto, E., Volz, A., Wolf, M. T., Sasmaz, G., Trauer, U., Reinhardt, R., et al. (2003). Mutations in a novel gene, nphp3, cause adolescent nephronophthisis, tapeto-retinal degeneration and hepatic fibrosis. *Nature genetics*, 34(4):455–459.

Ong, C.-T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293.

Ong, C.-T. and Corces, V. G. (2012). Enhancers: emerging roles in cell fate specification. *EMBO reports*, 13(5):423–430.

Onodera, T., Fukuhara, A., Jang, M. H., Shin, J., Aoi, K., Kikuta, J., Otsuki, M., Ishii, M., and Shimomura, I. (2015). Adipose tissue macrophages induce pparγ-high foxp3+ regulatory t cells. *Scientific reports*, 5:16801.

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284.

Pacheu-Grau, D., Rucktäschel, R., and Deckers, M. (2018). Mitochondrial dysfunction and its role in tissue-specific cellular stress. *Cell stress*, 2(8):184.

Pagliarini, D., Calvo, S., Change, B., Sheth, S., Vafai, S., Walford, G., Ong, S.-E., Sugiana, C., Boneh, A., Chen, W., Hill, D., Vidal, M., Evans, J., Thorburn, D., Carr, S., and Mootha, V. (2008). A mitochondrial protein compendium elucidates complex i disease biology. *Cell*, 134(1):112–23.

Palmer, N. D., Hester, J. M., An, S. S., Adeyemo, A., Rotimi, C., Langefeld, C. D., Freedman, B. I., Ng, M. C., and Bowden, D. W. (2011). Resequencing and analysis of variation in the tcf7l2 gene in african americans suggests that snp rs7903146 is the causal diabetes susceptibility variant. *Diabetes*, 60(2):662–668.

Palmer, N. D., McDonough, C. W., Hicks, P. J., Roh, B. H., Wing, M. R., An, S. S., Hester, J. M., Cooke, J. N., Bostrom, M. A., Rudock, M. E., et al. (2012). A genome-wide association search for type 2 diabetes genes in african americans. *PloS one*, 7(1).

Park, S. S. and Seo, Y.-K. (2020). Excess accumulation of lipid impairs insulin sensitivity in skeletal muscle. *International Journal of Molecular Sciences*, 21(6):1949.

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, 110(44):17921–17926.

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Morán, I., Gómez-Marín, C., Van De Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics*, 46(2):136.

Patti, M. E., Butte, A. J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., et al. (2003). Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of pgc1 and nrf1. *Proceedings of the National Academy of Sciences*, 100(14):8466–8471.

Pearson, K. and Lee, A. (1901). On the inheritance of characteristics not capable of exact quantitative measurements. *Philos. Trans R. Soc. Lond. Ser. A*, 34:27–42.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):381–385.

Pellatt, A. J., Slattery, M. L., Mullany, L. E., Wolff, R. K., and Pellatt, D. F. (2016). Dietary intake alters gene expression in colon tissue: possible underlying mechanism for the influence of diet on disease. *Pharmacogenetics and genomics*, 26(6):294.

Pengelly, R. J., Tapper, W., Gibson, J., Knut, M., Tearle, R., Collins, A., and Ennis, S. (2015). Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC genomics*, 16(1):666.

Perilhou, A., Tourrel-Cuzin, C., Kharroubi, I., Henique, C., Fauveau, V., Kitamura, T., Magnan, C., Postic, C., Prip-Buus, C., and Vasseur-Cognet, M. (2008). The transcription factor coup-tfii is negatively regulated by insulin and glucose via foxo1-and chrebp-controlled pathways. *Molecular and cellular biology*, 28(21):6568–6579.

Perry, J. R., McCarthy, M. I., Hattersley, A. T., Zeggini, E., Weedon, M. N., Frayling, T. M., Consortium, W. T. C. C., et al. (2009). Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes*, 58(6):1463–1467.

Perry, R. J., Samuel, V. T., Petersen, K. F., and Shulman, G. I. (2014). The role of hepatic lipids in hepatic insulin resistance and type 2 diabetes. *Nature*, 510(7503):84–91.

Persson, F. and Rossing, P. (2018). Diagnosis of diabetic kidney disease: state of the art and future perspective. *Kidney international supplements*, 8(1):2–7.

Petersen, K. F., Dufour, S., Befroy, D., Garcia, R., and Shulman, G. I. (2004). Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes. *New England Journal of Medicine*, 350(7):664–671.

Petersen, K. F., Dufour, S., and Shulman, G. I. (2005). Decreased insulin-stimulated atp synthesis and phosphate transport in muscle of insulin-resistant offspring of type 2 diabetic parents. *PLoS medicine*, 2(9).

Pettersen, A. K., Marshall, D. J., and White, C. R. (2018). Understanding variation in metabolic rate. *Journal of Experimental Biology*, 221(1).

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573.

Pihlajamäki, J., Boes, T., Kim, E.-Y., Dearie, F., Kim, B. W., Schroeder, J., Mun, E., Nasser, I., Park, P. J., Bianco, A. C., et al. (2009). Thyroid hormone-related regulation of gene expression in human fatty liver. *The Journal of Clinical Endocrinology & Metabolism*, 94(9):3521–3529.

Pihlajamäki, J., Lerin, C., Itkonen, P., Boes, T., Floss, T., Schroeder, J., Dearie, F., Crunkhorn, S., Burak, F., Jimenez-Chillaron, J. C., et al. (2011). Expression of the splicing factor gene sfrs10 is reduced in human obesity and contributes to enhanced lipogenesis. *Cell metabolism*, 14(2):208–218.

Polvani, S., Pepe, S., Milani, S., and Galli, A. (2020). Coup-tfii in health and disease. *Cells*, 9(1):101.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178.

Poulsen, P., Kyvik, K. O., Vaag, A., and Beck-Nielsen, H. (1999). Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145.

Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S., and Goldstein, D. B. (2019). Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics*, 20(12):747–759.

Prasad, R. B. and Groop, L. (2015). Genetics of type 2 diabetes-pitfalls and possibilities. *Genes*, 6(1):87–123.

Pravenec, M., Hyakukoku, M., Houstek, J., Zidek, V., Landa, V., Mlejnek, P., Miksik, I., Dudová-Mothejzikova, K., Pecina, P., Vrbackỳ, M., et al. (2007). Direct linkage of mitochondrial genome variation to risk factors for type 2 diabetes in conplastic strains. *Genome research*, 17(9):1319–1326.

Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant... or not? *Human molecular genetics*, 11(20):2417–2423.

Purcell, S., Cherny, S. S., and Sham, P. C. (2003). Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–150.

Qi, Q., Stilp, A. M., Sofer, T., Moon, J.-Y., Hidalgo, B., Szpiro, A. A., Wang, T., Ng, M. C., Guo, X., Chen, Y.-D. I., et al. (2017). Genetics of type 2 diabetes in us hispanic/latino individuals: results from the hispanic community health study/study of latinos (hchs/sol). *Diabetes*, 66(5):1419–1425.

Rabilloud, T., Kieffer, S., Procaccio, V., Louwagie, M., Courchesne, P. L., Patterson, S. D., Martinez, P., Garin, J., and Lunardi, J. (1998). Two-dimensional electrophoresis of human placental mitochondria and protein identification by mass spectrometry: Toward a human mitochondrial proteome. *Electrophoresis*, 19(6):1006–1014.

Rafelski, S. M. (2013). Mitochondrial network morphology: building an integrative, geometrical view. *BMC biology*, 11(1):1–9.

Ragolia, L., Palaia, T., Hall, C. E., Maesaka, J. K., Eguchi, N., and Urade, Y. (2005). Accelerated glucose intolerance, nephropathy, and atherosclerosis in prostaglandin d2 synthase knock-out mice. *Journal of Biological Chemistry*, 280(33):29946–29955.

Ramos-Rodríguez, M., Raurell-Villa, H., Colli, M. L., Alvelos, M. I., Subirana-Granes, M., Juan-Mateu, J., Norris, R., Turatsinze, J.-V., Nakayasu, E., Webb-Robertson, B.-J., et al. (2019). The impact of pro-inflammatory cytokines on the β-cell regulatory landscape provides new insights into the genetics of type 1 diabetes. *bioRxiv*, page 560193.

Randle, P. (1963). The glucose fatty acid cycle; its role in insulin sensitivity and the metabolic disturbances of diabetes mellitus. *Lancet*, 1:785–789.

Raulerson, C. K., Ko, A., Kidd, J. C., Currin, K. W., Brotman, S. M., Cannon, M. E., Wu, Y., Spracklen, C. N., Jackson, A. U., Stringham, H. M., et al. (2019). Adipose tissue gene expression associations reveal hundreds of candidate genes for cardiometabolic traits. *The American Journal of Human Genetics*, 105(4):773–787.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). Genecards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13(4):163.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510.

Ren, Y., Reddy, J. S., Pottier, C., Sarangi, V., Tian, S., Sinnwell, J. P., McDonnell, S. K., Biernacka, J. M., Carrasquillo, M. M., Ross, O. A., et al. (2018). Identification of missing variants by combining multiple analytic pipelines. *BMC bioinformatics*, 19(1):139.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894.

Reynisdottir, I., Thorleifsson, G., Benediktsson, R., Sigurdsson, G., Emilsson, V., Einarsdottir, A. S., Hjorleifsdottir, E. E., Orlygsdottir, G. T., Bjornsdottir, G. T., Saemundsdottir, J., et al. (2003). Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34–q35. 2. *The American Journal of Human Genetics*, 73(2):323–335.

Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., and Ting, A. Y. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, 339(6125):1328–1331.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268(5217):1584–1589.

Ritov, V. B., Menshikova, E. V., He, J., Ferrell, R. E., Goodpaster, B. H., and Kelley, D. E. (2005). Deficiency of subsarcolemmal mitochondria in obesity and type 2 diabetes. *Diabetes*, 54(1):8–14.

Ronquillo, C., Bernstein, P., and Baehr, W. (2012). Senior–løken syndrome: A syndromic form of retinal dystrophy associated with nephronophthisis. *Vision research*, 75:88–97.

Roomp, K., Kristinsson, H., Schvartz, D., Ubhayasekera, K., Sargsyan, E., Manukyan, L., Chowdhury, A., Manell, H., Satagopam, V., Groebe, K., et al. (2017). Combined lipidomic and proteomic analysis of isolated human islets exposed to palmitate reveals time-dependent changes in insulin secretion and lipid metabolism. *PloS one*, 12(4).

Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., and Hickey, J. M. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genetics Selection Evolution*, 50(1):64.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356–366.

Rouzier, C., Moore, D., Delorme, C., Lacas-Gervais, S., Ait-El-Mkadem, S., Fragaki, K., Burté, F., Serre, V., Bannwarth, S., Chaussenot, A., et al. (2017). A novel cisd2 mutation associated with a classical wolfram syndrome phenotype alters ca2+ homeostasis and er-mitochondria interactions. *Human molecular genetics*, 26(9):1599–1611.

Rovira-Llopis, S., Bañuls, C., Diaz-Morales, N., Hernandez-Mijares, A., Rocha, M., and Victor, V. M. (2017). Mitochondrial dynamics in type 2 diabetes: pathophysiological implications. *Redox biology*, 11:637–645.

Rowley, M. J. and Corces, V. G. (2018). Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, page 1.

Rubin, A. J., Barajas, B. C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M. R., Howard, I., Kim, D. S., Boxer, L. D., Cairns, J., Spivakov, M., et al. (2017). Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation. *Nature genetics*, 49(10):1522.

Rubio-Ruiz, M. E., Peredo-Escárcega, A. E., Cano-Martínez, A., and Guarner-Lans, V. (2015). An evolutionary perspective of nutrition and inflammation as mechanisms of cardiovascular disease. *International journal of evolutionary biology*, 2015.

Rytka, J. M., Wueest, S., Schoenle, E. J., and Konrad, D. (2011). The portal theory supported by venous drainage–selective fat transplantation. *Diabetes*, 60(1):56–63.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–934.

Sadlon, T. J., Wilkinson, B. G., Pederson, S., Brown, C. Y., Bresatz, S., Gargett, T., Melville, E. L., Peng, K., D'Andrea, R. J., Glonek, G. G., et al. (2010). Genome-wide identification of human foxp3 target genes in natural regulatory t cells. *The Journal of Immunology*, 185(2):1071–1081.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., et al. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabetes research and clinical practice*, 157:107843.

Sahin, K., Tuzcu, M., Orhan, C., Sahin, N., Kucuk, O., Ozercan, I. H., Juturu, V., and Komorowski, J. R. (2013). Anti-diabetic activity of chromium picolinate and biotin in rats with type 2 diabetes induced by high-fat diet and streptozotocin. *British Journal of Nutrition*, 110(2):197–205.

Sajuthi, S. P., Sharma, N. K., Chou, J. W., Palmer, N. D., McWilliams, D. R., Beal, J., Comeau, M. E., Ma, L., Calles-Escandon, J., Demons, J., et al. (2016). Mapping adipose and muscle tissue expression quantitative trait loci in african americans to identify genes for type 2 diabetes and obesity. *Human genetics*, 135(8):869–880.

Salem, H. H., Trojanowski, B., Fiedler, K., Maier, H. J., Schirmbeck, R., Wagner, M., Boehm, B. O., Wirth, T., and Baumann, B. (2014). Long-term ikk2/nf-κb signaling in pancreatic β-cells induces immune-mediated diabetes. *Diabetes*, 63(3):960–975.

Samocha-Bonet, D., Campbell, L. V., Mori, T. A., Croft, K. D., Greenfield, J. R., Turner, N., and Heilbronn, L. K. (2012). Overfeeding reduces insulin sensitivity and increases oxidative stress, without altering markers of mitochondrial content and function in humans. *PloS one*, 7(5):e36320.

Sandoval-Motta, S., Aldana, M., Martínez-Romero, E., and Frank, A. (2017). The human microbiome and the missing heritability problem. *Frontiers in genetics*, 8:80.

Sanghera, D. K. and Blackett, P. R. (2012). Type 2 diabetes genetics: beyond gwas. *Journal of diabetes & metabolism*, 3(198).

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vila, A. V., Võsa, U., Mujagic, Z., Masclee, A. A., Jonkers, D. M., Oosting, M., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature genetics*, 51(4):600–605.

Saxena, R., Elbers, C. C., Guo, Y., Peter, I., Gaunt, T. R., Mega, J. L., Lanktree, M. B., Tare, A., Castillo, B. A., Li, Y. R., et al. (2012). Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *The American Journal of Human Genetics*, 90(3):410–425.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504.

Scharfe, C., Zaccaria, P., Hoertnagel, K., Jaksch, M., Klopstock, T., Dembowski, M., Lill, R., Prokisch, H., Gerbitz, K., Neupert, W., Mewes, H., and Meitinger, T. (2000). Mitop, the mitochondrial proteome database: 2000 update. *Nucleic Acids Research*, 28(1):155–8.

Schoenberger, S. D., Kim, S. J., Sheng, J., Rezaei, K. A., Lalezary, M., and Cherney, E. (2012). Increased

prostaglandin e2 (pge2) levels in proliferative diabetic retinopathy, and correlation with vegf and inflammatory cytokines. *Investigative ophthalmology & visual science*, 53(9):5906–5911.

Schoenfelder, S. and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, page 1.

Schofield, E., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J. A., and Burren, O. S. (2016). Chicp: a web-based tool for the integrative and interactive visualization of promoter capture hi-c datasets. *Bioinformatics*, 32(16):2511–2513.

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219.

Schroeder, M. A., Atherton, H. J., Dodd, M. S., Lee, P., Cochlin, L. E., Radda, G. K., Clarke, K., and Tyler, D. J. (2012). The cycling of acetyl-coenzyme a through acetylcarnitine buffers cardiac substrate supply: a hyperpolarized 13c magnetic resonance study. *Circulation: Cardiovascular Imaging*, 5(2):201–209.

Schulman, I. G. (2017). Liver x receptors link lipid metabolism and inflammation. *FEBS letters*, 591(19):2978–2991.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., et al. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *science*, 316(5829):1341–1345.

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., Pervjakova, N., Pers, T. H., Johnson, A. D., Eicher, J. D., et al. (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, 66(11):2888–2902.

Sears, D., Hsiao, G., Hsiao, A., Yu, J., Courtney, C., Ofrecio, J., Chapman, J., and Subramaniam, S. (2009). Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization. *Proceedings of the National Academy of Sciences*, 106(44):18745–18750.

Seifert, W., Kühnisch, J., Tüysüz, B., Specker, C., Brouwers, A., and Horn, D. (2012). Mutations in the prostaglandin transporter encoding gene slco2a1 cause primary hypertrophic osteoarthropathy and isolated digital clubbing. *Human mutation*, 33(4):660–664.

Seong, H.-A., Manoharan, R., and Ha, H. (2018). Smad proteins differentially regulate obesity-induced glucose and lipid abnormalities and inflammation via class-specific control of ampk-related kinase mpk38/melk activity. *Cell death & disease*, 9(5):1–13.

Sergi, D., Naumovski, N. N., Heilbronn, L. H. K., Abeywardena, M., O'Callaghan, N., Lionetti, L., and Luscombe-Marsh, N. L.-M. (2019). Mitochondrial (dys) function and insulin resistance: From pathophysiological molecular mechanisms to the impact of diet. *Frontiers in physiology*, 10:532.

Sethi, J. K. and Vidal-Puig, A. J. (2007). Thematic review series: adipocyte biology. adipose tissue function and plasticity orchestrate nutritional adaptation. *Journal of lipid research*, 48(6):1253–1262.

Sharma, K. (2017). Mitochondrial dysfunction in the diabetic kidney. In *Mitochondrial Dynamics in Cardiovascular Medicine*, pages 553–562. Springer.

Shen, Y., Wu, L., Xi, B., Liu, X., Zhao, X., Cheng, H., Hou, D., Wang, X., and Mi, J. (2013). Gckr variants increase triglycerides while protecting from insulin resistance in chinese children. *PloS one*, 8(1):e55350.

Shiba, D., Nakata, K., Fukui, H., Kobayashi, D., and Yokoyama, T. (2012). A three-step process of nphp3 ciliary localization. *Cilia*, 1(1):P46.

Shimabukuro, M., Higa, M., Zhou, Y.-T., Wang, M.-Y., Newgard, C. B., and Unger, R. H. (1998). Lipoapoptosis in beta-cells of obese prediabeticfa/fa rats role of serine palmitoyltransferase overexpression. *Journal of Biological Chemistry*, 273(49):32487–32490.

Shriner, D., Adeyemo, A., Gerry, N. P., Herbert, A., Chen, G., Doumatey, A., Huang, H., Zhou, J., Christman, M. F., and Rotimi, C. N. (2009). Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS One*, 4(12).

Shtir, C., Aldahmesh, M. A., Al-Dahmash, S., Abboud, E., Alkuraya, H., Abouammoh, M. A., Nowailaty, S. R., Al-Thubaiti, G., Naim, E., ALYounes, B., et al. (2016). Exome-based case–control association study using extreme phenotype design reveals novel candidates with protective effect in diabetic retinopathy. *Human genetics*, 135(2):193–200.

Sieberts, S. K. and Schadt, E. E. (2007). Moving toward a system genetics view of disease. *Mammalian Genome*, 18(6-7):389–401.

Siersbæk, R., Madsen, J. G. S., Javierre, B. M., Nielsen, R., Bagge, E. K., Cairns, J., Wingett, S. W., Traynor, S., Spivakov, M., Fraser, P., et al. (2017). Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Molecular cell*, 66(3):420–435.

Silander, K., Mohlke, K. L., Scott, L. J., Peck, E. C., Hollstein, P., Skol, A. D., Jackson, A. U., Deloukas, P., Hunt, S., Stavrides, G., et al. (2004). Genetic variation near the hepatocyte nuclear factor-4α gene predicts susceptibility to type 2 diabetes. *Diabetes*, 53(4):1141–1149.

Simó-Servat, O., Hernández, C., and Simó, R. (2019). Diabetic retinopathy in the context of patients with diabetes. *Ophthalmic research*, 62(4):206–212.

Simons, N., Dekker, J. M., van Greevenbroek, M. M., Nijpels, G., Leen, M., van der Kallen, C. J., Schalkwijk, C. G., Schaper, N. C., Stehouwer, C. D., and Brouwers, M. C. (2016). A common gene variant in glucokinase regulatory protein interacts with glucose metabolism on diabetic dyslipidemia: The combined codam and hoorn studies. *Diabetes Care*, 39(10):1811–1817.

Simpson, N. E. (1969). Heritabilities of liability to diabetes when sex and age at onset are considered. *Annals of human genetics*, 32(3):283–303.

Sivitz, W. I. and Yorek, M. A. (2010). Mitochondrial dysfunction in diabetes: from molecular mechanisms to functional significance and therapeutic opportunities. *Antioxidants & redox signaling*, 12(4):537–577.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885.

Small, K. S., Todorčević, M., Civelek, M., Moustafa, J. S. E.-S., Wang, X., Simon, M. M., Fernandez-Tajes, J., Mahajan, A., Horikoshi, M., Hugill, A., et al. (2018). Regulatory variants at klf14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nature genetics*, 50(4):572–580.

Smith, A. V., Thomas, D. J., Munro, H. M., and Abecasis, G. R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. *Genome research*, 15(11):1519–1534.

Smith, C. J. and Ryckman, K. K. (2015). Epigenetic and developmental influences on the risk of obesity, diabetes, and metabolic syndrome. *Diabetes, metabolic syndrome and obesity: targets and therapy*, 8:295.

Smith, G. I., Mittendorfer, B., and Klein, S. (2019). Metabolically healthy obesity: facts and fantasies. *Journal of Clinical Investigation*, 129(10):3978–3989.

Soccio, R. E., Chen, E. R., Rajapurkar, S. R., Safabakhsh, P., Marinis, J. M., Dispirito, J. R., Emmett, M. J., Briggs, E. R., Fang, B., Everett, L. J., et al. (2015). Genetic variation determines pparγ function and anti-diabetic drug response in vivo. *Cell*, 162(1):33–44.

Soejima, A., Inoue, K., Takai, D., Kaneko, M., Ishihara, H., Oka, Y., and Hayashi, J.-I. (1996). Mitochondrial dna is required for regulation of glucose-stimulated insulin secretion in a mouse pancreatic beta cell line, min6. *Journal of Biological Chemistry*, 271(42):26194–26199.

Soeters, M. R. and Soeters, P. B. (2012). The evolutionary benefit of insulin resistance. *Clinical nutrition*, 31(6):1002–1007.

Sokolowska, E. and Błachnio-Zabielska, A. U. (2019). The role of ceramides in insulin resistance. *Frontiers in Endocrinology*, 10:577.

Solimena, M., Schulte, A. M., Marselli, L., Ehehalt, F., Richter, D., Kleeberg, M., Mziaut, H., Knoch, K.-P., Parnis, J., Bugliani, M., et al. (2018). Systems biology of the imidia biobank from organ donors

and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia*, 61(3):641–657.

Solon-Biet, S. M., Cogger, V. C., Pulpitel, T., Wahl, D., Clark, X., Bagley, E. E., Gregoriou, G. C., Senior, A. M., Wang, Q.-P., Brandon, A. E., et al. (2019). Branched-chain amino acids impact health and lifespan indirectly via amino acid balance and appetite control. *Nature metabolism*, 1(5):532–545.

Soronen, J., Laurila, P.-P., Naukkarinen, J., Surakka, I., Ripatti, S., Jauhiainen, M., Olkkonen, V. M., and Yki-Järvinen, H. (2012). Adipose tissue gene expression analysis reveals changes in inflammatory, mitochondrial respiratory and lipid metabolic pathways in obese insulin-resistant subjects. *BMC medical genomics*, 5(1):9.

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1):R111–R119.

Spanakis, E. K. and Golden, S. H. (2013). Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports*, 13(6):814–823.

Speakman, J. R. (2008). Thrifty genes for obesity, an attractive but flawed idea, and an alternative perspective: the 'drifty gene' hypothesis. *International journal of obesity*, 32(11):1611–1617.

Spencer, C. C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. *PLoS genetics*, 2(9).

Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J., and Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature genetics*, 39(2):226–231.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506.

Spracklen, C. N., Horikoshi, M., Kim, Y. J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C. H. T., Tabara, Y., Kwak, S.-H., Takeuchi, F., Long, J., Lim, V. J. Y., Chai, J.-F., Chen, C.-H., Nakatochi, M., Yao, J., Choi, H. S., Iyengar, A. K., Perrin, H. J., Brotman, S. M., van de Bunt, M., Gloyn, A. L., Below, J. E., Boehnke, M., Bowden, D. W., Chambers, J. C., Mahajan, A., McCarthy, M. I., Ng, M. C. Y., Petty, L. E., Zhang, W., Morris, A. P., Adair, L. S., Akiyama, M., Bian, Z., Chan, J. C. N., Chang, L.-C., Chee, M.-L., Chen, Y.-D. I., Chen, Y.-T., Chen, Z., Chuang, L.-M., Du, S., Gordon-Larsen, P., Gross, M., Guo, X., Guo, Y., Han, S., Howard, A.-G., Huang, W., Hung, Y.-J., Hwang, M. Y., Hwu, C.-M., Ichihara, S., Isono, M., Jang, H.-M., Jiang, G., Jonas, J. B., Kamatani, Y., Katsuya, T., Kawaguchi, T., Khor, C.-C., Kohara, K., Lee, M.-S., Lee, N. R., Li, L., Liu, J., Luk, A. O., Lv, J., Okada, Y., Pereira, M. A., Sabanayagam, C., Shi, J., Shin, D. M., So, W. Y., Takahashi, A., Tomlinson, B., Tsai, F.-J., van Dam, R. M., Xiang, Y.-B., Yamamoto, K., Yamauchi, T., Yoon, K., Yu, C., Yuan, J.-M., Zhang, L., Zheng, W., Igase, M., Cho, Y. S., Rotter, J. I., Wang, Y.-X., Sheu, W. H. H., Yokota, M., Wu, J.-Y., Cheng, C.-Y., Wong, T.-Y., Shu, X.-O., Kato, N., Park, K.-S., Tai, E.-S., Matsuda, F., Koh, W.-P., Ma, R. C. W., Maeda, S., Millwood, I. Y., Lee, J., Kadowaki, T., Walters, R. G., Kim, B.-J., Mohlke, K. L., and Sim, X. (2020). Identification of type 2 diabetes loci in 433,540 east asian individuals. *Nature*.

Stančáková, A., Civelek, M., Saleem, N. K., Soininen, P., Kangas, A. J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L. L., Morken, M. A., et al. (2012). Hyperglycemia and a common variant of gckr are associated with the levels of eight amino acids in 9,369 finnish men. *Diabetes*, 61(7):1895–1902.

Stanley, C. E. and Kulathinal, R. J. (2016). Neurogenomics and the role of a large mutational target on rapid behavioral change. *Biology direct*, 11(1):60.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., et al. (2007). A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nature genetics*, 39(6):770–775.

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H. T., Johannsdottir, H., Magnusson, O. T., Gudjonsson, S. A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature genetics*, 46(3):294.

Stephens, N. A., Xie, H., Johannsen, N. M., Church, T. S., Smith, S. R., and Sparks, L. M. (2015). A transcriptional signature of "exercise resistance" in skeletal muscle of individuals with type 2 diabetes mellitus. *Metabolism*, 64(9):999–1004.

Stoffers, D. A., Chèvre, J.-C., Durand, E., Stanojevic, V., Dina, C., Habener, J. F., Froguel, P., et al. (1999). Defective mutations in the insulin promoter factor-1 (ipf-1) gene in late-onset type 2 diabetes mellitus. *The Journal of clinical investigation*, 104(9):R41–R48.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. et al. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Stringer, S., Wray, N. R., Kahn, R. S., and Derks, E. M. (2011). Underestimated effect sizes in gwas: fundamental limitations of single snp analysis for dichotomous phenotypes. *PloS one*, 6(11).

Stroup, D. and Chiang, J. Y. (2000). Hnf4 and coup-tfii interact to modulate transcription of the cholesterol 7α-hydroxylase gene (cyp7a1). *Journal of lipid research*, 41(1):1–11.

Strycharz, J., Drzewoski, J., Szemraj, J., and Sliwinska, A. (2017). Is p53 involved in tissue-specific insulin resistance formation? *Oxidative medicine and cellular longevity*, 2017.

Stumvoll, M., Tschritter, O., Fritsche, A., Staiger, H., Renn, W., Weisser, M., Machicao, F., and Häring, H. (2002). Association of the tg polymorphism in adiponectin (exon 2) with obesity and insulin sensitivity: interaction with family history of type 2 diabetes. *Diabetes*, 51(1):37–41.

Su, Z., Nie, Y., Huang, X., Zhu, Y., Feng, B., Tang, L., and Zheng, G. (2019). Mitophagy in hepatic insulin resistance: Therapeutic potential and concerns. *Frontiers in pharmacology*, 10:1193.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Sun, P., Zhang, R., Jiang, Y., Wang, X., Li, J., Lv, H., Tang, G., Guo, X., Meng, X., Zhang, H., et al. (2011). Assessing the patterns of linkage disequilibrium in genic regions of the human genome. *The FEBS journal*, 278(19):3748–3755.

Suzuki, K., Akiyama, M., Ishigaki, K., Kanai, M., Hosoe, J., Shojima, N., Hozawa, A., Kadota, A., Kuriki, K., Naito, M., et al. (2019). Identification of 28 new susceptibility loci for type 2 diabetes in the japanese population. *Nature genetics*, 51(3):379–386.

Sved, J. A. and Hill, W. G. (2018). One hundred years of linkage disequilibrium. *Genetics*, 209(3):629–636.

Syeda, M. M., Jing, X., Mirza, R. H., Yu, H., Sellers, R. S., and Chi, Y. (2012). Prostaglandin transporter modulates wound healing in diabetes by regulating prostaglandin-induced angiogenesis. *The American journal of pathology*, 181(1):334–346.

Szendroedi, J., Phielix, E., and Roden, M. (2012). The role of mitochondria in insulin resistance and type 2 diabetes mellitus. *Nature Reviews Endocrinology*, 8(2):92–103.

Tabassum, R., Chauhan, G., Dwivedi, O. P., Mahajan, A., Jaiswal, A., Kaur, I., Bandesh, K., Singh, T., Mathai, B. J., Pandey, Y., et al. (2013). Genome-wide association study for type 2 diabetes in indians identifies a new susceptibility locus at 2q21. *Diabetes*, 62(3):977–986.

Tak, Y. G. and Farnham, P. J. (2015). Making sense of gwas: using epigenomics and genome engineering to understand the functional relevance of snps in non-coding regions of the human genome. *Epigenetics & chromatin*, 8(1):57.

Takeuchi, H., Yokota-Nakatsuma, A., Ohoka, Y., Kagechika, H., Kato, C., Song, S.-Y., and Iwata, M. (2013). Retinoid x receptor agonists modulate foxp3+ regulatory t cell and th17 cell differentiation with differential dependence on retinoic acid receptor activation. *The Journal of Immunology*, 191(7):3725–3733.

Tan, C., Chong, H., Tan, E., and Tan, N. (2012). Getting 'smad' about obesity and diabetes. *Nutrition & diabetes*, 2(3):e29–e29.

Tao, C., Sifuentes, A., and Holland, W. L. (2014). Regulation of glucose and lipid homeostasis by adiponectin: effects on hepatocytes, pancreatic β cells and adipocytes. *Best Practice & Research Clinical Endocrinology & Metabolism*, 28(1):43–58.

Tapper, W. (2007). Linkage disequilibrium maps and location databases. In *Linkage Disequilibrium and Association Mapping*, pages 23–45. Springer.

Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P., and Fraser, H. B. (2019). Fine-mapping cis-regulatory variants in diverse human populations. *Elife*, 8:e39595.

Teo, Y.-Y., Small, K. S., and Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in africa. *Nature Reviews Genetics*, 11(2):149–160.

Terwilliger, J. D. (2001). 23 on the resolution and feasibility of genome scanning approaches. *Advances in Genetics*, 42.

Terwilliger, J. D. and Ott, J. (1994). *Handbook of human genetic linkage*. JHU Press.

Thivolet, C., Vial, G., Cassel, R., Rieusset, J., and Madec, A.-M. (2017). Reduction of endoplasmic reticulum-mitochondria interactions in beta cells from patients with type 2 diabetes. *PLoS One*, 12(7):e0182027.

Tiemann-Boege, I., Schwarz, T., Striedner, Y., and Heissl, A. (2017). The consequences of sequence erosion in the evolution of recombination hotspots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160462.

Tong, L. (2013). Structure and function of biotin-dependent carboxylases. *Cellular and Molecular Life Sciences*, 70(5):863–891.

Tontonoz, P. and Spiegelman, B. M. (2008). Fat and beyond: the diverse biology of pparγ. *Annu. Rev. Biochem.*, 77:289–312.

Torkamani, A., Topol, E. J., and Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–272.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590.

Torres, J. M., Barbeira, A. N., Bonazzola, R., Morris, A. P., Shah, K. P., Wheeler, H. E., Bell, G., Cox, N. J., and Im, H. K. (2017). Integrative cross tissue analysis of gene expression identifies novel type 2 diabetes genes. *BioRxiv*, page 108134.

Torres, J. M., Gamazon, E. R., Parra, E. J., Below, J. E., Valladares-Salgado, A., Wacher, N., Cruz, M., Hanis, C. L., and Cox, N. J. (2014). Cross-tissue and tissue-specific eqtls: partitioning the heritability of a complex trait. *The American Journal of Human Genetics*, 95(5):521–534.

Traglia, M., Croen, L. A., Jones, K. L., Heuer, L. S., Yolken, R., Kharrazi, M., DeLorenze, G. N., Ashwood, P., Van de Water, J., and Weiss, L. A. (2018). Cross-genetic determination of maternal and neonatal immune mediators during pregnancy. *Genome medicine*, 10(1):1–17.

Trifunovic, A., Wredenberg, A., Falkenberg, M., Spelbrink, J. N., Rovio, A. T., Bruder, C. E., Bohlooly-Y, M., Gidlöf, S., Oldfors, A., Wibom, R., et al. (2004). Premature ageing in mice expressing defective mitochondrial dna polymerase. *Nature*, 429(6990):417–423.

Tsang, S. H., Aycinena, A. R., and Sharma, T. (2018). Ciliopathy: Alström syndrome. In *Atlas of Inherited Retinal Diseases*, pages 179–180. Springer.

Tsuruzoe, K., Araki, E., Furukawa, N., Shirotani, T., Matsumoto, K., Kaneko, K., Motoshima, H., Yoshizato, K., Shirakami, A., Kishikawa, H., et al. (1998). Creation and characterization of a mitochondrial dna-depleted pancreatic beta-cell line: impaired insulin secretion induced by glucose, leucine, and sulfonylureas. *Diabetes*, 47(4):621–631.

Tubbs, E., Chanon, S., Robert, M., Bendridi, N., Bidaux, G., Chauvin, M.-A., Ji-Cao, J., Durand, C., Gauvrit-Ramette, D., Vidal, H., et al. (2018). Disruption of mitochondria-associated endoplasmic reticulum membrane (mam) integrity contributes to muscle insulin resistance in mice and humans. *Diabetes*, 67(4):636–650.

Turner, N. and Heilbronn, L. K. (2008). Is mitochondrial dysfunction a cause of insulin resistance? *Trends in Endocrinology & Metabolism*, 19(9):324–330.

Udler, M. S. (2019). Type 2 diabetes: multiple genes, multiple diseases. *Current diabetes reports*, 19(8):55.

Udler, M. S., Kim, J., von Grotthuss, M., Bonàs-Guarch, S., Cole, J. B., Chiou, J., on behalf of METASTROKE, C. D. A., the ISGC, Boehnke, M., Laakso, M., Atzmon, G., et al. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS medicine*, 15(9):e1002654.

Ukropcova, B., McNeil, M., Sereda, O., De Jonge, L., Xie, H., Bray, G. A., Smith, S. R., et al. (2005). Dynamic changes in fat oxidation in human primary myocytes mirror metabolic characteristics of the donor. *The Journal of clinical investigation*, 115(7):1934–1941.

Umeno, J., Hisamatsu, T., Esaki, M., Hirano, A., Kubokura, N., Asano, K., Kochi, S., Yanai, S., Fuyuno, Y., Shimamura, K., et al. (2015). A hereditary enteropathy caused by mutations in the slco2a1 gene, encoding a prostaglandin transporter. *PLoS genetics*, 11(11).

Ung, C., Sanchez, A. V., Shen, L., Davoudi, S., Ahmadi, T., Navarro-Gomez, D., Chen, C. J., Hancock, H., Penman, A., Hoadley, S., et al. (2017). Whole exome sequencing identification of novel candidate genes in patients with proliferative diabetic retinopathy. *Vision research*, 139:168–176.

Unnikrishnan, R., Pradeepa, R., Joshi, S. R., and Mohan, V. (2017). Type 2 diabetes: demystifying the global epidemic. *Diabetes*, 66(6):1432–1442.

Vaag, A., Grunnet, L., Arora, G., and Brøns, C. (2012). The thrifty phenotype hypothesis revisited. *Diabetologia*, 55(8):2085–2088.

Valdés-Ramos, R., Ana Laura, G.-L., Beatriz Elina, M.-C., and Alejandra Donaji, B.-A. (2015). Vitamins and type 2 diabetes mellitus. *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, 15(1):54–63.

Van de Bunt, M., Cortes, A., Brown, M. A., Morris, A. P., McCarthy, M. I., Consortium, I., et al. (2015). Evaluating the performance of fine-mapping strategies at common variant gwas loci. *PLoS Genet*, 11(9):e1005535.

van de Bunt, M., Fox, J. E. M., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A. J., Johnson, P. R., Rajotte, R. V., Gaulton, K. J., et al. (2015). Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS genetics*, 11(12):e1005694.

van den Ouweland, J. M., Lemkes, H. H., Trembath, R. C., Ross, R., Velho, G., Cohen, D., Froguel, P., and Maassen, J. A. (1994). Maternally inherited diabetes and deafness is a distinct subtype of diabetes and associates with a single point mutation in the mitochondrial trna leu (uur) gene. *Diabetes*, 43(6):746–751.

Van der Kolk, B. W., Kalafati, M., Adriaens, M., Van Greevenbroek, M. M., Vogelzangs, N., Saris, W. H., Astrup, A., Valsesia, A., Langin, D., Van der Kallen, C. J., et al. (2019). Subcutaneous adipose tissue and systemic inflammation are associated with peripheral but not hepatic insulin resistance in humans. *Diabetes*, 68(12):2247–2258.

van Marken Lichtenbelt, W. D., Vanhommerig, J. W., Smulders, N. M., Drossaerts, J. M., Kemerink, G. J., Bouvy, N. D., Schrauwen, P., and Teule, G. J. (2009). Cold-activated brown adipose tissue in healthy men. *New England Journal of Medicine*, 360(15):1500–1508.

van Smeden, M., Harrell, F. E., and Dahly, D. L. (2018). Novel diabetes subgroups. *The Lancet Diabetes & Endocrinology*, 6(6):439–440.

Van Tienen, F., Praet, S. F., De Feyter, H., van Den Broek, N., Lindsey, P., Schoonderwoerd, K., de Coo, I., Nicolay, K., Prompers, J., Smeets, H., et al. (2012). Physical activity is the key determinant of skeletal muscle mitochondrial function in type 2 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 97(9):3261–3269.

van Tilburg, J., Van Haeften, T. W., Pearson, P., and Wijmenga, C. (2001). Defining the genetic contribution of type 2 diabetes mellitus. *Journal of medical genetics*, 38(9):569–578.

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic acids research*, 41(8):4378–4391.

Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew, I., Uhlén, M., Pedersen, B. K., and Nielsen, J. (2015). Proteome-and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell reports*, 11(6):921–933.

Varshney, A., Scott, L. J., Welch, R. P., Erdos, M. R., Chines, P. S., Narisu, N., Albanus, R. D., Orchard, P., Wolford, B. N., Kursawe, R., et al. (2017). Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proceedings of the National Academy of Sciences*, 114(9):2301–2306.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). Sift missense predictions for genomes. *Nature protocols*, 11(1):1.

Vasseur, F., Helbecque, N., Dina, C., Lobbens, S., Delannoy, V., Gaget, S., Boutin, P., Vaxillaire, M., Leprêtre, F., Dupont, S., et al. (2002). Single-nucleotide polymorphism haplotypes in the both proximal promoter and exon 3 of the apm1 gene modulate adipocyte-secreted adiponectin hormone levels and contribute to the genetic risk for type 2 diabetes in french caucasians. *Human molecular genetics*, 11(21):2607–2614.

Vecchio, I., Tornali, C., Bragazzi, N. L., and Martini, M. (2018). The discovery of insulin: an important milestone in the history of medicine. *Frontiers in endocrinology*, 9:613.

Vergara-Lope, A., Ennis, S., Vorechovsky, I., Pengelly, R. J., and Collins, A. (2019a). Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding rna and intergenic chromosome regions. *European Journal of Human Genetics*, 27(9):1436–1444.

Vergara-Lope, A., Jabalameli, M. R., Horscroft, C., Ennis, S., Collins, A., and Pengelly, R. J. (2019b). Linkage disequilibrium maps for european and african populations constructed from whole genome sequence data. *Scientific Data*, 6(1):1–4.

Vergotine, Z., Kengne, A. P., Erasmus, R. T., Yako, Y. Y., and Matsha, T. (2014). Rare mutations of peroxisome proliferator-activated receptor gamma: frequencies and relationship with insulin resistance and diabetes risk in the mixed ancestry population from south africa. *International journal of endocrinology*, 2014.

Viñuela, A., Varshney, A., Van De Bunt, M., Prasad, R. B., Asplund, O. B., Bennett, A., Boehnke, M. B., Brown, A. A., Erdos, M. R., Fadista, J., et al. (2019). Influence of genetic variants on gene expression in human pancreatic islets–implications for type 2 diabetes. *BioRxiv*, page 655670.

Vionnet, N., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Leprêtre, F., Lecoeur, C., Gallina, P., et al. (2000). Genomewide search for type 2 diabetes–susceptibility genes in french whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2–diabetes locus on chromosome 1q21–q24. *The American Journal of Human Genetics*, 67(6):1470–1480.

Virtanen, K. A., Lidell, M. E., Orava, J., Heglind, M., Westergren, R., Niemi, T., Taittonen, M., Laine, J., Savisto, N.-J., Enerbäck, S., et al. (2009). Functional brown adipose tissue in healthy adults. *New England Journal of Medicine*, 360(15):1518–1525.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012a). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Visscher, P. M., Goddard, M. E., Derks, E. M., and Wray, N. R. (2012b). Evidence-based psychiatric genetics, aka the false dichotomy between common and rare variant hypotheses. *Molecular psychiatry*, 17(5):474–485.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics*, 8(8).

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS biology*, 4(3).

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., Thorleifsson, G., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*, 42(7):579.

Vollmers, C., Gill, S., DiTacchio, L., Pulivarthy, S. R., Le, H. D., and Panda, S. (2009). Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression. *Proceedings of the National Academy of Sciences*, 106(50):21453–21458.

Volta, F. and Gerdes, J. M. (2017). The role of primary cilia in obesity and diabetes. *Annals of the New York Academy of Sciences*, 1391(1):71–84.

Volta, F., Scerbo, M. J., Seelig, A., Wagner, R., O'Brien, N., Gerst, F., Fritsche, A., Häring, H.-U., Zeigerer, A., Ullrich, S., et al. (2019). Glucose homeostasis is regulated by pancreatic β-cell cilia via endosomal epha-processing. *Nature communications*, 10(1):1–17.

von Engelhardt, D. (1989). Matthew dobson (1735–1784). clinical investigator of diabetes mellitus. In *Diabetes Its Medical and Cultural History*, pages 235–237. Springer.

Vujkovic, M., Keaton, J. M., Lynch, J. A., Miller, D. R., Zhou, J., Tcheandjieu, C., Huffman, J. E., Assimes, T. L., Lorenz, K., Zhu, X., et al. (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nature Genetics*, pages 1–12.

Wada, J. and Nakatsuka, A. (2016). Mitochondrial dynamics and mitochondrial dysfunction in diabetes. *Acta Medica Okayama*, 70(3):151–158.

Wagner, N.-M., Brandhorst, G., Czepluch, F., Lankeit, M., Eberle, C., Herzberg, S., Faustin, V., Riggert, J., Oellerich, M., Hasenfuss, G., et al. (2013). Circulating regulatory t cells are reduced in obesity and may identify subjects at increased metabolic and cardiovascular risk. *Obesity*, 21(3):461–468.

Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., et al. (2019). Recovery of trait heritability from whole genome sequence data. *bioRxiv*, page 588020.

Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H., McNeill, M., Smyth, D. J., Niblett, D., Cambien, F., Consortium, C., et al. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human molecular genetics*, 21(12):2815–2824.

Wang, C.-H., Chen, Y.-F., Wu, C.-Y., Wu, P.-C., Huang, Y.-L., Kao, C.-H., Lin, C.-H., Kao, L.-S., Tsai, T.-F., and Wei, Y.-H. (2014). Cisd2 modulates the differentiation and functioning of adipocytes by regulating intracellular ca2+ homeostasis. *Human molecular genetics*, 23(18):4770–4785.

Wang, C.-H., Tsai, T.-F., and Wei, Y.-H. (2015). Role of mitochondrial dysfunction and dysregulation of ca2+ homeostasis in insulin insensitivity of mammalian cells. *Annals of the New York Academy of Sciences*, 1350(1):66–76.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

Wang, K., Li, M., and Hakonarson, H. (2010a). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843.

Wang, K., Li, M., and Hakonarson, H. (2010b). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164.

Wang, Q. and Wu, H. (2018). T cells in adipose tissue: critical players in immunometabolism. *Frontiers in immunology*, 9:2509.

Wang, X., Chua, H.-X., Chen, P., Ong, R. T.-H., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.-C., Tay, W.-T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Human molecular genetics*, 22(11):2303–2311.

Wang, X., Vatamaniuk, M., Wang, S., Roneker, C., Simmons, R., and Lei, X. (2008). Molecular mechanisms for hyperinsulinaemia induced by overproduction of selenium-dependent glutathione peroxidase-1 in mice. *Diabetologia*, 51(8):1515–1524.

Wang-Sattler, R., Yu, Z., Herder, C., Messias, A. C., Floegel, A., He, Y., Heim, K., Campillos, M., Holzapfel, C., Thorand, B., et al. (2012). Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology*, 8(1):615.

Waters, K. M., Stram, D. O., Hassanein, M. T., Le Marchand, L., Wilkens, L. R., Maskarinec, G., Monroe, K. R., Kolonel, L. N., Altshuler, D., Henderson, B. E., et al. (2010). Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS genetics*, 6(8).

Watve, M. G. and Yajnik, C. S. (2007). Evolutionary origins of insulin resistance: a behavioral switch hypothesis. *BMC evolutionary biology*, 7(1):61.

Weatherill, A. R., Lee, J. Y., Zhao, L., Lemay, D. G., Youn, H. S., and Hwang, D. H. (2005). Saturated and polyunsaturated fatty acids reciprocally modulate dendritic cell functions mediated through tlr4. *The Journal of Immunology*, 174(9):5390–5397.

Webster, M. T. and Hurst, L. D. (2012). Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in genetics*, 28(3):101–109.

Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T., and Wright, C. F. (2019). Assessing the analytical validity of snp-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. *BioRxiv*, page 696799.

Weeks, D. E. and Lathrop, G. M. (1995). Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics*, 11(12):513–519.

Weikum, E. R., Liu, X., and Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Science*, 27(11):1876–1892.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006.

Wheeler, E. and Barroso, I. (2011). Genome-wide association studies and type 2 diabetes. *Briefings in functional genomics*, 10(2):52–60.

Wiederkehr, A. and Wollheim, C. B. (2008). Impact of mitochondrial calcium on the coupling of metabolism to insulin secretion in the pancreatic β-cell. *Cell calcium*, 44(1):64–76.

Wiese, S., Gronemeyer, T., Ofman, R., Kunze, M., Grou, C. P., Almeida, J. A., Eisenacher, M., Stephan, C., Hayen, H., Schollenberger, L., et al. (2007). Proteomics characterization of mouse kidney peroxi-

somes by tandem mass spectrometry and protein correlation profiling. *Molecular & cellular proteomics*, 6(12):2045–2057.

Willemsen, G., Ward, K. J., Bell, C. G., Christensen, K., Bowden, J., Dalgård, C., Harris, J. R., Kaprio, J., Lyle, R., Magnusson, P. K., et al. (2015). The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: the discordant twin (discotwin) consortium. *Twin Research and Human Genetics*, 18(6):762–771.

Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besan*cis*-con, S., Bommer, C., Esteghamati, A., Ogurtsova, K., Zhang, P., et al. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the international diabetes federation diabetes atlas. *Diabetes Research and Clinical Practice*, page 108072.

Woo, H. J. and Reifman, J. (2018). Genetic interaction effects reveal lipid-metabolic and inflammatory pathways underlying common metabolic disease risks. *BMC medical genomics*, 11(1):54.

Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., Beaumont, R. N., Tyrrell, J., Wood, A. R., Frayling, T. M., et al. (2019). Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *The American Journal of Human Genetics*, 104(2):275–286.

Wright, L., Brandon, A., Hoy, A., Forsberg, G.-B., Lelliott, C., Reznick, J., Löfgren, L., Oscarsson, J., Strömstedt, M., Cooney, G., et al. (2011). Amelioration of lipid-induced insulin resistance in rat skeletal muscle by overexpression of pgc-1β involves reductions in long-chain acyl-coa levels and oxidative stress. *Diabetologia*, 54(6):1417–1426.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942.

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., Yengo, L., Lloyd-Jones, L. R., Sidorenko, J., Wu, Y., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications*, 9(1):2941.

Yamada, T., Ida, T., Yamaoka, Y., Ozawa, K., Takasan, H., and Honjo, I. (1975). Two distinct patterns of glucose intolerance in icteric rats and rabbits. relationship to impaired liver mitochondria function. *The Journal of laboratory and clinical medicine*, 86(1):38–45.

Yan, R., Lai, S., Yang, Y., Shi, H., Cai, Z., Sorrentino, V., Du, H., and Chen, H. (2016). A novel type 2 diabetes risk allele increases the promoter activity of the muscle-specific small ankyrin 1 gene. *Scientific reports*, 6:25105.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565.

Yang, R. Y., Quan, J., Sodaei, R., Aguet, F., Segrè, A. V., Allen, J. A., Lanz, T. A., Reinhart, V., Crawford, M., Hasson, S., et al. (2018). A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. *bioRxiv*, page 311563.

Yang, Y., Ren, J., Tong, Y., Hu, X., Lv, Q., and Tong, N. (2016). Protective role of ppardelta in lipoapoptosis of pancreatic β cells. *Lipids*, 51(11):1259–1268.

Yaribeygi, H., Atkin, S. L., and Sahebkar, A. (2019). Mitochondrial dysfunction in diabetes and the regulatory roles of antidiabetic agents on the mitochondrial function. *Journal of cellular physiology*, 234(6):8402–8410.

Yazdanpanah, M., Chen, C., and Graham, J. (2013). Secondary analysis of publicly available data reveals superoxide and oxygen radical pathways are enriched for associations between type 2 diabetes and low-frequency variants. *Annals of human genetics*, 77(6):472–481.

Ye, R., Onodera, T., and Scherer, P. E. (2019). Lipotoxicity and β cell maintenance in obesity and type 2 diabetes. *Journal of the Endocrine Society*, 3(3):617–631.

Yeligar, S. M., Kang, B.-Y., Bijli, K. M., Kleinhenz, J. M., Murphy, T. C., Torres, G., San Martin, A., Sutliff, R. L., and Hart, C. M. (2018). Pparγ regulates mitochondrial structure and function and human pulmonary artery smooth muscle cell proliferation. *American journal of respiratory cell and molecular biology*, 58(5):648–657.

Yoo, Y. J., Sun, L., Poirier, J. G., Paterson, A. D., and Bull, S. B. (2017). Multiple linear combination (mlc) regression tests for common variants adapted to linkage disequilibrium structure. *Genetic epidemiology*, 41(2):108–121.

Young, A. I. (2019). Solving the missing heritability problem. *PLoS genetics*, 15(6):e1008222.

Yuan, S. and Larsson, S. C. (2020). Association of genetic variants related to plasma fatty acids with type 2 diabetes mellitus and glycaemic traits: a mendelian randomisation study. *Diabetologia*, 63(1):116–123.

Zaitlen, N., Pa*cis*-saniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *The American Journal of Human Genetics*, 86(1):23–33.

Zamponi, N., Zamponi, E., Cannas, S. A., Billoni, O. V., Helguera, P. R., and Chialvo, D. R. (2018). Mitochondrial network complexity emerges from fission/fusion dynamics. *Scientific reports*, 8(1):1–10.

Zaykin, D. V. and Zhivotovsky, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics*, 171(2):813–823.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., et al. (2007). Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341.

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M., Lloyd-Jones, L., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., Powell, J. E., Montgomery, G. W., Metspalu, A., Esko, T., Gibson, G., Wray, N. R., Visscher, P. M., and Yang, J. (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics*, 50(?):746–753.

Zhang, C., Qi, L., Hunter, D. J., Meigs, J. B., Manson, J. E., van Dam, R. M., and Hu, F. B. (2006). Variant of transcription factor 7-like 2 (tcf7l2) gene and the risk of type 2 diabetes in large cohorts of us women and men. *Diabetes*, 55(9):2645–2648.

Zhang, C., Xiao, C., Wang, P., Xu, W., Zhang, A., Li, Q., and Xu, X. (2014a). The alteration of th1/th2/th17/treg paradigm in patients with type 2 diabetes mellitus: relationship with diabetic nephropathy. *Human Immunology*, 75(4):289–296.

Zhang, C.-Y., Baffy, G., Perret, P., Krauss, S., Peroni, O., Grujic, D., Hagen, T., Vidal-Puig, A. J., Boss, O., Kim, Y.-B., et al. (2001). Uncoupling protein-2 negatively regulates insulin secretion and is a major link between obesity, β cell dysfunction, and type 2 diabetes. *Cell*, 105(6):745–755.

Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., and Hogenesch, J. B. (2014b). A circadian gene expression atlas in mammals: implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224.

Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., Bentley, D. R., and Morton, N. E. (2004). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proceedings of the National Academy of Sciences*, 101(52):18075–18080.

Zhang, W., Collins, A., Maniatis, N., Tapper, W., and Morton, N. E. (2002). Properties of linkage disequilibrium (ld) maps. *Proceedings of the National Academy of Sciences*, 99(26):17004–17007.

Zhao, L., Kwon, M.-J., Huang, S., Lee, J. Y., Fukase, K., Inohara, N., and Hwang, D. H. (2007). Differential modulation of nods signaling pathways by fatty acids in human colonic epithelial hct116 cells. *Journal of Biological Chemistry*, 282(16):11618–11628.

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.-J., Butterworth, A. S., Howson, J. M., Assimes, T. L., Chowdhury, R., Orho-Melander, M., Damrauer, S., et al. (2017). Identification of new susceptibility

loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature genetics*, 49(10):1450.

Zhenyukh, O., González-Amor, M., Rodrigues-Diez, R. R., Esteban, V., Ruiz-Ortega, M., Salaices, M., Mas, S., Briones, A. M., and Egido, J. (2018). Branched-chain amino acids promote endothelial dysfunction through increased reactive oxygen species generation and inflammation. *Journal of cellular and molecular medicine*, 22(10):4948–4962.

Zheutlin, A. B., Dennis, J., Karlsson Linnér, R., Moscati, A., Restrepo, N., Straub, P., Ruderfer, D., Castro, V. M., Chen, C.-Y., Ge, T., et al. (2019). Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. *American Journal of Psychiatry*, 176(10):846–855.

Zhou, W., Dai, J., Attanasio, M., and Hildebrandt, F. (2010). Nephrocystin-3 is required for ciliary function in zebrafish embryos. *American Journal of Physiology-Renal Physiology*, 299(1):F55–F62.

Zhou, Y. and Grill, V. (1995). Long term exposure to fatty acids and ketones inhibits b-cell functions in human pancreatic islets of langerhans. *The Journal of Clinical Endocrinology & Metabolism*, 80(5):1584–1590.

Zhou, Y.-P. and Grill, V. E. (1994). Long-term exposure of rat pancreatic islets to fatty acids inhibits glucose-induced insulin secretion and biosynthesis through a glucose fatty acid cycle. *The Journal of clinical investigation*, 93(2):870–876.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., et al. (2016). Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481.

Zimmet, P. (1982). Type 2 (non-insulin-dependent) diabetes-an epidemiological overview. *Diabetologia*, 22(6):399–411.

Zorzano, A., Liesa, M., and Palacín, M. (2009). Role of mitochondrial dynamics proteins in the pathophysiology of obesity and type 2 diabetes. *The international journal of biochemistry & cell biology*, 41(10):1846–1854.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.

A Appendix

A.1 Appendix 1: T2D *cis*–NEMG functions

Summary functions of the 50 T2D *cis*–NEMGs, as seen in Figure 19.

Gene	Summary of protein function
Mitochondrial Translation and transcription ●	
GATC	Glutamyl-TRNA(Gln) amidotransferase, subunit C. Allows the formation of correctly charged Gln-tRNA(Gln) through the transamidation of misacylated Glu-tRNA(Gln) in the mitochondria.
TRMT11	Catalytic subunit of an S-adenosyl-L-methionine-dependent tRNA methyltransferase complex that mediates the methylation of the guanosine nucleotide at position 10 (m2G10) in tRNAs.
MRPS33	Mitochondrial ribosomal protein S33
MTERFD3	Mitochondrial transcription termination factor.
NIF3L1	NGG1 interacting factor 3 like 1. May function as a transcription corepressor through its intersection with COPS2, negatively regulating the expression of genes involved in neuronal differentiation.
MARS2	Methionyl-tRNA synthetase 2, mitochondrial. This gene produces a mitochondrial methionyl-tRNA synthetase.
C12ORF65	Chr12 open reading frame 65. A mitochondrial matrix protein that appears to contribute to peptide chain termination in the mitochondrial translation machinery. May help rescuing stalled mitoribosomes during translation.
Oxidative phosphorylation ●	
COQ10B	Coenzyme Q10B. Required for the function of coenzyme Q in the respiratory chain. May serve as a chaperone or be involved in the transport of Q6 from its site of synthesis to the catalytic sites of the respiratory complexes.
NDUFB4	NADH:Ubiquinone Oxidoreductase subunit B4. Respiratory chain complex I accessory subunit (not catalytic).
COX7A2	Cytochrome C oxidase subunit 7A2. Encodes polypeptide 2 (liver isoform) of subunit Vlla of the terminal component of the mitochondrial respiratory chain.
COA6	Cytochrome C oxidase assembly factor 6. Involved in the maturation of the mitochondrial respiratory chain complex lV subunit MT-CO2/COX2.
NDUFV3	NADH dehydrogenase (ubiquinone) flavoprotein 3. Accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (complex l).
Lipid Metabolism ●	
GPAM	Glycerol-3-phosphate acyltransferase. Catalyses an essential step in glycerolipid biosynthesis (esterifies acyl-group from acyl-ACP to the sn-1 position of glycerol-3-phosphate).

PISD	Phosphatidylserine decarboxylase. Catalyses the formation of the phospholipid phosphatidylethanolamine (PE) and is involved in interorganelle trafficking of phosphatidylserine.
CYB5R2	NADH-cytochrome b5 reductase 2. Involved in multiple processes including desaturation and elongation of fatty acids, cholesterol biosynthesis, drug metabolism.
LACTB	Serine beta-lactamase-like protein. Regulates PISD levels and thus lipid metabolism (PMID:28329758). Forms stable filaments in the mitochondrial intermembrane space, promoting mitochondrial organization and micro-compartmentalization (PMID:19858488).
ACAD11	Acyl-CoA dehydrogenase family member 11. An acyl-CoA dehydrogenase enzyme involved in fatty acid β-oxidation.
ACADS	Acyl-CoA dehydrogenase short chain. Catalyses steps in the fatty acid β-oxidation and BCAA catabolism.
PCCA	Propionyl-coA carboxylase subunit alpha. Encodes the alpha subunit (the biotin binding subunit) of the heterodimeric mitochondrial enzyme propionyl-coA carboxylase, which catalyses the carboxylation of propionyl-coA (product of valine, isoleucine, methionine, threonine and odd-chain fatty acid metabolism) to methlmalonyl coA.
Amino Acid Metabolism ●	
ACSS1	Acyl-CoA synthetase short chain family member 1. Converts acetate to acetyl-CoA for entry to the TCA cycle and synthesizes propanoate from the product of isoleucine degradation (propanoyl-CoA). Important for maintaining normal body temperature during fasting and energy homeostasis. Essential for energy expenditure under ketogenic conditions.
ACADS	Acyl-CoA dehydrogenase short chain. Catalyses steps in the mitochondrial fatty acid β-oxidation pathway and branched chain amino acid catabolism.
PCCA	Propionyl-coA carboxylase subunit alpha. Encodes the alpha subunit (the biotin binding subunit) of the heterodimeric mitochondrial enzyme propionyl-coA carboxylase, which catalyses the carboxylation of propionyl-coA (product of valine, isoleucine, methionine, threonine and odd-chain fatty acid metabolism) to methlmalonyl coA. *MCCC1* paralog.
ABAT	4-Aminobutyrate aminotransferase. Involved in BCAA, alanine, aspartate and glutamate metabolism, plus catabolism of the neurotransmitter GABA into succinic semialdehyde. Can also convert delta- aminovalerate and beta-alanine.
MCCC1	Methylcrotonoyl-CoA carboxylase 1. Biotin-attachment subunit of the 3-methylcrotonyl-coA carboxylase, an enzyme that catalyzes the conversion of 3-methylcrotonyl-coA to 3-methylglutaconyl-coA, a critical step for leucine and isovaleric catabolism. *PCCA* paralog.
SUOX	Sulfite oxidase. Catalyses oxidation of sulphite to sulfate, the final reaction in cysteine and methionine degradation.

GLS2	Glutaminase 2. Catalyses the hydrolysis of glutamine to glutamate and ammonia. Promotes mitochondrial respiration and increases ATP generation in cells by catalysing the synthesis of glutamate and α- ketoglutarate. Increases cellular anti-oxidant function via NADH and glutathione production.
ALDH2	Aldehyde dehydrogenase 2 family member. The second enzyme of the major oxidative pathway of alcohol metabolism. Oxidize aldehydes to generate carboxylic acids for use in the muscle and heart.
CPS1	Carbamoyl-phosphate synthase 1. Catalyzes synthesis of carbamoyl phosphate from ammonia and bicarbonate, the first committed step of the urea cycle. May also be a core mitochondrial nucleoid protein.
Glycolysis / TCA cycle / Pyruvate Metabolism ●	
PDHA2	Pyruvate dehydrogenase E1 alpha 2 subunit. The pyruvate dehydrogenase complex catalyses conversion of pyruvate to acetyl-CoA and CO_2, linking the glycolytic pathway to the TCA cycle.
ACSS1	Acyl-CoA synthetase short chain family member 1. Converts acetate to acetyl-CoA for entry to the TCA cycle and synthesizes propanoate from the product of isoleucine degradation (propanoyl-CoA). Important for maintaining normal body temperature during fasting and for energy homeostasis. Essential for energy expenditure under ketogenic conditions.
IDH3A	Isocitrate dehydrogenase (NAD+) 3 alpha. Catalyses the decarboxylation of isocitrate to 2-oxoglutarate.
ALDH2	Aldehyde dehydrogenase 2 family member. The second enzyme of the major oxidative pathway of alcohol metabolism. Oxidize aldehydes to generate carboxylic acids for use in the muscle and heart.
Propanoate Metabolism ●	
ABAT	4-Aminobutyrate aminotransferase. Involved in BCAA, alanine, aspartate and glutamate metabolism. (Also responsible for catabolism of the neurotransmitter GABA into succinic semialdehyde). Can also convert delta-aminovalerate and beta-alanine.
ACSS1	Acyl-CoA synthetase short chain family member 1. Converts acetate to acetyl-CoA for entry to the TCA cycle and synthesizes propanoate from the product of isoleucine degradation (propanoyl-CoA). Important for maintaining normal body temperature during fasting and for energy homeostasis. Essential for energy expenditure under ketogenic conditions.
PCCA	Propionyl-coA carboxylase subunit alpha. Encodes the alpha subunit (the biotin binding subunit) of the heterodimeric mitochondrial enzyme propionyl-coA carboxylase, which catalyses the carboxylation of propionyl-coA (product of valine, isoleucine, methionine, threonine and odd-chain fatty acid metabolism) to methlmalonyl coA. *MCCC1* paralog.

ALDH2	Aldehyde dehydrogenase 2 family member. The second enzyme of the major oxidative pathway of alcohol metabolism. Oxidize aldehydes to generate carboxylic acids for use in the muscle and heart.
Butanoate Metabolism ●	
ACADS	Acyl-CoA dehydrogenase short chain. Catalyses steps in the mitochondrial fatty acid β-oxidation pathway and branched chain amino acid catabolism.
ABAT	4-Aminobutyrate aminotransferase. Involved in BCAA, alanine, aspartate and glutamate metabolism. (Also responsible for catabolism of the neurotransmitter GABA into succinic semialdehyde). Can also convert delta-aminovalerate and beta-alanine.
PDHA2	Pyruvate dehydrogenase E1 alpha 2 subunit. The pyruvate dehydrogenase complex catalyses conversion of pyruvate to acetyl-CoA and CO_2, linking the glycolytic pathway to the TCA cycle.
ALDH2	Aldehyde dehydrogenase 2 family member. The second enzyme of the major oxidative pathway of alcohol metabolism. Oxidize aldehydes to generate carboxylic acids for use in the muscle and heart.
Apoptosis ●	
PGAM5	PGAM family member 5, mitochondrial serine/threonine protein phosphatase). Involved in the TNF signalling pathway. May be regulator of mitochondrial dynamics. Acts as a central mediator for programmed necrosis induced by TNF, by reactive oxygen species and by calcium ionophore.
DIABLO	Diablo IAP-binding mitochondrial protein. Promotes apoptosis by activating caspases in the cytochrome c/Apaf-1/caspase-9 pathway.
Mitochondrial Organization/Dynamics ●	
MARCH5	Membrane associated ring-CH-type finger 5. Ubiquitin-protein ligase that plays a crucial role in the control of mitochondrial morphology by acting as a positive regulator of mitochondrial fission. May play a role in the prevention of cell senescence acting as a regulator of mitochondrial quality control.
LACTB	Serine beta-lactamase-like protein. Regulates PISD levels and thus lipid metabolism (PMID:28329758). Forms stable filaments in the mitochondrial intermembrane space, promoting mitochondrial organisation and micro-compartmentalization (PMID:19858488).
PGAM5	PGAM family member 5, mitochondrial serine/threonine protein phosphatase. May be regulator of mitochondrial dynamics. Acts as a central mediator for programmed necrosis induced by TNF, by reactive oxygen species and by calcium ionophore.
MFF	Mitochondrial fission factor. Encodes a protein that functions in mitochondrial and peroxisomal fission. Recruits dynamin-1-like protein (DNM1L) to the mitochondria.
MTFR1L	Mitochondrial fission regulator 1-like.

Mitochondrial Protein/Ion Transport ○	
ABCB9	ATP-binding cassette subfamily B member 9. ATP-dependent low-affinity peptide transporter which translocates a broad spectrum of peptide from the cytosol to the lysosomal lumen.
TOMM20	Translocase of outer membrane 20. Central component of the receptor complex responsible for the recognition and translocation of cytosolically synthesized mitochondrial preproteins, including tRNA.
HSPD1	Heat shock protein family D (Hsp60) member 1. Encodes a member of the chaperonin family. May act as a signalling molecule in the innate immune system. Essential for the folding and assembly of newly imported proteins in the mitochondria.
ABCA13	ATP binding cassette subfamily A member 13. Transmembrane transporter potentially involved in cholesterol transport.
CLIC4	Chloride intracellular channel 4. Forms a poorly selective ion channel that may also transport chloride ions (depending on the pH). Membrane insertion may only occur under oxidizing conditions. Promotes cell- surface expression of HRH3. Has potential roles in angiogenesis and maintaining apical-basolateral membrane polarity during mitosis and cytokinesis. Could also promote endothelial cell proliferation and regulate endothelial morphogenesis.
SFXN2	Sideroflexin 2. Cation transmembrane transporter activity.
SLC25A26	Solute carrier family 25 member 26. Involved in the transport of S-adenosylmethionine (SAM) into the mitochondria. Mutations in this gene are associated with combined OXPHOS deficiency 28.
DNA Damage Response ●	
FEN1	Flap structure-specific endonuclease 1. Removes 5' overhang in DNA repair (long patch base excision repair pathway) and processes the 5' ends of okazaki fragments in lagging strand DNA synthesis.
ALKBH3	AlkB homolog 3, alpha-ketoglutarate dependent dioxygenase. Dioxygenase that mediates demethylation of DNA and RNA containing 1-methyladenosine (m1A).
Autophagy & calcium homeostasis ●	
CCDC58	Coiled-coil domain containing 58. Potential GWAS hit associated with calcium measurement. *CISD2* CDGSH iron sulfur domain 2. Regulator of autophagy that contributes to antagonize BECN1-mediated cellular autophagy at the endoplasmic reticulum. Participates in the interaction of BCL1 with BECN1 and is required for BCL2-mediated depression of endoplasmic reticulum Ca2+ stores during autophagy.
MAIP1	Matrix AAA peptidase interacting protein 1. Promotes sorting of SMDT1/EMRE (the mitochondrial calcium unipoter complex) in mitochondria by ensuring its maturation.
HEBP1	Heme binding protein 1. An intracellular tetrapyrrole-binding protein. May bind free porphyrinogens that may be present in the cell and thus facilitate removal of these potentially toxic compounds. Promotes calcium mobilization and chemotaxis in monocytes and dendritic cells.

A.2 Appendix 2: Mitochondrial gene sets

All 41 gene sets were obtained from the Molecular Signatures Database (MSigDB) Liberzon et al. (2011), curated gene sets (C2). The gene sets have been collated from different sources which are referenced in brackets. Details for these gene sets can be found at https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=C2.

MSigDB C2 Gene Set (Source)

(**KEGG**)

Alanine aspartate and glutamate metabolism

Arginine and proline metabolism

Beta alanine metabolism

Butanoate metabolism

Citrate/TCA cycle

Fatty acid metabolism

Glycolysis and gluconeogenesis

Glycine, serine and threonine metabolism

Glyoxylate and dicarboxylate metabolism

Lysine degradation

Oxidative phosphorylation

PPAR signaling pathway

Propanoate metabolism

Pyruvate metabolism

Tryptophan metabolism

Valine leucine and isoleucine degradation

(**Reactome**)

Activated AMPK stimulates fatty acid oxidation in muscle

Branched chain amino acid catabolism

Fatty acid triacylglycerol and ketone body metabolism

Fatty-acyl CoA biosynthesis

Gluconeogenesis

Mitochondrial fatty acid β oxidation

Metabolism of amino acids and derivatives

Mitochondrial biogenesis

Mitochondrial calcium ion transport

Mitochondrial protein import

Mitochondrial tRNA aminoacylation

Purine catabolism

Pyrimidine catabolism

Pyrimidine metabolism

Pyruvate metabolism and citric acid TCA cycle

Release of apoptotic factors from the mitochondria

Respiratory electron transport

Synthesis and interconversion of nucleotide di- and triphosphates

Synthesis of very long chain fatty acyl CoAs

TCA cycle and respiratory electron transport

Tyrosine metabolism

(Others)

Mitochondria (Mootha)

Mitochondria pathway (Biocarta)

Mitochondria gene module (Wong)

Biotin carboxylases (manually defined)

(**Hallmark**)

Fatty acid metabolism

Oxidative phosphorylation (OXPHOS)

A.3 Appendix 3: GEO gene expression datasets

Table 20: NGT=normal glucose tolerance, IR=insulin resistant, T2D=Type 2 diabetes, FH+/-=positive/negative family history of T2D. M=Male, F=Female. For the arrays, HG-U133_Plus2=Affymetrix Human Genome; U133 Plus 2.0 Array (n=21276 genes); HG-U133A=Affymetrix Human Genome U133 Array (n=13931 genes); Hu6800=Affymetrix Human Full Length HuGeneFL Array (n=6021 genes); HTA-2_0 Affymetrix Human Transcriptome Array 2.0 (n=24254 genes); HG_U95A=Affymetrix Human Genome U95A Array (n=9666 genes); HuGene-1_0-st=Affymetrix Human Gene 1.0 ST Array (n=20212 genes); HG_U95Av2=Affymetrix Human Genome U95 Version 2 Array (n=9667 genes); U133_X3P, Affymetrix Human X3P Array (n=21157 genes); HuGene-1_1-st=Affymetrix Human Gene 1.1 ST Array (n=20212 genes). Age and BMI show the mean ± the standard deviation. *Individual sample level statistics were not available, summary data is presented.

Skeletal Muscle							
GEO ID tissue date submitted	Ca/Co (M/F)	Array	Medication	Disease duration	Co- variates	BMI	Age
GSE13070 Vastus lateralis Oct 2008 (Sears et al., 2009)	NGT 18 T2D 51 *Mostly male	HG- U133_Plus2	Discontinued 2 weeks prior	-	-	NGT 24.6±0.8* T2D 35.5±0.8* *Not matched*	NGT 25±0.8* T2D 36±0.8* *Not matched*
GSE25462 Vastus lateralis Mar 2011 (Jin et al., 2011)	FH- 7/8 FH+ 11/14 T2D 6/5	HG- U133_Plus2	None	T2D subjects >1 year.	Age & BMI	NGT 25.1±3.3 1 parent 27.7±5.6 2 parents 28.4±6.9 T2D 32.80±8.1	NGT 38±12.0 1 parent 36±10.0 2 parents 40±12.0 T2D 50±14.0

GEO ID tissue date submitted	Ca/Co (M/F)	Array	Medication	Disease duration	Co-variates	BMI	Age
GSE22435 Rectus abdominus Aug 2011 (Pihlajamäki et al., 2011)	NGT 0/7 T2D 0/10	HG-U133_Plus2	None	Diagnosed during the study.	Age & BMI	NGT 27.4±5.4 T2D 31.70±6.5	NGT 60±5.0 T2D 60±4.8
Subcutaneous Adipose							
GSE101492 May 2018 (Gao et al., 2018)	NGT 0/40 IR 0/40	HTA-2_0	None	-	Age	-	NGT 35.7±5.7* IR 36.4±6.3*
GSE26637 Jan 2011 (Soronen et al., 2012)	NGT 0/5 IGT 0/5	HG-U133_Plus2	None	N/A	-	NGT 22.0±0.7* IGT 32.5±1.7*	NGT 33±6* IGT 39±5*
GSE94752 Feb 2017 (Kulyte et al., 2017)	NGT 0/21 IR 0/18	HuGene-1_1-st	2 contraceptive pills, 4 SSRI, 1 Duloxetine, 2 thiazide and amiloride. One IGT subject on multiple medications	-	-	NGT 41±5* IGT 40±6*	NGT 36±7.8* IGT 40±8.1*
GSE20950 Mar 2010 (Hardy et al., 2011)	NGT 2/8 IGT 4/5	HG-U133_Plus2	-	Patients diagnosed during the study.	-	NGT 48±3* IGT 49±7* *Matched*	NGT 39±6* IGT 43±9*

GEO ID tissue date submitted	Ca/Co (M/F)	Array	Medication	Disease duration	Co-variates	BMI	Age
GSE27949 Mar 2011 (Keller et al., 2011)	NGT 11 IGT 10 T2D 12	HG-U133_Plus2	Discontinued 24h or one week (hypo-glycemic) prior.	-	Age & BMI	NGT 30.55±6.1 IGT 31.86±8.1 T2D 31.68±8.1	NGT 44±13.0 IGT 57±11.0 T2D 56±6.8
GSE13070 Oct 2008 (Sears et al., 2009)	NGT 6 T2D 28	HG-U133_Plus2	Discontinued two weeks prior.	-	-	NGT 24.6±0.8* T2D 35.5±0.8*	NGT 25±0.8* T2D 36±0.8*
Liver							
GSE64998 Jan 2015 (Kirchner et al., 2016)	NGT 7/0 T2D 8/0	HuGene-1_1-st	-	-	-	NGT 38.6±2.0* T2D 40.3±3.1*	NGT 46±7.6* T2D 41±11*
GSE15653 Apr 2009 (Pihlajamäki et al., 2009)	13/5	HG-U133A	None known	Diagnosed during the study	Age & BMI	NGT 51.5±4.4 T2D 52.4±6.6	NGT 39±10.9 T2D 46±12.4
Pancreatic islets							
GSE76894 Dec 2017 (Solimena et al., 2018) (Khamis et al., 2019)	NGT 83 T2D 19	HG-U133_Plus2	-	-	Age & BMI	NGT 26.5±3.6 T2D 25.8±4.2	NGT 60±16.2 T2D 72±7.5

GEO ID tissue date submitted	Ca/Co (M/F)	Array	Medication	Disease duration	Co- variates	BMI	Age
GSE25724 Nov 2010 (Dominguez et al., 2011)	NGT 4/3 T2D 3/3	HG- U133A	-	-	Age & BMI	NGT 24.8±2.5 T2D 26±2.2	NGT 58±17.3 T2D 71±9.2
GSE41762 Oct 2012 (Mahdi et al., 2012)	NGT 33/24 T2D 11/9	HuGene- 1_0-st	-	-	Age & BMI	NGT 25.4±3.1 T2D 28.5±4.6	NGT 56±10.6 T2D 59±9.3

A.4 Appendix 4: GSEA results, T2D *cis*–NEMGs vs the genomic background

GSEA results: T2D *cis*–NEMGs vs the genomic background

Dataset	Tissue	50 T2D *cis*–NEMGs vs ALL	50 random *cis*–NEMGs vs ALL (3 control gene sets)		
Meta-analysis	Muscle	↓<1.0*e*-04	↓<1.0*e*-04	↓<1.0*e*-04	↓<1.0*e*-04
Meta-analysis	Adipose	↓8.0*e*-04	↓0.002	↓2.0*e*-04	↓2.0*e*-04
Meta-analysis	Liver	n.s.	n.s.	↓0.036	n.s.
Meta-analysis	Pancreas	↓<4.0*e*-04	↓4.0*e*-04	↓0.001	↓6.7*e*-04
GSE13070	Muscle	↓<1.0*e*-04	↓<1.0*e*-04	↓<1.0*e*-04	↓<1.0*e*-04
GSE25462	Muscle	↓<4.0*e*-04	↓0.018	↓0.005	↓0.043
GSE22435	Muscle	↓<2.0*e*-04	↓2.0*e*-04	↓2.0*e*-04	↓2.0*e*-04
GSE101492	Adipose	↓0.049	n.s.	↓0.030	n.s.
GSE26637	Adipose	n.s.	n.s.	n.s.	n.s.
GSE94752	Adipose	↓0.014	↓0.006	n.s.	↓4.0*e*-04
GSE20950	Adipose	↓<1.7*e*-04	↓<1.3*e*-04	↓<1.3*e*-04	↓<1.3*e*-04
GSE27949	Adipose	n.s	n.s	n.s	n.s
GSE13070	Adipose	↓0.047	n.s.	↓0.047	↓0.001
GSE64998	Liver	↓0.008	n.s.	↓0.002	↓0.009
GSE15653	Liver	n.s.	n.s.	n.s.	n.s.
GSE76894	Pancreas	n.s.	↓0.001	↓0.012	↓0.001
GSE25724	Pancreas	↓<1.3*e*-04	↓<1.3*e*-04	↓2.0*e*-04	↓<1.3*e*-04
GSE41762	Pancreas	n.s.	n.s.	n.s.	n.s.
	Family history datasets				
GSE25462	Muscle	↓ 0.002	↓ 0.010	↓ 0.022	↓ 0.005

Table 21: Gene set enrichment analysis results comparing T2D *cis*-NEMGs to the genomic background. FDR-adjusted p-values $\leq$0.05 are shown, reflecting significant enrichment in the gene set for increased expression (↑), decreased expression (↓) and both increased and decreased, or 'mixed' expression (no arrow). GSEA used a wilcoxon statistic and 10,000 permutations. n.s. = not significant. As a comparison, GSEA was carried out for three random sets of 50 adipose *cis*-genes with eQTL > 2Mb away from a T2D location estimate. A p-value with < indicates that the GSEA returned the minimum FDR-adjusted p-value obtainable with 10,000 permutations.