From the Institute of Epidemiology II, Helmholtz Zentrum München
German Research Center for Environmental Health (GmbH)
Head: Prof. Dr. Annette Peters

and

the Institute of Medical Informatics, Biometry and Epidemiology,
Ludwig-Maximilians-Universität München
Chair: Prof. Dr. rer. nat. Ulrich Mansmann

Multi-omics of obesity and weight change in the post-genomic era

Thesis
submitted for a doctoral degree in natural sciences at the Faculty of Medicine,
Ludwig-Maximilians-Universität München

by
Simone Wahl

from
Mainz, Germany

2014

Printed with approval of the Faculty of Medicine
of the Ludwig-Maximilians-Universität München

Supervisor/Examiner:	Prof. Dr. Thomas Illig
Co-Examiners:	Prof. Dr. Gunnar Schotta
Dean:	Prof. Dr. med. Dr. h. c. Maximilian Reiser, FACR, FRCR
Date of oral examination:	13.05.2015

Abstract

The worldwide prevalence of obesity and comorbidities such as cardiovascular diseases and type 2 diabetes (T2D) is reaching epidemic proportions. In the past decade, a major effort has been made to elucidate the genetics behind obesity and related health problems. In large genome-wide association studies (GWAS), genetic variants have been identified that associate with anthropometric traits such as body mass index (BMI). "Post-genomic" obesity research now aims to understand the biological mechanisms underlying these associations, and to explain the large part of the heritability of BMI that could not be attributed to the identified genetic variants, the "missing heritability". This increasingly involves the integrated analysis of multiple omics data, referred to as "multi-omics".

This thesis comprises four studies that address the following post-genomic aims: (1) to metabolomically characterize selected genetic variants associated with obesity and T2D, (2) to explore the use of epigenomics for tackling the missing heritability of obesity and for better understanding the complex molecular processes linking obesity with metabolic disturbances such as insulin resistance and dyslipidemia, (3) to elucidate metabolomic and transcriptomic consequences of long-term weight change, and finally, (4) to identify metabolomic predictors of weight loss of obese children during lifestyle intervention.

To **characterize the strongest obesity risk locus, *FTO*, and the T2D risk locus *TCF7L2***, a novel strategy based on metabolomics measurements during different oral and intravenous challenge tests in healthy men was applied. This allowed the comprehensive description of physiological challenge responses, and the exploration of genotype effects on challenge responses. *TCF7L2* risk allele carriers showed a changed response of sphingomyelin and (lyso)phosphatidylcholine concentrations to intravenous glucose challenge. These perturbations could only be detected through the challenge test, demonstrating the utility of the approach in revealing early metabolic abnormalities prior to changes in conventional parameters of glucose homeostasis.

Next, an **epigenome-wide discovery and replication study of BMI and whole blood DNA methylation** was conducted based on more than 10,000 subjects of European and South Asian ancestry. This revealed solid associations for 187 methylation sites. Downstream analyses indicate an enrichment of these loci in open chromatin sites and an enrichment for genes involved in lipid- and insulin-related biological pathways. Fur-

thermore, a large part of these methylation sites were associated with gene expression at nearby genes and with genetic variants. Mendelian randomization and longitudinal analyses suggest that change methylation at the majority of loci was consequential rather than causal to change in BMI. Integration with clinical traits pinpoints selected methylation sites as potentially being involved in the development of obesity-related comorbidities.

An **integrated metabolomics and transcriptomics approach was applied for studying the metabolic consequences of long-term body weight change** in the general population. Serum metabolomics and whole blood transcriptomics measurements were available from a follow-up timepoint 7 years after initial assessment of weight status. Omics data were clustered into modules of tightly connected molecules using weighted correlation network analysis (WGCNA), followed by testing for association of the obtained modules with previous body weight change. This approach revealed six omics modules strongly associated with weight change. The four metabolite modules were centered around very low density lipoprotein (VLDL) subclasses and markers of energy metabolism, around high density lipoprotein (HDL) subclasses, around low density lipoprotein (LDL) subclasses, and around amino acids. The two gene expression modules reflected basophil/mast cell function and red blood cell development, respectively.

Finally, serum metabolomic, anthropometric and clinical data were used to **predict weight loss success over a 1-year lifestyle intervention program for obese children**. Using the regularized regression approach *least absolute shrinkage and selection operator* (LASSO), a sparse model for weight loss was built and carefully validated. The results point towards a significant role of abdominal adipose tissue and phospholipid metabolism in weight regulation.

The research efforts undertaken in this thesis not only deal with the challenges of multi-omics data, but also demonstrate their enormous potential for post-genomic research. Altogether, the findings of the studies in this thesis contribute to the completion of the complex mosaic of described molecular processes underlying obesity and weight change and relating them with comorbidities such as T2D and cardiovascular diseases. This improves the understanding of disease pathogenesis and presents a starting point for the development of individualized treatment and prevention strategies for obesity and its comorbidities.

Zusammenfassung

Die weltweite Prävalenz von Adipositas und Komorbiditäten wie kardiovaskulären Erkrankungen und Typ 2 Diabetes (T2D) hat epidemische Ausmaße erreicht. Im letzten Jahrzehnt wurde ein großer Aufwand in die Aufklärung der genetischen Grundlagen von Adipositas und der damit verbundenen Gesundheitsprobleme gesteckt. In großen genomweiten Assoziationsstudien (GWAS) wurden genetische Varianten identifiziert, die mit anthropometrischen Maßen wie dem *body mass index* (BMI) assoziiert sind. Die "post-genomische" Adipositasforschung zielt nun darauf ab, die biologischen Mechanismen zu verstehen, die diesen Assoziationen zugrunde liegen, sowie den großen Anteil der Heritabilität von BMI zu erklären, der den identifizierten genetischen Varianten nicht zugeschrieben werden konnte. Das beinhaltet zunehmend die integrierte Analyse verschiedener "omik"-Daten, als "Multi-omik" bezeichnet.

Diese Dissertation umfasst vier Studien mit den folgendenen post-genomischen Zielen: (1) Ausgewählte mit Adipositas und T2D assoziierte genetische Varianten sollen metabolomisch charakterisieren werden. (2) Die Rolle der Epigenomik soll im Hinblick auf die fehlende Heritabilität von Adipositas untersucht werden, sowie hinsichtlich der komplexen molekularen Prozesse, die Adipositas und Gewichtsveränderung mit Stoffwechselstörungen wie Insulinresistenz und Dyslipidämie verbinden. (3) Metabolomische und transkriptomische Folgen langfristiger Gewichtsveränderung sollen aufgeklärt werden. (4) Schließlich sollen metabolomische Prädiktoren der Gewichtsabnahme adipöser Kinder während einer Lebensstilintervention identifiziert werden.

Um den **stärksten Adipositas-Risikolokus, *FTO*, und den T2D-Risikolokus *TCF7L2* zu charakterisieren** wurde eine neue Strategie angewandt, die auf metabolomischen Messungen während verschiedener oraler und intravenöser Belastungstests in gesunden Männern beruht. Dies ermöglichte es, physiologische Reaktionen auf die metabolischen Belastungen umfassend zu beschreiben und Genotypeffekte auf diese Reaktionen zu erforschen. *TCF7L2*-Risikoallelträger wiesen eine veränderte Reaktion von Sphingomyelin- und (Lyso-)Phosphatidylcholin-Konzentrationen auf intravenöse Glukose-Belastung auf. Diese Veränderungen konnten erst durch den Belastungstest erkannt werden. Das zeigt, dass der Ansatz in der Lage ist, frühe Stoffwechselabnormalitäten zu erkennen, welche Veränderungen in konventionellen Parametern des Glukosehaushaltes vorangehen.

Eine **epigenom-weite Assoziationsstudie zu BMI und DNA-Methylierung im Vollblut** umfasste über 10.000 Individuen europäischer und südasiatischer Herkunft in einem Identifikations- und einem Replikationsschritt. Dabei wurden 187 Methylierungsstellen identifiziert, die solide Assoziationen mit BMI zeigten. In Folgeanalysen konnte eine Anreicherung dieser Methylierungsstellen in Regionen aktiven Chromatins beobachtet werden, sowie eine Anreichung für Gene, die in lipid- und insulin-assoziierten Stoffwechselwegen eine Rolle spielen. Darüber hinaus assoziierte ein großer Teil dieser Methylierungsstellen mit der Expression nahegelegener Gene, sowie mit genetischen Varianten. Die Ergebnisse von *Mendelian randomization* und longitudinalen Analysen deuten an, dass die Mehrzahl der Methylierungsveränderungen nicht eine Ursache, sondern eine Folge von BMI-Veränderungen darstellt. Wie die Integration mit klinischen Phänotypen zeigt, sind bestimmte Methylierungsstellen möglicherweise an der Entstehung von adipositas-assoziierten Komorbiditäten beteiligt.

Mithilfe eines **integrierten Metabolomik- und Transkriptomik-Ansatzes wurden die metabolischen Konsequenzen langfristiger Gewichtsveränderungen in der allgemeinen Bevölkerung untersucht**. Metabolomik-Messungen im Serum und Transkriptomik-Messungen im Vollblut waren zu einem Follow-up-Zeitpunkt 7 Jahre nach der ersten Erfassung des Gewichtsstatus verfügbar. Die omik-Daten wurden mithilfe einer gewichteten Korrelations-Netzwerkanalyse (*weighted correlation network analysis*, WGCNA) in Module stark korrelierter Moleküle gruppiert. Danach wurde die Assoziation dieser Module mit vorheriger Gewichtsveränderung untersucht. Dieser Ansatz brachte sechs omik-Module hervor, die stark mit Gewichtsveränderung assoziiert waren. Die vier Metaboliten-Module waren um Subklassen von Lipoproteinen sehr niedriger Dichte (*very low density lipoprotein*, VLDL) sowie Energiestoffwechsel-Metaboliten, um Subklassen von Lipoproteinen hoher Dichte (*high density lipoprotein*, HDL), um Subklassen von Lipoproteinen niedriger Dichte (*low density lipoprotein*, LDL), und um Aminosäuren gruppiert. Die beiden Genexpressions-Module reflektierten Basophile/Mastzell-Funktion respektive rote Blutzell-Entwicklung.

In der letzten Studie wurde mithilfe von metabolomischen, anthropometrischen und klinischen Daten der **Erfolg bei der Gewichtsabnahme während einer einjährigen Lebensstil-Intervention für adipöse Kinder prognostiziert**. Mithilfe eines regularisierten Regressionsansatzes, *least absolute shrinkage and selection operator* (LASSO), wurde ein sparsames Modell für die Gewichtsabnahme entwickelt und gründlich validiert. Die Ergebnisse weisen auf eine Rolle von abdominellem Fettgewebe sowie Phospholipid-Stoffwechsel bei der Gewichtsregulation hin.

Die Forschungsarbeiten, die im Zusammenhang mit der Dissertation durchgeführt wurden, zeigen Herausforderungen der Multi-omik-Daten sowie Lösungsansätze auf, und demonstrieren das enorme Potential dieser Daten für die post-genomische Forschung. Die Ergebnisse der Studien dieser Dissertation tragen gemeinsam zur Vervollständigung des kom-

plexen Mosaiks beschriebener molekularer Prozesse bei, welche Adipositas und Gewichtsveränderung zugrunde liegen und diese mit Komorbiditäten wie T2D und kardiovaskulären Erkrankungen verbinden. Das verbessert das Verständnis von Erkrankungsmechanismen und liefert einen Ausgangspunkt für die Entwicklung individualisierter Behandlungs- und Präventionsstrategien für Adipositas und damit verbundene Komorbiditäten.

Publications and contributions

A structured overview of publications by the thesis author within the context of the thesis is given in Figure 1.

Manuscripts that are part of the thesis

Section 4.1

- **Wahl S***, Krug S*, Then C*, Kirchhofer A, Kastenmüller G, Brand T, Skurk T, Claussnitzer M, Huth C, Heier M, Meisinger C, Peters A, Thorand B, Gieger C, Prehn C, Römisch-Margl W, Adamski J, Suhre K, Illig T, Grallert H, Laumen H, Seissler J, Hauner H (2014). "Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele." *Metabolomics*, **10**(3), 386-401.

  Contributions: I performed the data preprocessing and statistical analysis for this manuscript and wrote the manuscript. KS and TC conducted the challenge tests and blood sampling, and revised the manuscript.

- Then C*, **Wahl S***, Kirchhofer A, Grallert H, Krug S, Kastenmüller G, Römisch-Margl W, Claussnitzer M, Illig T, Heier M, Meisinger C, Adamski J, Thorand B, Huth C, Peters A, Prehn C, Heukamp I, Laumen H, Lechner A, Hauner H, Seissler J (2013). "Plasma metabolomics reveal alterations of sphingo- and glycerophospholipid levels in non-diabetic carriers of the Transcription Factor 7-Like 2 polymorphism rs7903146." *PLoS One*, **8**(10), e78430.

  Contributions: I performed the data preprocessing and statistical analysis concerning the metabolomics data for this manuscript, prepared the tables and wrote parts of the manuscript. CT conducted the challenge tests and the blood sampling, performed analyses concerning the clinical parameters and wrote large parts of the manuscript.

*contributed equally

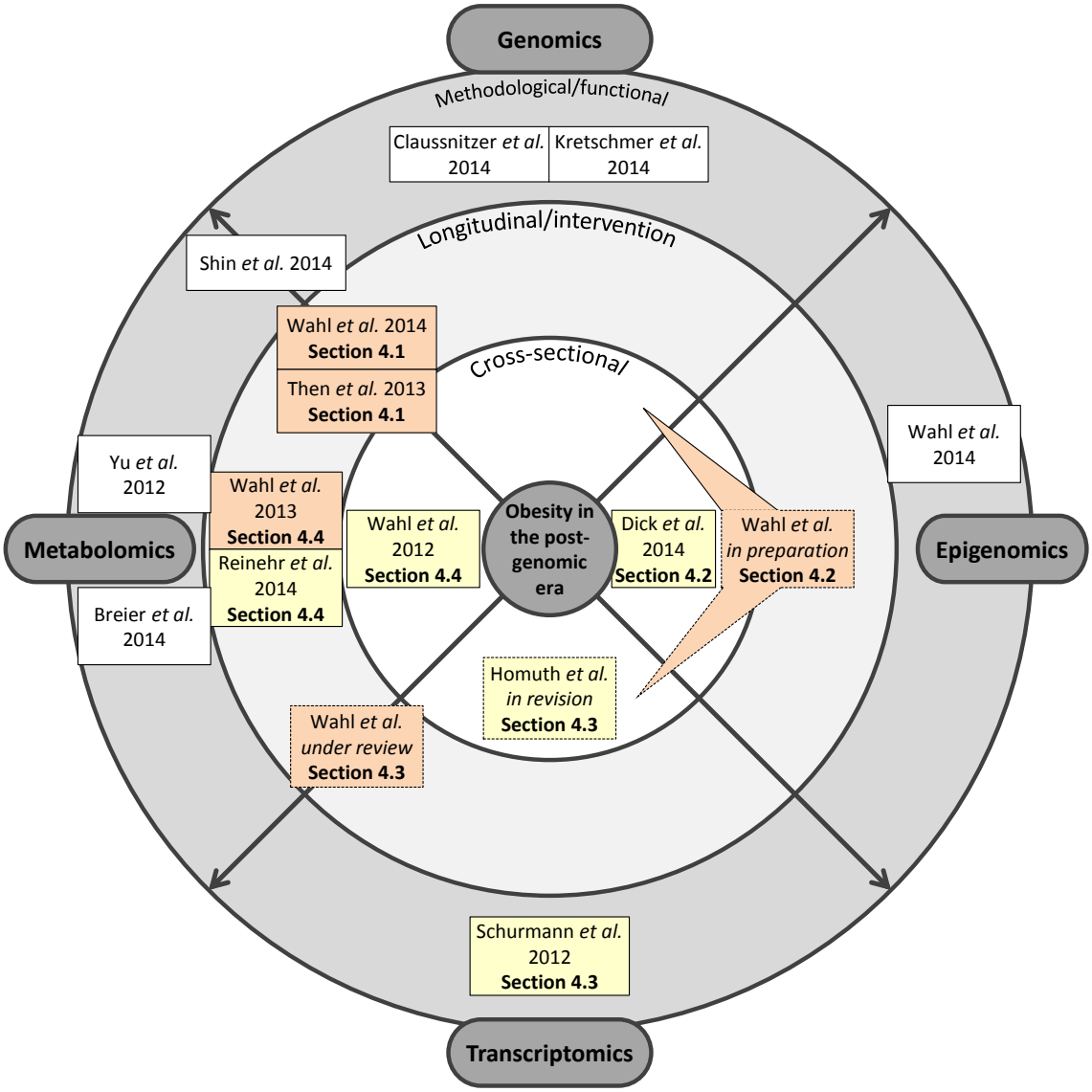

Figure 1: Multi-omics of obesity in the post-genomic era – an overview of publications by the thesis author in the context of the thesis. Color coding: orange, publications that are part of the thesis; yellow, publications that are closely related to the contents of the thesis and share aims with the thesis; white, additional publications that are not directly related to the thesis.

Section 4.2

- **Wahl S***, Lehne B*, Drong AW*, Loh M*, Zeilinger S, Fiorito G, Kasela S, Richmond R, Dehghan A, Franke L, Esko T, Milani L, Relton CL, Kriebel J, Prokisch H, Herder C, Peters A, Illig T, Waldenberger M, Bell JT, Franco OH, van der Harst P, Lindgren CM, McCarthy MI, Matullo G, Gieger C#, Kooner JS#, Grallert H#, Chambers JC#. "Epigenome-wide association study reveals extensive perturbations in DNA methylation associated with adiposity and its metabolic consequences." *in preparation.*

*,# contributed equally

Contributions: I performed the data preprocessing and all statistical analysis steps for the KORA F4 and F3 studies, set up the discovery and replication strategy together with BL, AD, ML, CG, HG and JC, conducted the meta-analysis in parallel with AD, and developed the analysis strategies for several downstream analyses (gene expression, clinical phenotypes, Mendelian randomization, longitudinal associations), for which I also conducted the meta-analysis. Other downstream analyses were developed by BL (SNP associations, *ad hoc* Mendelian randomization), AD (secondary signals, enrichment analyses) and ML (cross-tissue correlation, incident disease analysis), who also performed the respective meta-analyses. I wrote large parts of the manuscript.

Section 4.3

- **Wahl S***, Vogt S*, Stückler F, Krumsiek J, Bartel J, Schramm K, Carstensen M, Rathmann W, Roden M, Jourdan C, Kangas AJ, Soininen P, Ala-Korpela M, Nöthlings U, Boeing H, Theis F, Meisinger C, Waldenberger M, Suhre K, Gieger C, Kastenmüller G, Illig T, Linseisen J, Peters A, Prokisch H, Herder C, Thorand B$^{\#}$, Grallert H$^{\#}$. "Metabolic signature of weight change: an integrative metabolomics and transcriptomics approach." *under review.*

  Contributions: I performed the data preprocessing and statistical analysis for this manuscript and wrote the manuscript. Parts of the sensitivity analysis (preparation of lifestyle, medication and disease variables) and parts of the discussion were prepared by SV.

Section 4.4

- **Wahl S**, Holzapfel C, Yu Z, Breier M, Kondofersky I, Fuchs C, Singmann P, Prehn C, Adamski J, Grallert H, Illig T, Wang-Sattler R, Reinehr T (2013). "Metabolomics reveals determinants of overweight reduction during lifestyle intervention in obese children." *Metabolomics*, **9**(6), 1157-1167.

  Contributions: I performed the data preprocessing and statistical analysis for this manuscript and wrote the manuscript.

Additional manuscripts that were prepared during my time as a doctoral student

- Yu Z, Kastenmüller G, He Y, Belcredi P, Möller G, Prehn C, Mendes J, **Wahl S**, Roemisch-Margl W, Ceglarek U, Polonikov A, Dahmen N, Prokisch H, Xie L, Li Y, Wichmann HE, Peters A, Kronenberg F, Suhre K, Adamski J, Illig T, Wang-Sattler R

*,$^{\#}$ contributed equally

(2011). "Differences between human plasma and serum metabolite profiles." *PLoS One*, **6**(7), e21230.

- **Wahl S**, Yu Z, Kleber M, Singmann P, Holzapfel C, He Y, Mittelstrass K, Polonikov A, Prehn C, Römisch-Margl W, Adamski J, Suhre K, Grallert H, Illig T, Wang-Sattler R, Reinehr T (2012). "Childhood obesity is associated with changes in the serum metabolite profile." *Obesity Facts*, **5**(5), 660-670.

- Schurmann C*, Heim K*, Schillert A*, Blankenberg S, Carstensen M, Dörr M, Endlich K, Felix SB, Gieger C, Grallert H, Herder C, Hoffmann W, Homuth G, Illig T, Kruppa J, Meitinger T, Müller C, Nauck M, Peters A, Rettig R, Roden M, Strauch K, Völker U, Völzke H, **Wahl S**, Wallaschofski H, Wild PS, Zeller T, Teumer A#, Prokisch H#, Ziegler A# (2012). "Analyzing Illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the MetaXpress consortium." *PLoS One*, **7**(12), e50938.

- Reinehr T, Wolters B, Knop C, Lass N, Hellmuth C, Harder U, Peissner W, **Wahl S**, Grallert H, Adamski J, Illig T, Prehn C, Yu Z, Wang-Sattler R, Koletzko B (2014). "Changes in the serum metabolite profile in obese children with weight loss." *Eur J Nutr*, DOI: 10.1007/s00394-014-0698-8.

- Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, Lee H, Oskolkov N, Fadista J, Ehlers K, **Wahl S**, Hoffmann C, Qian K, Rönn T, Riess H, Müller-Nurasyid M, Bretschneider N, Schroeder T, Skurk T, Horsthemke B, DIAGRAM+ Consortium, Spieler D, Klingenspor M, Seifert M, Kern MJ, Mejhert N, Dahlman I, Hansson O, Hauck SM, Blüher M, Arner P, Groop L, Illig T, Suhre K, Hsu Y-H, Mellgren G, Hauner H, Laumen H (2014). "Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms." *Cell*, **156**(1-2), 343-358.

- Shin S-Y*, Petersen A-K*, **Wahl S**, Zhai G, Römisch-Margl W, Small KS, Döring A, Kato B, Peters A, Grundberg E, Prehn C, Wang-Sattler R, Wichmann H-E, Hrabé de Angelis M, Illig T, Adamski J, Deloukas P, Spector TD, Suhre K, Gieger C, Soranzo N (2014). "Interrogating causal pathways linking genetic variants, small molecule metabolites and circulating lipids." *Genome Medicine*, **6**(3), 25.

- Breier M, **Wahl S**, Prehn C, Fugmann M, Ferrari U, Weise M, Banning F, Seissler J, Grallert H, Adamski J, Lechner A (2014). "Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples." *PLoS ONE*, **9**(2), e89728.

- Kretschmer A, Möller G, Lee H, Laumen H, von Toerne C, Schramm K, Prokisch H, Eyerich S, **Wahl S**, Baurecht H, Franke A, Claussnitzer M, Eyerich K, Teumer A, Milani L, Klopp N, Hauck SM, Illig T, Peters A, Waldenberger M, Adamski J, Reischl E#,

*,# contributed equally

Weidinger S[#] (2014). "A common atopy-associated variant in the Th2 cytokine locus control region impacts transcriptional regulation and alters SMAD3 and SP1 binding." *Allergy*, **69**(5), 632-642.

- Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, **Wahl S**, Meduri E, Morange P-E, Gagnon F, Grallert H, Waldenberger M, Peters A, Erdmann J, Hengstenberg C, Cambien F, Goodall AH, Ouwehand WH, Schunkert H, Thompson JR, Spector TD, Gieger C, Trégouët D-A, Deloukas P, Samani NJ (2014). "Association of body mass index with epigenetic differences in the hypoxia-inducible factor alpha 3 gene: a genome-wide analysis of DNA methylation." *The Lancet*, **383**(9933), 1990-1998.

- **Wahl S**, Fenske N, Zeilinger S, Suhre K, Gieger C, Waldenberger M, Grallert H, Schmid M (2014). "On the potential of models for location and scale for genome-wide DNA methylation data." *BMC Bioinformatics*, **15**, 232.

- Homuth G[*], **Wahl S***, Müller C*, Schurmann C, Mäder U, Blankenberg S, Carstensen M, Dörr M, Endlich K, Englbrecht C, Felix SB, Gieger C, Grallert H, Herder C, Illig T, Kruppa J, Marzi CS, Mayerle J, Meitinger T, Metspalu A, Nauck M, Peters A, Rathmann W, Reinmaa E, Rettig R, Roden M, Schillert A, Schramm K, Steil L, Strauch K, Teumer A, Völzke H, Wallaschofski H, Wild PS, Ziegler A, Völker U[#], Prokisch P[#], Zeller T[#]. "Whole-blood transcriptome profiling of population-based cohorts reveals major gene expression changes correlated with body mass index." *in revision.*

[*,#] contributed equally

Contents

1 Introduction

1.1 Obesity – a global health problem

1.1.1 Definitions

Overweight and obesity are defined as conditions of excessive fat accumulation that may impair health (World Health Organization, 2000). With body fat mass being difficult to measure in a non-invasive way, the body mass index (BMI),

$$\text{BMI} = \frac{\text{Body weight (kg)}}{(\text{Body height (m)})^2},$$

has evolved as the traditional measure for body size in adults (World Health Organization, 2000). According to the WHO definition, normal weight is defined as BMI $\geq$ 18.5 - $<$ 25kg/m^2, overweight as BMI $\geq$ 25kg/m^2, and obesity as BMI $\geq$ 30kg/m^2.

Alternative non-invasive measures for body size include waist circumference, waist-hip ratio (WHR), and waist-height ratio. These measures reflect abdominal fat mass, which might be a better indicator for metabolic consequences than total body mass, owing to the metabolic activity of abdominal adipose tissue, specifically visceral adipose tissue (Després, 2006) (see Section 1.1.4). However, evidence for the superiority of these measures over BMI is not fully consistent (Huxley *et al.*, 2010, Taylor *et al.*, 2010, Janssen *et al.*, 2005, Pischon *et al.*, 2008). In addition, information on BMI is largely available in epidemiological studies, making it more suitable as a measure of obesity in large-scale omics meta-analyses and replication efforts.

In children, body mass highly depends on age, sex, pubertal state and ethnicity, so subgroup-specific percentile curves are required (Han *et al.*, 2010). BMI percentile curves can be calculated using the LMS method (Cole, 1990):

$$\text{BMI} = M(t) \cdot (1 + L(t)S(t)z_\alpha)^{\frac{1}{L(t)}},$$

where α represents the percentile, M the median BMI at age t, S the coefficient of variation of BMI at age t, and L a Box-Cox power transformation addressing the age-dependent skewness in BMI. z represents the z-score, or *standard deviation score* (SDS), of the stan-

dard normal distribution. Conversely, the BMI-SDS can be calculated from a given BMI:

$$\text{BMI-SDS}_{LMS} = \frac{\left(\frac{\text{BMI}}{M(t)}\right)^{L(t)} - 1}{L(t)S(t)}.$$

The International Obesity Taskforce (IOTF) provides international percentile curves (Cole *et al.*, 2000). In addition, reference curves based on a German population were published (Kromeyer-Hauschild *et al.*, 2001).

Due to the population specificity of BMI percentile curves, the definition of global cutoff points for overweight and obesity is challenging (Han *et al.*, 2010). According to the IOTF, the percentiles passing through a BMI of 25 and 30 kg/m^2 at the age of 18 are recommended as cutoff points (Cole *et al.*, 2000). In German reference data (Kromeyer-Hauschild *et al.*, 2001), these are approximately the 90$^{\text{th}}$ and 97$^{\text{th}}$ percentile (corresponding to BMI-SDS values of 1.282 and 1.881, respectively) (Wabitsch *et al.*, 2009). For the characterization of the weight of extremely obese children (BMI above the 99.5$^{\text{th}}$ percentile), use of the BMI-SDS is recommended (Wabitsch *et al.*, 2009).

1.1.2 Prevalence and trends in obesity

In the last few decades, the prevalence of overweight and obesity has increased dramatically worldwide, and is now at epidemic proportions (Ng *et al.*, 2014). Specifically, between 1980 and 2013, global prevalence of overweight and obesity increased by 27.5% among adults, reaching 36.9% and 10% in men, and 38.0% and 13.5% in women in 2013, respectively (Ng *et al.*, 2014). Prevalence was larger in developed countries throughout the observation period, although an increasing trend was observed for both developed and developing countries. In the German EPIC Postdam study, obesity prevalence increased from 16.6% in men and 15.8% in women in 1994-1998 to 24.6% in men and 22.2% in women in 2004-2008, with similar numbers observed in other European countries (von Ruesten *et al.*, 2011).

Obesity increasingly affects children. According to the IOTF cutoffs, the prevalence of overweight and obesity increased by 47.1% between 1980 and 2013 and reached on average 23.8% (boys) and 22.6% (girls) in developed countries in 2013 (Ng *et al.*, 2014). Lower, albeit increasing, numbers were observed for developing countries. In Germany, 20.5% of boys and 19.4% of girls (age $<$ 20 years) were overweight or obese in 2013, and 5.5% of boys and 5.3% of girls were obese (Ng *et al.*, 2014).

Predictions suggest that the rate of increase in obesity prevalence might decline in developed countries, including Germany (Ng *et al.*, 2014, Wabitsch *et al.*, 2014).

1.1.3 Etiology of excess body mass

The development of obesity is ultimately due to a chronic imbalance between energy intake and energy expenditure, which are mutually regulated through complex signaling mechanisms in intestine, brain, adipose tissue and further tissues (Woods and D'Alessio, 2008, Bell *et al.*, 2005). This imbalance originates from a multifactorial interplay between predisposing (epi-)genetic factors, *in utero* influences, and disadvantageous environmental and behavioral factors (Rhee *et al.*, 2012). An exception is monogenic forms of obesity, which arise from rare mutations (Loos and Bouchard, 2003).

In human evolution, a genetic setup has become prevalent that promotes parsimonious energy expenditure and rapid fat storage in times of plenty, improving survival during later food shortages. This development might be due to a positive selection of "thrifty" genotypes, i.e. genotypes promoting fat storage (Neel, 1962), or to random mutation and genetic drift (Speakman, 2007). In recent years, industrialization has promoted an "obesogenic" environment that is characterized by a sedentary lifestyle and the availability of high-caloric diets. In this environment, such genotypes tend to promote the development of obesity (Bell *et al.*, 2005, Loos and Bouchard, 2003).

The heritable component of BMI has been estimated at 40-70% from twin, adoption and family studies (Maes *et al.*, 1997, Atwood *et al.*, 2002, Salsberry and Reagan, 2010, Schousboe *et al.*, 2003, Stunkard *et al.*, 1986, Loos and Bouchard, 2003). However, despite considerable efforts to characterize the underlying genetic variants, the hitherto identified single nucleotide polymorphisms (SNPs) explain merely 1.45% of the variability in BMI (see Section 1.2.1, Speliotes *et al.* (2010)).

1.1.4 Health consequences of obesity

The high prevalence of overweight and obesity is greatly concerning, considering the serious health consequences of excess body mass. These include type 2 diabetes (T2D), cardiovascular diseases (including stroke, hypertension and coronary artery disease), different cancers, asthma, gallbladder disease, osteoarthritis and chronic back pain (Guh *et al.*, 2009). In addition, obesity is associated with increased all-cause mortality (Pischon *et al.*, 2008). Alarmingly, metabolic and cardiovascular risk factors such as insulin resistance, dyslipidemia, hypertension and chronic inflammation are already prevalent in obese children (Ebbeling *et al.*, 2002, Cook *et al.*, 2003), possibly establishing an increased adult risk of cardiovascular diseases (Owen *et al.*, 2009).

A causal relationship has been established between increased BMI and cardiometabolic traits, including T2D and insulin resistance, heart failure, dyslipidemia (increased triglyceride (TG) and decreased high density lipoprotein (HDL) cholesterol levels), hypertension, and the inflammatory marker C-reactive protein (CRP) (Fall *et al.*, 2013).

The underlying pathophysiology is not completely understood. Several mechanisms have

been described by which obesity increases T2D risk. They include the increased release of inflammatory cytokines – such as tumor necrosis factor-α (TNF-α), interleukin-6 (IL-6) and monocyte chemoattractant protein-1 (MCP-1) – from adipose tissue, which promote insulin resistance through different signal cascades, and through an inhibiting effect of TNF-α on the secretion of the insulin sensitizer adiponectin (Kahn *et al.*, 2006, Haslam and James, 2005). In addition, adipose tissue secretes non-esterified fatty acids (NEFAs), leading to increased NEFA concentrations in skeletal muscle and in the liver (de Ferranti and Mozaffarian, 2008). There, NEFAs and NEFA metabolites exert inhibiting effects on the insulin signaling cascade, e.g. through serine/threonine phosphorylation of insulin receptor substrates (IRS-1 and IRS-2) (Kahn *et al.*, 2006). Furthermore, chronically elevated NEFA levels might contribute to the development of β-cell dysfunction (Lupi *et al.*, 2002).

NEFAs and insulin resistance are also believed to be centrally involved in the obesity-related increased cardiovascular risk. For instance, increased NEFA concentrations trigger hepatic TG and very low density lipoprotein (VLDL) production (Klop *et al.*, 2013) and increase the activity of hepatic lipase (Brunzell and Hokanson, 1999). In the insulin resistant state, LDL receptor activity might also be impaired, resulting in a reduced clearance of TG-rich lipoproteins (VLDL, low density lipoprotein (LDL)) (Van Gaal *et al.*, 2006). Hepatic lipase hydrolyses TGs from TG-rich lipoproteins, producing small dense LDL particles, a process that also involves cholesterol ester transfer protein (CETP) (Van Gaal *et al.*, 2006, Klop *et al.*, 2013). Small dense LDL particles are specifically atherogenic due to their slow plasma clearance, their enhanced susceptibility to oxidation, their increased ability to enter the subendothelial space, and their lower affinity for the LDL receptor, causing them to be mostly taken up by macrophage scavenger receptor (Van Gaal *et al.*, 2006, Klop *et al.*, 2013). Furthermore, cytokines and hormones secreted by the adipose tissue contribute to the development of atherosclerotic lesions through their inflammatory and prothrombotic potential (Van Gaal *et al.*, 2006).

1.1.5 Effectiveness of treatment options

Considering the various health consequences of obesity, efficient prevention and treatment strategies are an urgent public health concern. Depending on the degree of obesity and the presence of comorbidities, lifestyle intervention, pharmacotherapy or surgical treatment might be indicated as treatment options (Haslam and James, 2005).

In children, lifestyle interventions based on dietary modifications, physical activity and behavioral therapy are the primary treatment strategy (Han *et al.*, 2010). In a systematic review, the majority of non-pharmacological lifestyle intervention approaches resulted in weight reduction, and some also in an improvement in cardiometabolic risk factors (Oude Luttikhuis *et al.*, 2009). However, not all children benefit equally from lifestyle intervention. Approximately 20-40% of children taking part in long-term lifestyle intervention

programs failed to reduce their BMI-SDS to a degree that is sufficient for an improvement in cardiovascular risk factors (Ford *et al.*, 2010, Reinehr *et al.*, 2004, Reinehr and Andler, 2004). For instance, during the lifestyle intervention program "Obeldicks", only about 20% of children achieved a BMI-SDS reduction of at least 0.5, the reduction necessary for improvements in insulin sensitivity, blood lipid profile and blood pressure (Reinehr *et al.*, 2004, Reinehr and Andler, 2004). A similar success rate was observed in other programs (Sabin *et al.*, 2007, Ford *et al.*, 2010).

So far, few determinants have been identified that reliably predict response to lifestyle intervention. Both environmental and genetic factors are likely to play a role. Familial environment, socio-economic status and psychosocial factors affect a child's adoption of behavior changes (Reinehr, 2011). At the same time, weight change in response to hypo- or hypercaloric challenge has a considerable heritable component, as observed in twin studies (Bouchard *et al.*, 1990, 1994). Specific genetic (Ghosh *et al.*, 2011, Reinehr, 2011) and epigenetic (Campión *et al.*, 2009, Bouchard *et al.*, 2010, Milagro *et al.*, 2011, Moleres *et al.*, 2013) factors were reported to associate with weight loss response. Furthermore, metabolic factors have been linked to weight loss in both adults and children, most prominently serum leptin concentration (Fleisch *et al.*, 2007, Reinehr *et al.*, 2009b).

1.2 Omics approaches in obesity research

Obesity adversely affects nearly all organ systems in the human body (Haslam and James, 2005, Han *et al.*, 2010). Hence, *systems biology* approaches provide important insights into the molecular basis of the etiology and metabolic consequences of obesity (Meng *et al.*, 2013). Traditionally, four functional levels of a biological system are distinguished: the *genome*, the *transcriptome*, the *proteome* and the *metabolome* (Cornelis and Hu, 2013, Somvanshi and Venkatesh, 2014). Recently, the *epigenome* has emerged as another major molecular player (Schnabel *et al.*, 2012, Meng *et al.*, 2013) (Figure 1.1).

1.2.1 Genomics

Genetic variation

Genomics is the study of the structure and function of genomes, i.e. the entire deoxyribonucleic acid (DNA) sequence. DNA is arranged in a double strand of complementary nucleotides (i.e., a base – cytosine (C), guanine (G), adenine (A) or thymine (T) – bound to the sugar deoxyribose and a phosphate group, which connects the nucleotides; Figure 1.1) (Brooker, 2005). C and G form a pair of complementary bases, as do A and T. The largest part ($> 99\%$) of the base sequence is shared by all human beings. It is the genetic variants in the remaining part that make each subject individual (Ziegler and König, 2010). These variants have developed through mutation during DNA replication, are inherited through generations and manifest in populations during evolution. The most

frequent and most often studied type of genetic variation are single base exchanges, *single nucleotide polymorphisms* (SNPs, Figure 1.1). For the majority of SNPs, three states, i.e. *genotypes*, are possible: a subject might be *homozygous* for the *major allele*, that is, both chromosome copies carry the base that is more frequent at this specific locus throughout the population, *homozygous* for the *minor allele*, where both chromosome copies carry the less frequent base, or *heterozygous*, where the chromosome copies carry different bases (Ziegler and König, 2010).

Genome-wide association studies (GWAS)

Technological advances have enabled the simultaneous determination of genotypes at hundreds of thousands of SNPs on microarrays. This gave rise to the era of *genome-wide association studies* (GWAS), that is, the univariate screening for statistical associations between common SNPs and a phenotype or disease. Since 2005, more than 1700 published studies on over 600 traits were included in the GWAS catalogue (Welter *et al.*, 2014). Very often, the measured SNPs data are complemented by imputed genotypes (Ziegler and König, 2010). These are estimated based on knowledge of the *linkage disequilibrium* (LD) structure, i.e. the association of alleles at nearby SNPs that developed due to the large probability of common inheritance. This requires the availability of fully sequenced reference data from which the LD structure can be inferred. The most comprehensive set of reference data is provided by the *1000 Genomes Project* (`http://www.1000genomes.org`).

GWAS on obesity

The first GWAS on BMI were conducted in 2007; they identified common variants in the *fat mass and obesity associated (FTO)* gene as being stably associated with obesity-related traits (Frayling *et al.*, 2007, Scuteri *et al.*, 2007, Dina *et al.*, 2007). Subsequently, large-scale meta-analyses were conducted that confirmed the association with *FTO*, while at the same time revealing associations at further loci (Loos *et al.*, 2008, Willer *et al.*, 2009, Thorleifsson *et al.*, 2009). In the largest meta-analytic effort to date, the *Genetic Investigation of ANthropometric Traits* (GIANT) consortium published 32 loci independently associated with BMI (Speliotes *et al.*, 2010). GWAS on other anthropometric traits, early onset obesity, extreme obesity and childhood obesity, as well as sex-stratified studies and GWAS on BMI variability complete the picture (Berndt *et al.*, 2013, Scherag *et al.*, 2010, Bradfield *et al.*, 2012, Randall *et al.*, 2013, Yang *et al.*, 2012). Of note, the identified associated variants do not necessarily have a causal role, but might be in LD with the causal variants.

The post-genomic era

Although GWAS efforts are still ongoing, their potential for identifying genetic variants with a large contribution to disease risk seems to be close to exhaustion. With this comes the "post-genomic" or "post-GWAS" era, which has the objective of elucidating the

causal variants underlying the observed associations (Claussnitzer *et al.*, 2014, Kretschmer *et al.*, 2014) and of understanding the biological mechanisms by which they predispose to obesity, which is a slow process (Speakman, 2013). Another post-genomic concern is tackling the so-called *missing heritability*. Despite their considerable power – the GWAS mentioned above comprised samples of up to 300,000 subjects – the hitherto identified loci explain only a small fraction of the variability in obesity-related traits (Speliotes *et al.*, 2010, Choquet and Meyre, 2011b). As mentioned above, the 32 identified genetic variants associated with BMI explain merely 1.45% of the variation in BMI (corresponding to 2-

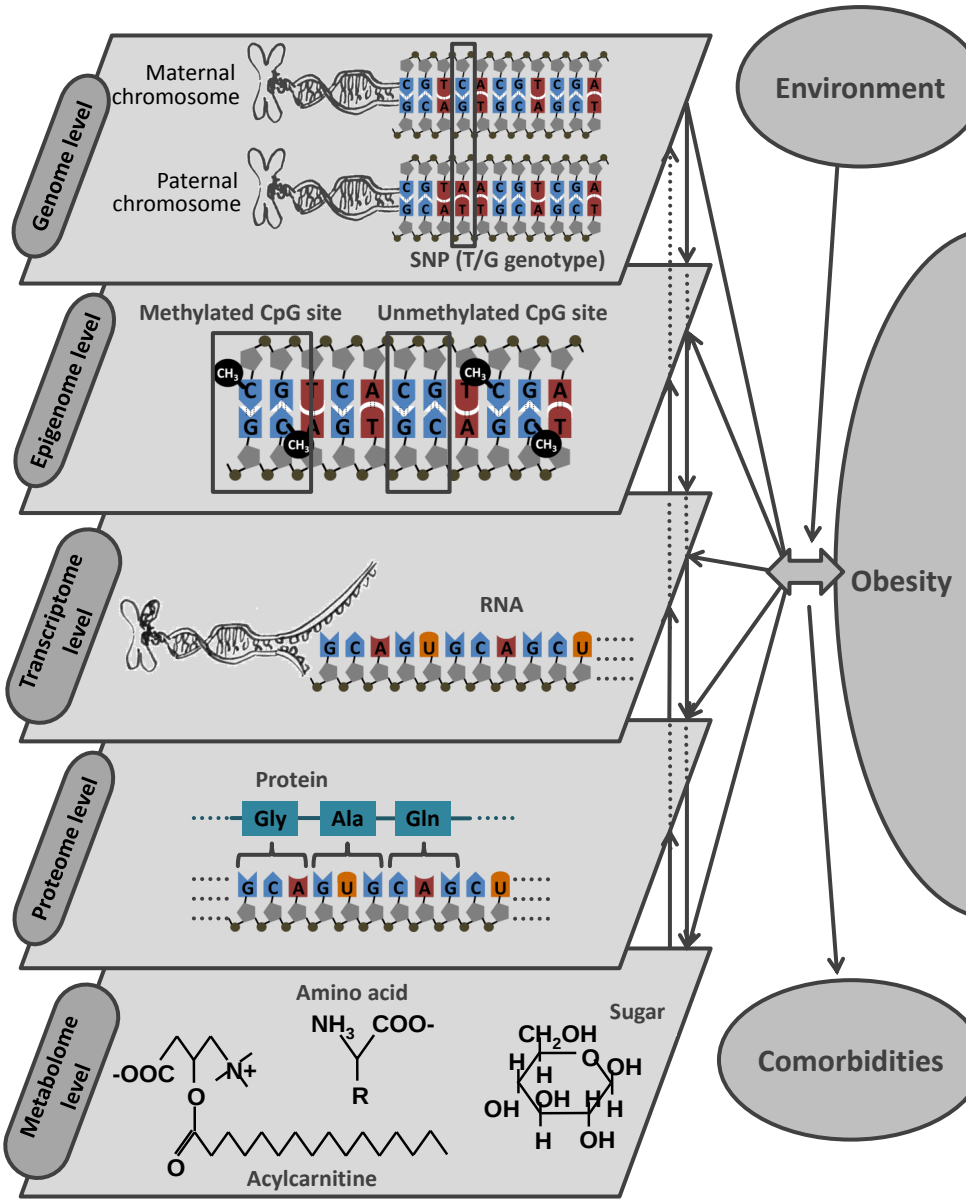

Figure 1.1: Scheme of the system levels and their relation with obesity and its comorbidities. A, adenosine; Ala, alanine; C, cytosine; CH_3, methyl group; G, guanine; Gln, glutamine; Gly, glycine; SNP, single nucleotide polymorphism; T, thymine; U, uracil.

4% of the heritability, estimated at 40-70%) (Speliotes *et al.*, 2010). A power analysis showed that a sample comprising 730,000 subjects might be sufficient to identify loci that together account for about 4.5% of phenotypic variation in BMI (corresponding to 6-11% heritability) (Speliotes *et al.*, 2010). This leaves at least 89% of the heritability unexplained. Factors potentially contributing to this missing heritability include rare genetic variants, common variants with a low penetrance, untagged structural variation including copy number variations and short insertion-deletion polymorphisms, imprinted genes, enthnicity-specific effects, as well as gene-environment and gene-gene interactions (Choquet and Meyre, 2011b). Finally, epigenetic mechanisms might play an important role (Anway *et al.*, 2005, Fraga *et al.*, 2005, Guerrero-Bosagna and Skinner, 2012).

Gene-environment interactions

Although the interaction of genes and environment in the etiology of excess body mass is an accepted concept (Bell *et al.*, 2005), less evidence exists for the interaction of specific genes and environmental or behavioral factors. It has been shown that the obesity-predisposing effect of the *FTO* locus is attenuated in physically active individuals (Andreasen *et al.*, 2008, Choquet and Meyre, 2011a, Ruiz *et al.*, 2010). Similarly, the effect of an obesity risk score comprising 12 SNPs was attenuated through high physical activity (Li *et al.*, 2010). Other examples include the effect of the *MC4R* risk genotype on treatment response during the lifestyle intervention program *Obeldicks* for obese children (Reinehr *et al.*, 2009a), and the effect of a variant in *TNF*α on post-challenge NEFA levels in obese diabetic subjects (Fontaine-Bisson *et al.*, 2007). Ordovas and Shen (2008), Bouchard (2008), Choquet and Meyre (2011b) provide further examples.

1.2.2 Epigenomics

Epigenetic variation

Epigenetics refers to heritable changes in gene function that are not caused by changes in the primary DNA sequence but by biochemical modification (Tollefsbol, 2011). Important epigenetic mechanisms include *DNA methylation*, *histone modification*, *chromatin remodeling* and *ribonucleic acid (RNA) inference*, which act in concert to regulate gene transcription and maintain genome stability (Portela and Esteller, 2010, Rakyan *et al.*, 2011).

The most frequently studied epigenetic modification is DNA methylation, the enzymatic attachment of a methyl (-CH_3) group to a DNA base (Tollefsbol, 2011), which is also the focus of this thesis. In humans, methylation occurs most frequently at the carbon-5 position of C nucleotides preceding a G nucleotide, referred to as *C-phosphate-G (CpG) sites* (Miller *et al.*, 1974, Tollefsbol, 2011). Thereby, a 5-methylcytosine is formed (Figure 1.2). Of note, the complementary strand consists of a CpG site as well, which generally

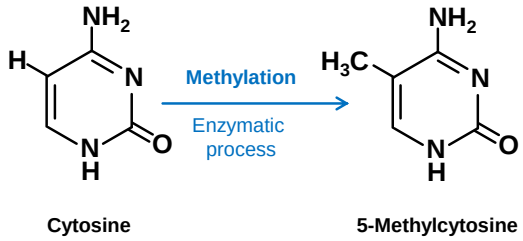

Figure 1.2: Enzymatic conversion of cytosine to 5-methylcytosine. In human DNA, cytosine bases can become methylated by the transfer of a methyl group from the molecule S-adenosylmethionine to cytosine. Thereby, 5-methylcytosine and S-adenosylhomocysteine are formed. The reaction is catalyzed by enzymes called DNA methyltransferases.

has the same methylation state (Suzuki and Bird, 2008). The human genome contains more than 10^7 CpGs (Rakyan *et al.*, 2011, Toperoff *et al.*, 2012).

The distribution of methylated and unmethylated CpG sites across the genome is not random. CpG sites tend to cluster in *CpG islands*, defined as genomic regions of >200 basepairs with a G+C content $\geq$ 50% and a ratio of observed vs. expected number of CpG sites > 0.6 (Gardiner-Garden and Frommer, 1987), that are often found in gene promoters and are mostly unmethylated (Portela and Esteller, 2010). On the other hand, the majority ($\sim$ 75%) of CpG sites throughout the genome are mostly methylated (Tost, 2010).

DNA methylation is involved in key processes including genome stability and the regulation of gene expression (Tollefsbol, 2011). It has a crucial role in imprinting, i.e. the parent-of-origin specific expression of genes (Reik and Walter, 2001), as well as X-chromosome inactivation, i.e. the inactivation of one of the X-chromosome copies in females (Portela and Esteller, 2010). An increased methylation in CpG islands or CpG island shores (regions of lower CpG density in proximity to CpG islands) is generally associated with transcriptional inactivation, for instance through interaction with other epigenetic mechanisms to reduce the accessibility of gene promoters to methylation-sensitive transcription factors (Portela and Esteller, 2010). Conversely, DNA methylation can also be associated with transcriptional activation, specifically at CpG sites in gene bodies (Portela and Esteller, 2010, Zilberman *et al.*, 2007).

Influences on DNA methylation

Through its effect on gene expression, DNA methylation can mechanistically affect individual phenotypes and disease risks (Portela and Esteller, 2010). Epidemiological studies are beginning to show a role of DNA methylation in human disease. However, in contrast to genetic variation, cause and consequence of epigenetic variation are hard to distinguish, and many factors are believed to act as confounders of the methylation - disease relationship (Martin *et al.*, 2011). Although the DNA methylation signature is in part genetically determined (Bell *et al.*, 2011) and animal studies suggest that it can be inherited across generations (Guerrero-Bosagna and Skinner, 2012), DNA methylation is subject to environmental and lifestyle influences, both *in utero* and throughout life (Bjornsson *et al.*,

2008, Rakyan *et al.*, 2011). In twin studies, phenotypic discordance of monozygotic twin pairs observed with increasing age was partly attributed to epigenetic changes occurring throughtout life (Fraga *et al.*, 2005, Wong *et al.*, 2010), possibly as the added effect of accumulating stochastic epigenetic events and environmental influences (Petronis, 2010). Specific environmental factors that have been shown to affect DNA methylation at specific CpG sites include tobacco smoking (Shenker *et al.*, 2013, Zeilinger *et al.*, 2013), alcohol intake (Philibert *et al.*, 2012, Zhu *et al.*, 2012), physical activity (Rönn *et al.*, 2013) and nutrition (Milagro *et al.*, 2011). Furthermore, many changes in DNA methylation are associated with increasing age (Bell *et al.*, 2012, Langevin *et al.*, 2011, Christensen *et al.*, 2009), and sex-specific DNA methylation patterns exist (Liu *et al.*, 2010, El-Maarri *et al.*, 2007, Boks *et al.*, 2009). In addition, methylation varies across different tissues and cell types. Consequently, methylation studies in biofluids representing a mixture of cell types are subject to confounding by cell proportions (Houseman *et al.*, 2012, Reinius *et al.*, 2012, Zhu *et al.*, 2012).

Epigenomics and obesity

Recent technological advances have allowed for the genome-wide analysis of DNA methylation (*epigenome-wide association studies*, EWAS). So far, few EWAS of obesity have been published (Wang *et al.*, 2010, Almén *et al.*, 2012, Xu *et al.*, 2013), all of which had small sample sizes and were focused on adolescents. Recently, Dick *et al.* (2014) conducted an EWAS of BMI in a cohort of 479 adults, revealing a BMI-associated CpG site in the *HIF3A* gene that was confirmed in two replication cohorts, including the KORA cohort that is the basis of this thesis.

Little evidence exists on the causality and direction of the observed associations. A candidate gene study showed a relation between umbilical cord methylation at *RXRA* and a child's fat mass at 9 years of age (Godfrey *et al.*, 2011), giving evidence for a causal effect of methylation at this site for the development of obesity. Furthermore, intervention studies have shown that DNA methylation signatures are predictive of weight reduction during caloric restriction (Milagro *et al.*, 2011, Campión *et al.*, 2009, Bouchard *et al.*, 2010, Moleres *et al.*, 2013), indicating a role of DNA methylation in weight regulation. On the other hand, changes in body mass affected DNA methylation (Milagro *et al.*, 2011, Bouchard *et al.*, 2010).

1.2.3 Transcriptomics

Transcriptomics is the study of the complete set of gene transcripts (messenger RNA molecules) in a cell or a tissue at a given time point (Cornelis and Hu, 2013), often also termed *gene expression analysis*. Gene transcripts are formed during the process of transcription, where the DNA corresponding to a specific gene is copied to RNA (Figure 1.1). During the subsequent process of translation, the base sequence coded by the transcript

is translated to a sequence of amino acids that makes up a protein. The human transcriptome is highly dynamic. It strongly varies between tissues (Petretto *et al.*, 2006) and cellular states (Gerrits *et al.*, 2009), and shows short-time responsiveness to environmental stimuli such as dietary changes (Bouwens *et al.*, 2010, Cornelis and Hu, 2013).

Microarrays are frequently employed to investigate genome-wide transcript levels (Butte, 2002). A number of cross-sectional genome-wide transcriptomics studies have been conducted on obesity-related traits (Emilsson *et al.*, 2008, Ghosh *et al.*, 2010, Zeller *et al.*, 2010, Naukkarinen *et al.*, 2010, Lee *et al.*, 2005, Das and Rao, 2007, Takamura *et al.*, 2008, Walley *et al.*, 2012, Pietiläinen *et al.*, 2008). Emilsson *et al.* (2008) were able to show a much stronger association of BMI with gene expression in subcutaneous adipose tissue rather than whole blood. Gene expression profiling has also been successful in identifying transcriptional markers associated with resistance of obese subjects to dietary intervention (Ghosh *et al.*, 2011), and changing in response to weight loss and weight maintenance (Johansson *et al.*, 2012, Larrouy *et al.*, 2008).

1.2.4 Metabolomics

The human metabolome

Metabolomics is the comprehensive study of – ideally – all metabolites within a biological system (e.g., a cell, tissue, biofluid or organism) under a given set of conditions (Pearson, 2007, Boccard *et al.*, 2010, Cornelis and Hu, 2013). The term *metabolite* refers to low molecular weight compounds (< 2000 Da) that are intermediate or end products of physiological processes (Wishart *et al.*, 2013). To date, the Human Metabolome Database comprises more than 40,000 entries of metabolites (`http://www.hmdb.ca`, accessed Mai 2014), including lipids, amino acids, peptides, amines, carbohydrates, organic acids, nucleic acids, vitamins, minerals, food additives and drugs (Boccard *et al.*, 2010, Wishart *et al.*, 2013).

The metabolome of a human biosample represents a mixture of endogenous and exogenous compounds. Thus, it can be seen as the combined product of (epi-)genetically determined molecular processes and their interactions with extrinsic (environmental) and intrinsic (pathophysiological) factors (Cornelis and Hu, 2013). This makes metabolomics a promising tool in studying the etiology of diseases through the capture of a large range of physiological pathways that are dysregulated in early disease states (Vinayavekhin *et al.*, 2010). Ultimately, the metabolome provides a source of potential diagnostic biomarkers of disease processes (Vinayavekhin *et al.*, 2010).

Metabolomics and obesity

Metabolomics has proven to be a useful tool in exploring the complex molecular disturbances associated with obesity in molecular epidemiological studies (Pietiläinen *et al.*,

2007, Newgard *et al.*, 2009, Mihalik *et al.*, 2010, Kim *et al.*, 2010, Oberbach *et al.*, 2011, Wang *et al.*, 2011, Wahl *et al.*, 2012, Szymanska *et al.*, 2012, Floegel *et al.*, 2014, Hanzu *et al.*, 2014). For instance, distinct differences in serum phopholipids, amino acids, acylcarnitines as well as sphingolipids were identified between 80 obese and 40 non-obese children (Wahl *et al.*, 2012), in agreement with other reports.

Besides cross-sectional studies, metabolomics has been useful in understanding the metabolic effect of weight loss in behavioral intervention trials. A part of the obesity-related metabolite changes showed improvement after participation in the 1-year intervention program *Obeldicks* for obese children (Reinehr *et al.*, 2014). A similar message is conveyed by Lien *et al.* (2009), Oberbach *et al.* (2011), Perez-Cornago *et al.* (2014). In addition, a potential of metabolomics to identify predictors of weight loss success during lifestyle intervention has been indicated (Pathmasiri *et al.*, 2012).

Metabolomics and challenge tests

The majority of metabolomics analyses are performed on fasting samples. However, assuming that under physiological conditions human metabolism is under tight homeostatic regulation, more insights into metabolic disturbances in early disease states might be obtained in conditions of perturbed homeostasis (van Ommen *et al.*, 2009). Challenge tests ranging from standardized glucose and lipid tolerance tests and mixed meals, to periods of fasting and physical activity challenge, allowed detailed characterization of postprandial metabolism (Ho *et al.*, 2013, Pellis *et al.*, 2012, Shaham *et al.*, 2008, Krug *et al.*, 2012, Skurk *et al.*, 2011) and uncovered changes in metabolic flexibility in early disease states such as overweight, impaired glucose tolerance and insulin resistance (Deo *et al.*, 2010, Ramos-Roman *et al.*, 2012, Shaham *et al.*, 2008).

1.2.5 Integration of omics approaches

The different levels of a biological system work together in maintaining the normal function of the system, and several if not all system levels are involved if the system is disturbed by environmental influences or pathophysiological processes (Cornelis and Hu, 2013, Somvanshi and Venkatesh, 2014) (Figure 1.1). Thus, the different omics approaches complement each other in the information they contribute to the understanding of disease-related processes (Cornelis and Hu, 2013, Somvanshi and Venkatesh, 2014). Genomic data provide a stable readout of the genetic predisposition to disease. Epigenomic data enable insights into transcriptional regulation as a consequence of inherited as well as environmental influences (Rakyan *et al.*, 2011). Transcriptomics reflects the combined regulatory processes of gene expression, being less stable than epigenetic data but closer to the level of proteins. Metabolomic signatures can be seen as the downstream product of the preceding omics processes, reflecting physiological processes most closely (Cornelis and Hu, 2013).

Few multi-omic analyses have been conducted in obesity research, most of them focusing on two or three system levels (Li *et al.*, 2008, Oberbach *et al.*, 2011, 2012, Kleemann *et al.*, 2010, Valcárcel *et al.*, 2014, Malpique *et al.*, 2014). For instance, Oberbach *et al.* (2011) conducted a cluster analysis based on combined serum proteomic and metabolomic profiles and observed a clear separation of lean and obese subjects, as well as of baseline and post-exercise timepoints with the combined data only. Valcárcel *et al.* (2014) integrated genotype and metabolomics data by performing a GWAS on correlation differences between pairs of metabolites that were differentially correlated in association with obesity. They show a greater power of their approach in identifying genetic variants involved in obesity-related processes. Malpique *et al.* (2014) studied transcriptomic signatures of pancreatic tissue, and proteomic and metabolomic signatures of peripancreatic adipose tissue of obese and lean rats. They integrated the different data sources through combined pathway analysis and network formation, thereby identifying obesity-related mechanisms potentially promoting the development of T2D.

1.3 Statistical analysis of omics data

Due to the special characteristics of omics data, their statistical analysis entails specific challenges. The development of suitable statistical methods for omics data and their appropriate application are active fields of research (Mayer, 2011).

Omics data are the result of complex sample preparation and measurement steps followed by computational translation of signals from microarrays, MS or NMR spectra into data points such as genotype calls, methylation proportions, RNA levels or metabolite concentrations. Hence, they are subject to a variety of technical influences that might "obscure" the biological variability (Hartemink *et al.*, 2001). Additional sources of technical variability arise from sample storage and from the manufacturing and processing of the arrays, plates or other facilities involved in the measurement (Hartemink *et al.*, 2001). Data preprocessing steps aim to exclude, reduce or otherwise account for technical variability to improve the identifiability of the relevant biological signals. Furthermore, the applied high-throughput techniques are unlikely to target all features, i.e., SNPs, CpG sites, transcripts or metabolites, with the same efficiency and reliability. Thus, there is a large need for data quality control and normalization prior to data analysis, a fact that has been recognized throughout the omics fields (Goodacre *et al.*, 2007, Dedeurwaerder *et al.*, 2013).

The high dimensionality and high correlatedness of most omics data requires specific considerations, both in univariate and multivariate data analysis. For instance, in the case of univariate models, correction for multiple testing is an important action to be taken (Dudoit *et al.*, 2003). In multivariate data analysis, many standard approaches fail in $p > n$ situations, i.e. when the number of variables p exceeds the sample size n. The appropriate multivariate approaches in turn require the careful choice of parameters and

model validation steps (Hastie *et al.*, 2009).

Omics data are biological data which do not always follow the distributions that standard statistical models assume (Fahrmeir *et al.*, 2013, Du *et al.*, 2010). Thus, one has to carefully weigh up the pros and cons of parametric and nonparametric approaches. Furthermore, due to the complex interrelationships between the different system levels as well as environmental factors, the issue of biological confounding has to be taken into account (Greenland and Morgenstern, 2001, Hernán *et al.*, 2002). In addition, associations of phenotypes with molecular features in the general population are often small, so power is limited. Thus, multiple epidemiological cohorts are frequently analyzed jointly, which requires meta-analysis and replication strategies (Normand, 1999, Ioannidis, 2007).

A further aspect is the frequent presence of missing values within omics data sets. These arise for various reasons including technical effects and values below the detection limit being deliberately set to missing during the step of signal translation. Depending on type and origin of missing values in both omics and phenotype data, different approaches are appropriate to handle these missing values (Raessler *et al.*, 2008).

Finally, a specific characteristic of omics data is that the primary data analysis results in potentially large sets of (multi)-omics markers associated with some kind of state, phenotype or disease. Specific statistical and bioinformatical tools are required to extract biological knowledge from these, e.g. to identify enriched biological pathways or functional genomic features, to explore causality (Didelez and Sheehan, 2007) and to understand the relation between the features (Krumsiek *et al.*, 2011).

1.4 Aims of this thesis

In large genome-wide association studies (GWAS), obesity-related genetic variants were identified. However, for most variants, the mechanisms underlying their association with obesity-related traits are not yet understood, and a large part of the heritability of obesity remains unexplained by these variants.

The overall objective of this thesis is to contribute to the post-genomic era of obesity by means of analyzing multiple omics data to explore the molecular mechanisms underlying obesity, weight change and related metabolic disturbances. Four studies were conducted to pursue the following specific research questions.

In the first study (Section 4.1), the overall aim was to learn more about the early metabolic derangements linking the strongest obesity risk locus, *FTO*, and the T2D risk locus *TCF7L2* to disease risk. Using a novel strategy based on metabolomics measurements during different oral and intravenous challenge tests, the specific goals were (1) to comprehensively characterize physiological challenge responses in healthy subjects, and (2) to study the feasibility of the approach for studying gene-environment interactions, specifically, (3) to explore the effect of the *FTO* and *TCF7L2* genotypes on challenge response.

In the second study (Section 4.2), the overall research question was whether DNA methylation at specific sites could be identified as a factor contributing to obesity risk (thereby potentially explaining a part of the missing heritability of obesity), or to the development of obesity-related comorbidities. Specific goals were (1) to study the associations between BMI and whole blood DNA methylation in a large meta-analysis of EWAS, (2) to characterize the identified CpG sites with regard to their genomic location and proximity to functional genomic features, their enrichment for biological pathways and for loci previously reported in GWAS for obesity and related diseases, (3) to explore association of methylation at these sites with gene expression at nearby genes, (4) to study the relation of methylation to genetic variation at nearby sites to understand the genetic basis of methylation, (5) to explore causality and direction of the observed associations, and (6) to investigate whether the identified CpG sites accounted for the association of BMI with clinical traits and incident T2D in order to explore the clinical relevance of the findings.

The third study (Section 4.3) was aimed at studying the metabolomic and transcriptomic consequences of body weight changes over a 7-year period in the general population. Specific goals were (1) to investigate associations of weight change with serum metabolite concentrations and blood cell gene expression, (2) to study the interrelationship of the identified omics signatures, and the stability of the findings in relevant subgroups, e.g. of subjects with weight gain versus weight reduction, and (3) to study the relation of the identified omics signatures with clinical traits.

Finally, the aim of the fourth study (Section 4.4) was to explore the potential of combined serum metabolomics and anthropometric and clinical data to predict weight loss success over a 1-year lifestyle intervention program for obese children.

Throughout the thesis, a particular emphasis was placed on the optimized choice of statistical methodology and its careful implementation. In the first study, a statistical goal was to find a way of identifying joint trends in challenge response trajectories of single metabolites. The primary statistical aims of the second study were to develop MR approaches to study causality and compare the results with those of longitudinal association analyses, to appropriately account for cell type confounding, and to use resampling procedures e.g. to assess the proportion of BMI-trait associations accounted for by methylation. The methodological tasks in the third study were the careful imputation of missing values, and the implementation of a cluster approach that would improve the clearness and interpretability of of metabolomic and transcriptomic signatures of weight change. Finally, in the fourth study, the statistical aim was to employ a multivariate regularized regression approach to form a sparse predictive model of weight loss success, to evaluate this model within a nested cross-validation approach, and to compare the results obtained with this approach to standard univariate regression results.

2 Study populations and data retrieval

In this chapter, the *Cooperative Health Research in the Region of Augsburg* (KORA) study, the *Virtual Institute Diabetes* (VID) study and the *Obeldicks* study are described, and details on the assessment of phenotypic and molecular data are given. For the cooperation cohorts *London Life Sciences Prospective Population Study* (LOLIPOP)/EpiMigrant, EPICOR, the *Rotterdam* studies, *LifeLines Deep*, *Avon Longitudinal Study of Parents and Children* (ALSPAC), *Leiden Longevity*, *TwinsUK* and the *Estonian Genome Center of the University of Tartu* (EGCUT) studies (all part of the DNA methylation analysis in Section 4.2), subject characteristics and analysis details are provided in Appendix Tables A.1 to A.4.

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

2.1.1 Study population

KORA is a research platform of independent population-based health surveys and subsequent follow-up examinations of individuals of German nationality resident in the region of Augsburg in Southern Germany (Holle *et al.*, 2005). Written informed consent was obtained from all participants in accordance with institutional requirements and the Declaration of Helsinki principles. The studies were approved by the ethics committee of the Bavarian Medical Association (Bayerische Landesärztekammer). Study design, sampling method and data collection have been described in detail elsewhere (Holle *et al.*, 2005). The surveys S3 and S4 were conducted in 1994/1995 and 1999-2001, respectively, and comprised independent samples of 4856 and 4261 subjects aged 25-74 years (Wolfenstetter *et al.*, 2012). Both cohorts were reinvestigated in the follow-up examinations F3 and F4 in 2004/2005 and 2006-2008, respectively, with 2974 and 3080 participants. KORA S4/F4 stands out as a phenotypically, biochemically and molecularly well-characterized population cohort allowing for extensive and integrative omics analysis of obesity and weight change.

2.1.2 Anthropometric measurements and interviews

Body weight, height, waist and hip circumference, and systolic and diastolic blood pressure were measured using standard protocols as described elsewhere (Rathmann *et al.*, 2003). For the study of metabolic consequences of weight change (Section 4.3), weight change between KORA S4 and F4 was defined as percentage body weight change (ΔBW) in kg per follow-up year, where weight gain was coded as ΔBW, and weight loss as negative ΔBW:

$$\Delta\text{BW} = 100\% \cdot \frac{\frac{\text{BW (F4) [kg]} - \text{BW (S4) [kg]}}{\text{BW (S4) [kg]}}}{\text{Follow-up time [years]}}.$$

Information on lifestyle factors and diseases are based on self-report during a standardized interview conducted by trained interviewers. Ascertainment was comparable in KORA S4 and F4, allowing plausible categories of changes to be formed. To categorize physical activity level, participants were classified as "active" if they spent at least one hour of moderate and vigorous physical activity per week during leisure time in summer and winter, and as "inactive", else (Meisinger *et al.*, 2007). Changes in physical activity between KORA S4 and F4 were categorized as "no change", "became active" and "became inactive". Smoking categories were formed as "current", "former" and "never" smokers (in Section 4.3 as "ever" and "never" smokers), and changes in smoking status were categorized as "no change", "started smoking" and "quit smoking". Sleeping behavior was assessed as problems to fall asleep or to sleep through the night, with the categories "often" and "sometimes/almost never", and changes were categorized as "no change", "improvement", "worsening". Nutrition habits were assessed using a food frequency questionnaire. Based on how often participants reported to consume 15 different food categories, their nutrition habits were categorized as "disadvantageous", "normal" or "advantageous" (based on recommendations of the German Nutrition Society). Disease information, including myocardial infarction, stroke, T2D, and cancer, is based on self-reported physician's diagnosis. Change in disease was categorized as "incident disease during follow-up" and "no incident disease during follow-up". Intake of medication within seven days prior to examination was recorded with the IDOM-Software (Mühlberger *et al.*, 2003), and categorized according to the *Anatomical Therapeutic Chemical* (ATC) classification index. Change in medication intake between KORA S4 and F4 was denoted as "no change", "stopped intake" and "started intake".

2.1.3 Blood sampling and biochemical measurements

Blood samples were collected during study center visits between 8 and 11 a.m., after participants were instructed to fast overnight for at least 8 h. Whole blood was collected using PAXgene Blood RNA tubes (BD, Heidelberg, Germany) and stored at -80°C until analysis. To obtain plasma, blood was immediately centrifuged and plasma frozen at -80°C until measurments. Serum collection in KORA F4 samples has been described by Illig *et al.* (2010). Briefly, blood was drawn into serum gel S-Monovette tubes (Sarstedt,

Nümbrecht, Germany), gently inverted two to three times and rested for 30 min at room temperature to obtain complete coagulation. This was followed by centrifugation for 10 min (2,750g at 15°C). Serum was aliquoted into synthetic straws which were kept for a maximum of 6 h at 4°C before storage at -80°C until analysis.

In KORA F4, red blood cell count, hemoglobin concentration, hematocrit, mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC) and cell volume of erythrocytes (MCV) were determined in a small hemogram (Coulter LH-750, Beckman, Germany). HbA1c was determined from EDTA blood using HPLC (HA 8160, Menarini). Fasting and 2-hour oral glucose tolerance test (OGTT) glucose levels were assessed with the hexokinase method (GLU Flex; Dade Behring, Marburg, Germany). For fasting plasma insulin measurements, a microparticle enzyme immunoassay (MEIA, IMX Insulin, Abbott Laboratories, Wiesbaden, Germany) was used. Total, high density lipoprotein (HDL) and low density lipoprotein (LDL) cholesterol levels were determined in serum using the CHOD-PAP method (CHOL Flex; AHDL and ALDL Flex, Dade-Behring, Marburg, Germany). Triglyceride (TG) levels were determined with the GPO-PAP method (TGL Flex; Dade-Behring, Marburg, Germany). C-reactive protein (CRP) was measured with nephelometry on a BN II using reagents from Siemens (Eschborn, Germany).

2.1.4 Genotyping

Genotyping of the KORA F3 and F4 samples was performed on the *Illumina Omni Express* and *Affymetrix Axiom* platforms, respectively. Genotypes were called with the Affymetrix software and Genome Studio, respectively, and annotated to NCBI build 37. For both cohorts, genotype imputation was performed based on the 1000G phase 1 reference panel using IMPUTE v2.3.0 (Howie *et al.*, 2009), with SHAPEIT v2 (O'Connell *et al.*, 2014) as a pre-phasing tool.

2.1.5 DNA methylation measurement

Genome-wide DNA methylation measurement at 485,577 genomic sites was performed using the *Infinium HumanMethylation450K BeadChip*® (Illumina, Inc., CA, USA, Bibikova *et al.* (2011)) in 1814 KORA F4, 500 KORA F3 (comprising smokers and never smokers, Zeilinger *et al.* (2013)) and 1535 KORA S4 samples. The laboratory process has been described previously (Zeilinger *et al.*, 2013, Petersen *et al.*, 2014). Briefly, denaturated single-stranded genomic DNA was subjected to bisulfite treatment using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA). Bisulfite-converted samples were subjected to whole genome amplification, followed by enzymatic fragmentation and application to the BeadChips. The arrays were fluorescently stained and scanned with the Illumina HiScan SQ scanner.

As readout, a methylated and an unmethylated signal count per CpG site are obtained.

Counts are commonly combined to *β-values*, defined as the ratio of the methylated signal intensity divided by the overall signal intensity (Bibikova *et al.*, 2011, Du *et al.*, 2010):

$$\beta\text{-value} = \frac{M}{M + U + \alpha}.$$

Illumina recommends the inclusion of an offset $\alpha = 100$ as a regularization for the situation when both M and U are low (Du *et al.*, 2010). Since the number of signal intensities mostly exceeds 1000, the offset does not induce much bias. The methylation β-value can be interpreted as the proportion of methylation at a given CpG site.

2.1.6 Gene expression measurement

RNA preparation and gene expression measurement in the KORA F4 data ($n = 993$ aged 62-81 years with genotype data available) has been described in detail elsewhere (Schurmann *et al.*, 2012). Briefly, RNA was isolated from whole blood stored for 856 ± 179 days at -80°C using the PAXgene^{TM} Blood miRNA kit (Qiagen, Hilden, Germany). Purity and concentration of RNA were determined using a NanoDrop ND-1000 UV-Vis Spectrophotometer (Thermo Scientific, Henningsdorf, Germany). To ensure a constantly high quality of the RNA preparations, all samples were analyzed using RNA 6000 Nano LabChips (Agilent Technologies, Germany) on a 2100 Bioanalyzer (Agilent Technologies, Germany) according to the manufacturer's instructions. Samples exhibiting an *RNA integrity number* less than seven confirmed by manual adjustment were excluded from further analysis. The Illumina TotalPrep-96 RNA Amplification Kit (Ambion, Darmstadt, Germany) was used for reverse transcription of 500 ng RNA into double-stranded cDNA and subsequent synthesis of biotin-UTP-labeled antisense-cRNA using this cDNA as the template. 3,000 ng of cRNA were hybridized to the *Illumina HumanHT-12 v3 Expression BeadChip* arrays followed by washing and detection steps in accordance with the Illumina protocol. BeadChips were scanned using the *Illumina Bead Array Reader*.

2.1.7 Metabolomics measurement on the *Metabolon* platform

Metabolomics measurements for 1768 KORA F4 subjects were performed on a commercial mass spectrometry (MS)-based platform at the company Metabolon Inc. (Durham, NC, USA). The analytical process, including metabolite quantification and identification as well as quality control, has been described in detail elsewhere (DeHaven *et al.*, 2010, Evans *et al.*, 2009, Suhre *et al.*, 2011). Briefly, the analytical platform is based on two ultrahigh-performance liquid chromatography/tandem mass spectrometry (UHPLC/MS/MS^2) injections and one gas chromatography/mass spectrometry (GC/MS) injection per sample. The two UHPLC injections were optimized for basic and acidic species, respectively. In the KORA F4 data, relative quantification for a total of 517 compounds was provided, 325 of which could so far be identified based on a standard library of MS/MS spectra.

The identified molecules cover a large spectrum of metabolites classes including amino acids, peptides, carbohydrates, fatty acids, glycerophospholipids, acylcarnitines, sphingolipids, steroids, ketone bodies, bile acid metabolites, nucleotide metabolites, vitamins and xenobiotics. The full list of identified metabolites is provided in Appendix Table A.5.

2.1.8 Metabolomics measurement on the *NMR* platform

For 1788 KORA F4 samples, metabolite measurement was conducted using an nuclear magnetic resonance (NMR)-based platform. The precise experimental methodology has been described elsewhere (Soininen *et al.*, 2009, Inouye *et al.*, 2010a). Briefly, serum samples were thawed in a refrigerator at 4°C overnight. The samples were mixed gently and centrifuged at 3400g. 300 μl of each serum sample were mixed with 300 μl of sodium phosphate buffer by three times slow aspiration. Sample preparation was done automatically with a Gilson Liquid Handler215 in 5 mm outer-diameter SampleJet NMR tubes (Bruker BioSpin, Germany). During NMR spectroscopy, samples were kept at 6°C, and measurement took place at 37°C. ^{1}H NMR spectroscopy was conducted on a Bruker AVANCE III spectrometer operating at 500.36 MHz. NMR spectroscopy was applied within three molecular windows, to improve the detection of metabolites from different molecular weight classes. Two of these were applied to native serum and provide quantification of lipoprotein subclass concentration and composition, as well as of low-molecular-weight compounds including amino acids, ketone bodies and carbohydrate metabolites. The third molecular window was applied to serum lipid extracts and provides information on the composition of serum lipids including fatty acids, cholesterol and sphingomyelins (Inouye *et al.*, 2010a). Computational strategies of metabolite identification and quantification from the NMR spectra are described by Inouye *et al.* (2010a). A total of 130 metabolic readouts were obtained from the NMR platform.

2.2 Virtual Institute Diabetes (VID)

2.2.1 Study participants

The *Virtual Institute Diabetes* (VID) is a cooperative project of the Helmholtz Zentrum München, the Ludwig-Maximilians-Universität München and the Technische Universität München aiming to understand the molecular basis of glucose regulation and T2D. To investigate the dynamics of the serum metabolome during nutritional challenges in healthy men dependent on the *FTO* rs9939609 and *TCF7L2* rs7903146 genotypes, male KORA S4/F4 participants aged 18-65 years were re-invited based on existing genotype information. 25 men carrying the *FTO* rs9939609 obesity risk variant (AA genotype), 22 men carrying the *TCF7L2* rs7903146 T2D risk allele (11 TT and 11 CT genotype) and 31 subjects carrying none of the risk alleles (*FTO* TT genotype and *TCF7L2* CC genotype) were recruited. Individuals with known or apparent diabetes, immune suppressive therapy,

clinical cardiovascular disease, liver disease (GOT, GPT > 3 fold above normal range), kidney disease (creatinine > 1.2 mg/dl) and psychiatric disease were excluded from participation. Participants did not take any medication known to affect insulin sensitivity or secretion. All subjects were of Caucasian origin. The study was conducted in accordance with the Declaration of Helsinki principles. Written informed consent was given by all participants and the study was approved by the ethics committee of the Bavarian Medical Association (Bayerische Landesärztekammer). Neither the participants nor any of the attending physicians or assistants knew the genotype of the probands at the time of the interventions.

2.2.2 Genotyping

Genome-wide SNP data were available for 1814 KORA S4/F4 subjects from the *Affymetrix GeneChip array 6.0* (Kolz *et al.*, 2009). Targeted genotype data were obtained for the remaining 2,235 KORA S4/F4 participants using MALDI-TOF with the *Sequenom i-Plex Gold Assay* (Holzapfel *et al.*, 2010). Of 1,134 male participants, 93 homozygous *FTO* risk allele carriers and 207 non-risk allele carriers met the inclusion criteria. Only 27 subjects carried the *TCF7L2* risk allele. Thus, both homozygous and heterozygous subjects were included. Of these subjects, 25, 31 and 22 could be recruited for the present study, respectively.

2.2.3 Metabolic challenge tests

Challenge tests were conducted at two different study centers in two separate visits. Intravenous challenges took place at the Medizinische Klinik and Poliklinik IV, Ludwig-Maximilians-Universität München, oral challenges at the human study center of the Else Kröner-Fresenius Center for Nutritional Medicine at the Technische Universität München.

Intravenous challenges

At the first study visit, participants underwent an intravenous glucose tolerance test (IVGTT; 0.33 g glucose/kg body weight of a 50% (vol/vol) glucose solution within 2 min) between 8 and 9 a.m. after overnight fasting (Figure 2.1). 35 min after the glucose load, an euglycemic-hyperinsulinemic (EH) clamp was conducted. An insulin (Actrapid, Novo Nordisk, Copenhagen, Denmark) bolus was given until stable blood glucose values of 70–80 mg/dl were reached, followed by a continuous infusion (1.05 mU/kg/h) of short-acting human insulin and a variable infusion of a 20% glucose solution to maintain plasma glucose concentration at 80 mg/dl for 120 min. Blood glucose was determined in 6 min intervals throughout the clamp and measured using a bedside glucose analyzer (Super-GL ambulance, HITADO, Möhnesee, Germany). Steady state was reached 3–4 h after the IVGTT on average. Venous blood samples were taken from the opposite arm at baseline, 1, 3, 5, 10, 15, 25 and 35 min after IVGTT, as well as 0, 15, 30 and 45 min after

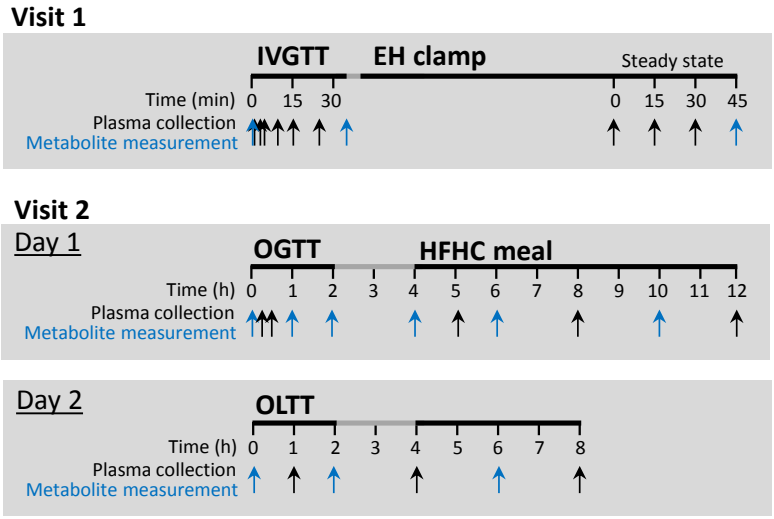

Figure 2.1: Scheme of challenges conducted at the two study visits. Arrows indicate times of plasma collection, with metabolite concentrations determined in plasma samples taken at times indicated by blue color. Clamp steady state was reached about 3-4 h after the start of the IVGTT. EH clamp, euglycemic-hyperinsulinemic clamp; HFHC meal, high-fat high-carbohydrate meal; IVGTT, intravenous glucose tolerance test; OGTT, oral glucose tolerance test; OLTT, oral lipid tolerance test.

the clamp steady state was reached. Blood samples for biochemical and metabolomics measurements were immediately cooled to 4°C and centrifuged (10 min at 3,000 g, 4°C). Aliquots of plasma samples were stored at -80°C until assayed.

Oral challenges

At the second study visit, oral challenge tests were conducted in the metabolic ward of the study center (Figure 2.1). Participants were carefully instructed to consume only the food and drinks provided by the study personnel and to refrain from any physical activity. In order to control nutrient intake and activity levels, all participants stayed at the study center over the two-day study period. All participants were compliant to the following protocol: After overnight fasting, on day 1 at 8 a.m., subjects underwent a standardized oral glucose tolerance test (OGTT) (75g glucose, Dextro O.G.T., Roche Diagnostics, Mannheim, Germany) within 5 min. At 12 a.m., a high-fat high-carbohydrate (HFHC) meal, comprising a Big Mac burger, 0.4 l Fanta and 114 g French fries (McDonald's, Freising, Germany) were consumed within 15 min. On day 2 at 8 am, a standardized oral lipid drink, corresponding to 35 g fat/m^2 body surface (on average, 422 ml), was consumed within 5 min. The drink was prepared at room temperature from three parts of Fresubin Energy drink chocolate (Fresenius Kabi, Bad Homburg, Germany) and one part of Calogen (Nutricia, Pfrimmer, Germany). Besides the intervention meals, subjects received a standardized supper on day one and *ad libitum* mineral water and unsweetened fruit or herbal tea. The macronutrient composition of the challenge tests is shown in Supplementary Table 1 of the original publication (Wahl *et al.*, 2013b). During OGTT, venous blood was taken at baseline, 15, 30, 60 and 120 min after the glucose load, with

baseline, 1 and 2 h samples taken for metabolomics measurement. During the HFHC meal and the oral lipid tolerance test (OLTT), samples were taken at baseline, 1, 2, 4, 6 and 8 h after the challenge, with metabolomics measurements conducted in the baseline, 2 and 6 h samples. Time points were chosen based on earlier investigations on postprandial metabolite changes (Krug *et al.*, 2012, Skurk *et al.*, 2011). Blood samples for biochemical and metabolomics measurements were immediately centrifuged (10 min at 3,000 g, 20°C), afterwards kept frozen at -80°C and thawed only once directly before measurement.

2.2.4 Anthropometric and biochemical measurements

Anthropometric examination included measurements of body weight, height, waist circumference and blood pressure according to standard procedures. Glucose levels were assessed in plasma (first study visit) using the hexokinase method (GLU Flex, Dade Behring, Marburg, Germany) and in venous blood (second study visit) by enzymatic amperometric technique (Super GL easy+, Dr. Müller Geräte Bau, Freital, Germany). Blood glucose values were converted to plasma equivalents by multiplication with the recommended factor 1.11 (D'Orazio *et al.*, 2005). Plasma insulin levels were measured by enzyme-linked immunosorbent assay (ELISA) (first study visit: LINCO research, St. Charles, USA; second study visit: Dako #K6219, Glostrup, Denmark). Insulin sensitivity index was calculated as glucose infusion rate per kg body weight necessary to maintain euglycemia during the last 45 min of the clamp steady state per unit of plasma insulin concentration. Lactate was measured by enzymatic amperometric technique (Super GL easy+ , Dr. Müller Geräte Bau, Freital, Germany). Total cholesterol concentrations, HDL cholesterol and LDL cholesterol levels were determined with enzymatic methods (CHOD-PAP, Dade Behring). TGs were measured by an enzymatic color test (first study visit: GPO-PAP-method, TGL Flex, Dade Behring; second study visit: Fluitest TG, Analyticon Biotechnologies AG, Lichtenfels, Germany). High sensitive CRP was determined by IRMA (Dade Behring). HbA1c was measured using the HPLC method. Total levels of non-esterified fatty acids (NEFAs) were measured by an enzymatic colorimetric method assay (NEFA-HR, Wako Chemicals GmbH, Neuss, Germany). Serum creatinine concentrations were assessed with a modified Jaffe test (Krea Flex, Dade Behring). Plasma proinsulin and insulin concentrations were quantified with ELISA Kits (LINCO research, St. Charles, USA) as described recently (Anzeneder *et al.*, 2011). Serum C-peptide was determined with the radioimmunoassay from Radim Diagnostics (Pomezia, Italy). All assays were conducted according to the manufacturers' guidelines.

2.2.5 Metabolomics measurement on the *Biocrates* platform

Concentrations of 163 metabolites were determined in the plasma samples using the targeted AbsoluteIDQ^{TM} kit p150 (Biocrates Life Sciences AG, Innsbruck, Austria), following the instructions described in the manufacturer‘s manual. The procedure has been

described in detail elsewhere (Illig *et al.*, 2010, Römisch-Margl *et al.*, 2011). Briefly, liquid handling of plasma samples was performed with a Hamilton Microlab STAR^{TM} robot (Hamilton Bonaduz AG, Bonaduz, Switzerland). Samples were analyzed on an API4000 LC/MS/MS system (AB Sciex Deutschland GmbH, Darmstadt, Germany) equipped with an HTC PAL autosampler (CTC Analytics, Zwingen, Switzerland) and an electrospray ionization (ESI) source which was used in both positive and negative mode. MS/MS analysis was run in the Multiple Reaction Monitoring mode. The entire analytical process was controlled by the Analyst 1.4 software and the Met*IQ*^{TM} software package. Metabolite concentrations were determined with the Met*IQ* software. The metabolite panel targeted by the kit comprises amino acids, hexose, free carnitine (C0), conjugated carnitines (acylcarnitines (Cx:y), hydroxylacylcarnitines (C(OH)x:y), and dicarboxylacylcarnitines (Cx:y-DC)), diacyl phosphatidylcholines (PC aaCx:y), acyl-alkyl phosphatidylcholines (PC ae Cx:y), lysophosphatidylcholines (LPC a Cx:y) as well as sphingomyelins (SM Cx:y) and hydroxysphingomyelins (SM (OH) Cx:y). See Wahl *et al.* (2012) for a full list of metabolites. Cx:y abbreviates the lipid side chain composition and x and y denote the sum of carbons and double bonds, respectively. Importantly, the analytical technique applied here is not capable of determining the precise position of the double bonds and – in the case of PCs – the distribution of carbon atoms between the two fatty acid side chains. All Biocrates metabolite concentrations are reported in μmol/l.

2.3 Obeldicks

2.3.1 Study design

Obeldicks is a one-year weight loss intervention program based on physical activity, nutritional education and behavior therapy that includes individual psychological care of the child and his/her family. The program is tailored to obese children aged 6–15 years and is conducted at the outpatient clinic for obesity of the Vestische Kinder- und Jugendklinik Datteln, Germany. All participating children were born in Germany. Children with syndromal obesity, psychiatric or endocrine disorders including T2D were excluded. A detailed description of the program can be found elsewhere (Reinehr *et al.*, 2006). Written informed consent was obtained from all parents and all children from the age of 12 years. The study was approved by the ethics committee of the University of Witten/Herdecke. Of the children who had completed the Obeldicks program in 2008 or 2009, 40 were randomly selected who had reduced their BMI-SDS substantially during their one-year participation, as defined by a BMI-SDS reduction of ≥ 0.5, and 40 with a BMI-SDS reduction of < 0.1 and a similar distribution of sex, baseline age, pubertal stage and BMI-SDS. The cut-off at a BMI-SDS of 0.5 was chosen based on the finding of previous studies that this amount of BMI-SDS reduction is approximately required to achieve a considerable improvement in the cardiovascular risk profile (Reinehr *et al.*, 2004, Ford *et al.*, 2010). Compliance was

given for all 80 children by participation in at least 90% of the meetings.

2.3.2 Anthropometric measures

Body height was measured to the nearest centimeter using a rigid stadiometer. Undressed body weight was measured to the nearest 0.1 kg using a calibrated balance scale. BMI percentiles and BMI-SDS were calculated according to Cole's LMS-method (Cole, 1990), applied to German reference data (see Section 1.1.1, Kromeyer-Hauschild *et al.* (2001)). All children's BMI was above the 97th percentile. Waist circumference was measured halfway between lower rib and iliac crest (Kromeyer-Hauschild *et al.*, 2008). Pubertal stage was assessed according to Marshall and Tanner (1969, 1970) and categorized into three stages based on pubic hair and genital stages: prepubertal = boys/girls with pubic hair stage I and gonadal/breast stage I; pubertal/postpubertal = boys/girls with pubic hair stage $\geq$ II and gonadal/breast stage $\geq$ II and boys with change of voice and girls with menarche. Systolic and diastolic blood pressure was measured twice according to a validated protocol and the two measurements were averaged (National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents, 2004).

2.3.3 Biochemical measurements

Blood samples were taken at 8 a.m. after overnight fasting for at least 10 h. Following coagulation at room temperature, blood samples were centrifuged for 10 min at 8,000 rpm and aliquoted. Biochemical measurements were conducted directly on the fresh serum samples. TGs, total cholesterol and glucose concentrations were determined with a colorimetric test using the VitroTM analyzer (Ortho Clinical Diagnostics, Neckargemünd, Germany). LDL and HDL cholesterol were measured with an enzymatic test using the LDL-C and HDL-C-PlusTM assays (Roche Diagnostics, Mannheim, Germany), respectively. Insulin concentrations were determined with a microparticle-enhanced immunometric assay (MEIATM, Abbott, Wiesbaden, Germany). Intra- and interassay coefficients of variation were $< 5\%$ for all tests. As a measure of insulin resistance, the homeostasis model assessment of insulin resistance (HOMA-IR) was calculated as serum insulin (mU/l) $\times$ serum glucose (mmol/l)/22.5 (Matthews *et al.*, 1985). This index has been validated in healthy children (Gungor *et al.*, 2004). Aliquoted serum samples were stored at -80°C and thawed only once at room temperature for the metabolomics assay.

2.3.4 Metabolomics measurement on the *Biocrates* platform

Metabolite measurement in serum samples was performed in two batches, as described above for the VID study (Section 2.2.5).

3 Statistical methods

In this chapter, the principles and technical details of the applied statistical methods are described. A conceptual visualization is given in Figure 3.1.

3.1 Data preprocessing and quality control

3.1.1 SNP data

To ensure data quality, observations were removed for which a sex discordance (i.e., discordance between phenotypic and genetic sex), or a discordance of genotypes available from different genotyping platforms was observed, pointing towards sample swapping. In addition, observations that represented population outliers (i.e., that did not cluster with the HapMap CEU population in a joint plot of the first two principal components) or heterozygosity outliers (i.e., observations that deviated by at least 5 standard deviations from the mean heterozygosity rate) were excluded from the data set.

Furthermore, the observation-wise *callrate* was defined as the number of SNPs for which a reliable genotype "call", i.e. genotype assignment, could be made based on the fluorescence signals obtained for the two alleles of the SNP. Observations with a callrate below 97% were excluded from the data set, to ensure data reliability. Similarly a callrate threshold of 98% was applied in a SNP-wise fashion.

Another important quality measure is the *Hardy-Weinberg equilibrium* (HWE), which states that allele and genotype frequencies remain constant across the generations of a population under certain conditions (e.g., random mating, no migration) and that therefore there should be a fix relationship between allele and genotype frequencies (Ziegler and König, 2010). Violation of the HWE, that is, inconsistency of the genotype distribution of a SNP across the investigated population with the distribution expected from the allele frequencies, in most cases points towards laboratory issues. Therefore, SNPs were excluded when the HWE p-value was below 5×10^{-6}, indicating strong deviation from the HWE.

In KORA S3/F3 but not S4/F4, SNPs with a minor allele frequency (MAF) below 1% were also excluded. After imputation, SNPs of both studies with an imputation information score below 0.5 were excluded, and data were transformed to dosages, i.e. expected allele counts of the non-reference allele. In total, $>$ 31 Mio measured or imputed SNPs were

available for analysis.

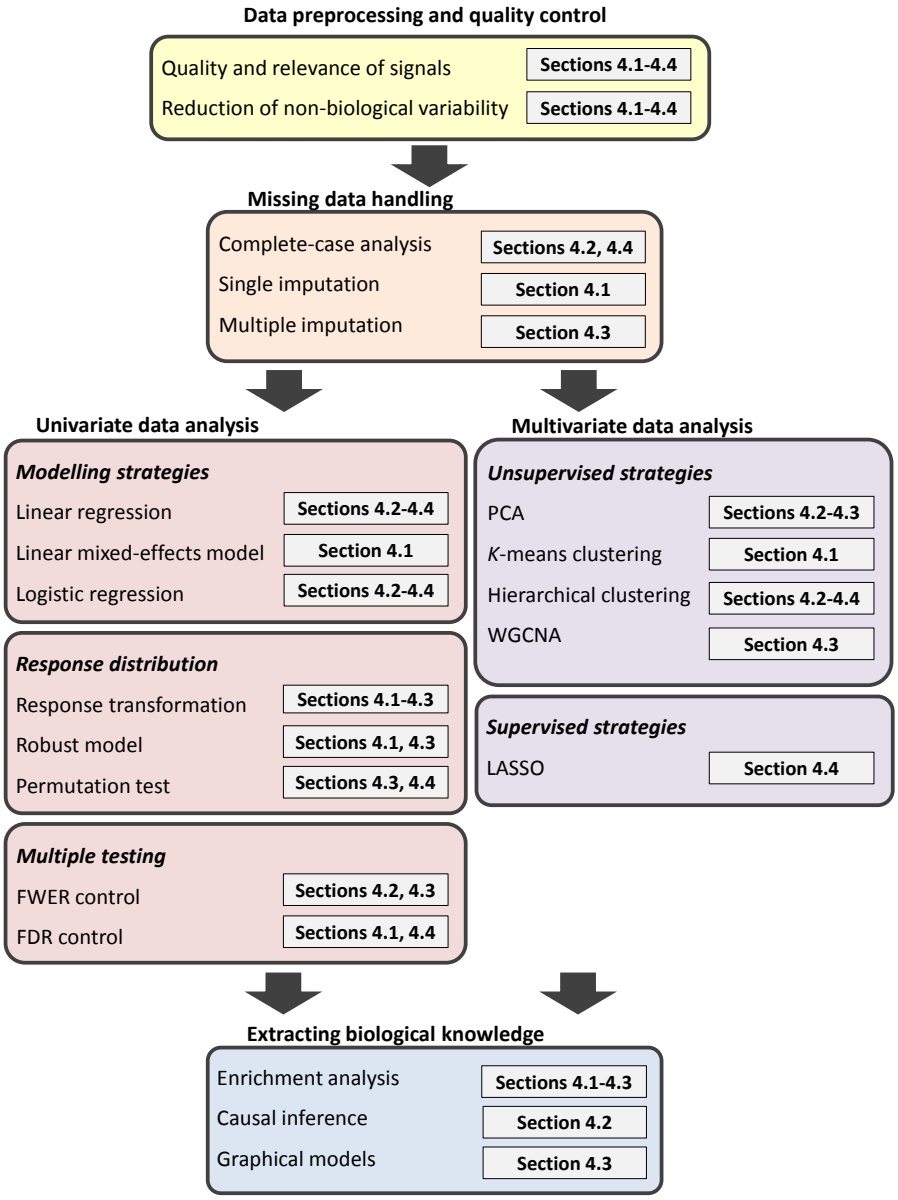

Figure 3.1: Statistical methods applied in this thesis. Colors represent the different blocks of statistical methods and are used consistently throughout the thesis. Grey boxes indicate references to results and discussion sections where the respective methods are applied.

3.1.2 DNA methylation data

DNA methylation data were preprocessed as follows: First, 65 probes that represent SNPs were excluded. Second, background correction was performed using the R package *minfi*, version 1.6.0 (Aryee *et al.*, 2014). Third, *detection p-values* were defined as the probability of a signal being detected above the background signal level, as estimated from negative

control probes. Consequently, signals with detection p-values ≥ 0.01 were removed, incidating putatively unreliable signals. Similary, signals summarized from less than three functional beads on the chip were characterized as potentially unreliable and removed from the data set. Observations with less than 95% CpG sites providing reliable signals (72 in KORA F4, 15 in KORA F3, 0 in KORA S4) were excluded.

To avoid spurious results, CpG sites were flagged that were targeted by cross-reactive probes (information provided by Price *et al.* (2013)). These are probes that co-hybridize to highly homologous genomic sequences other than the target sequence, causing ambiguous signals at the targeted sites (Price *et al.*, 2013, Chen *et al.*, 2013). Similarly, probes with genetic variants located in the binding sequence were flagged (Price *et al.*, 2013). The genetic variants might affect probe-binding efficiency and thereby the detected methylation signals. In addition, CpG sites that themselves contained a SNP were flagged since in that case the obtained signal is likely predominantly derived from the SNP rather than the methylation state (Dedeurwaerder *et al.*, 2013). In Section 4.2, the stability of the identified methylation-BMI associations to the SNPs in the probe or in the CpG was investigated.

To reduce the non-biological variability between observations, data were normalized. For gene expression data, well-established normalization methods such as *quantile normalization* (QN) exist (Bolstad *et al.*, 2003, Irizarry *et al.*, 2003) (see Appendix A.1.1 for algorithm). However, re-evaluation of normalization strategies was necessary for methylation data, since these differ considerably from gene expression data in their data structure, and QN on methylation β-values does not show a good performance (Touleimat and Tost, 2012). Several competing normalization methods for DNA methylation data were proposed (Dedeurwaerder *et al.*, 2011, Maksimovic *et al.*, 2012, Touleimat and Tost, 2012, Teschendorff *et al.*, 2013, Pidsley *et al.*, 2013). Some of these strategies, including QN on the raw signal intensities (Pidsley *et al.*, 2013) and beta-mixture quantile normalization (BMIQ) (Teschendorff *et al.*, 2013), outperformed the others in terms of reduction of technical variation and subsequent detection of true signals (Marabita *et al.*, 2013, Pidsley *et al.*, 2013). Both methods performed similarly well in KORA using the criteria proposed by Pidsley *et al.* (2013), with the former performing slightly better. Therefore, QN on the raw signal intensities was chosen to normalize the KORA data. Precisely, QN was stratified to six probe categories based probe type and color channel (i.e., Infinium I signals from beads targeting methylated CpG sites obtained through the red and the green color channels, Infinium I signals from beads targeting unmethylated CpG sites obtained through the red and the green color channels, and Infinium II signals obtained through the red and the green color channels, see Bibikova *et al.* (2011)) using the R package *limma*, version 3.16.5 (Smyth, 2005).

Data normalization may partly reduce technical effects, but might not fully do so (Marabita *et al.*, 2013). Thus, another novel strategy was applied to avoid plate effects as well as

other technical effects in the methylation data (Lehne *et al.*, personal communication). The method is based on the 235 positive control probes that are present on every single position of the Infinium HumanMethylation450K BeadChip and serve as a quality control for different data preparation and measurement steps. Based on the assumption that the control probe intensities differ purely due to technical influences rather than biological differences of the samples, principal components (PCs, see Section 3.4.1) of the control probes are thought to capture the technical variability in the experiment. Including the first 20 control probe PCs as covariates in the model considerably removed technical biases (Lehne *et al.*, pesonal communication). Therefore, the first 20 control probe PCs were included as covariates in all models involving DNA methylation data in this thesis (see also Section 3.3.2 for covariates).

3.1.3 Gene expression data

Sample quality control and imputation of missing values was performed with GenomeStudio^{TM}, version 2010.1. Thereby, 4 samples with less than 6000 detected probes (detection p-value > 0.01) were excluded, leaving 993 KORA F4 subjects in the data set.

Data were quantile normalized using the R package *lumi*, version 2.8.0 (Du *et al.*, 2008). *Quantile normalization* (QN) achieves identity of the feature distributions of all observations (see Appendix A.1.1 for algorithm, Bolstad *et al.* (2003)). Although this method has the strong assumption that quantile values, including the tails, are equal across all subjects, it seems to work well for gene expression data (Bolstad *et al.*, 2003).

For gene expression data, no control probes were available. A simple alternative was chosen, namely the inclusion of known technical factors as a covariate in the statistical models. For the KORA data, assignment to one of 30 amplification plates, sample storage time between blood sampling and RNA isolation, and RNA integrity number were important technical aspects explaining a large proportion of variability in the gene expression data (Schurmann *et al.*, 2012).

3.1.4 Metabolomics data

Metabolon platform

Technical effects were controlled for by dividing the metabolite concentration values by the median value of samples measured on the same day for each metabolite. In addition, outlier values of > 4 standard deviations from the mean of the respective metabolite on the $\log_{10}$ scale were set to missing. Finally, 81 metabolites (42 identified, 39 unidentified) with more than 50% missing values were excluded, and 5 observations with more than 20% missings, leaving a total of 1763 observations of 436 metabolites (283 identified, 153 unidentified) for analysis.

NMR platform

Preprocessing of NMR data was similar to Metabolon data in terms of outlier exclusion and detection rate thresholds. None of the metabolite traits had more than 50% missing values. 4 observations with more than 20% missings were excluded from the data set, leaving a total of 1784 observations for analysis.

Combined analysis of the Metabolon and NMR metabolomics platforms in KORA

For 1658 KORA F4 subjects, both Metabolon and NMR data were available. They were combined in a multi-platform approach in Section 4.3. Data from different metabolomics platforms have been jointly analyzed before, and selected ratios of two metabolites determined on different platforms have provided a disease-related readout (Suhre *et al.*, 2010). The Metabolon and NMR data are based on different underlying techniques with different characteristics (Vinayavekhin *et al.*, 2010, Issaq *et al.*, 2009). Whereas MS-based technology (Metabolon) is more sensitive and requires a small sample volume, it is usually coupled to chromatographic separation, which often requires derivatization of metabolites, preventing sample re-use. In contrast, NMR is nondestructive and also outperforms MS-based technology in terms of its reproducibility: NMR is less susceptible to technical effects due to less complex sample preparation and measurement steps as compared to MS (Suhre and Gieger, 2012). Since both platforms also differ in covered metabolite spectrum, their combined usage has been recommended (Suhre and Gieger, 2012).

20 metabolites were covered by both techniques, comprising amino acids, two fatty acids, total serum cholesterol, 3-hydroxybutyrat, glycerol, citrate, creatinine, lactate, pyruvate and urea. For these, the cross-platform correlations were explored. The median Pearson correlation coefficient was 0.72, ranging from 0.19 (linoleate) to 0.91 (3-hydroxybutyrate), and being above 0.45 for all metabolites but linoleate. Disentangling the reasons behind the low correlation observed for linoleate is not the objective of this thesis, however, results concerning linoleate should be interpreted with care. [M] and [N] is added to the metabolites in this work to indicate measurement on the Metabolon and the NMR platform, respectively. Metabolites of both platforms were assigned to super-pathways and sub-pathways in accordance with the pathways proposed by Metabolon on the basis of *Kyoto Encyclopedia of Genes and Genomes* (KEGG) pathways (Appendix Table A.5).

Biocrates platform (VID study)

To determine measurement stability, the coefficient of variation of repeatedly measured reference plasma metabolite concentrations was defined as standard deviation divided by mean metabolite concentration. 31 metabolites showing impaired measurement stability (coefficient of variation > 0.25 on any plate), or with $>95\%$ observations having zero concentration, were excluded, leaving 132 metabolites for analysis.

Slightly different quality control criteria were used for the analysis of *TCF7L2* genotype presented in Section 4.1.3. These are described in detail in the corresponding publication (Then *et al.*, 2013).

Measurements took place on 9 plates, with 5 reference plasma samples measured repeatedly on each plate. This enabled the usage of the *geometric ratio-based method* to correct for plate effects (Luo *et al.*, 2010). For each metabolite j,$j = 1, \ldots, p$, values were multiplied with a plate-specific correction factor derived through

1. computing the geometric means of reference sample values on each plate k, denoted as g_{jk},
2. computing the geometric mean of all means g_{jk}, denoted as g_j,
3. deriving the plate-specific correction factor as $f_{jk} = \frac{g_j}{g_{jk}}$, and
4. multiplying all values of metabolite j on plate k with f_{jk}.

This method has shown good performance when applied to gene expression data in the evaluation by Luo *et al.* (2010), although another investigation by Chen *et al.* (2011) has found it to be outperformed by more sophisticated methods. However, since it is based on reference samples rather than on biological samples, the geometric ratio-based method might perform specifically well when technical and biological variation are indistinguishable (Chen *et al.*, 2011). This was the case in this thesis, where biologically distinct groups of samples (i.e., from obese and non-obese subjects, as in the Obeldicks study in Section 4.4, or from different challenge time points, as in the VID study in Section 4.1) were not randomly assigned to the plates.

Biocrates platform (Obeldicks study)

Quality control of the Obeldicks data was conducted in a slightly different way. Three criteria for measurement reliability were used:

(1) The concentration of the metabolite should be above the limit of detection specified by the manufacturer in $\geq 60\%$ of the samples, since values below the limit of detection might represent non-reliable signals.

(2) The Pearson's correlation coefficient of the metabolite concentrations in the 43 repeatedly measured samples should be ≥ 0.5, since lower correlations might point towards large technical variability.

(3) On each batch, the coefficient of variation for the metabolite concentration in a reference sample that was measured five times on each batch should be ≤ 0.2, to ensure measurement stability.

33 Metabolites that failed to meet at least two of these criteria were excluded from the analysis. The majority of these were also characterized by concentrations below or marginally above the limit of detection.

Data were then normalized using the geometric ratio-based method, based on the 43 samples measured repeatedly in both batches.

3.1.5 Phenotype data

For all analyses of the KORA F4 data, subjects were excluded that had a fasting duration of less than 8 h. For the study in Section 4.2, one subject with BMI>50 kg/m^2 was excluded. For Section 4.3, 5 subjects with outlying values in body weight change, defined as values outside mean $\pm$ 5 standard deviations, were excluded. In addition, in Sections 4.2 and 4.4, few subjects with missing information in relevant phenotypes were excluded (see Section 3.2 below for details on missing value handling).

3.2 Missing data handling

3.2.1 The problem

The majority of statistical methods, including those applied in this thesis, require a complete data matrix. However, missing values occur frequently in omics data. They arise from technical reasons, from the exclusion of unreliable signals during quality control, or from the unavailability of biosamples. Dependent on the origin of the missing values, they might be categorized into three groups (Little and Rubin, 2002, Raessler *et al.*, 2008):

- *Missing completely at random (MCAR)*: Their missingness is completely independent of the missing values themselves or any other values in the data set. This applies to values missing for purely technical reasons, e.g. missings in methylation data when not enough DNA has bound to the beads on the methylation chip to determine a reliable methylated or unmethylated signal.

- *Missing at random (MAR)*: Their missingness might depend on observed data, but not on the missing values themselves, i.e. these values are missing randomly, given the observed data. This applies to values missing from the systematic unavailability of biosamples for a specified subset of subjects.

- *Missing not at random (MNAR)*: Their missingness might depend on the missing values themselves, even given the observed data. This applies to values that were set to missing for being smaller than the detection limit of a laboratory machine. Note, however, that sufficient correlation with other variables can render their missingness MAR, since their missingness might then to a large degree be explainable by observed values.

This categorization, together with the extent of missingness in the data set, fundamentally determines how missing values should be handled. If missings are rare, and most observations are complete, and/or missings are MCAR, and univariate models are pursued, simple approaches such as *complete-case analysis* (where each molecular feature is modeled using the available observations only) and *single imputation* (where a single complete data set is generated by filling up the missing values through mean imputation or regression techniques) might be acceptable (and even favored for computational reasons). This was the case in the majority of the projects in this thesis, where complete-case analysis (Sections 4.2 and 4.4) or single imputation (Section 4.1) were applied. Specifically, in Section 4.4, children were excluded that had missing values in waist circumference, which was only determined in a random part of participating children. These missing values were considered MCAR, and despite potential loss of efficiency when applying the multivariate LASSO approach, complete-case analysis was chosen to avoid the problem of applying LASSO within the context of more complex missing data handling methods.

With an increasing number of missing values, complete-case analysis becomes inefficient (or impossible), specifically in multivariate analyses, where only complete observations can be included. Even more, if MCAR is not given (which is the case for the *Metabolon* and *NMR* metabolomics data analyzed in Section 4.3), both complete-case analysis and mean imputation can be associated with serious bias in the estimated effects and p-values (Little and Rubin, 2002). More sophisticated single imputation approaches are also likely invalid, since standard analyses on a single imputed data set underestimate the variance of estimates since they ignore the uncertainty of the imputed values (Little and Rubin, 2002, Raessler *et al.*, 2008).

3.2.2 Multiple imputation

A solution to this problem is *multiple imputation*, which has first been introduced by (Rubin, 1978). It provides valid results if the MAR assumption is plausible, and might give a good approximation in MNAR situations (Raessler *et al.*, 2008). It involves three steps (van Buuren *et al.*, 1999):

1. *Imputation*: Repeated application of single imputation to generate multiple (M) imputed data sets. This can be achieved through *Multiple imputation by chained equations* (MICE) (van Buuren *et al.*, 1999, Raghunathan *et al.*, 2001, van Buuren, 2007); see Appendix A.1.2 for algorithm.

2. *Complete-data analysis* on each imputed data set.

3. *Combination* of the analysis results from the M data sets, thereby taking into account the uncertainty that is due to imputation of missing values and which is reflected in the variability of the multiply imputed values; see Appendix A.1.2 for combination rules by Rubin (1987).

Description of missingness and correlation structure in the data set

Multiple imputation using MICE was applied to the Metabolon and NMR metabolomics data in Section 4.3. Prior to analysis, missingness was closely examined. Of the data set comprising 1631 observations of 582 variables (27 phenotypes, 436 Metabolon metabolites and 119 NMR metabolites after removal of metabolic traits representing sums, differences or ratios), 7.2% entries were missing. The median number of missing entries among the observations was 40 (6.9%), ranging from 11 (1.9%) to 83 (13.3%). Phenotypes had at most 8 (0.5%) missing values. Of the Metabolon metabolites, 19 (4.4%) were completely observed, while among the remaining 417 metabolites, 23 had more than 40% missing values. The median number of missing observations was 25 (1.5%), ranging from 0 to 815 (50.0%). Of the NMR variables, 7 (5.9%) were completely observed. The median number of missing entries was 6 (0.4%), ranging from 0 to 517 (32.3%). The overall missingness pattern was unstructured (Figure 3.2), indicating that missingness did not co-occur in

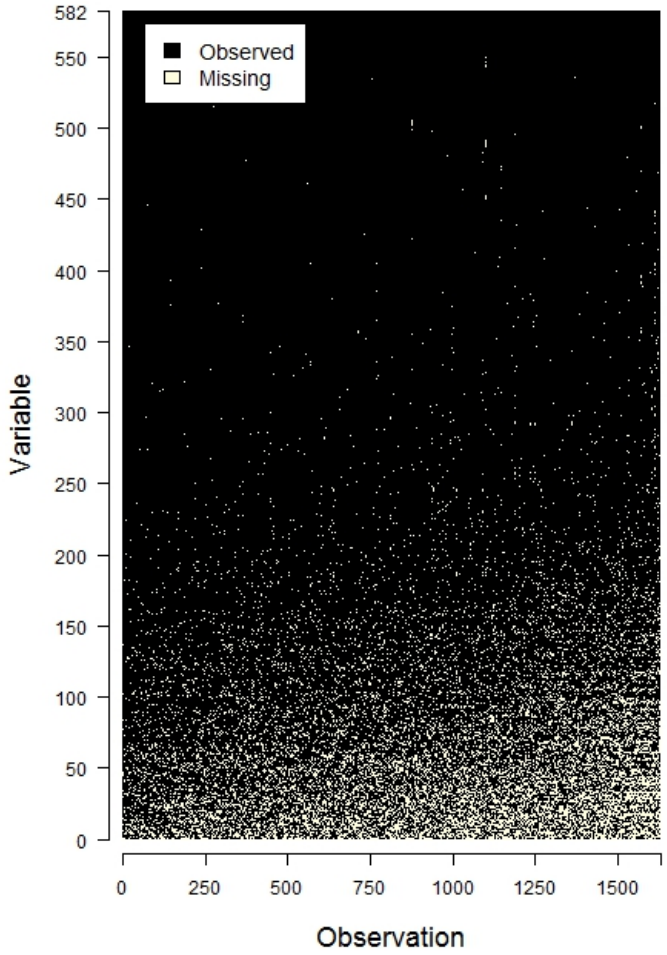

Figure 3.2: Missingness pattern in the metabolomics data set. Plot of missingness indicators (black, observed; light yellow, missing) for the 582 variables against the 1631 observations, both sorted by percentage of missing values.

large blocks of variables, which is beneficial to the imputation process in that for the missing values of a specific variable, values of correlated variables have been observed and can be used to improve imputation of the missing values.

For the 184 variables (177 Metabolon and 7 NMR metabolites) with more than 5% missing entries, more detailed descriptive analyses were performed. First, correlation of these variables with all other variables was visualized in heatmaps to get an impression of how much information for their imputation could be borrowed from other variables (example for correlation of the 7 NMR metabolites with all NMR metabolites is shown in Figure 3.3A). Since few categorical (phenotypic) variables were included in the data set, Kendall's rank correlation coefficients τ were used. Each of the 184 variables showed absolute correlation of $|\tau| > 0.1$ with at least one other variable in the data set (exactly one for the Metabolon metabolites leucylleucine, thymol sulfate and X-12443, and up to 187 for 1-palmitoylglycerol). On the other hand, 561 of the 582 variables in the data set provided information for at least one of the 184 variables. The strongest correlations were observed within rather than between the two metabolomics platforms.

Second, to explore the MAR assumption, correlation heatmaps of missingness indicators of the 184 variables with values of all variables were drawn (example for correlation of the 7 NMR metabolites with all NMR metabolites is shown in Figure 3.3B). Specifically for 4 NMR metabolites, XXL_VLDL_P, XXL_VLDL_PL, XXL_VLDL_L and XXL_VLDL_TG, missingness showed strong negative and positive correlations with VLDL (up to $\tau = -0.52$) and HDL (up to $\tau = 0.31$) metabolites, respectively. Missingness of Metabolon metabolites showed less pronounced correlations with variable values.

Together, these descriptive insights showed that the degree of missingness in the metabolomics data set was moderate, that the MCAR assumption was unlikely, whereas the MAR assumption was satisfied due to the strong interrelatedness of the variables. Thus, multiple imputation was chosen as an appropriate solution.

Imputation settings

Prior to imputation, the distribution of the continous variables was investigated. Raw, natural log transformed, cubic root and square root transformed variables were tested for normality using Shapiro-Wilk tests. The transformation that showed the smallest deviation from normality was chosen, and also kept for all subsequent statistical analyses. 350, 123 and 49 variables were log, cubic root and square root transformed, respectively, and 16 variables were not transformed.

Data were imputed with the R package *mice*, version 2.21 (van Buuren and Groothuis-Oudshoorn, 2011). Both Bayesian linear regression and predictive mean matching (PMM) were used to impute continuous variables (see Appendix A.1.2). Specifically, although PMM might be preferable to Bayesian linear regression for being more robust, it might

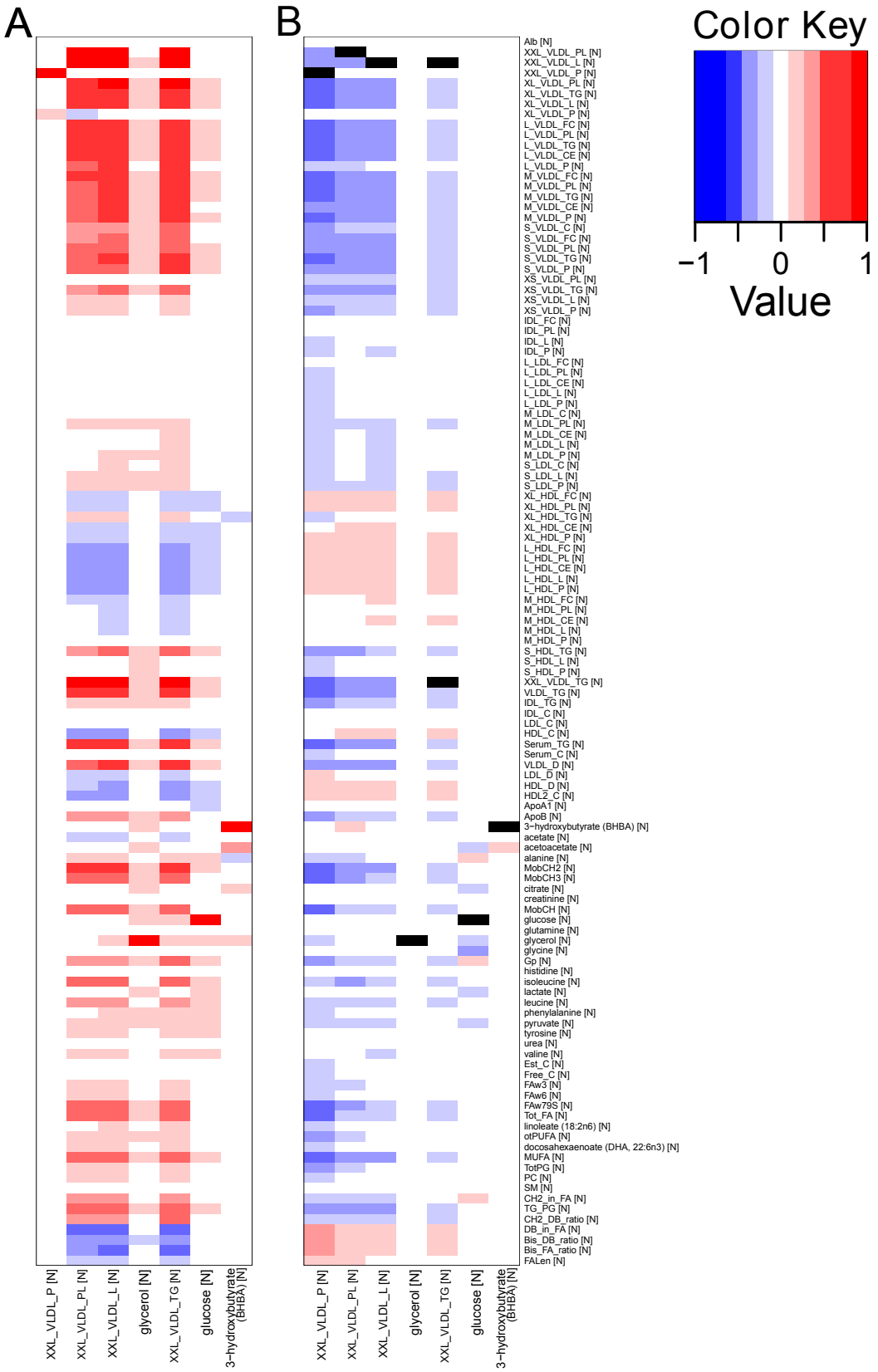

Figure 3.3: Correlation among variables and missingness. **A** Kendall's correlation between the 7 NMR variables with more than 5% missing values against all NMR variables. **B** Kendall's correlation between the missingness indicators (0, observed; 1, missing) of the 7 NMR variables with more than 5% missing values against all NMR variables. White, $|\tau| < 0.15$; red, $\tau \geq 0.15$; blue, $\tau \leq 0.15$.

have undesirable properties in the case of MAR or MNAR values arising from values below the detection limit (i.e., actually plausible low values will be imputed with higher values). Thus, PMM was only applied to the phenotypes and to two metabolites (XXL_VLDL_P and X-12544) showing strongly asymmetric distributions even after transformation, whereas Bayesian linear regression was applied to the remaining metabolites. To avoid the occurence of negative metabolite values generated through Bayesian linear regression, the *squeeze* function was used as a postprocessing step wherever variables were not log-transformed. Dichotomous and categorical variables were imputed using logistic and generalized logistic regression, respectively.

Covariates for the imputation models were chosen according to the recommendations described in Appendix A.1.2. Specifically, auxiliary variables were included as covariates if they correlated with the value or missingness of incomplete variables at $|\tau| > 0.1$ and were observed for at least 20% of the subjects missing the incomplete variable. The number of auxiliary variables was restricted to 30. Unidentified Metabolon metabolites (see Section 2.1.7) were only imputed, if they represented auxiliary variables for identified variables, or for unidentified auxiliary variables. After imputation, unidentified metabolites were removed from the data set.

Imputation diagnostics

20 imputed data sets were generated with 10 iterations each, which ensured convergence of the imputation algorithm sufficiently for all variables. Example plots are shown in Figure 3.4 for XXL_VLDL_P [N] (32.3% missing values) and and 1,7-dimethylurate [M] (42.7% missing values). At 20 imputations, *relative efficiency* (RE, see Appendix A.1.2) was above 0.97 for all analyses. Distributions of imputed and observed values of each variable were compared by means of kernel density plots, revealing, as expected, by trend lower imputed than observed values for a number of metabolites (extreme example 1,7-dimethylurate [M] in Figure 3.5).

Combination of single imputation estimates

Where appropriate, i.e. for combination of linear regression estimates, Rubin's rules were applied (see Appendix A.1.2, Rubin (1987)). To cluster metabolites, WGCNA was applied to each imputed data set. Since clustering solutions differed only marginally, a single clustering solution from one imputed data set was chosen. As an *ad hoc* solution, the clustering solution that assigned the majority of metabolites to a module, leaving the lowest number of metabolites unassigned, was chosen. All subsequent models involving the module eigengenes were again combined using Rubin's rules.

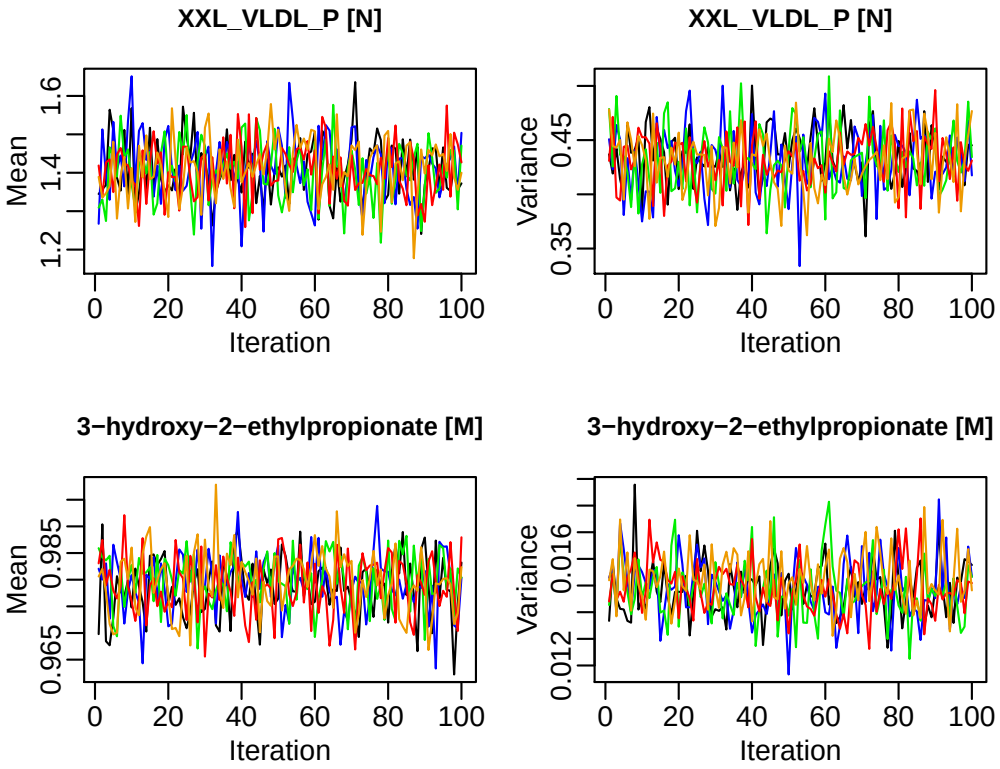

Figure 3.4: Convergence of the MICE algorithm for two selected variables. Plotted is mean (left) and variance (right) of imputed values for each of 5 imputation chains across 100 iterations, exemplarily for two variables with large numbers of missing values.

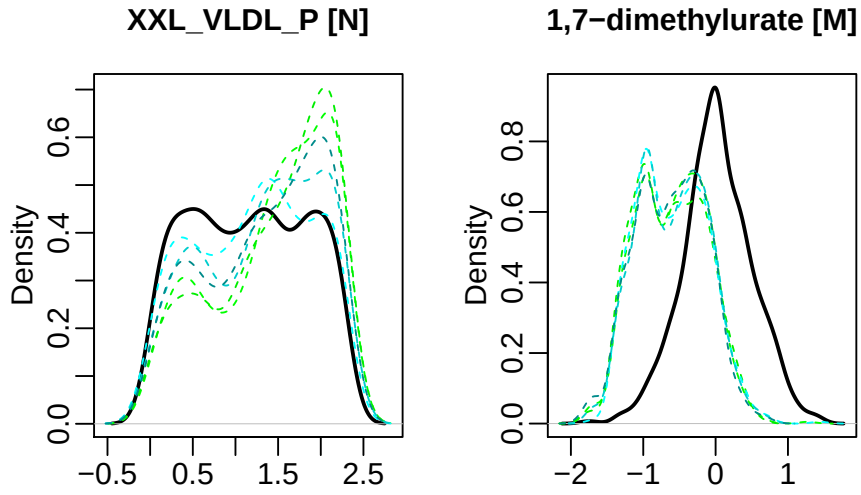

Figure 3.5: Imputation diagnostics for two selected variables. Kernel density plots of observed (black solid line) vs. imputed (dashed lines in different shades of green) values, shown for the first five imputations.

3.3 Univariate data analysis

3.3.1 Modeling the relation between a phenotype and a matrix of molecular variables

In this thesis, univariate screening for associations between molecular features and a specific phenotype/state was frequently performed as a key explorative step.

The linear regression model

The *linear regression model* was most frequently applied (in Sections 4.2 to 4.4). It is given as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right),$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ denotes an $n \times 1$ continuous response vector corresponding to n independent observations, $\mathbf{X}$ an $n \times (p+1)$ covariate matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ a $(p+1) \times 1$ vector of unknown regression coefficients (including an intercept), $\boldsymbol{\eta}$ the linear predictor, and $\boldsymbol{\epsilon}$ an $n \times 1$ error term with the common variance σ^2 (Fahrmeir *et al.*, 2013, Faraway, 2002) (see details on estimation of the parameters $\boldsymbol{\beta}$ and σ^2 and hypothesis testing in Appendix A.1.3). Note that p is used to abbreviate both the number of covariates and the p-value in this thesis.

In Section 4.3, interactions between covariates were included in the linear predictor in order to perform subgroup analyses. Specifically, the basis model

$$\boldsymbol{y}_i = \beta_0 + \beta_1 \Delta\mathrm{BW}_i + \beta_2 \mathrm{sex}_i + \beta_3 \mathrm{age}_i + \beta_4 \mathrm{BW}_{\mathrm{baseline,i}} + \boldsymbol{\epsilon}_i$$

was extended to incoporate a subgroup indicator $\mathrm{Sub}_i = I(\text{subject i outside subgroup})$ (where the subgroup was consecutively specified as weight gain/weight loss, obese (BMI > 30) /non-obese, central obese (waist-hip ratio (WHR) > 1 in males and > 0.85 in females)/not central obese, male/female, and age $> 55/\leq 55$ years) and its interaction with the remaining covariates:

$$\begin{aligned} y_i \;=\; & \beta_0 + \beta_1 \Delta\mathrm{BW}_i + \beta_2 \mathrm{sex}_i + \beta_3 \mathrm{age}_i + \beta_4 \mathrm{BW}_{\mathrm{baseline,i}} + \beta_5 \mathrm{Sub}_i \\ & + \beta_6 \mathrm{Sub}_i\ \Delta\mathrm{BW}_i + \beta_7 \mathrm{Sub}_i\ \mathrm{sex}_i + \beta_8 \mathrm{Sub}_i\ \mathrm{age}_i + \beta_9 \mathrm{Sub}_i \mathrm{BW}_{\mathrm{baseline,i}} + \boldsymbol{\epsilon}_i. \end{aligned}$$

The main effects β_1 to β_4 were then interpreted as effects of the respective covariates within the subgroup, whereas the interaction effects β_6 to β_9 reflect the difference in covariate effects between the groups of subjects outside vs. within the subgroup.

The linear mixed-effects model (LME)

To study the effect of different challenges on metabolite concentrations in the VID study (Section 4.1), the linear model could not be used. Due to multiple measurements per subject, the assumption of independent observations with uncorrelated error terms ϵ_i would have been violated. The *linear mixed-effects model* allows for correlation among the observations of a subject by explicitly modeling subject-specific effects. A simple representation is (Fahrmeir *et al.*, 2013):

$$\begin{aligned} \boldsymbol{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, n_i, \text{ with} \\ \boldsymbol{b}_i &\sim N(0, \mathbf{D}) \\ \boldsymbol{\epsilon}_i &\sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}\right), \end{aligned}$$

where

- $\boldsymbol{y}_i$ the $n_i \times 1$ vector of the response at n_i measurements of subject i
- $\mathbf{X}_i$ a $n_i \times (p+1)$ matrix of covariates for subject i including an intercept
- $\mathbf{Z}_i$ a $n_i \times q$ matrix of covariates for subject i including an intercept
- $\boldsymbol{b}_i$ a $q \times 1$ vector of subject-specific effects
- $\boldsymbol{\epsilon}_i$ the $n_i \times 1$ vector of error terms corresponding to subject i.

Specifically, in Section 4.1, a *random intercept* model was fitted, where $q = 1$, so $Z_i = \mathbf{1}_{n_i}$ and b_i is a scalar with scalar variance d^2. It assumes that all measurement corresponding to a subject i have a subject-specific response level, which is expressed as the random intercept b_i that adds to the global intercept β_0. The model assumes that all measurements corresponding to a subject i show the same correlation with each other (Fahrmeir *et al.*, 2013). Estimation, testing and parameter interpretation are very similar as for the linear model described above. More details, for instance on restricted maximum likelihood estimation of the variance components, can be found in Fahrmeir *et al.* (2013).

Random intercept LMEs were fitted using the R package *nlme*, version 3.1-103 (Laird and Ware, 1982). Prior to analysis, best-fitting random effects were determined using the Akaike information criterion, and random intercept was consequently used in all models. Separate models were fitted for each of the four metabolic challenges (see Section 2.2.3). The specific model formulation was (for each of the 132 metabolites):

$$y_{ij} = \beta_0 + \beta_1 I(t_j = 1) + \beta_2 I(t_j = 2) + \beta_3 \text{BMI}_i + \beta_4 \text{Age}_i + \beta_5 I(\text{Genotype}_i = \text{AA}) + b_{i0} + \epsilon_{ij},$$

where t_j represents the time point with baseline as reference (see Figure 2.1 for an explanation of the two post-challenge timepoint for each challenge), y_{ij} the scaled log-transformed concentration of the respective metabolite for subject i at timepoint t_j, and $I(\cdot)$ the indicator function.

The effect of challenge on metabolite concentration was investigated by testing the two

linear hypotheses with t tests: $\beta_1 = 0$ and $\beta_2 - \beta_1 = 0$, referring to the challenge effect between baseline and the first post-challenge time point, and between the first and second post-challenge time point, respectively. Similarly, challenge-induced changes in clinical traits were investigated, including all time points for which measurements were available (Figure 2.1). To assess genotype effects on challenge response of metabolites or clinical traits, interactions between time points and genotype were added to the LMEs, additionally including interactions of BMI and age with time point to adjust for BMI or age effects on challenge response. Genotype main and interaction effects of these models were interpreted as genotype effect on baseline concentrations and on challenge response, respectively.

The logistic regression model

Finally, in the case of binary responses, such as disease status (Sections 4.2 and 4.3) or weight loss success (Section 4.4), *logistic regression* was used (Fahrmeir *et al.*, 2013). See Appendix A.1.4 for model formulation.

3.3.2 The choice of covariates

The appropriate choice of covariates in the above described models was given particular attention in this thesis. When the association between a response variable and an independent variable is studied, additional covariates need to be correctly specified (1) to to ensure that model assumptions are met, (2) to reduce the "noise" in the response variable, thereby increasing the power to detect true associations with the independent variable of interest, and (3) to avoid *confounding* of the investigated association.

Prior knowledge on variables that relate to both response and the independent variable of interest, as well as knowledge on the direction of these associations is crucial for an appropriate choice of covariates (Hernán *et al.*, 2002). Here, the definition of *confounders* and *colliders* comes into play.

Confounders can be defined as variables that represent a common cause of response and independent variable (Greenland and Morgenstern, 2001). It is widely acknowledged that confounders need to be adjusted for (e.g., by including them as covariates in the model) to avoid spurious false associations between response and independent variable. The problem is visualized in Figure 3.6A. An example is a model with a specific age-related metabolite as response, BMI as independent variable, and age as confounder, where age is known to have an effect on BMI. If age is not included in the model, the metabolite would be found to be associated with BMI due to their common association with age. Note that in some instances, it might be reasonable also to include variables that might be on the causal path between independent variable and response (i.e., potential *mediators*) as covariates to abolish indirect effects.

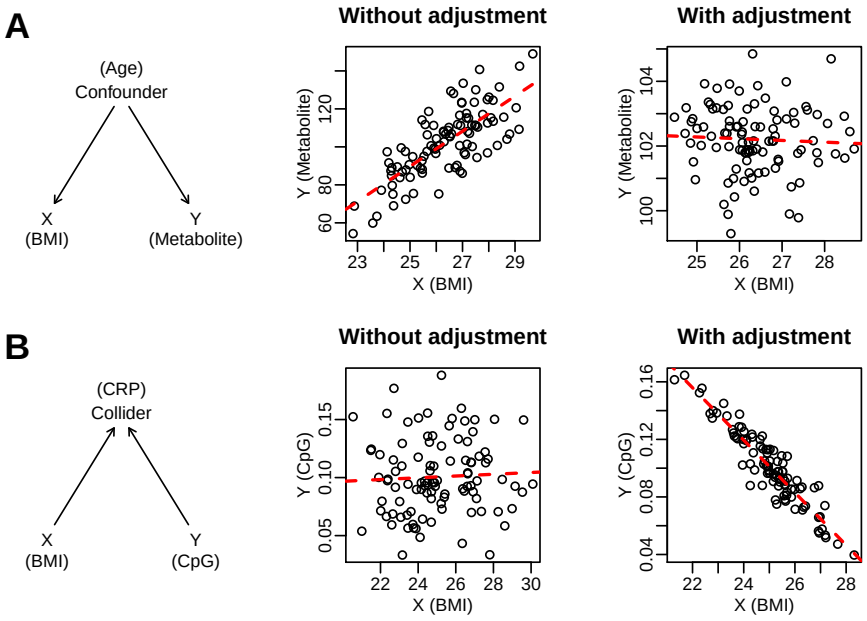

Figure 3.6: Confounder versus collider. A Missing adjustment for a confounder can cause spurious associations between variables X and Y. **B** Adjustment for a collider can cause spurious associations between variables X and Y.

Associations with genetic variants are least prone to confounding, since alleles of a genetic variant are inherited randomly and should not be affected by phenotypes. However, there are few exceptions. These include *population stratification*, which refers to the situation where confounding occurs through an admixture of subpopulations that differ in both their allele frequencies and in the phenotype (Ziegler and König, 2010).

Particularly challenging confounders for methylation and expression data analysis from whole blood samples are the proportions of blood cells. Whole blood is a heterogeneous mixture of different cell types that differ strongly in DNA methylation and expression of specific genes (Houseman *et al.*, 2012, Reinius *et al.*, 2012, Zhu *et al.*, 2012). At the same time, phenotypes and diseases might be associated with changes in blood cell proportions. Evidence for an effect of body mass on blood cell proportions is given by Bellows *et al.* (2011) and Trottier *et al.* (2012). Thus, cell proportions are potentially strong confounders of methylation (or gene expression) - phenotype relationships that need to be accounted for (Jaffe and Irizarry, 2014). Since cell type measurements were not available for the KORA data in Section 4.2, proportions of selected cell types (i.e., granulocytes, monocytes, B cells, $CD4^+$ T cells, $CD8^+$ T cells and natural killer cells) were estimated using the procedure proposed by Houseman *et al.* (2012). This is a two-step projection procedure where in the first step, the 500 most cell type-specific CpG sites are determined from pure cell data, and in the second step this information is used to infer cell proportions from the target whole blood data (i.e., KORA). The obtained estimates of the six cell proportions were

included as covariates when modeling the methylation data. A slightly different approach was used for the gene expression data analysis in Section 4.3. Transcripts associated with specific relevant cell types, that is, reticulocytes and basophils, were obtained (Whitney *et al.*, 2003) and subjected to PCA (see Section 3.4.1). The resulting PCs were used as covariates in the model of interest.

Colliders are variables that represent a common consequence of response and independent variable (Janszky *et al.*, 2010, Cole *et al.*, 2010). Conditioning on a collider, e.g. by including it as a covariate in the model, can cause spurious associations even in absence of a true association between response and the independent variable of interest (Figure 3.6B). An example is given with a specific methylation site as response, which is involved in inflammation (say, methylation is positively associated with CRP levels) but not related to body weight, and BMI (which causes CRP levels to rise, Fall *et al.* (2013)) as the independent variable. If CRP is included as a covariate in the model, a spurious negative association between methylation and BMI might results. This is due to the fact that given high CRP levels and high methylation levels, large BMI is less likely to be the reason for the high CRP levels than if methylation levels had been lower.

Thus, it is important to know potential confounders and colliders of a specific research question before model building. Specifically for omics data, associated factors are just beginning to be understood, and even more basic is the knowledge on the causal relations (see Section 3.5 below). Thus, to abolish confounding while avoiding the inclusion of a collider, it might be a good approach to start with a sparse model comprising only few covariates, and potentially include more covariates in additional models. If the inclusion of covariates diminishes the association between a molecular features and a phenotype, confounding may have been present. If it profoundly strengthens their association, or indicates an association that was not present without these covariates, the covariates might have been colliders and the association should not be trusted.

The specific covariates included in models in this thesis are given in the respective sections in Chapter 4.

3.3.3 Violation of the distribution assumption

The models described in Section 3.3.1 are *parametric* models, which might be loosely translated to "models with a distribution assumption". Specifically, the linear and linear mixed-effects models assume a normally distributed error term, which is equivalent to assuming a normally distributed response conditioning on the covariates:

$$\begin{aligned} \boldsymbol{\epsilon} &\sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right) \\ \boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} &\sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right). \end{aligned}$$

Both omics data and phenotypic traits were not always normally distributed. Phenotypes often assumed a right-skewed distribution, as did metabolite concentrations and transcript levels. Methylation values represent proportions that are defined on the unit interval $(0, 1)$ (Bibikova *et al.*, 2011) and do not show a normal distribution (Wahl *et al.*, 2014). Three different strategies were applied in this thesis to address this issue.

Response transformation

As a simple remedy to achieve at least approximate normality and ensure the applicability of simple parametric models, the response variable was frequently transformed. Gene expression data in Sections 4.2 and 4.3 were $\log_2$-transformed, Biocrates metabolite concentrations in Section 4.1 were natural log-transformed, as were clinical trait levels in Sections 4.2 and 4.3. For Metabolon and NMR metabolite concentrations in Section 4.3, the transformation achieving the smallest deviation from normality was chosen for each metabolite (compare Section 3.2.2).

Robust effect estimation

An alternative solution is to use more robust approaches. Ordinal regression (Harrell, Jr., 2006) was applied in Section 4.1 to assess differences in baseline characteristics between genotype groups, and rank correlation (Kendall, 1938) was applied to the Metabolon and NMR metabolomics data in Section 4.3 to investigate missingness (see Section 3.2.2).

Permutation tests

A third option that was frequently applied in this thesis is the use of a standard parametric model followed by a non-parametric test such as a *permutation test*. The idea behind permutation tests is that the distribution of a test statistic obtained from fitting the model to B permuted data sets with a randomly shuffled response vector resembles its distribution under the null hypothesis that there is no effect (Moore *et al.*, 2003, Knijnenburg *et al.*, 2009, Radmacher *et al.*, 2002). The proportion of resampling folds where the test statistic $\theta_{perm}^{(b)}$ is at least as extreme as the observed test statistic $\hat{\theta}_{\text{obs}}$ in the original data can therefore be interpreted as a p-value:

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} I(|\theta_{perm}^{(b)}| \geq |\hat{\theta}_{\text{obs}}|).$$

Note that permutation tests are computationally intensive, and that the number B of required permutations increases with the required resolution of the p-value (and consequently, with the number of tests performed, see Section 3.3.4 on multiple testing) (Knijnenburg *et al.*, 2009). A solution can be to normal approximate the null distribution of the test statistic, and then to derive the p-value in a parametric way.

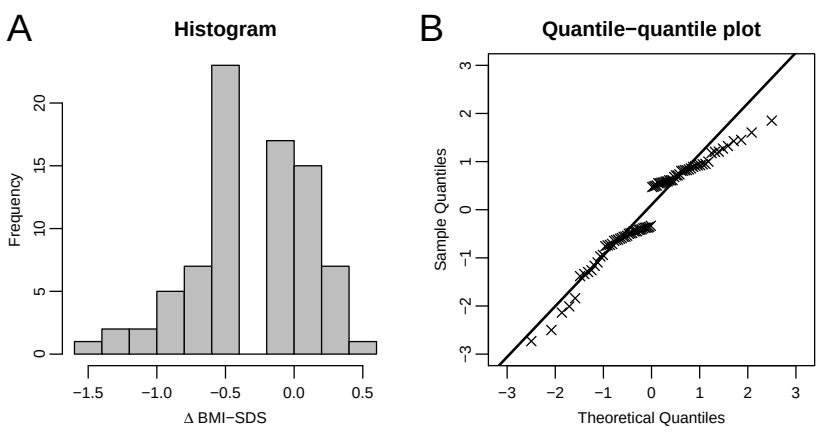

Figure 3.7: Distribution of the response variable "BMI-SDS change during intervention" (ΔBMI-SDS) in the Obeldicks study. A Histogram. B Normal quantile-quantile plot. The distribution is not normal according to Shapiro-Wilk test (p-value = 0.0019).

A permutation test with $B = 10,000$ permutations was applied in the VID study (Section 4.1), where the response variable ΔBMI-SDS by design did not follow a normal distribution, since only children with a substantial BMI-SDS reduction, and children without BMI-SDS reduction, were included in the study (Figure 3.7). Permutation testing was also applied several times in this thesis to determine the significance of test statistics the distribution of which is not known. These include the measure of predictive performance of the LASSO model, Q^2, in Section 4.4 (see Section 3.4.2 for the statistical method), and the inter- and intra-module connectivity measures in Section 4.3 (see Section 3.4.1 for the statistical method). In Section 4.2, a permutation test with normal approximation was used in the context of correlation analysis of average methylation signals of 187 CpG sites between different tissues. In that analysis, independence of the "observations" (i.e., CpG sites) is not necessarily given, so the conditions for standard inference are not met. Finally, permutation tests were applied within the context of enrichment analyses in Sections 4.1, 4.2 and 4.3 (see Section 3.5 for methodology).

Bootstrap

In Section 4.2, non-parametric bootstrap was chosen as a mode of inference, which is also based upon resampling. In contrast to permutation testing, observations are sampled as a whole, traditionally with replacement to obtain bootstrap samples of size n. Bootstrap p-values for H_0: $\theta = 0$ can be derived as follows, according to the guidelines by Hall and Wilson (1991):

- Estimate $\hat{\theta}_{\text{obs}}$ from the observed data.
- Generate B bootstrap samples. In each sample, estimate $\theta^{(b)}$.
- Define p-value $= \frac{1}{B}\sum_{b=1}^{B} I(|\theta^{(b)} - \hat{\theta}_{\text{obs}}| \geq |\hat{\theta}_{\text{obs}}|)$. Alternatively, normal approximate the null distribution of $|\theta^{(b)} - \hat{\theta}_{\text{obs}}|$.

Specifically, bootstrap p-values were obtained to assess significance of the part of the association between BMI and a clinical trait that was explained by a CpG site (or several

CpG sites), i.e., the "indirect" effect. First, to determine "total" and "direct" (i.e., remaining after adjustment for CpG(s)) association of BMI and trait, two linear/logistic regression models were fitted in the case of continuous/binary traits: Clinical trait was modeled as a function of BMI and discovery covariates (Section 4.2), without and with CpG(s) as additional covariate(s), respectively. BMI effect on clinical trait estimated in both models was defined as $\hat{\beta}_{\text{total}}$ and $\hat{\beta}_{\text{direct}}$, and indirect effect as $\hat{\beta}_{\text{indirect}} = \hat{\beta}_{\text{total}} - \hat{\beta}_{\text{direct}}$. Then, H_0: $\beta_{\text{indirect}} = 0$ was tested, using the bootstrap procedure described above with normal approximation, after a very good normal distribution fit of $|\beta^{(b)}_{\text{indirect}} - \hat{\beta}_{\text{indirect}}|$ was observed at $B = 10,000$.

A potential upward bias in the estimation of β_{indirect} using all CpGs was anticipated, arising from the fact that β_{direct} was estimated in a model including 187 additional covariates. Thus, 187 additional covariates were also added in the other model as randomly permuted CpGs. Subsequently, $\beta^{(b)}_{\text{total}}$ was estimated from this model and the average $\frac{1}{B}\sum \beta^{(b)}_{\text{indirect}}$ across all bootstrap samples was compared to $\hat{\beta}_{\text{indirect}}$. Bias was negligible.

3.3.4 Multiple testing

With the high dimensionality of omics data comes the challenge of multiple hypothesis testing. The larger the number of tests (i.e., of molecular features) p, the larger becomes the probability of false discoveries. Assume that for each molecular feature j, $j = 1, 2, \ldots, p$, a null hypothesis H_j is tested and rejected at $p_j < \alpha$, where α is commonly chosen as 0.05 in single-test scenarios. Then the probability of a false rejection would be

$$\begin{aligned} P(\text{any false rejection}) &= 1 - P(\text{no false rejection}) = 1 - \prod_{j=1}^{p} P(H_j \text{ not falsely rejected}) \\ &= 1 - \prod_{j=1}^{p}(1-\alpha) = 1 - (1-\alpha)^p, \end{aligned}$$

which assumes values above 0.5 from $p \geq 14$.

In this thesis, two different correction procedures were applied to avoid this problem: the *Bonferroni procedure* (Sections 4.2 and 4.3) and the *Benjamini-Hochberg procedure* (Benjamini and Hochberg (1995), Sections 4.1 and 4.4). The Bonferroni procedure controls the *family-wise error rate* (FWER), i.e. the probability of any false rejection, whereas the Benjamini-Hochberg procedure controls the *false discovery rate* (FDR), i.e. the expected proportion of false rejections (Dudoit *et al.*, 2003). See Appendix A.1.5 for details.

3.3.5 Power calculation

The *power* of a statistical model is defined as the probability of detecting a true effect (Walters, 2004). Power is tightly interrelated with the sample size n, the significance level

α, and the effect size. For simple statistical methods, such as the standard t test, analytical formula exist to determine the power of a model at a given n, $\alpha = 0.05$ and effect size.

As models become more complex, as in the case of Section 4.1, where the power for identifying an interaction effect in an LME is wanted, analytical power formula are difficult to derive, and bootstrap methods might be used instead (Efron and Tibshirani, 1994, Walters, 2004). $B = 10,000$ bootstrap samples (Section 3.3.3) were drawn stratified by genotype, the model of interest is fitted to each bootstrap sample and a p-value $p^{(b)}$ determined. Then, power was determined as $\frac{1}{B}\sum_{b=1}^{B} I(|p^{(b)}| < \alpha^*)$, where α^* represents the significance level of interest (e.g., α/p in the case of Bonferroni adjustment, see Section 3.3.4). In addition, the sample size needed in a future study to achieve a power of 80%, given that the observed effects were true effects, was calculated. Therefore, the procedure described above was repeated with increasing sample size, and the sample size, for which the estimated power for a given significance threshold exceeded 80%, was recorded.

3.3.6 Meta-analysis and external validation

In Section 4.2, results from several independent cohorts were combined to increase the power to detect small BMI-methylation effects. Results of discovery, replication and downstream analyses were combined by *meta-analysis*. Generally, *fixed-effects* and *random-effects* meta-analyses are distinguished (see Appendix A.1.6 for details). The former assumes a common true effect underlying all studies, an assumption that implies that studies are similar in phenotypic and technical characteristics (Borenstein *et al.*, 2010). This assumption was assessed in Section 4.2 by means of the heterogeneity measures Q and I^2 (Borenstein *et al.*, 2010, Higgins and Thompson, 2002). Sufficient homogeneity was observed for all meta-analyses (after genomic control), so fixed-effects meta-analyses were used. All meta-analyses were conducted using METAL, version 2011-03-25 (`http://www.sph.umich.edu/csg/abecasis/Metal/`).

For the epigenome-wide discovery step in Section 4.2, *genomic control* was applied to both p-values of the individual studies and the meta-analysis to account for population structure within and between the studies. Genomic control was developed for genome-wide association testing of SNPs to control for an inflated magnitude and variability of test statistics (associated with an increased number of false positives), termed *genomic inflation*, that results e.g. from the presence of population stratification (Devlin *et al.*, 2001). The genomic inflation factor λ is derived as

$$\lambda = \underset{j=1,\ldots,p}{median}\left(\frac{1-\chi_1^2(p_j)}{1-\chi_1^2(0.5)}\right),$$

where p_j denotes the individual p-values across p methylation sites, and χ_1^2 the quantile of the χ^2 distribution with one degree of freedom. λ is a robust estimate for genomic inflation, since – assuming that the majority of sites does not associate with the phenotype

of interest – the p-value distribution should be approximately uniform on [0,1] and thus centered around 0.5. Thus $\lambda > 1$ indicates deflation of p-values, i.e. inflation of test statistics. If $\lambda > 1$, genomic control can be achieved through

$$p_{j,GC} = P\left(X^2 > \frac{\chi_1^2(p_j)}{\lambda}\right),$$

where $p_{j,GC}$ are the corrected p-values and X^2 a random χ_1^2 distributed variable.

Besides meta-analysis, it has become common practice in genome-wide studies to confirm the results obtained from an initial meta-analysis in a *replication* stage, i.e. to perform external validation of the identified associations. The need for validation has arisen from the observation that several initially reported GWAS findings failed to be replicated, and that inconsistency of effect sizes between studies led to the overestimation of effect sizes in most initial investigations ("winner's curse" phenomenon) (Ioannidis, 2007). In Section 4.2, a two-stage approach was applied, where significant findings from the discovery stage where put forward to replication, and associations reaching $p < 0.05$ in the replication stage and $p < 10^{-7}$ in a joint meta-analysis of both discovery and replication cohorts where declared significant.

3.4 Multivariate data analysis

In the context of this thesis, the term *multivariate analysis* is used to refer to statistical approaches that simultaneously model a large number of molecular features. *Unsupervised* and *supervised* multivariate strategies can be distinguished. Unsupervised strategies are aimed at describing the relations among the features, without considering the phenotype, and to potentially reduce dimension of the feature matrix, whereas supervised strategies intend to describe the relation of the features with the phenotype. More specifically, their aim is to predict the phenotype (as the response) based on the features (Hastie *et al.*, 2009).

3.4.1 Unsupervised statistical approaches

Principal component analysis (PCA)

PCA aims to reduce dimensionality by transforming the p features into orthogonal *principal components* (PCs) resembling linear combinations of the features and successively explaining maximum variance in the data (Jolliffe, 2002) (See Appendix A.1.7 for details). Depending on the data, the first few principal components explain a large proportion of the variance, so dimension is reduced by focusing on the first few principal components without great loss of information. PCA was frequently applied in this thesis, e.g. within the context of adjustment for technical confounding via control probe PCs (see Section 3.1.2 for method, Section 4.2 for application), for cell type confounding (see Section 3.3.2

for problem, Section 4.3 for application), and within the context of WGCNA to summarize the signal of feature modules (see below for method, Section 4.3 for application).

Cluster analysis

To group features according to their similarity among the observations, *cluster analysis* was applied in different applications in this thesis.

As a prerequisite, the similarity, or dissimilarity, between features needs to be defined. Frequently chosen definitions are the *Euclidean distance*, and dissimilarity measures based on correlations between features. The Euclidean distance between two $n \times 1$ feature vectors $\boldsymbol{x}_j$ and $\boldsymbol{x}_l$ is defined as

$$d(\boldsymbol{x}_j, \boldsymbol{x}_l) = ||\boldsymbol{x}_j - \boldsymbol{x}_l|| = \sqrt{\sum_{i=1}^{n} |\boldsymbol{x}_{ij} - \boldsymbol{x}il|^2}.$$

It can be extended to cluster time trajectories of features $\boldsymbol{x}_{j..}$ (Genolini *et al.*, 2013), i.e. $n \times T$ matrices of concentrations of features j and l, observed for n subjects at T time points:

$$d(\boldsymbol{x}_{j..}, \boldsymbol{x}_{l..}) = \sqrt{\sum_{i=1}^{n} \sum_{t=1}^{T} |\boldsymbol{x}_{ijt} - \boldsymbol{x}_{ilt}|^2}.$$

Furthermore, correlation-based dissimilarity measures can be defined that do not require the features to be on the same scale, e.g.:

$$d(\boldsymbol{x}_j, \boldsymbol{x}_l) = \frac{1 - \text{cor}\,(\boldsymbol{x}_j, \boldsymbol{x}_l)}{2} = 1 - \frac{1 + \text{cor}\,(\boldsymbol{x}_j, \boldsymbol{x}_l)}{2} = 1 - s_{jl}, \tag{3.1}$$

where $\text{cor}(\cdot)$ corresponds to Pearson's correlation coefficient. Within the context of WGCNA (see below), another improved correlation-based dissimilarity measure is proposed (Zhang and Horvath, 2005).

Two clustering concepts can be distinguished, which were both applied in this thesis: *K-means clustering* and *hierarchical clustering* (Hastie *et al.*, 2009) (see Appendix A.1.8 for definition). The former aims to assign all features to a pre-specified number of K clusters. The latter does not achieve a hard grouping of features into a pre-specified number of clusters, but rather generates a hierarchical tree of features based on similarity, which can potentially be followed by tree cutting to obtain defined clusters.

In Section 4.1, K-means clustering with the longitudinal Euclidean distance measure was applied in order to cluster 132 investigated metabolites according to similar challenge-induced concentration changes rather than similar concentrations *per se*. The implementation in the R package *kml3d*, version 2.1, was used (Genolini *et al.*, 2013). Prior to clustering, log-transformed metabolite concentrations were mean-centered and scaled

across all time points, and differences between adjacent time points within each challenge (IVGTT/EH clamp, OGTT, OLTT, and HFHC meal, see Section 2.2.3) were used as basis for the clustering, yielding $T = 8$ challenge responses. In *kml3d*, missing values were imputed using the *copyMean* method described by Genolini and Falissard (2011). To choose the optimal number of clusters, three criteria were considered, all of which are measures of the similarity within the clusters in relation to the similarity between clusters: the *Calinski & Harabasz*, the *Ray & Turi* and the *Davies & Bouldin* criteria (Appendix A.1.8). Since superiority of any of these criteria above the others is controversially discussed, the three criteria were fused (Kryszczuk and Hurley, 2010). To make them comparable, negative values of the Ray & Turi and Davies & Bouldin criteria were used and all criteria were min–max normalized to a $[0, 1]$ range. Kryszczuk and Hurley (2010) showed the best accuracy for a decision-level fusion method, DF-A, where the fused criterion is defined as the arithmetic mean of the optimal number of clusters according to each of the criteria. They did not, however, consider the case of repeated runs yielding multiple clustering solutions per cluster number. Therefore, two methods were applied to find the optimal cluster number, first, a modified decision-level method, where first, for each number of clusters the "best" clustering solution was selected according to the Calinsky and Harabasz criterion, subsequently, the DF-A method was applied on the resulting nine partitions and finally, the partition with the largest fused criterion was chosen as the optimal partition. As a sensitivity analysis, the score fusion-based method SF-A was applied, where the arithmetic mean of the three normalized criteria was calculated for each of the 9,000 clustering solutions and the solution with the largest arithmetic mean was selected. Both methods identified very similar optimal clustering solutions, and partitions identified by the decision-based method are reported.

Hierarchical clustering was applied in Sections 4.2 and 4.4 mainly for visualization purposes. In Section 4.2, clinical phenotypes were clustered based on their association with methylation at selected CpG sites using the R package *gplots*, version 2.13.0 (Warnes *et al.*, 2014). The Euclidean distance was used to define distance between pairs of traits, and cluster distance was defined through complete linkage (see Appendix A.1.8). In Section 4.4, a matrix of pairwise Pearson's correlation coefficients of selected metabolites was subjected to agglomerative hierarchical clustering using the R package *Heatplus*, version 2.1.0 (Ploner, 2011). Distance between pairs of metabolites was defined as the Pearson's correlation-based distance, and cluster distance was defined through complete linkage.

Weighted gene co-expression network analysis (WGCNA)

Within the context of gene expression data, a framework for clustering features, and potentially visualize them as a correlation network, has been introduced by Zhang and Horvath (2005) and termed *weighted gene co-expression network analysis* or *weighted correlation network analysis* (WGCNA). Although the method was developed for gene expression

data, it has been successfully applied to metabolomics data (see e.g., Zhang *et al.* (2013)).

Zhang and Horvath (2005) propose to define dissimilarity between two features $\boldsymbol{x}_j$ and $\boldsymbol{x}_l$ using the *topological overlap dissimilarity measure*, i.e. a measure of the interconnectedness of features:

$$d_{jl}^{\omega} = 1 - \omega_{jl} = 1 - \frac{\sum_{u=1,\ldots,p} a_{ju} a_{ul} + a_{jl}}{\min(c_j, c_l) + 1 - a_{jl}},$$

where a_{jl} is the adjacency between $\boldsymbol{x}_j$ and $\boldsymbol{x}_l$ defined as

$$a_{jl} = |s_{jl}|^{\lambda} = |\frac{1 + \text{cor}\,(\boldsymbol{x}_j, \boldsymbol{x}_l)}{2}|^{\lambda},$$

with s_{jl} the correlation-based similarity as described in Equation 3.1, and c_j the connectivity of feature j defined as

$$c_j = \sum_{u=1,\ldots,p} a_{ju}.$$

Clustering based on the topological overlap measure was shown in an application example to give rise to more distinct modules than the standard correlation-based dissimilarity measure (Zhang and Horvath, 2005).

It is recommended to choose the tuning parameter λ such that the *scale-free topology criterion* is satified. Scale-free topology of a network is given when the connectivity c of the features follows the power law $p(c) \sim c^{-\gamma}$, which is approximately given for most biological networks (Jeong *et al.*, 2000) and can therefore be used as a plausibility/quality criterion for constructed networks. The scale-free topology criterion implies

$$\log(p(c)) = \log(c^{-\lambda}) = -\lambda \log(c) \;\Leftrightarrow\; \text{cor}\,(\log(p(c)), \log(c)) = -1.$$

Thus, Zhang and Horvath (2005) propose to choose λ such that the signed squared correlation

$$\text{signed } R^2 = (-1) \cdot sign\left(\frac{\log(p(c))}{\log(c)}\right) \text{cor}^2\,(\log(p(c)), \log(c))$$

is close to 1. In practice, a trade-off exists between scale-free topology and mean connectivity: the larger λ, the larger R^2 tends to be (although no monotonic relationship exists), but the smaller becomes the mean connectivity of the features, which is undesirable for cluster formation (Zhang and Horvath, 2005). Thus, it is proposed to chose the smallest λ meeting a minimum R^2 threshold such as 0.85 (see Figure 3.8 for the application in this thesis).

The matrix $\left(d_{jl}^{\omega}\right)$ can subsequently be subjected to e.g., hierarchical clustering. Clear-cut clusters (also referred to as network modules) can be simply obtained by cutting the tree at a specified height. Alternatively, a dynamic tree cutting algorithm can be used (Langfelder *et al.*, 2008).

To obtain a representative summary signal from each module, Horvath and Dong (2008)

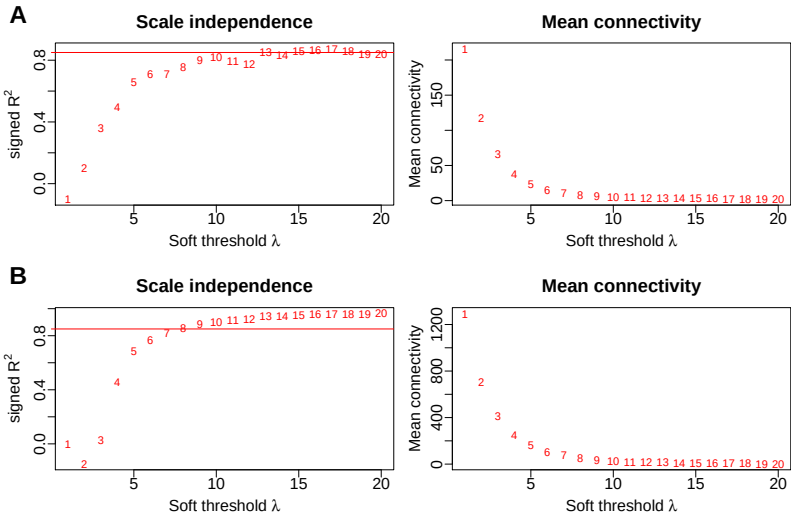

Figure 3.8: Choice of the scale-free topology parameter λ in weighted correlation network analysis (WGCNA). A Metabolomics data. **B** Gene expression data. The smallest λ (13 and 8, respectively) with $R^2 \geq 0.85$ was chosen.

show that a sensible measure is the first principal component of a PCA on the scaled matrix of features of the respective module k, $\mathbf{X}^{(k)}$. This is equivalent with determining the first eigengene from a singular value decomposition of $\mathbf{X}^{(k)}$. Thus, $\mathbf{ME}_k$ is used to denote the *module eigengene* (ME) of module k (Langfelder and Horvath, 2008). Subsequently, a measure of *module membership strength* can be derived for each feature of a module k as

$$c_j^{(k)} = \mathrm{cor}\left(\boldsymbol{x}_j, \mathbf{ME}_k\right),$$

and *intramodular connectivity* can subsequently be defined as the average $c_j^{(k)}$ across all module members j.

In Section 4.3, WGCNA was performed to cluster metabolites and transcripts, using the R package *WGCNA*, version 1.34 (Langfelder and Horvath, 2008). The 411 identified metabolites (281 from the Metabolon platform and 130 from the NMR platform) were jointly subjected to WGCNA. For WGCNA on transcriptomics data, the intention was to keep the focus on genes relevant for blood metabolism. Therefore, transcripts were preselected prior to clustering based on their association with metabolite concentrations. Prior to modeling, $\log_2$-transformed transcriptomics data were adjusted for the three technical variables RNA integrity number, amplification plate indicator as well as sample storage time (see Section 3.1.3 and Schurmann *et al.* (2012)). This was achieved by modeling transcript levels as a function of these variables to obtain the model residuals for use as "adjusted transcript levels" in subsequent analyses. Association between transcripts and metabolites was then determined using univariate linear models with transformed metabolite as response (see Section 3.2), and adjusted transcript as covariate, adjusting for age and sex, and a linear model additionally adjusted for BW and ΔBW to avoid the selection of transcripts related to metabolites due to their common association with these variables (see Section 3.3.2 on confounders and colliders). 2537 transcripts with at least a suggestive association ($p < 10^{-5}$) in both models where selected for WGCNA.

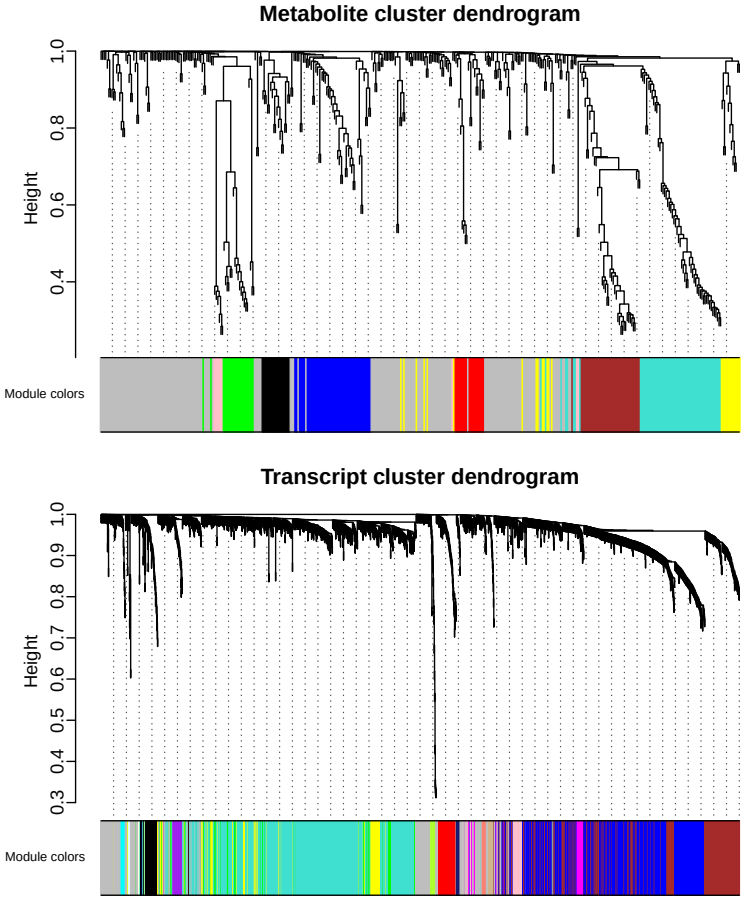

Figure 3.9: Weighted correlation network analysis (WGCNA) cluster dendrograms. Metabolomics data and gene expression data were clustered into 8 and 19 modules, respectively. Grey color represents features showing a low connectivity that are not assigned to a cluster.

Soft-thresholding powers of $\lambda = 13$ (metabolite network) and 8 (transcript network) were chosen (Figure 3.8). Figure 3.9 visualizes the cluster dendrograms derived from hierarchical clustering of the features based on topological overlap dissimilarity, with distance between clusters defined through average linkage (Appendix A.1.8). Modules obtained through dynamic tree cutting, followed by merging closely correlated modules at a dendrogram height of 0.25, are visualized as colors.

Relationships between modules and phenotypes/disease were determined by modeling the association of the ME's with external information. Relationships among modules, i.e. *inter-module connectivity* (within and between the metabolite and gene expression networks), were determined through Pearson's correlation between the respective ME's. The effect of ΔBW status (i.e., weight gain versus weight loss) on inter-module connectivity were studied using permutation testing (Section 3.3.3), defining the test statistic as difference in inter-module connectivity, and shuffling weight change status randomly 10^5 times. Similarly, the effect on intra-module connectivity was investigated.

3.4.2 Supervised statistical approaches

The $p >> n$ problem

In Section 4.4, the aim is to form a predictive model of weight loss during intervention. Since the number of features $p = 144$ exceeds the number of observations $n = 80$, and in addition, features (metabolites) show strong correlation wiht each other, standard methods to model weight loss as a function of all features, such as multiple linear regression, fail (Hastie and Tibshirani, 2004, Hastie *et al.*, 2009). Consider the definition of the least squares estimate $\hat{\boldsymbol{\beta}}$ in Appendix Equation A.2. The rank of the $n \times (p+1)$ matrix $\mathbf{X}$ is

$$rg(\mathbf{X}) = rg(\mathbf{X}^T) \leq min(n, p+1) = n,$$

and of the $(p+1) \times (p+1)$ matrix $\mathbf{X}^T\mathbf{X}$

$$rg(\mathbf{X}^T\mathbf{X}) \leq min(rg(\mathbf{X}^T), rg(\mathbf{X})) = n.$$

Thus, $\mathbf{X}^T\mathbf{X}$ does not have full rank and is therefore not invertible, so the least squares solution does not exist. Or rather, an infinite number of perfect solutions to the least squares criterion exists. In addition, if the features are strongly correlated, multicollinearity can occur even in $p < n$ scenarios. In that case, $\mathbf{X}^T\mathbf{X}$ might be close to singular and the parameter estimation becomes instable (Faraway, 2002).

Methods that cope with the $p >> n$ scenario can be categorized into approaches based on (1) explicit variable selection (univariate or multivariate), (2) dimension reduction (e.g., PC regression or *Partial Least Squares* (PLS) regression, as applied in Wahl *et al.* (2012)) and (3) methods handling a high number of variables directly (Boulesteix *et al.*, 2008). The latter include (a) regularization methods (b) ensemble methods (Hastie *et al.*, 2009). In Section 4.4, a regularized regression method, the *least absolute shrinkage and selection operator* (LASSO) was chosen, which was considered specifically elegant since it combined regularization with internal variable selection, achieving also a good interpretation of the coefficients of the selected variables as effect strengths (Hastie *et al.*, 2009). The R package *glmnet*, version 1.7.3, was used (Friedman *et al.*, 2010).

The *least absolute shrinkage and selection operator* (LASSO)

In LASSO regression, a regularization term is added to the least squares criterion (Equation A.1) that penalizes large coefficients (in absolute terms):

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \lambda \sum_{j=1}^{p} |\beta_j|,$$

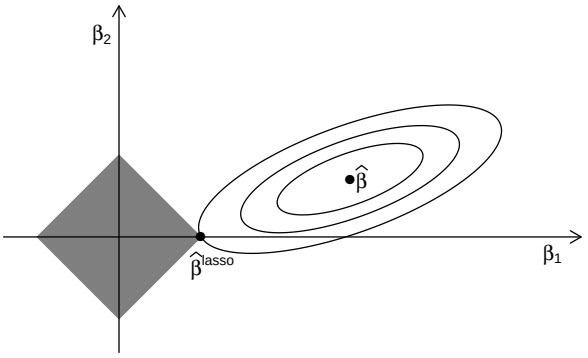

Figure 3.10: Visualization of the LASSO estimation (adopted from Tibshirani (1996)). Solid grey area, constraint region $|\beta_1|+|\beta_2|\leq t$; ellipses, contours of the residual sum of squares; $\hat{\beta}$, full least squares estimate.

where λ is the penalization parameter. The minimization problem can also be written as

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}, \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq t.$$

The penalization yields coefficient estimates shrunken towards zero, dependent on the size of λ, thereby allowing for a solution of the minimization problem even in the $p >> n$ case. There is no closed form for $\hat{\boldsymbol{\beta}}^{\text{LASSO}}$ as there is for the least squares estimate, but efficient algorithms are available for computing the coefficient paths $\hat{\boldsymbol{\beta}}^{\text{LASSO}}(\lambda)$ (Tibshirani, 1996). A specific property of LASSO is the intrinsic variable selection: The most informative features are selected into the model, whereas the coefficients of the remaining features are shrunken to exactly zero. Why this is the case might be understood when looking at the visualization of the LASSO minimization problem in the $p = 2$ scenario in Figure 3.10 (Tibshirani, 1996, Hastie *et al.*, 2009). The residual sum of squares has elliptical contours, with the full least squares estimate $\hat{\boldsymbol{\beta}}$ in the center. The constraint region resembles a diamond ($|\beta_1|+|\beta_2|\leq t$). Minimizing the residual sum of squares under this constraint means finding the first point where the elliptical contours hit the constraint region. This might occur at a corner of the constrain region, resulting in setting one parameter, here β_2, to exactly zero.

Model validation and performance assessment

To assess the predictive performance of a multivariate statistical approach, different measures are available. In the situation with a continuous response, as in Section 4.4, an appropriate measure is the equivalent to the coefficient of determination in linear regression, i.e. 1 minus the residual sum of squares divided by the total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i represents the observed response for subject i, $\hat{y}_i$ the response predicted from the model (in the case of LASSO: $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^{\text{LASSO}}$), and $\bar{y}$ the average observed response. In

metabolomics applications, the terms R^2 and Q^2 are frequently use to distinguish the measures evaluated on the data used for model fitting and on independent data, respectively (Broadhurst and Kell, 2006). Importantly, R^2 and Q^2 cannot, unlike in unregularized regression, be interpreted as the percentage of total variance of the response variable explained by the model. Still, they might serve as goodness-of fit measures with respect to the fit of the present data and to the prediction of independent data, respectively.

Model performance should not (only) be evaluated in the data involved in the model fitting, since these would (1) overestimate model performance and (2) favor complex models that fit the data at hand extremely well but perform poorly on independent data. This issue is referred to as *overfitting*. Thus, independent data are required. This applies to two steps during model building and validation:

- **Step 1: Appropriate tuning of the model hyperparameter(s):** The tuning parameter(s) of a model (e.g., λ or t in the case of LASSO) control(s) the tradeoff between complexity/variance and sparseness/smoothness/bias. Strong penalization of the coefficient estimates (e.g., large values of λ; small values of t) yields a sparse model (with few selected features) that might underfit the data and have a large bias (e.g., a smaller accuracy in the data set). On the other hand, weak penalization yields a complex model that might overfit the data and have a larger variance. For more details on the bias-variance tradeoff see Hastie *et al.* (2009). During parameter tuning, the hyperparameter(s) is/are chosen that optimize(s) model performance (with regard to a certain criterion such as Q^2) in independent data.
- **Step 2: Model evaluation:** The data on which the model is evaluated must be completely blind to all previous model fitting and validation steps in order to obtain an unbiased estimate of predictive model performance (Dupuy and Simon, 2007, Varma and Simon, 2006).

When large data sets are available, data might be divided into a *training data set*, to which the model is fitted, a *validation data set* in which model performance is determined according to different hyperparameter values chosen during model fit, and a *test data set*, in which the final model is evaluated (Hastie *et al.*, 2009). Very often, as in Section 4.4, large data sets are not available for omics data, since measurement is costly, and biosamples might be rare. Thus, strategies for efficient data re-use might be applied (Hastie *et al.*, 2009, Boulesteix *et al.*, 2008). These include *cross-validation* (CV), where (1) the data set is split into k non-overlapping folds (e.g., $k = 10$ might be chosen, Ambroise and McLachlan (2002)), (2) model fitting is iteratively performed in all folds but one left-out fold, and (3) model performance is then assessed in this left-out fold, which serves as validation data in this step. Finally, model performance is averaged across all folds.

In Section 4.4, CV was used for both parameter tuning (step 1) and model evaluation (step 2) (Figure 3.11). This was achieved by using a nested CV procedure, where the

hyperparameter λ was tuned in the inner CV loop and predictive performance estimated in the outer CV loop (Varma and Simon, 2006). The procedure was repeated 10 times to increase stability (Braga-Neto and Dougherty, 2004). Finally, the significance of the estimated model performance was assessed using permutation tests (Radmacher *et al.*, 2002) with 10,000 permutations (see Section 3.3.3 for permutation tests).

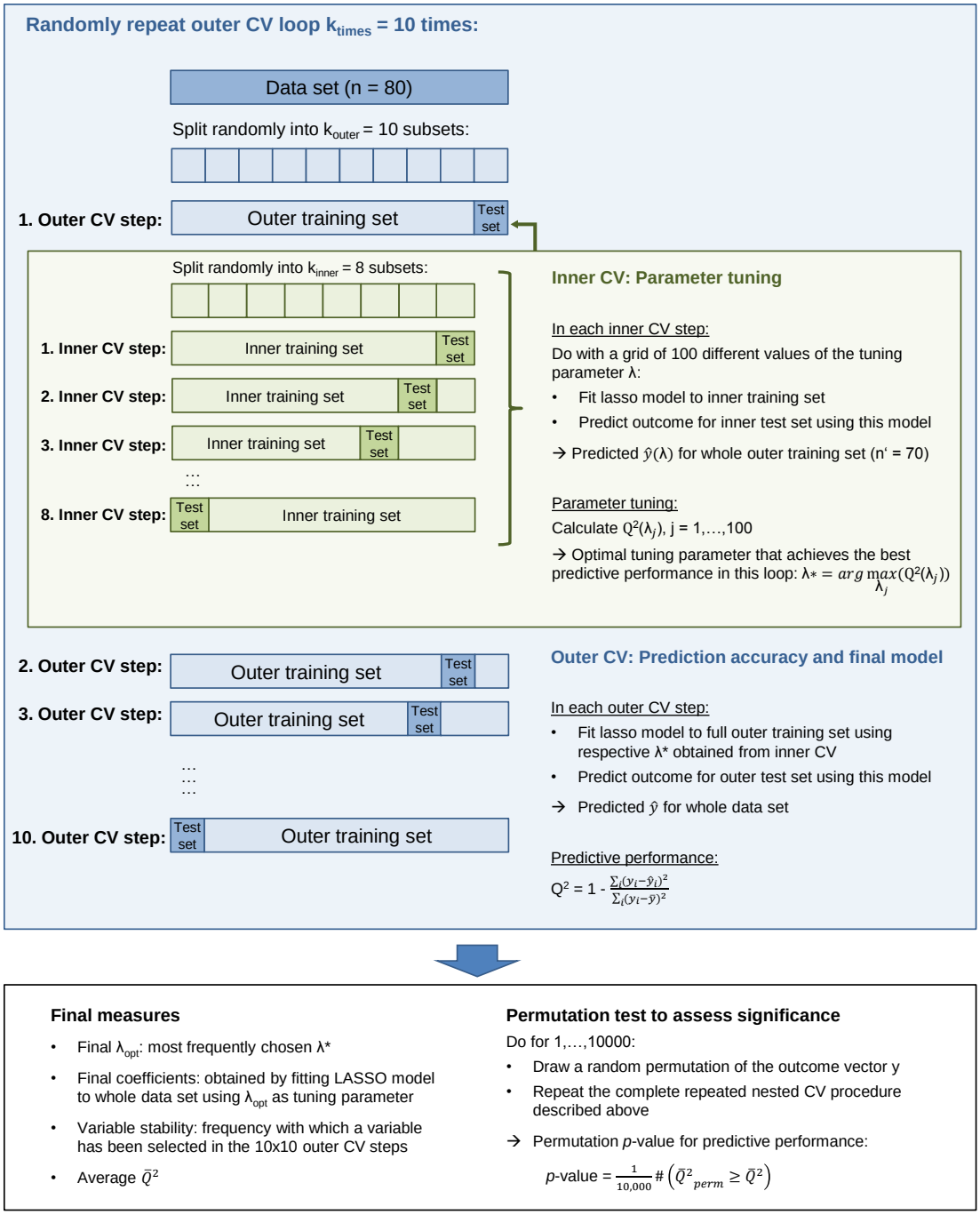

Figure 3.11: Repeated nested cross-validation and permutation scheme.

3.5 Extracting biological knowledge and integrating omics data

3.5.1 Enrichment analysis

It is plausible to believe that if a phenotype, disease or state is associated with changes in biological processes, then this should be reflected by a change in more than one molecular feature (although this depends on the coverage of the omics data). Thus, potentially long lists of association signals might be analyzed for *enrichment/overrepresentation* of features in known biological pathways (external knowledge), in clusters of features defined through data-driven methods such as cluster analysis (Section 3.4.1) or graphical models (Section 3.5.3), in genomic locations (e.g., location to CpG islands, location in/near genes, or location with regard to functional features) or in sites previously associated with some other trait.

A simple test of enrichment is *Fisher's exact test* (Fisher, 1922). Imagine that of p features, the numbers of significant/non-significant features, and of features belonging/not belonging to a certain pathway are given in Table 3.1.

Table 3.1: Fisher's exact test for pathway enrichment.

	# Signif. associated features	**# Not associated features**	**Total**
# Features in pathway	k	$R-k$	R
# Features not in pathway	$C-k$	$p-C-R+k$	$p-R$
Total	C	$p-C$	p

Then, the probability of observing this matrix given the row and column sums and assuming the null hypothesis "no enrichment" is given by the hypergeometric distribution:

$$P(X=k)=\frac{\left(\begin{array}{c}R\\k\end{array}\right)\left(\begin{array}{c}p-R\\C-k\end{array}\right)}{\left(\begin{array}{c}p\\C\end{array}\right)}.$$

Thus, a parametric enrichment p-value, i.e. the probability of observing a matrix "as extreme" as the one observed given the null hypothesis, can be obtained by adding the probabilities of the matrices with an even smaller probability.

This test is the basis of the majority of enrichment analyses in Section 4.2, including enrichment of the identified CpG sites in specific genomic locations and functional features provided by Illumina (Bibikova *et al.*, 2011) and downloaded from the UCSC database (Ram *et al.*, 2011), as well as pathway enrichment using the *gene set enrichment analysis* (GSEA) MSigDB platform (`http://www.broadinstitute.org/gsea/msigdb`). Fisher's

exact test also underlies the core analysis module of the commercial *Ingenuity Pathway Analysis* (IPA) software tool, applied in Section 4.3, where a test for enrichment of genes in biological pathways from the Ingenuity Konwledge Base is performed (see `http://www.ingenuity.com`). Analyses were conduced with IPA build version 308606M; content version 18488943; release date 2014-03-23. The reference set of features was defined as genes represented on the Illumina HumanHT-12 v3 BeadChip, and only human annotations were considered. In case multiple probes mapped to one gene, the probe exhibiting the largest module membership was considered. Finally, Fisher' exact test is the basis of the enrichment analysis for gene ontology (GO) terms performed in Section 4.3 using the R packages *GO.db*, version 2.9.0, *AnnotationDbi*, version 1.22.6, and *org.Hs.eg.db*, version 2.9.0.

Fisher's exact test is based on a significance threshold that dichotomizes features into "associated" versus "not associated". Tests taking into account the ranking or the test statistics of the features might be more powerful. An example is *weighted enrichment analysis* as applied by Krumsiek *et al.* (2012b). For each cluster (or pathway) c an enrichment statistic S_c is defined as the weighted sum of (absolute) test statistics across all features belonging to that cluster. A permutation test can be conducted to determine significance by randomly permuting assignment of the features to the cluster, determining $S_c^{(b)}$ in each permutation b, $b = 1, 2, \ldots, B$ and obtaining a p-value as $\frac{1}{B}\sum_{b=1}^{B} I\left(S_c^{(b)} \geq S_c\right)$.

This procedure was applied in Section 4.1 with $B = 10^7$ permutations to determine enrichment of associations in data-driven metabolite clusters. It was further applied in Section 4.3 with $B = 10^5$ to determine enrichment of co-clustering metabolites, weighted by their module membership strengths (see Section 3.4.1) in pre-specified metabolite super- and sub-pathways. A permutation test also underlies the enrichment analysis for genes previously published in GWAS on different traits, using *Meta-Analysis Gene-set Enrichment of variaNT Associations* (MAGENTA, `http://www.broadinstitute.org/mpg/magenta`, (Segrè *et al.*, 2010)) in Section 4.2. Of note, permutation tests are more appropriate than Fisher's exact test when analyzing enrichment specifically for methylation data, where the latter might induce bias due to differences in the coverage of different genes and genomic locations by the measured CpGs. This was demonstrated for GO enrichment analysis recently (Geeleher *et al.*, 2013). Consequently, the enrichment results based on Fisher's exact test in Section 4.2 should be interpreted with care.

3.5.2 Causal inference

A major issue with the observational studies in this thesis is that it is not known, which of two associated variables is the cause and which is the consequence, or whether both variables are common effects of a (possibly unknown) confounder (Didelez and Sheehan, 2007). Longitudinal studies help to elucidate temporal relationships, which are a prerequisite but not a sufficient condition for causality (Hill, 1965). In randomized controlled

trials, subjects are randomly assigned to the treatment groups, so the relation between treatment and effect is unconfounded. However, ethical or technical reasons very often prohibit the conduction of randomized controlled trials. For instance, it is not possible to modify DNA methylation in humans in order to investigate its effect on BMI.

In an approach called *Mendelian randomization* (MR), data on genetic variation is utilized to infer causality (Smith and Ebrahim, 2003, Didelez and Sheehan, 2007, Bochud and Rousson, 2010). It relies on the fact that the alleles of a genetic variant are inherited randomly from parents to offspring. Thus, the relation of a genetic variant with a phenotype should not be confounded (neglecting e.g. the case of population stratification). The principle of MR, which goes back to the more general concept of *instrumental variable estimation*, is visualized in Figure 3.12, where X and Y correspond to a molecular feature - phenotype pair for which an association has been observed, and Z corresponds to the *instrument(al) variable*, i.e. the genetic variant (or a score of several genetic variants) known to affect X. If the effect of X on Y is causal and the study has enough power, Z should also associate with Y. If it does not, the hypothesis of a causal relation might be rejected (Didelez and Sheehan, 2007).

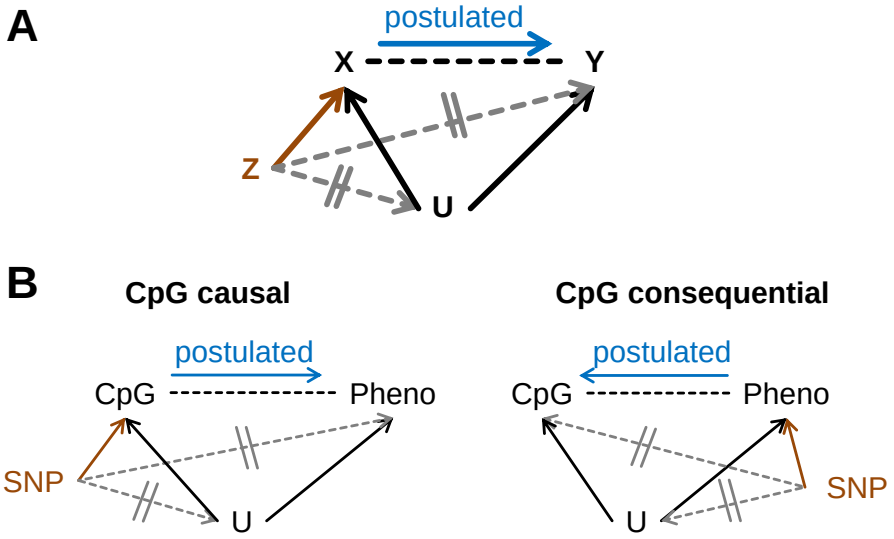

Figure 3.12: Mendelian Randomization (MR). A General setup: Z represents a genetic variant (score) that serves as an instrumental variable to estimate the causal effect of X on Y. **B** Application example of a two-directional MR approach: The causal effect of methylation at a specific CpG site on a phenotype (Pheno) can be estimated using single nucleotide polymorphisms (SNPs) known to affect methylation. Vice verse, the causal effect of the phenotype on methylation can be estimated.

An *ad hoc* MR approach

Thus, an *ad hoc* "MR" approach is to compare the observed association of Z and Y with the association predicted from the product of the observed effect sizes between Z and X as

well as X and Y. This approach was developed to explore causality of BMI-methylation associations in Section 4.2.

For the causal direction (CpG (X) causal to BMI (Y), Figure 3.12B left-hand side), for each CpG the strongest associated *cis*-SNP ($\leq$1 Mb distance) was used as instrument Z. Consequently, linear models were used to assess the associations between Z and X as well as X and Y, adjusted for the discovery covariates of this study (see Section 4.2). To increase power, SNP (Z)-BMI (Y) associations were retrieved from a large GWAS previously published by the GIANT consortium ($n >$ 100,000, Speliotes *et al.* (2010)). Effect sizes and standard errors for the predicted association between Z and Y were derived as follows:

$$\begin{aligned} \beta_{\text{pred}} &= \beta_{ZX} \cdot \beta_{XY}, \text{ and} \\ \text{SE}_{\text{pred}} &= \sqrt{\text{SE}_{ZX}^2 \cdot \text{SE}_{XY}^2 + \text{SE}_{ZX}^2 \cdot \beta_{XY}^2 + \text{SE}_{XY}^2 \cdot \beta_{ZX}^2}. \end{aligned}$$

Inverse normal transformed BMI was used to make results consistent with those of Speliotes *et al.* (2010).

For the consequential direction (CpG (Y) consequential to BMI (X), Figure 3.12B right-hand side), Z was defined as *genomic risk score* (GRS) $Z = \sum_k w_k Z_k$, $k = 1, \ldots, 31$, i.e. the sum of expected risk alleles Z_k from the 31 SNPs previously reported to associate with BMI (Speliotes *et al.*, 2010) (32 less one SNP, rs7359397, which showed an association with one of the CpGs independent of BMI), weighted with w_k, the effect sizes derived from the same publication. Both observed and predicted Z-Y associations were determined from the available data, similar as described for the causal direction above.

Evidence for causality was declared when the observed effect was significant and had the same direction as the predicted effect. This approach does not attempt to assess the conditions underlying MR, and does not provide an actual estimation of the causal effects. Therefore, it was complemented by a more formal MR approach.

A formal MR approach

In a more formal MR approach, the causal (part of the) effect of X on Y was estimated using a *two-stage least squares* (TSLS) approach:

$$\begin{aligned} X &= \beta_{0X} + \beta_X Z + \epsilon_X \;\Rightarrow\; \hat{X}(Z) = \hat{\beta}_{0X} + \hat{\beta}_X Z \\ Y &= \beta_{0Y} + \beta_Y \hat{X}(Z) + \epsilon_Y. \end{aligned}$$

It can be shown that the MR estimate ($\hat{\beta}_Y$) is an approximately unbiased estimate of the causal effect (Bochud and Rousson, 2010). This approach can be extended to include several genetic variants $Z_1, Z_2, \ldots$, or a GRS (Palmer *et al.*, 2012). Also, known confounders can easily be incorporated as additional covariates. Tests of the null hypothesis $\beta_Y = 0$

can be based on the asymptotic normality of β_Y.

This MR approach is subject to several limitations. First, the MR estimate is biased in finite samples, with the relative bias (relative to the bias of the one-stage least squares estimate) being inversely proportional to the coefficient of determination R^2 of the step 1 model: $\frac{\text{bias(MR)}}{\text{bias(LS)}} \approx \frac{K}{nR^2}$, where n is the sample size and K the number of instruments. Thus, if Z explains only little of the variance in X, bias is introduced. A rule of thumb is that an F statistic of less than 10 indicates a *weak instrument*, since $F = \frac{R^2/K}{(1-R^2)/(n-K-1)}$ is approximately inversely proportional to $\frac{\text{bias(MR)}}{\text{bias(LS)}}$ and thus, relative bias is approximately 10% when $F = 10$ (Bound *et al.*, 1995, Staiger and Stock, 1997, Palmer *et al.*, 2012). Since SNP effects are typically small for common diseases and phenotypes, resulting in low R^2 values, MR studies intending to estimate the association between a phenotype X and a molecular feature Y require very large sample sizes to avoid weak instrument bias.

Second, instrumental variables need to meet three core conditions (Didelez and Sheehan, 2007):

(1) $Z \not\perp X$, i.e. Z must be associated with X,

(2) $Z \perp Y|(X,U)$, i.e. Y must be independent from Z conditionally on X and any (un)observed confounders U, that is, Y is only affected by Z through X, and

(3) $Z \perp U$, i.e. Z must be independent of U.

Conditions (2) and (3) require that no genetic variants that are in *linkage disequilibrium* with Z are associated with Y and U, respectively, that no *population stratification* is present, and that the genetic variant does not have *pleiotropic* effects, i.e. effects on several phenotypes associated with Y or U. More conditions and assumptions are discussed by Didelez and Sheehan (2007) and Bochud and Rousson (2010). Note that conditions (2) and (3) are not possible to be tested directly, due to the fact that U is unknown.

Application of the TSLS approach to methylation data

In Section 4.2, TSLS was applied to determine the causal part of the observed methylation-BMI associations. Within the EWAS discovery cohorts, SNP data from five (sub-)cohorts (three EpiMigrant cohorts based on three different genotyping platforms, KORA F4 and KORA F3) were available.

Prior to MR, genetic confounding, i.e. confounding of the methylation-BMI association by known BMI SNPs (Speliotes *et al.*, 2010), was explored using linear models with and without adjustment for each SNP. None of the associations were subject to genetic confouding. Next, for the causal direction (CpG (X) causal to BMI (Y), Figure 3.12B left-hand side), the same SNPs were used as described for the *ad hoc* approach above. Prior to TSLS, the assumption $Z \perp Y|X$ was tested as a surrogate for $Z \perp Y|(X,U)$ using an equivalence test

with the null hypothesis "no independence" ($Z \not\perp Y|X$) as proposed by (Millstein *et al.*, 2009). Precisely, the following algorithm was used with $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ representing vectors of observed values of Z, X, and Y, and $\mathbf{C}$ being a matrix of covariates:

1. Fit the model $\boldsymbol{x} = \beta_{0x} + \boldsymbol{z}\beta_{zx} + \mathbf{C}\boldsymbol{\beta}_{Cx} + \boldsymbol{\epsilon}_x$, permute $\boldsymbol{\epsilon}_x$ randomly to obtain $\boldsymbol{\epsilon}_x^{(b)}$, $b = 1, \ldots, B$, then define $\boldsymbol{x}^{(b)} = \hat{\beta}_{0x} + \boldsymbol{z}\hat{\beta}_{zx} + \mathbf{C}\hat{\boldsymbol{\beta}}_{Cx} + \boldsymbol{\epsilon}_x^{(b)}$ using the estimated parameters.

2. For $b = 1, \ldots, B$, fit the model $\boldsymbol{y} = \beta_{0y} + \boldsymbol{x}^{(b)}\beta_{xy} + \boldsymbol{z}\beta_{zy} + \mathbf{C}\boldsymbol{\beta}_{Cy} + \boldsymbol{\epsilon}_y$ and compute the F statistic $F^{(b)}$ for the test with H_0: $\beta_{zy} = 0$.

3. Obtain degrees of freedom df_1 and df_2 and compute non-centrality parameter $\lambda = \frac{F^{(b)}\mathrm{df}_1(\mathrm{df}_2-2)}{\mathrm{df}_2} - \mathrm{df}_1$, then transform $F^{(b)}$ to normally distributed $U^{(b)} = \psi^{-1}\left(P(F < F^{(b)})\right)$, where F represents an $F_{\mathrm{df}_1,\mathrm{df}_2,\lambda}$ distributed random variable. Then, $U^{(b)}$ is normally distributed with mean 0 and standard deviation derived empirically from the distribution of the $U^{(b)}$, $b = 1, \ldots, B$.

4. Determine p-value for H_0: β_{SNP} (β_{zy}) $\neq 0$ as the probability of observing U given its null distribution derived above.

Results from the five cohorts were meta-analyzed using a z-score based fixed-effects meta-analysis. 69 CpGs had a combined p-value below 0.05/184 (Bonferroni) and were subsequently submitted to MR analysis.

To avoid weak instrument bias (see above, Palmer *et al.* (2012)), in each single cohort only CpGs with an F statistic above 10 for the test for the SNP-CpG association were submitted to TSLS. Since single cohorts were small, $F < 10$ was frequently observed and only 52 CpGs could be analyzed in TSLS in at least one cohort. TSLS was performed using the function *ivreg* from the R package *AER*, version 1.2-2 (Kleiber and Zeileis, 2008). Were results from more than one cohort were available, they were meta-analyzed using a z-score based fixed-effects meta-analysis.

For the consequential direction (CpG (Y) consequential to BMI (X), Figure 3.12B right-hand side), Z was defined as GRS as described above for the *ad hoc* approach, with the difference that only SNPs with a significant result in the equivalence test (H_0: β_{SNP} (β_{zy}) $\neq 0$ rejected) were included in the GRS. In addition, only cohorts with $F > 10$ were considered. For 134 of the 187 CpGs, at least one study met instrument strength requirements.

3.5.3 Graphical models

To understand the interrelationship of feature modules defined through WGCNA in Section 4.3, *Gaussian graphical models* were applied.

In Gaussian graphical models, pairwise *partial* correlations between the features (here: module eigengenes, see Section 3.4.1) are visualized as connective edges between nodes

representing the features. The partial correlation coefficient between two variables X_j and X_k is defined as their correlation, conditioned on all other variables. It can be computed as the standard Pearson correlation coefficient between the residuals of the regression of X_j on the remaining variables $X_{-(j,k)}$ and the residuals of X_k on $X_{-(j,k)}$. In case of a full-ranked Pearson correlation matrix $\mathbf{\Sigma}$, the matrix $\mathbf{Z}$ of partial correlations ζ_{jk} can be derived in a single matrix inversion step (Lauritzen, 1996, Krumsiek *et al.*, 2011):

$$\mathbf{Z} = \left(\zeta_{jk}\right) = \left(-\frac{\omega_{jk}}{\sqrt{\omega_{jj}\omega_{kk}}}\right), \text{ with } \left(\omega_{jk}\right) = \mathbf{\Sigma}^{-1}.$$

Krumsiek *et al.* (2011) and Krumsiek *et al.* (2012a) provide solid proof that graphical models based on partial correlations of metabolite concentrations indeed recover metabolic pathway reactions.

Sex, age, body weight and previous weight change (ΔBW) were included as covariates. Since multicollinearity among the MEs might result in spurious negative partial correlations (see *collider* problem described in Section 3.3.2), pairwise marginal correlation (i.e. Pearson's correlation, uncorrected for any other variables) was required to have the same sign for a network edge to be drawn, and partial correlation was cut at the magnitude of marginal correlation prior to network formation.

3.6 Software

The majority of statistical analyses was performed using R (R Core Team, 2013). Specifically, versions 2.14.2 and 2.15.1 were used to preprocess and analyze the VID data (Section 4.1), version 3.0.1 was used to preprocess and analyze the DNA methylation data in Section 4.2 and the metabolomics and transcriptomics data in Section 4.3. Version 2.14.2 was used to analyze the Obeldicks data (Section 4.4).

4 Results and Discussion

4.1 Characterization of obesity and type 2 diabetes risk loci by metabolic challenge tests

As described in Section 1.2.1, obesity has a large heritable component. The *fat mass and obesity associated (FTO)* gene constitutes the strongest obesity risk locus identified so far. It accounts for approximately 25% of the variability in BMI explained by the variants identified in the most recent GWAS (Speliotes *et al.*, 2010). For type 2 diabetes (T2D), the strongest genetic risk in Caucasians is conferred by variants in the *transcription factor 7-like 2 (TCF7L2)* locus (Morris *et al.*, 2012).

The mechanisms behind the increased disease risk associated with *FTO* and *TCF7L2* remain largely unknown. For *TCF7L2* SNPs, an effect on insulin secretion was observed (Lyssenko *et al.*, 2007), whereas *FTO* might act through an effect on food intake (Tung and Yeo, 2011). Metabolic challenge tests show promise in revealing early metabolic dysregulation associated with risk genotypes that are not observed in the fasting state (see Section 1.2.4, van Ommen *et al.* (2009)). They might give rise to new hypotheses concerning the mechanisms underlying disease risk associated with genetic variants.

The aim of this study was to provide a comprehensive comparison of the plasma metabolomics response to five different challenges in healthy men (Figure 4.1). This was then used as a basis for studying the feasibility of these challenge tests for the identification of genotype-challenge interactions, using the examples of *FTO* and *TCF7L2*.

The results reported in this section are part of the publications

- **Wahl S**[*], Krug S[*], Then C[*], Kirchhofer A, Kastenmüller G, Brand T, Skurk T, Claussnitzer M, Huth C, Heier M, Meisinger C, Peters A, Thorand B, Gieger C, Prehn C, Römisch-Margl W, Adamski J, Suhre K, Illig T, Grallert H, Laumen H, Seissler J, Hauner H (2014). "Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele." *Metabolomics*, **10**(3), 386-401.

- Then C[*], **Wahl S***, Kirchhofer A, Grallert H, Krug S, Kastenmüller G, Römisch-Margl W, Claussnitzer M, Illig T, Heier M, Meisinger C, Adamski J, Thorand B, Huth C,

[*,#] contributed equally

Peters A, Prehn C, Heukamp I, Laumen H, Lechner A, Hauner H, Seissler J (2013). "Plasma metabolomics reveal alterations of sphingo- and glycerophospholipid levels in non-diabetic carriers of the Transcription Factor 7-Like 2 polymorphism rs7903146." *PLoS One*, **8**(10), e78430.

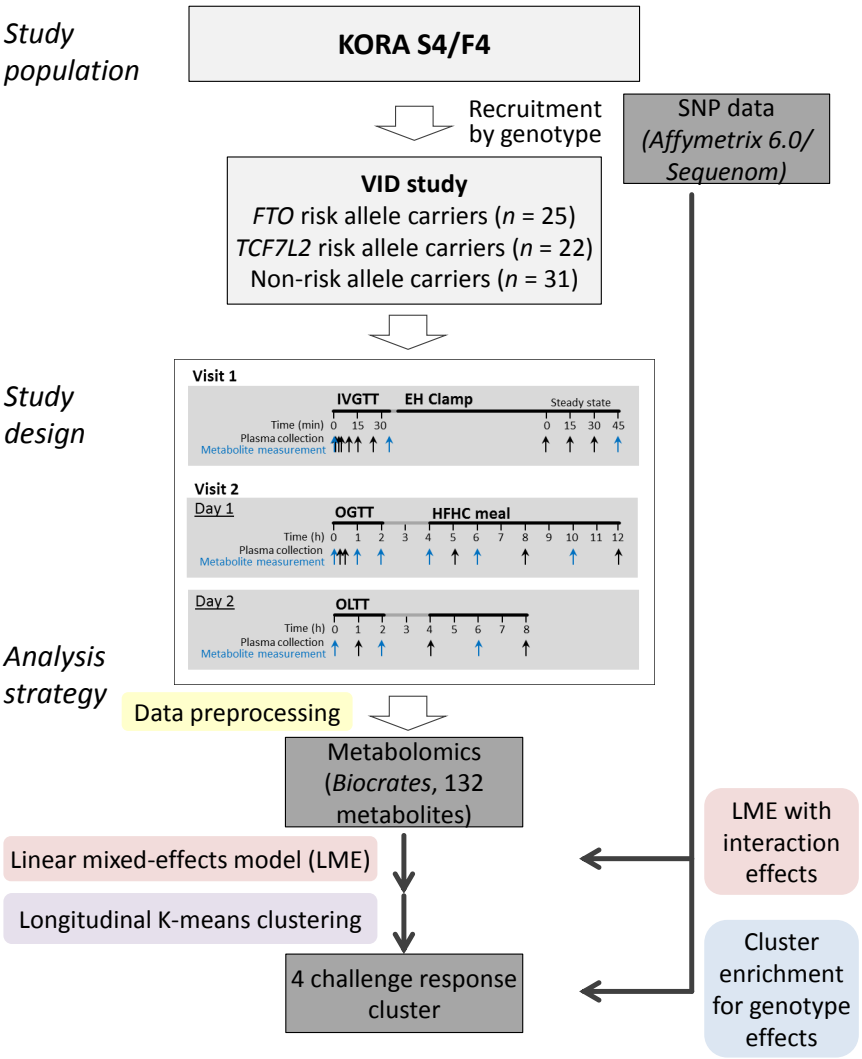

Figure 4.1: Virtual institute diabetes (VID): Study design and analysis strategy. Details on recruitment and challenge tests are given in Section 2.2. Color coding of statistical methods: yellow, data preprocessing and quality control (Section 3.1); red, univariate data analysis (Section 3.3); violet, multivariate data analysis (Section 3.4); blue, extraction of biological knowledge (Section 3.5). EH clamp, euglycemic-hyperinsulinemic clamp; *FTO*, *fat mass and obesity associated*; HFHC meal, high-fat high-carbohydrate meal; IVGTT, intravenous glucose tolerance test; LME, linear mixed-effects model; OGTT, oral glucose tolerance test; OLTT, oral lipid tolerance test; *TCF7L2*, *transcription factor 7-like 2*.

4.1.1 Metabolic challenge response

The study population comprised 25 *FTO* risk allele carriers, 22 *TCF7L2* risk allele carriers and 31 subjects carrying none of the risk alleles (Figure 4.1). After an initial screen for genotype effects, *FTO* effects were found to be marginal. Slightly stronger *TCF7L2* effects were observed (see below). Therefore, the 31 non-risk allele carriers and the 25 *FTO* risk allele carriers, but not the 22 *TCF7L2* risk allele carriers were included in the general investigation of metabolic challenge responses. Together, metabolite challenge responses were explored in a total sample of 56 subjects aged 24-66 years (see Table 4.1 for baseline characteristics). Of these, 48 took part in the intravenous challenge tests and 39 in the oral challenge tests, with an overlap of 31 participating in both.

Table 4.1: Baseline characteristics of the VID study population, overall and per *FTO* genotype.

Variable	Overall (n=56)	*FTO* risk allele carriers (n = 25)	Non-carriers (n = 31)	p-value
Age (years)	50.5 (10.3)	50.1 (10.2)	50.9 (10.6)	0.832
BMI (kg/m^2)	26.7 (2.9)	26.3 (3.3)	27.1 (2.4)	0.309
Waist circumference (cm)	97.9 (8.9)	96.9 (10.3)	98.7 (7.8)	0.940
Systolic blood pressure (mmHg)	136.0 (18.9)	137.0 (18.1)	135.3 (19.8)	0.742
Diastolic blood pressure (mmHg)	81.4 (11.4)	81.3 (10.6)	81.5 (12.2)	0.641
Total cholesterol (mg/dl)	207.1 (36.3)	203.0 (35.2)	210.4 (37.5)	0.723
LDL cholesterol (mg/dl)	124.3 (35.2)	117.9 (30.1)	129.4 (38.5)	0.540
HDL cholesterol (mg/dl)	55.7 (16.2)	60.3 (18.3)	52.0 (13.4)	0.152
Triglycerides (mg/dl)	131.1 (62.3)	121.8 (49.6)	138.5 (70.9)	0.461
Non-esterified fatty acids (mmol/l)[a]	0.50 (0.27)	0.47 (0.04)	0.51 (0.29)	0.513
Insulin (mU/l)	78.4 (41.3)	92.1 (27.9)	67.4 (47.2)	1.000
Glucose (mg/dl)	24.3 (33.3)	12.0 (20.6)	34.1 (38.4)	0.359
HbA1c (%)	5.6 (0.26)	5.6 (0.27)	5.5 (0.25)	0.271
Insulin sensitivity index[b]	0.11 (0.05)	0.11 (0.05)	0.11 (0.05)	0.894
Lactate (mmol/l)[a]	7.3 (2.2)	6.7 (0.14)	7.4 (2.4)	0.513
C-reactive protein (mg/dl)	0.26 (0.39)	0.18 (0.15)	0.33 (0.50)	0.452
Thyroid-stimulating hormone (μU/ml)	1.5 (0.81)	1.5 (0.86)	1.5 (0.79)	0.431

Data are shown as mean (standard deviation) and refer to the value at the first study visit. p-values were derived from ordinal regression models adjusted for BMI, age, and study center. [a] Non-esterified fatty acids and lactate measurements were available for 2 *FTO* risk allele carriers and 11 non-carriers only. [b] Insulin sensitivity index was calculated for the 24 carriers and 24 non-carriers that participated in the EH clamp challenge. BMI, body mass index; HDL, high density lipoprotein; LDL, low density lipoprotein.

Challenge responses of clinical parameters (glucose, insulin, triglycerides (TGs), non-esterified fatty acids (NEFAs) and lactate) and metabolite concentrations (132 metabolites) were investigated by means of linear mixed-effects models (LMEs) adjusted for age, BMI and genotype (see Section 3.3.1). After correction for multiple testing using the Benjamini-Hochberg procedure (see Section 3.3.4), a significant change of plasma concentrations in response to at least one of the challenges was observed for all clinical traits and metabolites (see detailed results in Supplementary Tables 2 and 3 of the original publication (Wahl *et al.*, 2013b)).

Next, K-means clustering was applied to group metabolites according to similarities in the observed metabolite challenge responses (see Section 3.4.1 for methodology). The optimal clustering solution is shown in Figure 4.2, comprising four clusters of metabolites with a similar challenge response profile.

Cluster 4, comprising only hexose, can be used to explain the visualization. The mean of standardized log-transformed hexose concentrations is shown for each time point. Because of standardization, the curve is centered around zero for each metabolite. Hexose is the only metabolite with a strong concentration increase in response to intravenous glucose tolerance test (IVGTT), followed by a decrease during euglycemic-hyperinsulinemic (EH) clamp, and with a strong increase in response to oral glucose tolerance test (OGTT), followed by a decrease at 2 h, which likely explains its separate clustering. In addition, in the figure, significant changes determined using LMEs are indicated as solid red lines. The hexose response during the different challenges compares well to that of glucose determined by an enzymatic assay (Figure 4.3A,C), validating the observed response.

Of note, the cluster approach tended to group metabolites with similar biological structures together. Whereas hexose formed a separate cluster (cluster 4), the largest cluster (cluster 1) comprised all measured phosphatidylcholines (PCs), lysophosphatidylcholines (LPCs) and sphingomyelins (SMs) as well as carnitine and the acylcarnitines C4, C5:1, C8:1, C10:2 and C18:0. For visualization, separate time course plots are shown for the biological groups within cluster 1 (Figure 4.4). Cluster 2 contained all investigated amino acids as well as the acylcarnitines C3 and C5, whereas cluster 3 comprised the remaining 14 acylcarnitines. Both the results of the LMEs and of the clustering remained stable when the analysis of metabolite levels was restricted to the subjects who participated in all challenge tests (data not shown).

OGTT

In response to the oral glucose challenge, glucose levels increased significantly with a peak at 30 min, decreasing to baseline level 2 h after the glucose load (Figure 4.3C). A correspondingly delayed increase of insulin and lactate was observed with a peak at 1 h (Figure 4.3D,G).

Using LMEs, significant concentration changes were observed for 116 metabolites in the first hour and for 29 metabolites in the second hour (Figure 4.2). Grouping of metabolites by K-means clustering showed that the majority of metabolites decreased in the first hour post-OGTT, including amino acids (cluster 2), with the biggest fold change observed for leucin/isoleucine (mean fold change 0.76, corrected $p = 6.7 \times 10^{-17}$), tyrosine (0.84, $p = 2.4 \times 10^{-9}$), and methionine (0.86, $p = 1.3 \times 10^{-7}$), acylcarnitines (cluster 1 and 3) and NEFAs (Figure 4.3F) as well as phospholipids (cluster 1). Amino acids and acylcarnitines showed a further decrease between 1 and 2 h after challenge.

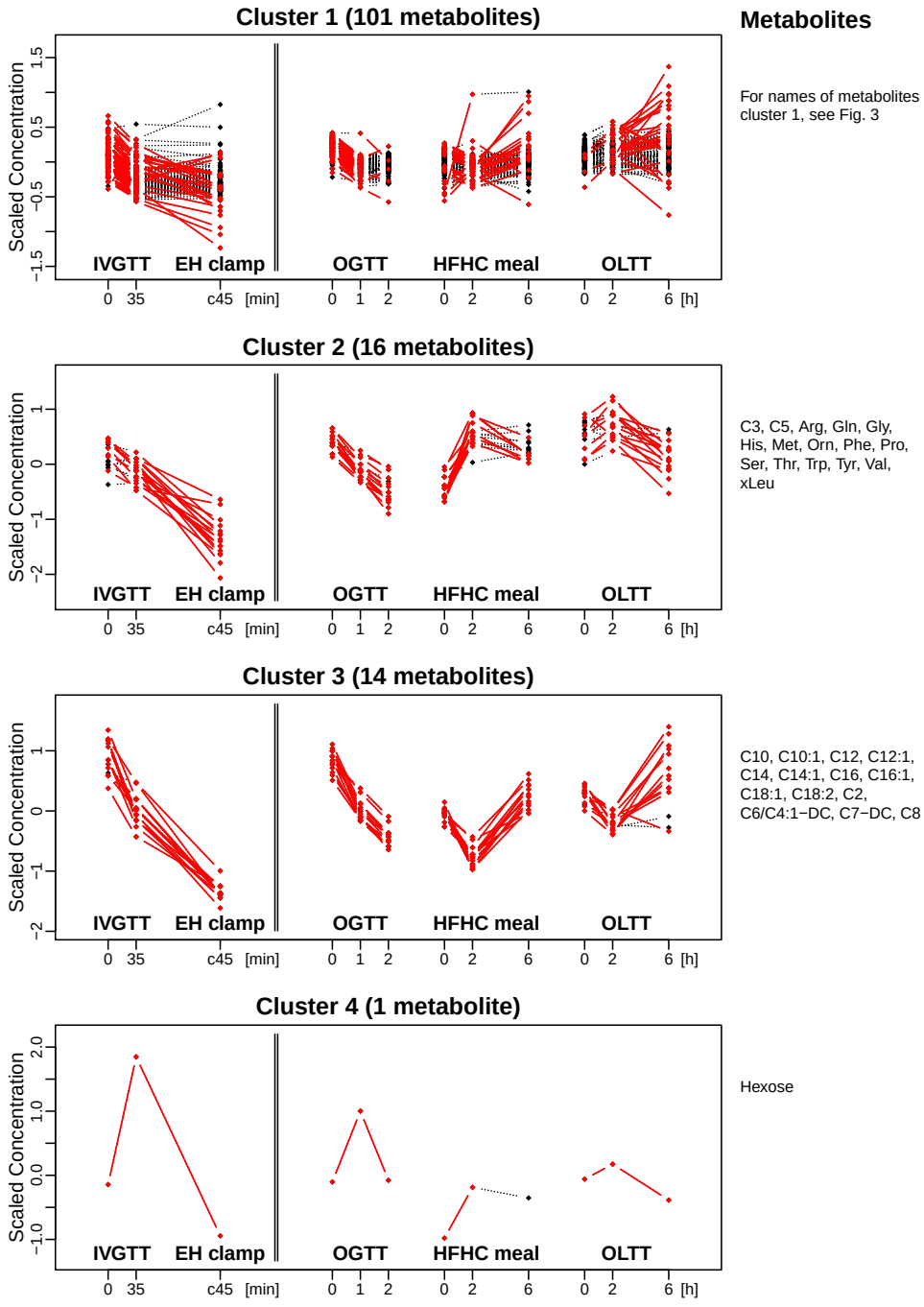

Figure 4.2: K-means clustering of challenge response profiles. Mean scaled concentrations of each metabolite at the different time points are shown, connected through lines. See Section 3.4.1 for details on clustering procedure and choice of the number ($K = 4$) of clusters. Solid red lines, significant concentration changes as identified in linear mixed-effects models (LMEs), after correction for multiple testing. Dotted black lines, not significant. Metabolites belonging to the respective cluster are specified on the right-hand side of the graphs, the number of metabolites in each cluster is included in the graph titles. c45, 45 min after clamp steady state; EH clamp, euglycemic-hyperinsulinemic clamp; HFHC meal, high-fat high-carbohydrate meal; IVGTT, intravenous glucose tolerance test; OGTT, oral glucose tolerance test; OLTT, oral lipid tolerance test.

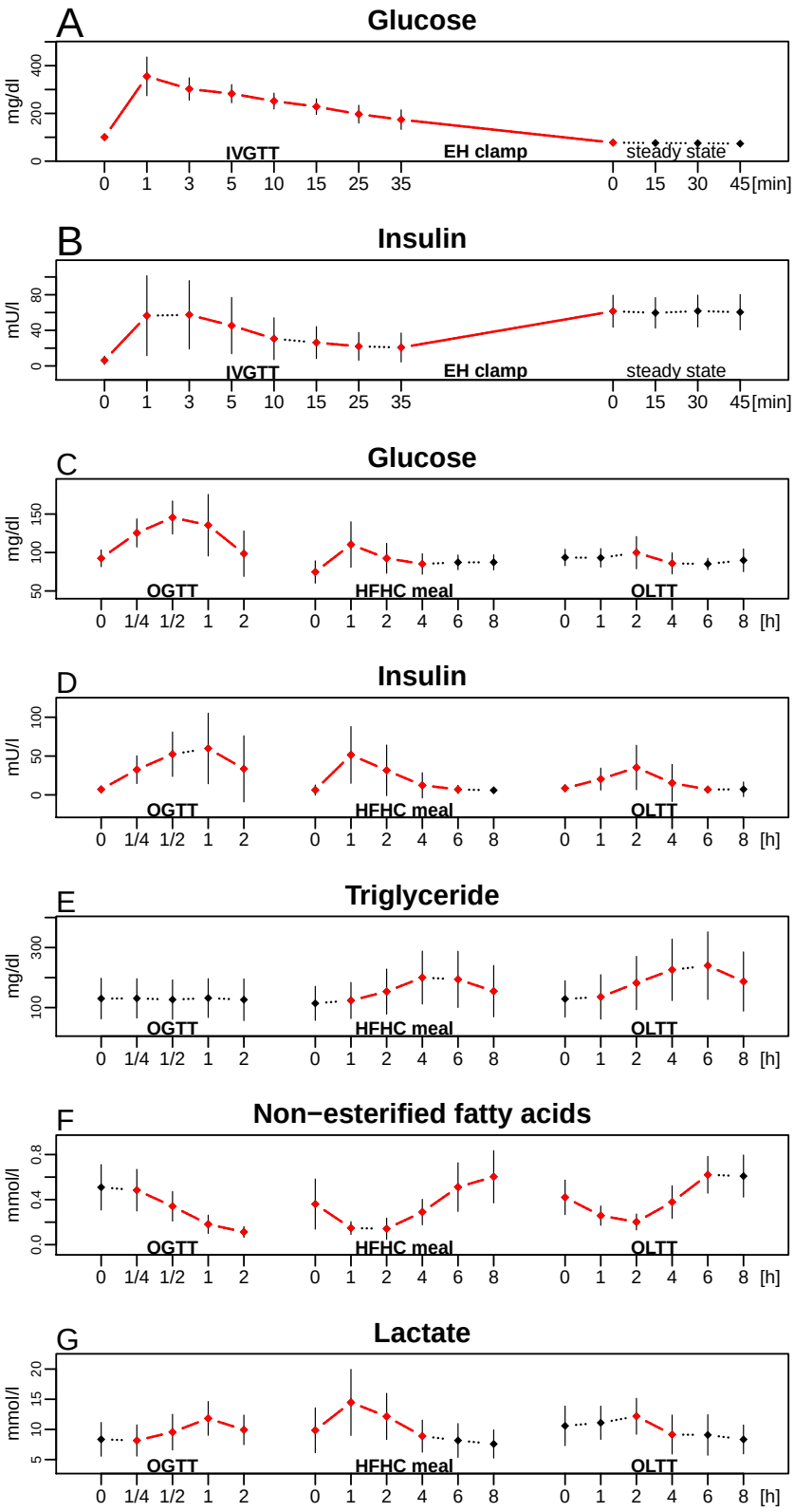

Figure 4.3: Time course of challenge responses of clinical traits. **a** and **b** Response to intravenous challenges, **c** to **g** response to oral challenges. Mean and standard deviation of plasma concentrations at the different time points are shown, connected through lines. Solid red lines, significant concentration changes as identified in LMEs, after correction for multiple testing. Dotted black lines, not significant. EH clamp, euglycemic-hyperinsulinemic clamp; HFHC meal, high-fat high-carbohydrate meal; IVGTT, intravenous glucose tolerance test; OGTT, oral glucose tolerance test; OLTT, oral lipid tolerance test.

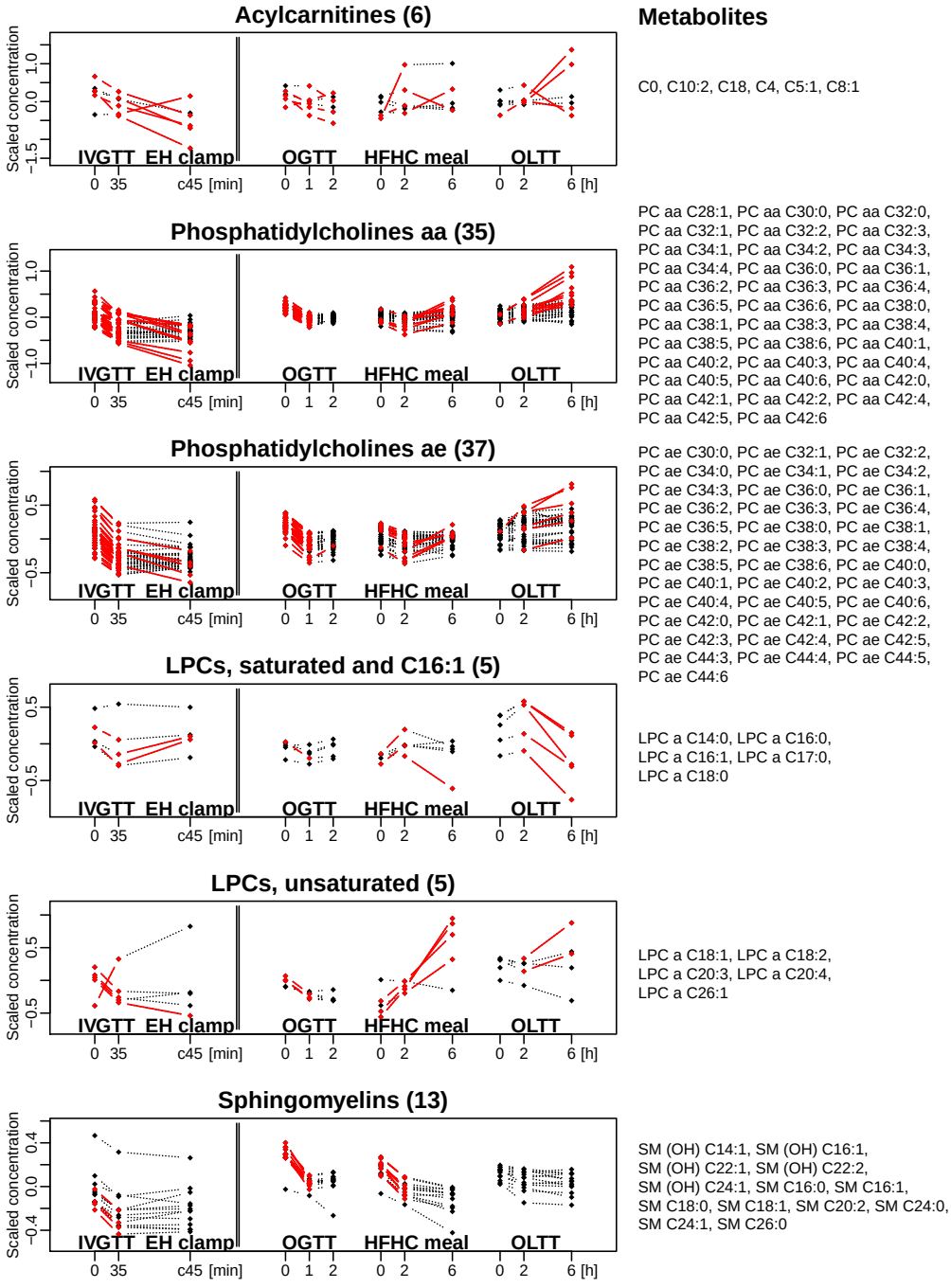

Figure 4.4: Time courses of challenge response for different biological metabolite groups within cluster 1. Mean scaled concentrations of each metabolite at the different time points are shown, connected through lines. Solid red lines, significant concentration changes as identified in linear mixed-effects models (LMEs), after correction for multiple testing. Dotted black lines, not significant. Metabolites belonging to the respective biological group are specified on the right-hand side of the graphs, the number of metabolites in each group is included in the graph titles. c45, 45 min after clamp steady state; EH clamp, euglycemic-hyperinsulinemic clamp; HFHC meal, high-fat high-carbohydrate meal; IVGTT, intravenous glucose tolerance test; OGTT, oral glucose tolerance test; OLTT oral lipid tolerance test.

IVGTT and EH clamp

After intravenous glucose injection, glucose and insulin plasma concentrations increased immediately, with a peak at 1 min (Figure 4.3A,B). As shown by LMEs and K-means clustering, the metabolite response to IVGTT at 35 min post-challenge was largely similar to the 1 h response to oral glucose, with decreases in amino acids (cluster 2), acylcarnitines (cluster 1 and 3), and phospholipids (cluster 1) (Figure 4.2). By trend, a weaker amino acid decrease (not significant for Gln, His, Pro, Ser and Thr), and a slightly stronger acylcarnitine decrease were observed during 35 min IVGTT as compared to 1 h OGTT. Decreases in PCs were similar, whereas decreases in SMs were slightly weaker in response to intravenous glucose (Figure 4.4). Whereas most LPCs decreased in response to both the oral and intravenous glucose challenge, the very-long-chain species LPC a C26:1 showed a significant increase in response to IVGTT (corrected $p = 9.1 \times 10^{-3}$).

Between the 35 min IVGTT measurement and the EH clamp steady state measurement, 51 significant metabolite changes were observed in LMEs. K-means clustering illustrated that hyperinsulinemia led to further significant decreases of amino acid levels that were stronger than the decrease induced by the glucose load (Figure 4.2). Also, levels of all 21 measured acylcarnitines continued to decrease significantly during the EH clamp, whereas carnitine (cluster 1) increased. Concentrations of PCs, predominantly of those with up to 36 carbon atoms, decreased significantly during EH clamp, whereas no significant change was observed for longer-chain PCs and for SMs (Figure 4.4). Also, for most LPCs, no significant change was observed in response to EH clamp, with the exception of two saturated LPCs, LPC C16:0 (corrected $p = 5.7 \times 10^{-3}$) and C17:0 (corrected $p = 7.9 \times 10^{-3}$), which showed concentration increases, and the unsaturated LPC C18:2 (corrected $p = 2.8 \times 10^{-2}$), which decreased.

Mixed-nutrient challenges

Metabolic response to mixed-nutrient challenges (OLTT, HFHC meal) was assessed 2 and 6 h post-challenge. During both challenges, insulin increased with a peak at 1 h (HFHC meal) and 2 h (OLTT) (Figure 4.3D), whereas glucose and lactate showed only a slight increase during the HFHC meal, which was not significant during the OLTT (Figure 4.3A,G). TG concentrations increased in response to both challenges, with a peak between 4 and 6 h (Figure 4.3E).

Metabolite response to OLTT was as follows: amino acids largely increased at 2 h (significant in the case of Val, Leu/Ile, Met, Pro, Tyr, Orn, corrected p-values ranging from 4.5^{-2} to 4.5×10^{-13}), followed by a decrease at 6 h (significant in the case of Val, Leu/Ile, Met, Pro, Tyr, His, Phe, Thr, Arg and Gly, corrected p-values ranging from 2.4×10^{-2} to 1.7×10^{-19}) (Figure 4.2). Acylcarnitines (cluster 3) and NEFAs decreased during the first 2 h and increased thereafter. Two sets of acylcarnitines, C3 and C5 (cluster 2) as well as C4, C5:1, C8:1, C10:2 and C18 (cluster 1) clustered separately from the majority of

acylcarnitines (cluster 3). This separate clustering was largely attributable to a diverging response of acylcarnitines from clusters 1, 2 and 3 to mixed-nutrient challenges, whereas the response of acylcarnitines from the three different clusters to glucose challenges was similar. For instance, during OLTT, the acylcarnitines C3, C4 and C5 (cluster 2) increased during the first two hours, decreasing thereafter, similar to the response observed for free carnitine (C0).

Phospholipid response to OLTT was divergent, with SMs not showing a significant response at all (Figure 4.4). For some PCs, a significant increase at 2 h (PC aa C34:2/3, C36:2 and PC ae C38:2) and at 6 h (additionally PC aa C36:1, C36:3/4/6, C40:2/3, C42:2 and PC ae C34:2, C36:2, C38:0, C40:0/1) was observed. LPCs differed in their response to OLTT depending on their chain length and degree of saturation. The long-chain unsaturated LPCs C18:2 (corrected $p = 9.7 \times 10^{-7}$) and C20:3 (corrected $p = 2.2 \times 10^{-2}$) significantly increased at 6 h post-OLTT, whereas saturated LPCs (corrected p-values ranging from 9.5×10^{-3} to 3.3×10^{-12}) and LPC C16:1 (corrected $p = 4.6 \times 10^{-3}$) significantly decreased.

The response to the HFHC meal was largely comparable to the response observed to the OLTT. However, amino acid concentrations showed a more pronounced increase in the first 2 h, and a less pronounced decrease between 2 and 6 h, as compared to the OLTT (Figure 4.2). Moreover, acylcarnitines (cluster 3) showed a more pronounced increase at 2 h and a less pronounced decrease between 2 and 6 h, indicating a weaker β-oxidative response to HFHC meal as compared to OLTT. Similarly to the OLTT, the acylcarnitines C3 and C5 (cluster 2) behaved oppositely to the majority of acylcarnitines (cluster 3). In contrast to 2 h OLTT, where apart from four PCs no significant responses were observed for phospholipids, SMs tended to decrease in the first 2 h during HFHC meal, as did some PCs, whereas the LPCs C16:0, C17:0, C18:2, C20:3/4 increased (Figure 4.4).

4.1.2 Effect of the *FTO* rs9939609 risk allele on challenge responses

Next, it was explored whether the *FTO* risk allele modified challenge responses, thereby revealing early metabolic disturbances associated with the risk allele. In terms of baseline clinical traits (Table 4.1) and fasting metabolite concentrations, no differences were observed between the 25 carriers of the *FTO* rs9939609 risk allele (AA genotype) and the 31 carriers of the TT genotype. Likewise, when the effect of the *FTO* genotype on metabolite challenge responses as well as response of glucose, insulin, TGs, NEFAs and lactate was explored by means of interaction terms in LMEs (see Section 3.3.1) adjusted for BMI and age and the respective interactions with time point, no significant *FTO* genotype effects on responses to IVGTT, OLTT and HFHC meal were observed. During OGTT, a tendency to a weaker decrease of concentrations of 15 long-chain PCs and one SM (cluster 1) could be observed in the first hour post-challenge in *FTO* risk allele carriers compared to non-carriers at an uncorrected p-value of < 0.05 (Figure 4.5, Supplementary Table 4 of the original publication (Wahl *et al.*, 2013b)). Furthermore, for acylcarnitine C12:1

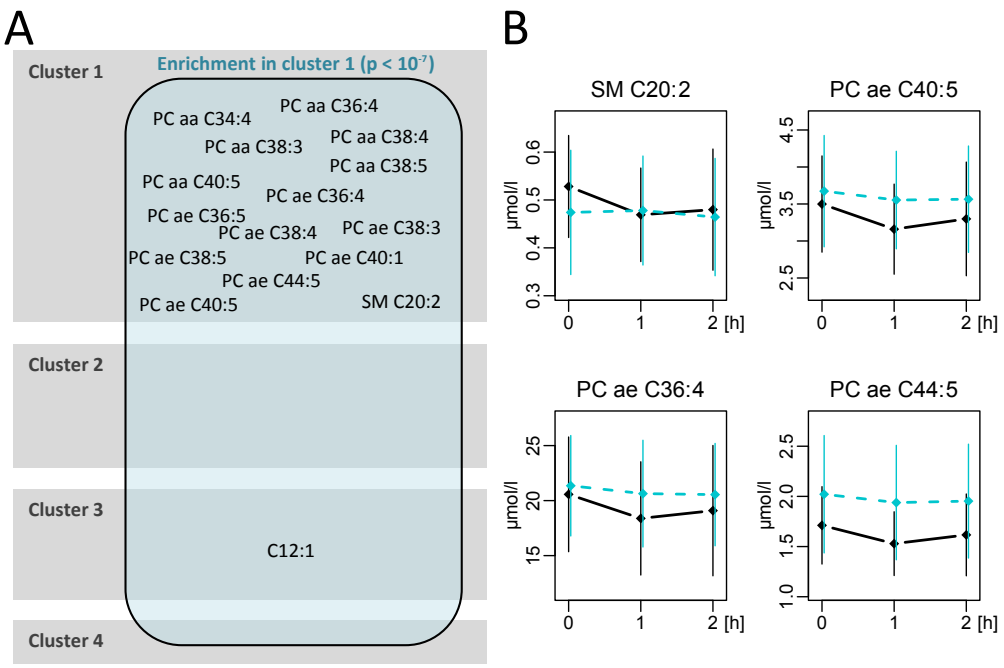

Figure 4.5: Investigation of *FTO* genotype effect on OGTT-induced metabolite changes. **a** Significant enrichment of *FTO*-OGTT interaction at 1 h post-challenge in cluster 1. **b** Time course plots of OGTT-induced metabolite changes in *FTO* risk allele carriers (dashed green line) and non-carriers (solid black line). Metabolite abbreviations are explained in Section 2.2.5.

(cluster 3), a slightly stronger decrease was observed during the first hour post-challenge in *FTO* risk allele carriers compared to non-carriers (uncorrected $p = 0.011$). A weighted enrichment analysis (see Section 3.5.1) showed that the observed *FTO* effects were significantly enriched for cluster 1, that is, *FTO* effects on OGTT response of phospholipids were observed more often than expected by chance ($p < 1 \times 10^{-7}$) (Figure 4.5). Results were similar when BMI and BMI-time point interaction were omitted from the model (data not shown).

To investigate the power of the chosen LME interaction models for analyzing *FTO* genotype effects on challenge response, *post hoc* power analyses were performed for two representative metabolites, SM C20:2 ($\beta = 0.66$ (95% CI: 0.16, 1.16), $p = 1.1 \times 10^{-2}$) and PC ae C44:5 ($\beta = 0.25$ (0.03, 0.47), $p = 2.5 \times 10^{-2}$) (see Section 3.3.5). Assuming that the observed genotype effects on the 1 h OGTT response for these metabolites resemble the respective true effects, the power to observe p-values lower than 10^{-3} was 23.2% and 12.1%, and to observe Bonferroni-significant p-values, 5.1% and 1.9%, respectively. Thus, the present study was largely underpowered to detect genotype interaction effects. To observe p-values lower than 10^{-3} with a reasonable power of 80%, a sample size of 90 and 116 would have be needed, and to observe Bonferroni-significant p-values lower than 5.0×10^{-5}, a sample size of 126 and 160, respectively.

4.1.3 Effect of the *TCF7L2* rs7903146 risk allele on intravenous challenge response

TCF7L2 risk allele effects on IVGTT and EH clamp response were reported in a separate publication (Then *et al.*, 2013). The underlying study population comprised 17 *TCF7L2* risk allele carriers (8 homozygous, TT genotype; 9 heterozygous, CT genotype), as well as the 24 non-carriers that participated in the intravenous challenge tests. See Table 1 of the original publication Then *et al.* (2013) for detailed characteristics. No significant genotype effects were observed on metabolite concentrations in the fasting state.

For 17 phospholipid metabolites, concentration changes during IVGTT were significantly modified by the *TCF7L2* risk allele (Table 4.2). Note that this table also comprises sums of metabolite concentrations, since in this publication, 45 metabolite sums were jointly analyzed with the single metabolite concentrations, as proposed in the Biocrates

Table 4.2: Metabolites and sums with significant effect of the *TCF7L2* genotype on IVGTT response.

	IVGTT response (non-carriers)		**IVGTT response (*TCF7L2* risk allele carriers)**		**Genotype interaction effect**		
Metabolite	β	p-**value**	β	p-**value**	β	p-**value**	**corrected** p-**value**
Sphingomyelins							
SM (OH) C14:1	0.05	6.2E-01	-0.45	9.0E-04	-0.50	4.8E-03	3.9E-02
SM (OH) C22:1	0.02	8.4E-01	-0.49	7.3E-04	-0.51	7.1E-03	5.0E-02
SM (OH) C22:2	0.04	7.2E-01	-0.57	1.6E-04	-0.61	2.0E-03	3.6E-02
SM (OH) C24:1	0.09	4.6E-01	-0.57	1.3E-04	-0.66	8.3E-04	3.6E-02
SM C16:0	0.06	7.0E-01	-0.64	5.7E-04	-0.70	4.3E-03	3.9E-02
SM C16:1	0.02	8.8E-01	-0.64	2.1E-04	-0.66	3.5E-03	3.6E-02
SM C18:0	0.00	9.9E-01	-0.57	2.1E-04	-0.57	4.7E-03	3.9E-02
SM C18:1	0.02	8.9E-01	-0.52	5.2E-04	-0.53	6.5E-03	4.8E-02
SM C24:0	0.05	6.8E-01	-0.61	2.7E-04	-0.67	2.6E-03	3.6E-02
SM C24:1	0.07	6.1E-01	-0.55	7.4E-04	-0.61	4.2E-03	3.9E-02
SMs	0.05	7.2E-01	-0.65	3.7E-04	-0.70	3.5E-03	3.6E-02
SM C	0.05	7.2E-01	-0.64	3.9E-04	-0.70	3.6E-03	3.6E-02
SM (OH)	0.04	7.4E-01	-0.53	3.5E-04	-0.57	3.5E-03	3.6E-02
long-chain SMs	0.06	6.4E-01	-0.60	4.7E-04	-0.66	3.2E-03	3.6E-02
long-chain SM C	0.06	6.4E-01	-0.60	5.0E-04	-0.66	3.4E-03	3.6E-02
long-chain SM (OH)	0.09	4.6E-01	-0.57	1.3E-04	-0.66	8.3E-04	3.6E-02
Lysophosphatidylcholines							
LPC a C14:0	0.39	1.3E-02	-0.42	2.6E-02	-0.81	1.6E-03	3.6E-02
LPC a C16:0	-0.04	7.5E-01	-0.64	3.4E-05	-0.60	2.6E-03	3.6E-02
LPC a C16:1	-0.07	3.6E-01	-0.47	1.4E-06	-0.40	1.2E-03	3.6E-02
LPC a C17:0	-0.04	7.6E-01	-0.57	1.4E-04	-0.53	6.1E-03	4.7E-02
LPCs	-0.12	2.4E-01	-0.63	2.8E-06	-0.51	3.0E-03	3.6E-02
saturated LPCs	-0.03	8.1E-01	-0.65	3.2E-05	-0.62	2.2E-03	3.6E-02
Phosphatidylcholines							
PC aa C28:1	-0.13	2.4E-01	-0.60	1.4E-05	-0.47	7.3E-03	5.0E-02
PC aa C40:4	-0.05	5.6E-01	-0.48	8.4E-06	-0.43	1.8E-03	3.6E-02
PC ae C40:5	-0.10	4.9E-01	-0.84	1.1E-05	-0.74	2.7E-03	3.6E-02

Metabolite abbreviations are explained in Section 2.2.5.

MetaDis*IDQ*^{TM} manual. They are assumed to indicate a certain metabolic state or process. Specifically, *TCF7L2* risk allele carriers showed a significantly stronger decline of metabolite concentrations during IVGTT as compared to non-carriers for 10 out of 13 SM species as well as for the total sum of SMs and different partial sums. A similar observation was made for four out of 14 individual LPC species as well as for the sum of all measured LPC species, and the sum of measured saturated LPCs. Finally, of 74 analysed PCs, three, namely PC aa C28:1, PC aa C40:4 and PC ae C40:5, showed a stronger concentration decrease in *TCF7L2* risk allele carriers as compared to non-carriers during IVGTT.

When metabolic effects to EH clamp were investigated, no significant genotype effects were observed on metabolite concentration changes between the post-IVGTT timepoint (35 min) and the clamp steady state timepoint. Accordingly, significant effects were observed when baseline and clamp steady state concentrations were compared, which were largely similar to the effects on IVGTT response described above (see Table 4 of the original publication (Then *et al.*, 2013).

4.1.4 Discussion

Using a targeted metabolomics approach, changes in metabolite plasma levels elicited by the challenge tests IVGTT, EH clamp, OGTT as well as OLTT and HFHC meal were characterized. Furthermore, it was investigated whether challenge responses were modified by known obesity and T2D risk alleles at the *FTO* and *TCF7L2* loci, respectively.

Unsupervised clustering identified four metabolite clusters with a distinct response profile across all challenge tests. Even though no biological knowledge was used for cluster formation, metabolites were largely clustered into groups of biological and structurally related molecules. Moreover, these metabolite groups (amino acids, hexose, acylcarnitines, and phospholipids) also formed the most clearly defined sub-networks in a Gaussian graphical modeling approach based on partial correlations of fasting metabolite concentrations in a large population representative sample (Krumsiek *et al.*, 2011). Here, it could be demonstrated that these groups of structurally related metabolites also show comparable postprandial responses, as it has been observed for other metabolite panels before (Ho *et al.*, 2013, Pellis *et al.*, 2012).

Metabolic response to glucose and insulin challenges

Both glucose challenges, OGTT and IVGTT, triggered a switch from catabolism to anabolism in the three major nutrient axes, as reflected by increase in lactate levels (stimulation of glycolysis), decrease in amino acid concentrations (inhibition of proteolysis) as well as decrease in acylcarnitine and NEFA concentrations (inhibition of fat oxidation). Similar observations during OGTT have been reported before (Ho *et al.*, 2013, Shaham *et al.*,

2008, Skurk *et al.*, 2011, Zhao *et al.*, 2009). Moreover, it could be replicated that the fold changes during the 2 h OGTT were most prominent for the amino acids leucine/isoleucine, tyrosine and methionine (Ho *et al.*, 2013, Shaham *et al.*, 2008, Skurk *et al.*, 2011). These neutral amino acids share common insulin-dependent transporters that mediate amino acid shuttling into peripheral tissues (Deo *et al.*, 2010, Skurk *et al.*, 2011), in exchange for glutamine, which did not decrease during OGTT and IVGTT in this study.

It was observed that the intravenous glucose challenge exerted a weaker anti-proteolytic effect at 35 min as compared to the 1 h oral glucose challenge. Oral glucose triggers the release of incretin hormones such as glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide 1 (GLP-1) which stimulate glucose-dependent insulin release (Ranganath, 2008). This incretin effect accounts for the higher insulin secretion in response to oral as compared to intravenous glucose, and can contribute to the amplification of insulin-induced metabolic effects such as reduced proteolysis. Prolonged decrease during EH clamp further supports the notion that the observed amino acid decrease is mediated by insulin.

The anti-β-oxidative response to IVGTT, measured by a decrease in acylcarnitines at 35 min, was stronger as compared to the 1 h OGTT response. Various *in vitro* studies have shown a stimulatory effect of GLP-1 and GIP on lipolysis in human and murine adipocytes (Getty-Kaushik *et al.*, 2006, He *et al.*, 2010, Ruiz-Grande *et al.*, 1992, Sancho *et al.*, 2005, Timper *et al.*, 2013, Vendrell *et al.*, 2011, Villanueva-Peñacarrillo *et al.*, 2001). Furthermore, a positive association between fasting plasma GLP-1 concentration and fat oxidation has been observed in humans (Pannacciulli *et al.*, 2006). An incretin-induced prolipolytic effect might potentially attenuate insulin-induced suppression of lipolysis. In contrast to this hypothesis, a recent clamp study has shown an abolishment of the lipolytic effect of GLP-1 in the presence of insulin (Seghieri *et al.*, 2013). In addition, these differences between OGTT and IVGTT might also be attributable to a more immediate plasma insulin response to intravenous than to oral glucose.

In this study, a significant downregulation of plasma levels was observed for most phospholipids, including PCs, LPCs and SMs, during both OGTT (1 h) and IVGTT (35 min). The decrease of PCs may be explained by insulin triggering hydrolysis of PCs by the activation of specific phospholipases, as previously reported in rat myotubes (Standaert *et al.*, 1996a), rat adipocytes (Standaert *et al.*, 1996b), rat hepatocytes (Donchenko *et al.*, 1994) and human hepatoma cells (Novotná *et al.*, 2003). In addition, insulin triggers both LDL receptor activity (Duvillard *et al.*, 2003, Nägele *et al.*, 1997) and adipose tissue lipoprotein lipase protein expression (McTernan *et al.*, 2002), thereby stimulating clearance of phospholipid containing lipoproteins from the circulation (Ogita *et al.*, 2008). Apart from few PCs with up to 36 carbon atoms, no further reduction of PC and SM plasma levels during EH clamp was observed, suggesting a selective saturation of the described processes.

Of note, the very-long-chain LPC C26:1 increased in response to IVGTT but not OGTT,

whereas most mediumand long-chain LPCs decreased in response to the glucose challenge tests. Metabolism of very-long-chain fatty acids is unique with an initial oxidation in peroxisomes before mitochondrial β-oxidation (Lee *et al.*, 2012). The findings of this study are consistent with peroxisomal fat oxidation being inhibited by insulin (Hamel *et al.*, 2001) and induced by incretins (Lee *et al.*, 2012, Svegliati-Baroni *et al.*, 2011).

Metabolic response to meal challenges

The observed metabolite changes induced by OLTT and HFHC meal may be attributed to both the specific nutrient contents of the respective challenges and to putative insulin effects. For instance, the protein contained in the test meals provoked increases in amino acid plasma concentrations in the first 2 h of both challenges, whereas decreases in amino acid levels at 6 h are likely attributable to anti-proteolytic insulin effects. Plasma amino acid concentrations showed a stronger increase at 2 h and a weaker decrease at 6 h in response to the HFHC meal, potentially attributable to the fact that during HFHC meal, subjects consumed twice the amount of protein (32 g) as compared to the OLTT (17.8 g) on average. In agreement with a putative insulin effect on lipolysis and lipid clearance, decreases in acylcarnitine, NEFA, PC and SM concentrations were observed at 2 h, more pronounced during HFHC meal, where a stronger and earlier (at 1 h as compared to 2 h during OLTT) insulin peak was observed. By 6 h, a peak in TGs was observed (as by Lopez-Miranda *et al.* (2007), Westphal *et al.* (2000)), insulin reached baseline concentrations and acylcarnitine, NEFA and PC concentrations increased, consistent with dietary lipids being oxidized and incorporated into phospholipids.

A group of acylcarnitines, including the short- and medium-chain species C3, C4, C5, C5:1, C8:1 and C10:2, showed a diverging response to mixed nutrient challenges as compared to the majority of acylcarnitines. This is in agreement with the postprandial increase of C3 and C4 (by trend also of C5 and C10:2) observed by Ramos-Roman *et al.* (2012). Krug *et al.* (2012) found the short-chain acylcarnitines (C3, C4, C5 and C5:1) to respond similarly to diverse anabolic and catabolic challenges as the branched-chain amino acids (BCAAs), tyrosine and methionine. Short-chain fatty acids, including isovalerate (C5), α-methylbutyrate (C5), isobutyrate (C4), and proprionate (C3), are byproducts of BCAA metabolism (Luís *et al.*, 2011). Thus, the corresponding acylcarnitines may be derived from a triggered metabolism of BCAAs after oral protein intake, which is additionally supported by the here observed clustering of these acylcarnitines with amino acids.

6 h after OLTT, concentrations of the majority of metabolites differed significantly from concentrations in fasting plasma samples. Thus, analyzing fasting and nonfasting samples together in epidemiological studies should be avoided to prevent confounding.

Taken together, the combination of different challenge tests applied in this study allowed a thorough characterization of the physiological behavior of distinct metabolite subclasses.

Assessment of genotype-challenge interactions

Metabolic challenge tests have assisted the detection of early metabolic changes associated with risk-conferring genotypes (Fontaine-Bisson *et al.*, 2007, Franks *et al.*, 2007, Tan *et al.*, 2006, Weickert *et al.*, 2007, Wybranska *et al.*, 2007). Such genotype-associated metabolic effects may be hidden behind tight homeostatic regulation when solely fasting state conditions are analyzed (van Ommen *et al.*, 2009). Here, targeted metabolomics was used as a hypothesis free approach to investigate whether metabolite responses to defined challenges may contribute in unraveling novel genotype effects at the known obesity risk locus *FTO* rs9939609 (Frayling *et al.*, 2007, Speliotes *et al.*, 2010) and at the known T2D risk locus *TCF7L2* rs7903146 (Morris *et al.*, 2012).

The results suggest alterations in post-IVGTT sphingolipid and phospholipid metabolism in subjects carrying the *TCF7L2* risk genotype that were not observed in genome-wide metabolomics association studies in the fasting state (Gieger *et al.*, 2008, Illig *et al.*, 2010, Suhre *et al.*, 2011). Specifically, most sphingomyelins and few phosphotidylcholine and lysophosphatidylcholine showed a stronger reduction of plasma concentration in carriers of the *TCF7L2* risk allele as compared to non-carriers upon IVGTT, with concentrations remaining low during EH clamp. Although metabolite differences were not significant in the fasting state, these metabolite showed by trend higher levels in risk allele carriers.

Sphingomyelins and phosphatidylcholines are important structural components of plasma lipoprotein and cell membranes and are involved in the regulation of cell function, membrane protein trafficking and inflammation (Gault *et al.*, 2010). Increased sphingomyelin levels have been reported in subjects with T2D (Zhu *et al.*, 2011). A possible explanation of the stronger sphingomyelin decrease in risk allele carriers compared to non-carriers upon IVGTT is an increased action of the enzyme sphingomyelinase. Sphingomyelinase degrades sphingomyelins to ceramide, thereby potentially contributing to β-cell apoptosis (Zhang *et al.*, 2009). Inhibition of ceramide synthesis decreased β-cell apoptosis and defective protein trafficking in β-cells exposed to lipotoxicity (Boslem *et al.*, 2011). In addition, mitochondrial dysfunction might be involved in decreased insulin secretion upon altered membrane sphingomyelin and ceramide content (Yano *et al.*, 2011). Together, these findings provide a potential link between TCF7L2-induced changes in sphingolipid metabolism and reduced β-cell function. Importantly, an impaired first-phase insulin response in *TCF7L2* risk allele carriers as compared to non-carriers was also observed in the VID study (Then *et al.*, 2013). Experimental studies may lead to a closer understanding of the underlying processes.

Strengths and limitations

Plasma metabolomic responses to five different metabolic challenges were comprehensively investigated. Thereby, previously unknown postprandial effects on the metabolite profile

were identified, including previously unreported metabolic effects of IVGTT and EH clamp. A more thorough comparison of responses to different challenges is provided than has been reported before in a similarly sized sample. The combined statistical approach of LMEs and K-means clustering allowed to elaborate similarities and divergences in responses to the five challenges for distinct metabolite groups. Using interaction models, the utility of the chosen approach for investigating genotype-specific effects could be explored. The present study provides a framework for further analysis of additional risk variants.

A limitation of this study is the small sample size and the lack of replication. *Post hoc* power analyses showed that a two- to threefold sample size would have been needed to detect the potential small *FTO*-OGTT interaction effects as statistically significant. Accordingly, a similarly larger genotyped cohort would have been needed as recruitment base. Thus, the presented data indicate the limitations of hypothesis-free explorative investigations in experimental settings (Bouchard, 2008), where the identification of genotype effects most likely strongly depends on the appropriate choice of challenges and the comprehensiveness of the analyzed metabolomics panel. Furthermore, only male subjects were included to increase the homogeneity of the study population, which is advantageous for the investigation of gene-environment interactions. However, it should not be ignored that men and women differ in their postprandial metabolic response (Ho *et al.*, 2013). Thus, generalization to both sexes should happen with care. In addition, metabolomics measurements in shorter time intervals during the challenges would have allowed a more detailed characterization of metabolic response profiles. Also, the study was limited to anabolic challenges, which might be complemented by catabolic challenges such as prolonged fasting or exercise in future investigations.

Conclusions

This study contributes to the understanding of the physiological plasma metabolomics response to different metabolic challenges, and assesses the utility of the chosen approach for unraveling genotype-challenge interactions. The obtained results confirm established effects of oral glucose or mixed nutrient intake on carbohydrate and protein metabolism. Previously unreported responses in different phospholipid metabolites are presented and metabolite changes in response to IVGTT as compared to OGTT are reported for the first time. A *post hoc* power analysis on FTO-challenge interactions demonstrates the limited feasibility of such an experimental approach for large-scale hypothesis-free testing of genotype effects. At the same time, early *TCF7L2*-conferred perturbations of sphingo- and phospholipid metabolism were observed that could only be detected through challenge tests and that occured in a stage when conventional parameters of glucose homeostasis were not yet affected, thereby improving the understanding of the molecular mechanisms underlying the development of T2D in subjects with the *TCF7L2* risk allele.

4.2 Methylome-wide association study of body mass index

Studying body mass-related DNA methylation signatures is promising from two perspectives. First, epigenetic mechanisms might be involved in the development of obesity, potentially explaining a part of the missing heritability. Second, epigenetic regulation is a potential pathway underlying obesity-related pathogenic processes including disturbed lipid and glucose metabolism. To examine these two perspectives, an EWAS of BMI was conducted in a two-stage discovery-replication analysis comprising more than 10,000 subjects to identify and validate the perturbations in DNA methylation associated with obesity (Figure 4.6). The identified loci were integrated with other omics data as well as clinical data to obtain new insights into the role of these loci in the development of obesity and its cardiometabolic consequences.

The contents of this section are mainly based on the manuscript

- **Wahl S**[*], Lehne B[*], Drong AW[*], Loh M[*], Zeilinger S, Fiorito G, Kasela S, Richmond R, Dehghan A, Franke L, Esko T, Milani L, Relton CL, Kriebel J, Prokisch H, Herder C, Peters A, Illig T, Waldenberger M, Bell JT, Franco OH, van der Harst P, Lindgren CM, McCarthy MI, Matullo G, Gieger C[#], Kooner JS[#], Grallert H[#], Chambers JC[#]. "Epigenome-wide association study reveals extensive perturbations in DNA methylation associated with adiposity and its metabolic consequences." *in preparation.*

4.2.1 Epigenome-wide association and replication

In the first stage, epigenome-wide association testing was performed in four large population-based studies comprising 5387 subjects of European ($n = 2707$) and South Asian ($n = 2680$) origin. Characteristics and analysis details are summarized in Appendix Tables A.1 and A.2. In each cohort, associations of DNA methylation and BMI were determined using linear models with BMI as response and methylation at a single CpG site as the covariate, adjusting for age, sex, physical activity, smoking status, alcohol intake, estimated white blood cell proportions and technical variables (see Section 2.1.2 for exact definition of behavioral factors, Section 3.1.2 for technical effects, Section 3.3.2 for choice of covariates). There was little evidence for heterogeneity of effects between the ethnic groups. Therefore, results could be meta-analyzed using inverse-variance weighted fixed-effects meta-analysis (see Section 3.3.6). The genomic control inflation factor ranged from 0.98 to 1.29 in the individual studies, and was 1.11 in meta-analysis. Genomic control correction was applied before and after meta-analysis.

278 CpG sites showed genome-wide significant ($p < 10^{-7}$) association with BMI. These CpGs were distributed between 207 genetic loci. The lead marker at each of these loci, defined as the CpG site with the lowest p-value for BMI association, was selected for further

[*,#] contributed equally

analysis. Conditional analyses suggest the presence of multiple CpG sites independently associated with BMI at 23 of these loci ($p < 10^{-7}$ after conditioning on the lead CpG site of the respective locus). There were no additional loci showing association of methylation with BMI at $p < 10^{-7}$ on either of the sex chromosomes.

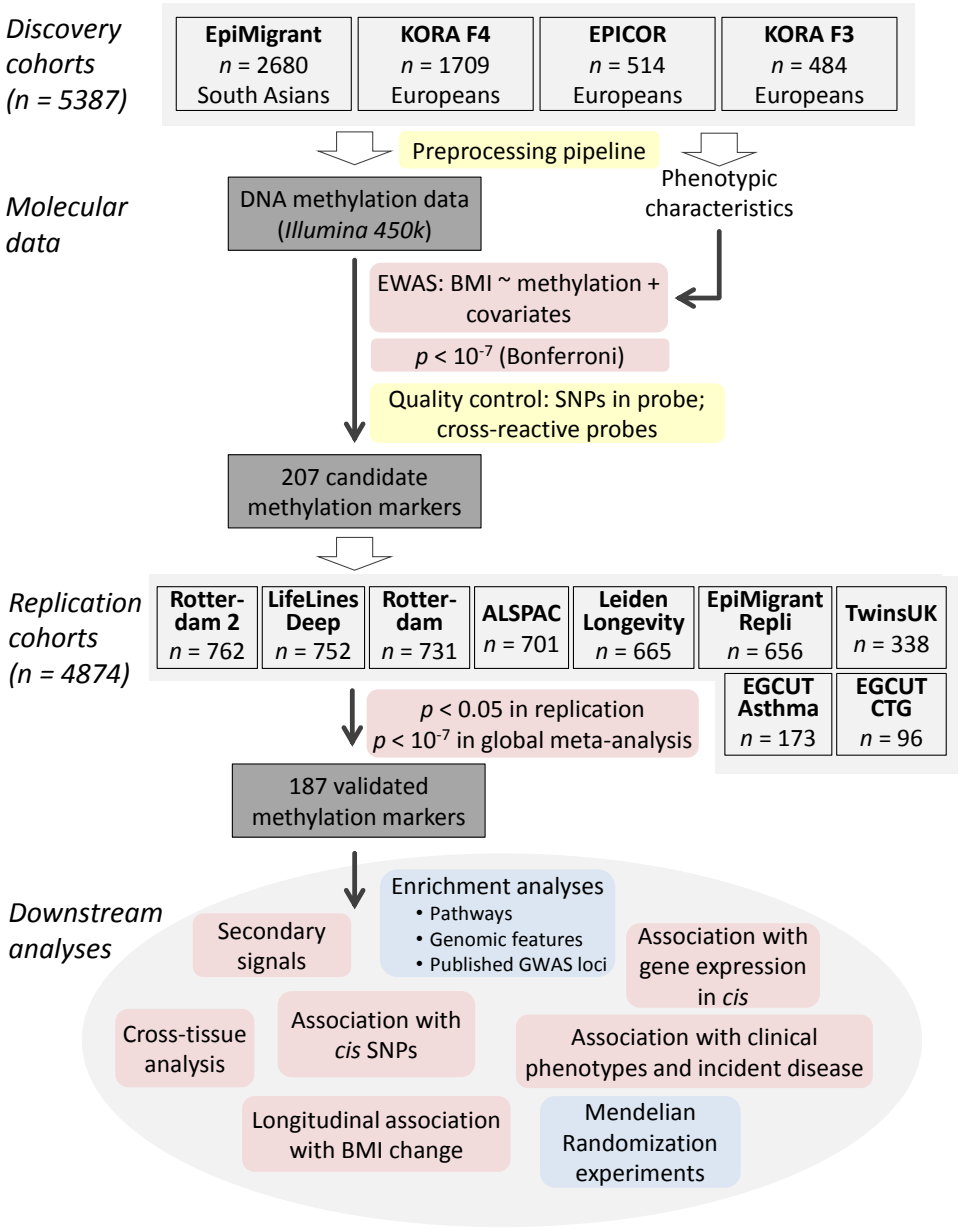

Figure 4.6: Epigenomics of BMI: Study design and analysis strategy. Details on studies and omics measurements are given in Section 2.1 and in Appendix Tables A.1 to A.3. Color coding of statistical methods: yellow, data preprocessing and quality control (Section 3.1); red, univariate data analysis (Section 3.3); blue, extraction of biological knowledge (Section 3.5). eQTL, expression quantitative trait locus; EWAS, epigenome-wide association study; ENCODE, Encyclopedia of DNA Elements.

43 of the 207 CpGs contained one or more known SNPs within their probe sequence, which may impair hybridization of the probe (see Section 3.1.2). To ensure that these SNPs – 58 in total – are not causing spurious association results, 1000G-imputed SNP data were obtained from a subset of $n = 3961$ subjects from the discovery cohorts (Appendix Table A.4). Of the 58 SNPs, 38 were non-monomorphic in at least one of the studies, and 13 had minor allele frequencies (MAFs) above 1%. No genetic confounding by any of these 38 SNPs was observed (Figure 4.7). However, 11 SNPs showed significant association with methylation at the respective CpG site ($p < 1.3 \times 10^{-3}$, corresponding to Bonferroni correction), including 3 low-frequency SNPs located within the CpG site itself (Table 4.3). Since it cannot be excluded that these associations are due to hybridization artefacts rather than biological mechanisms, the corresponding 11 CpGs were excluded from downstream analyses involving SNPs. Furthermore, for four CpGs, cross-hybridization of the probe had been reported (Price *et al.*, 2013), including cg19373099 (*CRYGFP* locus), cg25096107 (*IGHA2* locus), cg13097800 (*RPL10L* locus) and cg10505902 (*PDE4DIP* locus).

The 207 lead CpGs were put forward to the replication stage, where association with BMI was tested among 4874 subjects from 9 studies (see Appendix Tables A.1 and A.3 for characteristics and analysis pipelines). For 187 loci, BMI-associated perturbations in DNA methylation could be validated (at $p < 0.05$ in the replication stage and $p < 10^{-7}$ in a combined meta-analysis) (Table 4.4, Figure 4.8). The strongest effects were observed for cg06500161 (*ABCG1* locus, z-score = 18.4, $p = 2.0 \cdot 10^{-75}$), cg00574958 (*CPT1A* locus, z-score = -15.4, $p = 1.2 \cdot 10^{-53}$), cg11024682 (*SREBF1* locus, z-score = 15.0, $p = 1.3 \cdot 10^{-50}$), cg17501210 (*RPS6KA2* locus, z-score = -13.4, $p = 6.5 \cdot 10^{-41}$), and cg18181703 (*SOCS3* locus, z-score = -12.9, $p = 3.6 \cdot 10^{-38}$).

4.2.2 Cross-tissue patterns of DNA methylation

To address the question of whether the observed whole blood methylation signatures are representative of methylation in metabolically relevant tissues, correlation of methylation across different tissues was studied. To this end, publicly available data from the *Gene Expression Omnibus* (GEO) database (accession number GSE48472, Slieker *et al.* (2013)) were used, comprising genome-wide methylation data for 41 samples from blood and six metabolically relevant tissues. Correlation analysis of the 187 BMI-related CpG sites revealed high correlation of blood methylation with that in spleen as well as omental and subcutaneous fat (Figure 4.9), whereas correlation with liver, pancreas and muscle was lower but still significant.

4.2.3 Association with gene expression

Next, *cis*-associations between DNA methylation at the 187 validated methylation markers and gene expression were analyzed in a subgroup of 1785 subjects from KORA F4 (n = 703) and EpiMigrant (907 South Asians, 175 Europeans, partial overlap with discovery

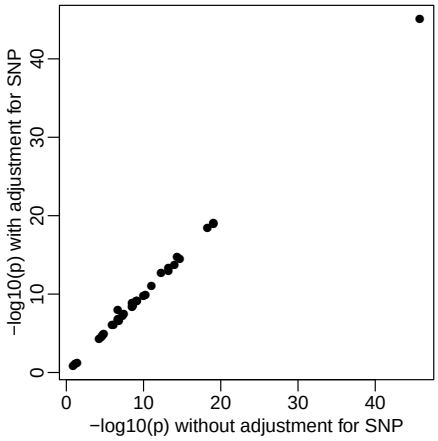

Figure 4.7: Investigation of genetic confounding by SNP located in Infinium 450k probe sequence. Adjustment for the SNP had no material impact on the p-values for association between BMI and methylation. SNPs were included in the models as dosages, i.e. expected allele counts of the non-reference allele. Linear models were adjusted for the discovery covariates and results were combined by inverse-variance weighted fixed-effects meta-analysis. See Appendix Table A.4 for details on the SNP data.

Table 4.3: CpGs with associated SNP in the probe sequence.

						p-values				
CpG ID	**Chr.**	**Position**	**SNP**	**MAF**	**Dist.**	BMI~CpG	BMI~CpG|SNP	CpG~SNP	BMI~SNP	***n***
cg22700686	1	153538764	rs2478147	0.006	1	4.6E-15	1.8E-15	8.9E-11	5.6E-01	3959
cg07202479	1	159174162	rs3027012	0.161	39	5.3E-11	1.1E-10	1.3E-17	2.5E-01	3943
cg09222732	6	466893	rs73374982	0.105	14	2.4E-09	2.4E-09	4.5E-04	7.0E-01	3958
cg10975897	6	15504844	rs56186721	0.049	31	1.5E-07	2.8E-07	5.0E-24	2.9E-01	3932
cg09697999	7	733198	rs73047920	0.276	23	2.1E-07	9.6E-09	3.5E-133	7.3E-01	3959
cg21037180	8	82276987	rs11784442	0.064	41	3.0E-09	1.3E-09	6.1E-59	8.5E-01	3821
cg16611584	17	19809078	rs56403226	0.010	0	9.7E-15	1.7E-14	7.1E-55	2.5E-01	3955
cg04524040	19	4153364	rs350880	0.243	36	1.0E-10	1.9E-10	5.6E-21	3.7E-01	3957
cg26836479	19	42706353	rs16975684	0.141	43	2.0E-07	1.6E-07	4.5E-09	9.8E-01	3956
cg02711608	19	47287964	rs76693964	0.005	0	5.3E-13	2.1E-13	3.5E-17	7.6E-01	3959
cg06500161	21	43656587	rs9982016	0.042	7	1.7E-46	8.8E-46	5.5E-31	6.2E-02	3960

SNPs were included in the models in an additive genetic mode (dosage format). Linear models were adjusted for the discovery covariates and results were combined by inverse-variance weighted fixed-effects meta-analyses. See Appendix Table A.4 for details on the SNP data. Chr., chromosome; Dist., distance between CpG and SNP position; MAF, minor allele frequency.

cohort). Models were adjusted for the discovery covariates and technical covariates relevant for gene expression data (see Section 3.1.3). Associations were determined in each of the three studies separately, followed by inverse-variance weighted fixed-effects meta-analysis.

A total of 5569 transcripts were located in *cis* (±500 Mb) to the 187 CpG sites. Of these, 44 transcripts of 38 genes associated with DNA methylation at 31 CpGs without adjustment for BMI (Table 4.5), with minor changes after adjustment for BMI (not shown). The majority of the observed associations were negative. The strongest *cis*-signals were observed for cg09315878 with *TNFRSF4* expression ($p = 7.2 \times 10^{-86}$), cg09152259 with *MAP3K2* expression ($p = 1.6 \times 10^{-67}$) and cg14476101 with *PHGDH* expression ($p = 1.0 \times 10^{-64}$).

Table 4.4: EWAS of BMI – Discovery and replication results. Results are shown for the 187 CpG sites that were validated in the replication step ($p < 0.05$ in the replication step and $p < 10^{-7}$ in the joint meta-analysis).

						p-values			
CpG ID	**Chr.**	**Position**	**Nearest gene**	**# CpGs**	**Dir.**	Discovery	Replication	Joint	n
cg09315878	1	1152580	*SDF4*	1	-	3.2E-08	3.9E-04	5.7E-12	8843
cg11832534	1	3563998	*WRAP73*	1	+	1.3E-08	1.6E-06	7.2E-15	9587
cg08648047	1	11028561	*C1orf127*	0	+	7.2E-10	1.3E-03	1.1E-12	9593
cg03885055	1	16723232	*SPATA21*	0	-	1.3E-09	3.8E-03	1.4E-11	10260
cg12484113	1	27898757	*AHDC1*	0	+	1.2E-14	2.2E-09	1.8E-24	9565
cg16815882	1	35908609	*KIAA0319L*	0	+	6.0E-08	1.8E-02	2.6E-09	9593
cg17971578	1	36852463	*STK40*	0	-	4.6E-08	4.4E-04	1.4E-11	9511
cg27547344	1	43765617	*TIE1*	0	+	6.2E-08	3.7E-03	2.7E-10	9555
cg17901584	1	55353706	*DHCR24*	0	-	1.4E-10	4.2E-10	9.8E-21	9595
cg16594806	1	59473943	*PHBP3*	0	-	2.3E-18	1.3E-02	2.7E-18	9579
cg25001190	1	61668835	*NFIA*	0	-	8.1E-08	4.6E-05	1.8E-12	9509
cg03050965	1	101705237	*S1PR1*	0	-	3.0E-08	2.4E-02	3.6E-09	10260
cg03725309	1	109757585	*SARS*	0	-	3.8E-13	1.8E-07	1.6E-20	10251
cg14476101	1	120255992	*PHGDH*	1	-	2.1E-17	9.4E-13	3.7E-31	9554
cg10505902	1	144892111	*PDE4DIP*	0	-	1.4E-08	3.2E-06	1.4E-14	8073
cg22700686	1	153538764	*S100A2*	3	-	3.2E-12	1.4E-04	1.4E-16	8832
cg12593793	1	156074135	*LMNA*	1	-	2.9E-27	3.9E-09	9.3E-37	9582
cg25217710	1	156609523	*BCAN*	0	+	4.4E-13	3.4E-03	1.4E-14	10261
cg07202479	1	159174162	*DARC*	1	-	2.2E-09	4.0E-08	2.3E-17	10243
cg09554443	1	167487762	*CD247*	0	-	2.5E-10	2.5E-04	4.1E-14	9593
cg22534374	1	201511610	*RPS10P7*	0	-	1.6E-08	2.4E-05	2.1E-13	10254
cg10717869	1	205780912	*SLC41A1*	0	+	1.8E-10	9.1E-08	3.0E-18	9586
cg15323828	1	226053673	*TMEM63A*	0	-	4.0E-10	8.5E-03	9.7E-12	9577
cg01101459	1	234871477	*LINC00184*	1	+	3.1E-10	1.3E-05	2.1E-15	10258
cg02560388	2	11969958	*LPIN1*	0	-	4.7E-08	1.3E-04	4.3E-12	10260
cg04011474	2	28904455	*RNA5SP89*	0	-	3.7E-09	2.7E-06	4.0E-15	10257
ch.2.30415474F	2	30561970	*LBH*	0	-	7.7E-09	2.3E-04	6.9E-13	8824
cg16163382	2	37938640	*CDC42EP3*	0	-	5.8E-13	1.2E-03	1.5E-15	9579
cg26253134	2	70751721	*TGFA*	0	-	3.2E-12	2.3E-07	1.7E-19	10258
cg25570328	2	108903952	*SULT1C2*	0	-	6.0E-09	5.6E-03	9.8E-11	10260
cg09152259	2	128156114	*MAP3K2*	0	-	4.7E-08	4.1E-15	6.0E-22	9590
cg15357118	2	128927972	*UGGT1*	0	+	1.6E-08	1.3E-08	6.7E-17	9593
cg03327570	2	145304883	*ZEB2*	0	-	1.0E-11	1.5E-08	2.6E-20	10258
cg17178175	2	178109973	*NFE2L2*	0	-	5.1E-10	2.9E-06	4.0E-16	9587
cg09613192	2	181388538	*FTH1P20*	0	+	7.3E-09	2.9E-05	8.7E-14	9567
cg19373099	2	210008092	*CRYGFP*	0	+	5.7E-10	1.1E-05	3.4E-15	10118
cg00634542	2	219254588	*SLC11A1*	0	+	2.1E-08	5.8E-03	3.2E-10	10225
cg00144180	2	240294362	*HDAC4*	0	+	1.5E-08	1.5E-08	7.4E-17	10190
cg23032421	3	3152038	*IL5RA*	0	-	1.3E-08	1.4E-04	1.5E-12	10249
cg15681239	3	38080203	*DLEC1*	0	-	5.5E-08	3.3E-04	1.2E-11	9596
cg00138407	3	47386505	*KLHL18*	0	+	4.4E-08	8.4E-03	1.0E-09	10261
cg00108715	3	52565015	*NT5DC2*	0	+	4.0E-10	2.4E-04	1.0E-13	10259
cg22012981	3	58522689	*ACOX2*	0	+	3.4E-10	5.8E-04	3.1E-13	10250
cg10549088	3	64277154	*PRICKLE2*	0	+	7.7E-08	4.1E-02	3.1E-09	8077
cg12992827	3	101901234	*ZPLD1*	0	-	2.9E-10	2.2E-06	1.6E-16	9595
cg23232188	3	121556543	*EAF2*	0	+	4.0E-12	1.9E-05	5.0E-17	10255
cg16846518	3	128062608	*EEFSEC*	0	-	6.7E-08	1.8E-03	1.1E-10	9588
cg25197194	3	128758787	*EFCC1*	0	-	3.1E-08	4.4E-03	2.9E-10	10259
cg00673344	3	156807691	*LINC00880*	0	-	5.3E-08	1.5E-03	1.1E-10	10259
cg18098839	3	167742700	*GOLIM4*	0	-	2.7E-11	1.7E-09	6.0E-21	9581
cg15721584	3	181326755	*SOX2-OT*	2	+	4.9E-14	1.6E-08	6.9E-23	9541
cg10513161	3	183705727	*ABCC5*	0	+	1.5E-08	1.2E-03	1.7E-11	9581
cg06164260	3	187454439	*BCL6*	0	-	8.1E-18	4.9E-09	4.6E-27	10253
cg18513344	3	195531298	*MUC4*	0	-	6.0E-12	1.1E-07	1.4E-19	10252
cg10438589	4	14531493	*LINC00504*	0	+	6.5E-08	2.4E-04	9.5E-12	9585
cg26542660	4	56813860	*CEP135*	0	-	8.5E-08	4.4E-03	7.2E-10	10228
cg06690548	4	139162808	*SLC7A11*	0	-	3.0E-10	1.0E-12	6.7E-23	10247
cg11080651	5	10445523	*ROPN1L*	3	-	2.9E-09	9.1E-03	6.2E-11	9591
cg10179300	5	14147618	*TRIO*	0	+	9.1E-10	4.4E-07	9.7E-17	9593
cg04232128	5	138861241	*TMEM173*	0	-	3.0E-08	5.0E-05	9.6E-13	10257
cg26403843	5	158634085	*RNF145*	0	+	2.9E-15	5.0E-14	5.3E-30	10218
cg11927233	5	170816542	*NPM1*	0	+	1.3E-09	4.3E-06	1.7E-15	9577
cg02286155	5	176826262	*SLC34A1*	0	+	1.2E-10	9.3E-05	9.9E-15	10257
cg22590032	5	180050565	*FLT4*	0	+	2.5E-09	1.1E-04	2.1E-13	10258
cg09222732	6	466893	*EXOC2*	0	-	5.0E-08	2.1E-05	3.8E-13	8841
cg10975897	6	15504844	*JARID2*	0	-	8.2E-09	1.1E-04	4.5E-13	9566
cg00094412	6	29592854	*GABBR1*	0	-	1.1E-08	3.0E-06	1.0E-14	9513
cg13123009	6	31681882	*LY6G6F*	1	+	3.4E-09	1.2E-05	1.4E-14	9590
cg03957124	6	37016869	*COX6A1P2*	0	-	1.6E-09	8.5E-03	5.9E-11	10252
cg18120259	6	43894639	*C6orf223*	1	-	1.6E-09	3.2E-09	1.2E-18	10254
cg06012428	6	157477204	*ARID1B*	0	-	3.7E-09	2.1E-02	5.2E-10	10258
cg03940776	6	158490013	*SYNJ2*	0	-	3.0E-09	2.0E-08	1.7E-17	10253
cg17501210	6	166970252	*RPS6KA2*	0	-	8.2E-19	6.4E-21	6.5E-41	9594
cg05095590	7	2139259	*MAD1L1*	1	+	9.6E-10	4.1E-04	7.0E-14	6933
cg26804423	7	8201134	*ICA1*	0	+	2.4E-10	2.6E-03	2.0E-12	10261
cg24469729	7	27160520	*HOXA-AS2*	0	+	2.3E-08	8.6E-05	1.3E-12	10256
cg21429551	7	30635762	*GARS*	1	-	6.2E-10	2.8E-07	4.8E-17	10258
cg04577162	7	73667397	*RFC2*	0	+	2.6E-09	2.7E-04	7.1E-13	10260
cg19566658	7	100466241	*TRIP6*	0	+	5.9E-09	3.0E-03	3.8E-11	10254
cg22103219	7	101934892	*SH2B2*	1	-	7.0E-12	2.0E-04	2.1E-15	10256
cg05720226	7	116786597	*ST7*	0	+	1.4E-08	2.1E-04	2.5E-12	10261
cg27269962	7	127540997	*SND1*	0	+	1.5E-09	4.0E-03	9.8E-12	9532
cg25435714	7	157083381	*RN7SL142P*	0	+	5.3E-10	1.5E-02	6.0E-11	10255
cg24531955	8	23154691	*LOXL2*	0	-	3.2E-10	9.1E-09	6.0E-19	10260
cg19589396	8	103937374	*RPL5P24*	0	-	2.3E-10	2.2E-07	1.3E-17	10214
cg07471614	8	125855152	*LINC00964*	0	+	9.4E-09	3.2E-03	6.6E-11	10256
cg26952928	8	142230233	*SLC45A4*	2	+	3.6E-09	2.1E-02	4.9E-10	10257
cg26361535	8	144576604	*ZC3H3*	0	+	1.4E-10	8.7E-06	5.6E-16	10257
cg02716826	9	33447032	*AQP3*	0	-	1.6E-12	3.5E-10	5.9E-23	10248
cg13591783	9	75768868	*ANXA1*	0	-	9.0E-08	2.3E-05	9.5E-13	9592
cg14264316	9	134280803	*PRRC2B*	0	+	1.1E-08	3.9E-04	3.0E-12	9587
cg13781414	9	138951648	*NACC2*	1	-	4.2E-08	5.5E-04	1.7E-11	9570
cg19695507	10	13526193	*BEND7*	0	+	7.8E-10	9.3E-03	3.8E-11	10254
cg26033520	10	74004071	*ANAPC16*	0	+	1.6E-10	2.9E-09	8.3E-20	10191

Table 6.2 continued.

CpG ID	Chr.	Position	Nearest gene	# CpGs	Dir.	*p*-values Discovery	Replication	Joint	*n*
cg04126866	10	85932763	*C10orf99*	0	+	3.0E-08	1.3E-03	5.3E-11	10220
cg16578636	10	92987457	*PCGF5*	0	-	8.8E-08	3.9E-02	1.2E-08	9589
cg07504977	10	102131012	*LINC00263*	0	+	4.5E-09	1.7E-10	2.2E-19	10257
cg00431050	10	103985730	*ELOVL3*	1	-	1.1E-10	2.9E-03	1.1E-12	10257
cg26878209	10	112375475	*SMC3*	0	+	4.3E-08	8.0E-04	4.0E-11	10251
cg00244001	10	126336805	*FAM53B*	0	-	2.0E-12	3.4E-04	1.4E-15	10259
cg00238353	10	129785537	*PTPRE*	0	+	9.2E-08	2.4E-03	3.5E-10	10234
cg10927968	11	1807333	*CTSD*	0	+	3.9E-11	6.0E-06	6.7E-17	9590
cg06603309	11	2724144	*KCNQ1*	1	-	1.3E-11	5.3E-03	5.1E-13	10253
cg07136133	11	36422377	*PRR5L*	1	-	6.5E-11	2.3E-07	3.8E-18	10261
cg21108085	11	44591098	*CD82*	0	-	5.0E-08	2.3E-04	9.4E-12	10253
cg05648472	11	45232364	*PRDM11*	1	+	6.1E-11	3.4E-04	3.0E-14	10254
cg11376147	11	57261198	*SLC43A1*	0	-	5.8E-15	1.6E-04	1.0E-18	9591
cg03433986	11	62477624	*BSCL2*	0	-	4.5E-09	2.0E-04	8.2E-13	10260
cg00574958	11	68607622	*CPT1A*	1	-	3.3E-33	2.7E-18	1.2E-53	10252
cg09777883	11	112093696	*BCO2*	0	+	1.3E-08	1.1E-02	3.2E-10	9593
cg17260706	11	118782879	*BCL9L*	0	-	1.2E-15	4.7E-06	1.7E-21	9590
cg26894079	11	122954435	*CLMP*	0	-	2.2E-09	3.8E-10	2.2E-19	10254
cg22488164	12	14716910	*PLBD1*	0	+	2.5E-10	5.2E-06	5.7E-16	10238
cg06898549	12	41083590	*CNTN1*	0	+	1.8E-08	7.0E-06	5.9E-14	10241
cg06559575	12	53490352	*IGFBP6*	0	-	2.5E-08	2.0E-07	1.8E-15	10249
cg05845030	12	91573247	*DCN*	0	-	1.1E-09	5.9E-05	4.3E-14	10254
cg27117792	12	102330180	*DRAM1*	0	-	4.5E-08	5.1E-06	1.1E-13	10261
cg01511901	13	31004719	*UBE2L5P*	0	-	7.9E-08	3.3E-04	2.4E-11	10251
cg19750657	13	38935967	*UFM1*	0	+	2.2E-16	4.7E-11	5.2E-28	10258
cg26687842	13	41055491	*LINC00598*	0	+	2.6E-08	3.5E-05	5.3E-13	10260
cg11650298	13	44690989	*SMIM2-AS1*	0	-	1.7E-10	1.3E-04	1.2E-14	9584
cg19881557	14	20967426	*RNASE10*	0	+	3.7E-08	9.4E-07	1.4E-14	10250
cg03523676	14	24540235	*CPNE6*	0	+	3.5E-13	1.7E-03	3.8E-15	10261
cg13097800	14	47104140	*RPL10L*	0	-	1.4E-08	1.3E-02	4.3E-10	9592
cg26357885	14	65006204	*HSPA2*	0	-	5.7E-10	2.4E-02	1.5E-10	10251
cg10919522	14	74227441	*ELMSAN1*	0	-	7.5E-10	1.4E-06	2.7E-16	9584
cg19998073	14	89078443	*ZC3H14*	1	+	1.3E-09	9.2E-03	5.9E-11	10259
cg10814005	14	91711041	*GPR68*	0	-	5.1E-08	5.4E-03	3.9E-10	9584
cg25096107	14	106037781	*IGHA2*	0	-	7.0E-10	2.5E-03	6.5E-13	8077
cg10734665	15	26107410	*ATP10A*	0	-	1.3E-08	2.3E-03	3.2E-11	9594
cg27184903	15	29285727	*APBA2*	0	+	6.8E-08	8.0E-04	2.9E-11	8724
cg06192883	15	52554171	*MYO5C*	0	+	3.9E-09	8.7E-11	1.0E-19	10258
cg07037944	15	64290807	*DAPK2*	0	-	4.9E-11	1.3E-07	1.5E-18	10220
cg02119938	15	78505051	*ACSBG1*	0	-	2.9E-11	1.1E-09	4.9E-21	10251
cg07728579	15	83475013	*FSD2*	0	+	1.7E-09	2.0E-06	9.5E-16	9593
cg11183227	15	91455407	*MAN2A2*	0	+	8.0E-10	1.7E-03	1.6E-12	9587
cg27614723	15	92399897	*SLCO3A1*	0	+	2.7E-10	2.1E-02	2.9E-11	9559
cg00973118	16	374570	*AXIN1*	1	+	2.3E-08	1.6E-05	1.9E-13	10256
cg05063895	16	2073518	*SLC9A3R2*	0	-	6.2E-09	5.8E-03	1.1E-10	10258
cg06946797	16	11422409	*RMI2*	0	-	1.5E-13	6.0E-07	2.7E-20	10249
cg26663590	16	28959310	*NFATC2IP*	0	+	1.3E-11	3.3E-05	3.2E-16	10248
cg00711896	16	30410051	*ZNF48*	1	+	3.1E-09	2.0E-04	5.8E-13	10255
cg01243823	16	50732212	*NOD2*	2	-	1.7E-12	2.2E-12	3.3E-25	10259
cg00863378	16	56549757	*BBS2*	0	+	2.4E-08	5.7E-03	3.4E-10	10257
cg10922280	16	68034227	*DUS2L*	0	+	6.1E-13	4.2E-05	1.2E-17	9583
cg08305942	16	79692354	*MAF*	0	-	3.3E-08	1.3E-11	3.1E-19	9594
cg03159676	16	85600536	*GSE1*	0	+	3.5E-08	4.4E-07	6.0E-15	10261
cg07021906	16	87866833	*SLC7A5*	1	+	3.2E-10	4.4E-04	2.0E-13	10260
cg08443038	16	89006877	*CBFA2T3*	0	-	4.2E-08	3.2E-02	4.5E-09	9592
cg08726900	16	89550474	*ANKRD11*	1	-	2.0E-09	3.0E-02	3.2E-10	9476
cg09664445	17	2612406	*CLUH*	1	+	2.4E-16	1.2E-04	3.9E-20	9594
cg01798813	17	3906674	*ZZEF1*	0	+	6.1E-09	8.7E-06	2.5E-14	10251
cg19217955	17	7123994	*ACADVL*	0	-	4.3E-08	2.4E-04	8.8E-12	10258
cg22695339	17	7791630	*CHD3*	0	-	6.3E-08	2.0E-04	1.0E-11	10254
cg11024682	17	17730094	*SREBF1*	0	+	1.9E-23	3.0E-25	1.3E-50	9592
cg16611584	17	19809078	*AKAP10*	0	+	1.0E-11	6.3E-04	7.4E-15	9588
cg25649826	17	20938740	*USP22*	0	+	8.9E-09	4.2E-04	4.0E-12	10260
cg13274938	17	38493822	*RARA*	0	+	9.1E-09	6.1E-06	2.0E-14	9567
cg08857797	17	40927699	*VPS25*	3	+	3.3E-14	2.3E-03	3.3E-16	9556
cg18219562	17	41773643	*MEOX1*	1	+	5.7E-09	3.5E-04	2.1E-12	10260
cg27050612	17	46133198	*NFE2L1*	0	-	4.3E-09	3.3E-04	1.4E-12	10254
cg02650017	17	47301614	*PHOSPHO1*	0	-	1.9E-16	2.0E-08	5.7E-25	10250
cg24174557	17	57903544	*VMP1*	0	-	2.4E-12	3.0E-08	1.3E-20	10217
cg08813944	17	71258589	*CPSF4L*	0	+	7.6E-08	2.4E-03	8.1E-11	8043
cg14020176	17	72764985	*SLC9A3R1*	0	+	4.6E-08	3.7E-03	2.1E-10	9593
cg21486834	17	74477542	*RHBDF2*	0	+	6.4E-12	6.2E-03	3.5E-13	10259
cg18181703	17	76354621	*SOCS3*	3	-	2.8E-20	1.6E-16	3.6E-38	10258
cg11202345	17	76976057	*LGALS3BP*	4	+	2.3E-11	1.9E-06	9.6E-18	9586
cg11969813	17	79816559	*P4HB*	5	+	1.6E-11	2.6E-05	1.7E-16	9580
cg18608055	19	1130866	*SBNO2*	5	-	5.9E-13	4.7E-06	1.4E-18	10231
cg04524040	19	4153364	*CREB3L3*	0	-	4.1E-08	7.4E-03	4.6E-10	9505
cg07769588	19	10655622	*ATG4D*	2	+	1.3E-12	2.2E-05	1.2E-17	9566
cg13922488	19	14545201	*PKN1*	0	+	4.8E-08	2.5E-06	4.5E-14	9591
cg24679890	19	17246356	*MYO9B*	0	+	8.3E-11	1.0E-04	7.9E-15	10260
cg07682160	19	18959935	*UPF1*	0	+	1.5E-09	1.0E-03	1.5E-12	9575
cg26836479	19	42706353	*DEDD2*	0	-	1.5E-08	6.5E-03	1.5E-10	9504
cg27087650	19	45255796	*BCL3*	1	-	2.1E-09	3.2E-02	6.8E-10	10243
cg02711608	19	47287964	*SLC1A5*	1	-	5.6E-11	6.0E-09	5.9E-20	10258
cg11614585	20	897050	*ANGPT4*	0	+	8.9E-08	1.1E-02	1.8E-09	9586
cg18217136	20	36157651	*PPIAP3*	0	+	3.8E-08	1.1E-03	5.0E-11	10259
cg24403644	20	42574624	*TOX2*	0	+	5.9E-08	2.7E-10	7.9E-18	10260
cg08309687	21	35320596	*LINC00649*	1	-	1.1E-13	2.9E-05	1.5E-18	9595
cg06500161	21	43656587	*ABCG1*	4	+	6.0E-46	1.1E-25	2.0E-75	10260
cg08548559	22	31686097	*PIK3IP1*	0	-	3.4E-09	4.9E-07	4.6E-16	9570
cg27115863	22	37921640	*CARD10*	0	-	1.1E-08	7.6E-10	2.8E-18	10260
cg03318904	22	39801522	*TAB1*	2	+	2.2E-12	2.6E-08	9.5E-21	10254
cg09349128	22	50327986	*CRELD2*	1	-	1.7E-20	7.3E-13	1.2E-34	9594

Linear models were adjusted for age, sex, physical activity, smoking status, alcohol intake, estimated white blood cell proportions as well as the first 20 control probe PCs, and meta-analyzed using *z*-score based fixed-effects meta-analysis. # CpGs, number of further significant CpGs in locus; Chr., chromosome; Dir., effect direction.

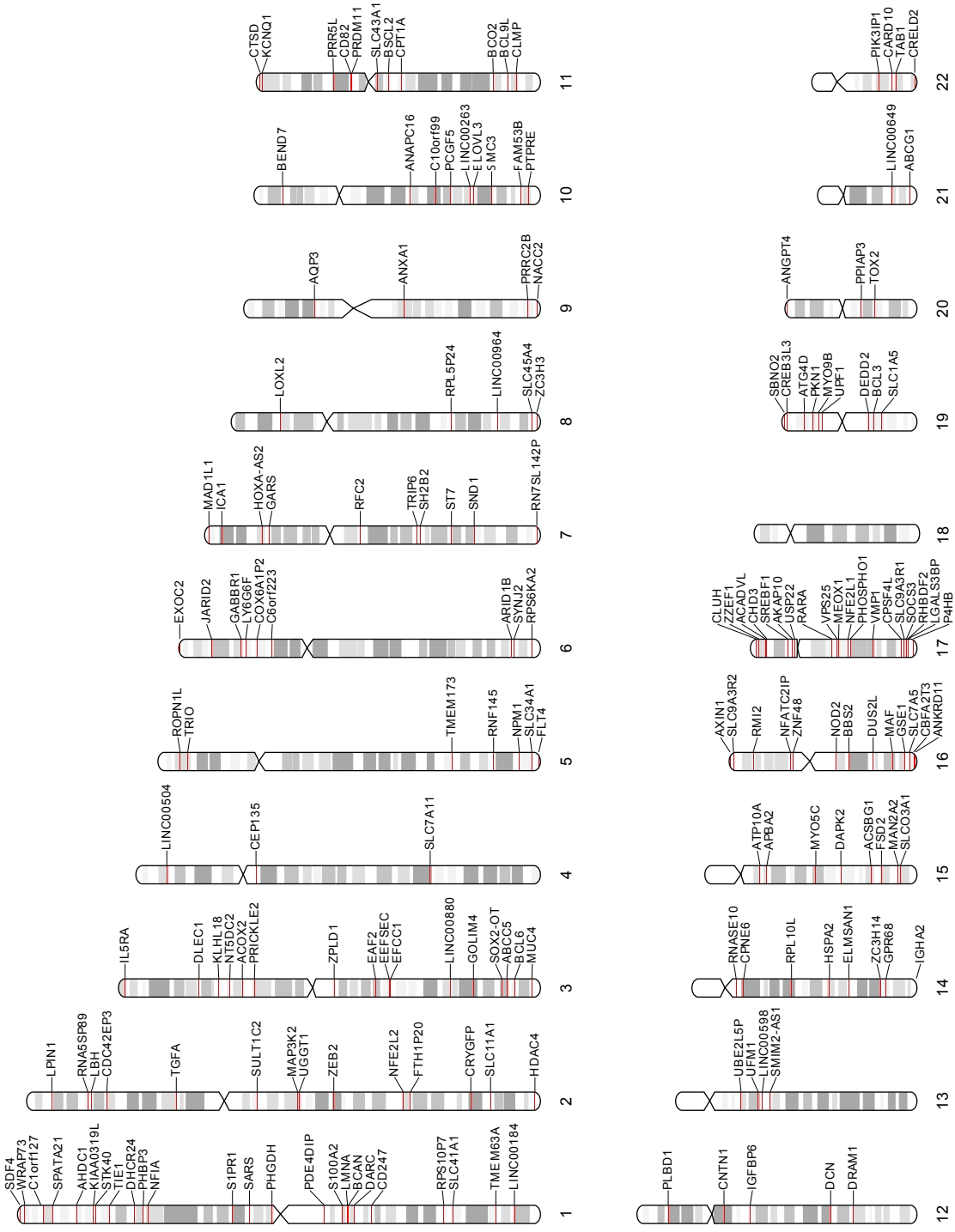

Figure 4.8: Karyogram visualizing the genomic positions of the 187 validated CpGs associated with BMI.

4.2.4 Functional genomics

The identified markers were studied with regard to (1) their location in relation to functional genomic features, (2) gene sets defining known pathways, and (3) published GWAS

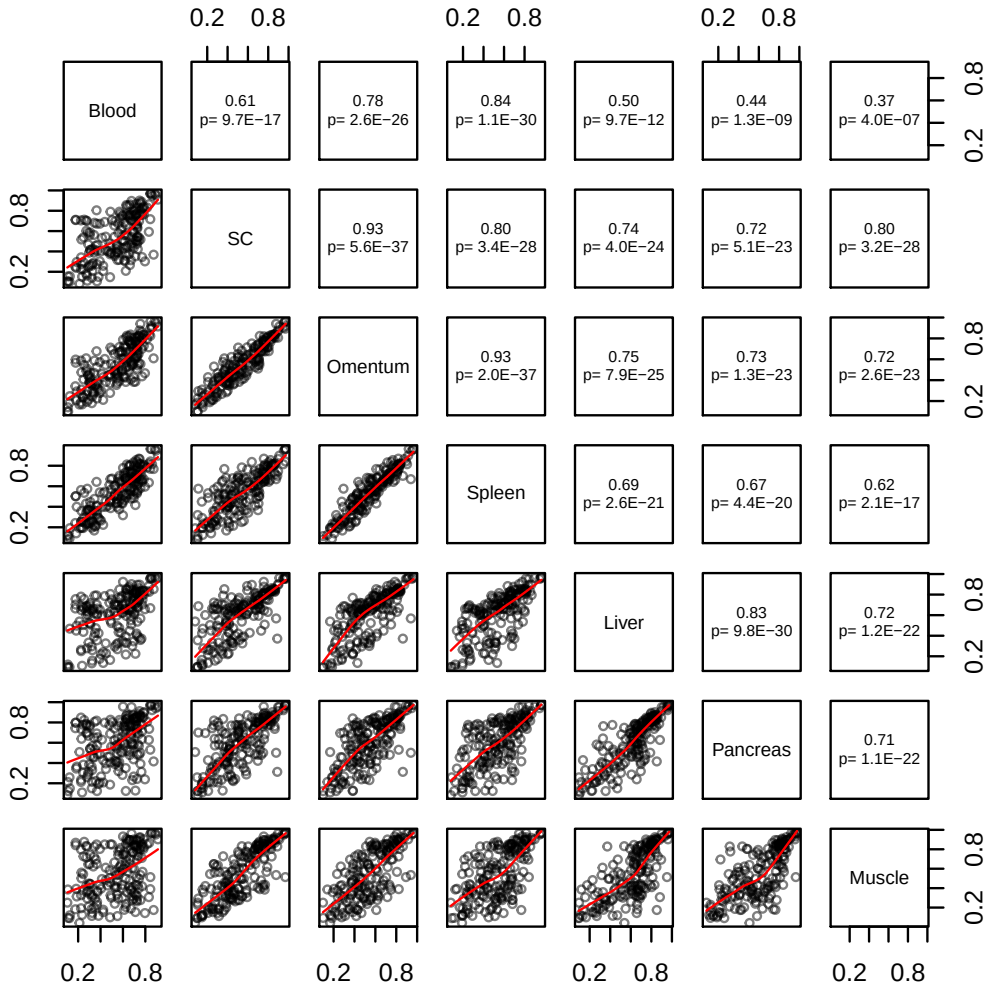

Figure 4.9: Cross-tissue correlation of DNA methylation. Matrix of scatter plots (lower triangle) and Pearson's correlation (upper triangle; with permutation p-value using 10,000 permutations, see Section 3.3.3) of DNA methylation at the 187 validated BMI-associated CpGs between different tissues (diagonal). Tissue type is denoted in the diagonal.

loci. Enrichment analysis (see Section 3.5.1) for functional genomic features included relationship to regulatory genomic sites (retrieved from the UCSC database, Ram *et al.* (2011)), and to CpG islands and location in/near genes (retrieved from Bibikova *et al.* (2011)).

The 187 loci were significantly enriched with respect to open chromatin sites (Figure 4.10A), including DNase hypersensitive sites ($p = 7.1 \times 10^{-8}$), enhancers ($p = 2.2 \times 10^{-10}$) and the histone modifications *histone H3 lysine 4 monomethylation* (h3k4me1) ($p = 2.6 \times 10^{-19}$) and *histone H3 lysine 27 acetylation* (h3k27ac) ($p = 2.3 \times 10^{-5}$), which mark open chromatin at active promoters and enhancers (Ram *et al.*, 2011). Only a trend of

enrichment was observed for the activating histone mark h3k4me3 ($p = 0.015$). The CpGs were also enriched in "open sea" locations ($p = 4.3 \times 10^{-10}$) but markedly depleted in CpG islands ($p = 5.6 \times 10^{-26}$) (Figure 4.10B). In addition, a significant enrichment within gene bodies ($p = 1.0 \times 10^{-8}$ and depletion at transcriptional start sites (TSS200, $p = 2.2 \times 10^{-9}$;

Table 4.5: Significant *cis*-associations with gene expression. Significance level was 9.0×10^{-6}, corresponding to Bonferroni correction for 5569 transcripts in *cis* (± 500 Mb) to the 187 BMI-associated CpGs. Only the strongest association per CpG-gene pair is shown.

CpG ID	Chr.	Position	Nearest gene	Transcript ID	Gene	Dir.	*p*-value	Relation of CpG to gene**
cg09315878	1	1152580	*SDF4*	ILMN_2112256	*TNFRSF4*	-	7.2E-86	downstream
cg09315878	1	1152580	*SDF4*	ILMN_2349633	*TNFRSF18*	-	7.2E-15	downstream
cg17901584	1	55353706	*DHCR24*	ILMN_1725510	*DHCR24*	-	2.7E-09	promoters
cg14476101	1	120255992	*PHGDH*	ILMN_1704537	*PHGDH*	-	1.0E-64	introns
cg09152259	2	128156114	*MAP3K2*	ILMN_1765122	*MAP3K2*	-	1.6E-67	downstream
cg09613192	2	181388538	*FTH1P20*	ILMN_2390338	*UBE2E3*	-	5.3E-12	upstream
cg23032421	3	3152038	*IL5RA*	ILMN_1756455	*IL5RA*	-	1.1E-18	exons/promoters
cg04232128	5	138861241	*TMEM173*	ILMN_1745256	*CXXC5*	+	1.4E-09	upstream
cg00094412	6	29592854	*GABBR1*	ILMN_2395375	*GABBR1*	-	7.5E-06	introns
cg13123009	6	31681882	*LY6G6F*	ILMN_1654566	*HSPA1L*	-	5.4E-06	upstream
cg03957124	6	37016869	*COX6A1P2*	ILMN_2123450	*FLJ43093*	+	1.4E-06	downstream
cg24469729	7	27160520	*HOXA-AS2*	ILMN_1657129	*SKAP2*	-	1.3E-10	downstream
cg24469729	7	27160520	*HOXA-AS2*	ILMN_1753613	*HOXA5*	-	7.4E-18	upstream
cg19589396	8	103937374	*RPL5P24*	ILMN_1656682	*AZIN1*	+	2.4E-06	downstream
cg13591783	9	75768868	*ANXA1*	ILMN_2184184	*ANXA1*	-	1.4E-08	introns
cg07136133	11	36422377	*PRR5L*	ILMN_1697491	*PRR5L*	-	9.2E-07	introns/promoters
cg21108085	11	44591098	*CD82*	ILMN_1699980	*TSPAN18*	-	7.8E-06	upstream
cg25096107	14	106037781	*IGHA2*	ILMN_1707491	*KIAA0125*	-	7.0E-06	upstream
cg25096107	14	106037781	*IGHA2*	ILMN_3239445	*ZBTB42*	+	4.0E-06	downstream
cg07037944	15	64290807	*DAPK2*	ILMN_1791847	*DAPK2*	+	3.5E-07	introns
cg11183227	15	91455407	*MAN2A2*	ILMN_1693650	*FES*	-	3.8E-06	downstream
cg00973118	16	374570	*AXIN1*	ILMN_1741371	*TMEM8A*	+	2.5E-10	upstream
cg00711896	16	30410051	*ZNF48*	ILMN_2125747	*LOC606724*	-	4.4E-06	downstream
cg00711896	16	30410051	*ZNF48*	ILMN_2179726	*C16ORF93*	-	1.8E-11	upstream
cg00863378	16	56549757	*BBS2*	ILMN_2230035	*BBS2*	-	2.5E-27	introns
cg10922280	16	68034227	*DUS2L*	ILMN_1689160	*DPEP2*	-	1.7E-15	promoters
cg10922280	16	68034227	*DUS2L*	ILMN_1741736	*DDX28*	-	6.8E-06	upstream
cg10922280	16	68034227	*DUS2L*	ILMN_1811650	*DUS2L*	-	2.2E-11	upstream
cg01798813	17	3906674	*ZZEF1*	ILMN_1668984	*SPNS3*	-	2.0E-30	upstream
cg01798813	17	3906674	*ZZEF1*	ILMN_1807719	*CTNS*	-	5.2E-08	downstream
cg11024682	17	17730094	*SREBF1*	ILMN_1663035	*SREBF1*	-	6.0E-07	introns
cg16611584	17	19809078	*AKAP10*	ILMN_1718808	*AKAP10*	-	2.2E-07	exons
cg08813944	17	71258589	*CPSF4L*	ILMN_1745223	*CDC42EP4*	-	8.0E-07	upstream
cg14020176	17	72764985	*SLC9A3R1*	ILMN_2112357	*CD300LF*	-	1.1E-09	downstream
cg11202345	17	76976057	*LGALS3BP*	ILMN_1659688	*LGALS3BP*	-	8.6E-17	exons/promoters
cg07769588	19	10655622	*ATG4D*	ILMN_2073184	*S1PR5*	-	4.1E-06	downstream
cg06500161	21	43656587	*ABCG1*	ILMN_1794782	*ABCG1*	-	1.2E-18	introns
cg08548559	22	31686097	*PIK3IP1*	ILMN_1651429	*SELM*	-	9.1E-07	downstream

*Nearest gene according to ensembl annotation. **Genomic features according to RefSeq annotation retrieved from the UCSC database. Overlaps with the CpG site positions were determined using the R package *GenomicRanges*, version 1.14.4. Pairwise association testing of a transcript and a CpG site was performed using linear models with gene expression as response variable, and DNA methylation as well as discovery covariates and technical factors as independent variables (see Section 3.1.3). Results were combined using inverse-variance weighted fixed-effects meta-analysis.

TSS1500, $p = 7.2 \times 10^{-3}$) was observed (Figure 4.10B).

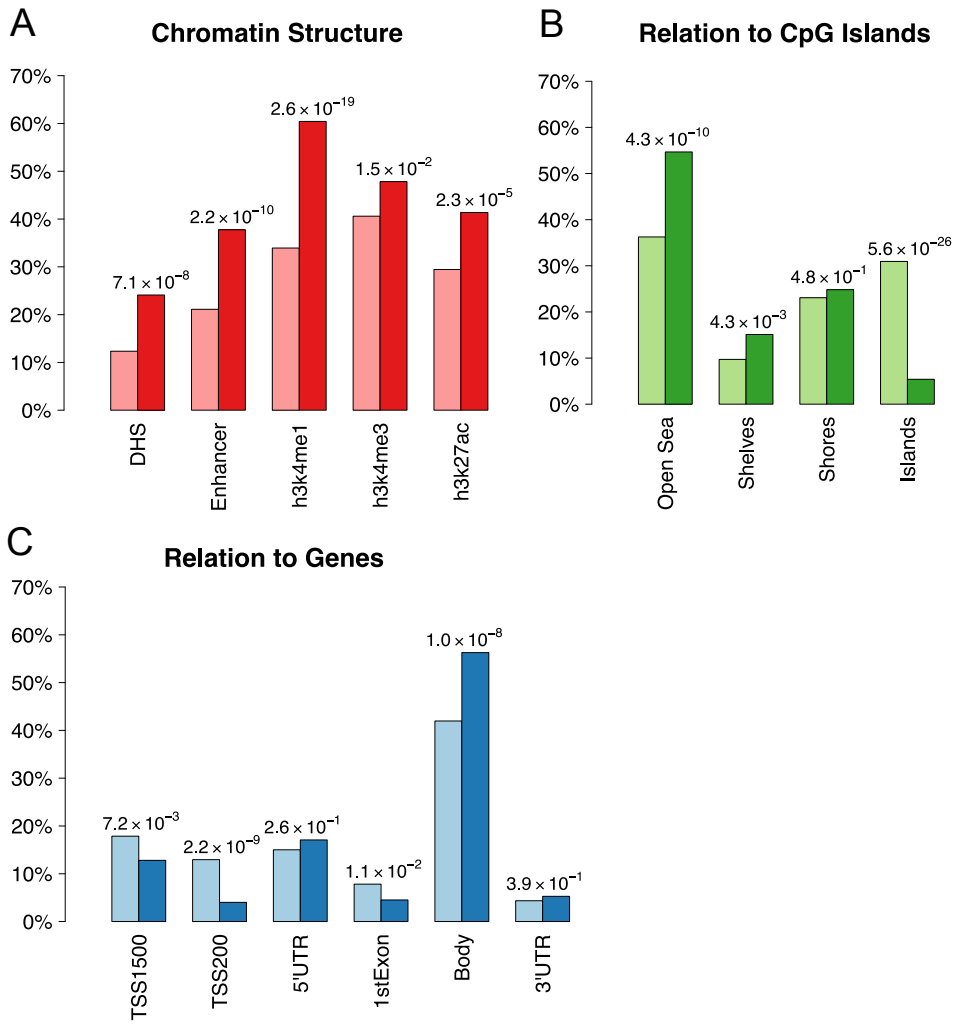

Figure 4.10: Enrichment of the BMI-associated CpGs within functional genomic sites. Enrichment was determined using Fisher's exact test (see Section 3.5). Left and right bars represent sites not associated and associated with BMI, respectively. DHS, DNase hypersentive sites; h3k4me1, histone H3 lysine 4 monomethylation; h3k4me3, histone H3 lysine 4 trimethylation; h3k27ac, histone H3 lysine 27 acytylation; TSS, transcription start site; UTR, untranslated region.

4.2.5 Candidate genes at the identified loci

Genes were prioritized as likely candidates underlying the observed methylation-BMI associations at the 187 loci using the following criteria: (1) distance $\leq$ 40 kb to the CpG position, since associations of methylation with expression have been reported to typically range about $\pm$ 40 kb distance between gene and CpG position (Liu *et al.*, 2013), and (2) distance $\leq$ 500 kb to the CpG position with expression associated with methylation (see Section 4.2.3 above). This yielded a combined list of 546 unique genes.

The *gene set enrichment analysis* (GSEA) MSigDB platform (see Section 3.5.1) was used to explore enrichment of these genes against a set of curated pathway sets, including Bio-

Carta, KEGG, Pathway Interaction Database, Reactome, SigmaAldrich, Signaling Gateway, Signal Transduction KE and SuperArray. This revealed significant enrichment for genes involved in metabolic signaling (lipid metabolism, insulin signaling) and transmembrane transport (solute carrier protein, binding cassettes) as well as hemostasis and autophagy (Table 4.6).

Table 4.6: Enrichment of the BMI-associated CpGs for biological pathways.

Pathway	Origin	p-value	FDR-corrected p-value
Transmembrane transport of small molecules	REACTOME	1.9E-06	2.5E-03
Direct p53 effectors	PID	4.4E-05	2.0E-02
Metabolism of lipids and lipoproteins	REACTOME	4.6E-05	2.0E-02
GPCR ligand binding	REACTOME	1.3E-04	3.4E-02
Hemostasis	REACTOME	1.3E-04	3.4E-02
Amino acid transport across the plasma membrane	REACTOME	1.8E-04	3.7E-02
Transport of inorganic cations, anions and amino acids/oligopeptides	REACTOME	2.4E-04	3.7E-02
Fc-epsilon receptor I signaling in mast cells	PID	2.7E-04	3.7E-02
Insulin signaling pathway	KEGG	3.0E-04	3.7E-02
Regulation of autophagy	KEGG	3.0E-04	3.7E-02

Enrichment analysis was conducted using the gene set enrichment analysis (GSEA) MSigDB platform. p-values were derived from Fisher's exact test (Section 3.5.1). KEGG, Kyoto Encyclopedia of Genes and Genomes; PID, pathway interaction database.

A separate enrichment analysis was conducted for genes related to clinical traits in published GWAS using *Meta-Analysis Gene-set Enrichment of variaNT Associations* (MAGENTA, Segrè *et al.* (2010), see Section 3.5.1). This revealed a significant enrichment for genes related to LDL cholesterol ($p = 3.0 \times 10^{-3}$) and waist-hip ratio (WHR) adjusted for BMI ($p = 8.6 \times 10^{-3}$) but not for genes related to BMI ($p = 0.678$) (Table 4.7).

4.2.6 DNA methylation is influenced by DNA sequence variation

Using genome-wide SNP data imputed to the 1000G reference panel from a subset of 3961 individuals from the EpiMigrant and KORA F3 as well as F4 cohorts (see Appendix Table A.4), a search for *cis*-located (±1 Mb) genetic variants influencing methylation at the identified CpG sites was performed. Associations between SNPs and methylation were determined using linear models with methylation as response and SNP, discovery covariates and the first 5 PCs derived from the genomic data in EpiMigrant to correct for population stratification as independent variables. Results were combined by inverse-variance weighted fixed-effects meta-analysis. BMI as a covariate did not affect the results and was therefore omitted. A total number of 867,921 CpG-SNP pairs, corresponding to 175 CpGs and 825,286 SNPs, was retrieved after QC (see Section 3.1 and Appendix Table A.4; 9 CpGs were excluded due to associated SNPs in the probe-binding area, and 3 due to lack of common *cis*-SNPs in the cohorts), of which 29,807 pairs, corresponding to 125

Table 4.7: Enrichment of the BMI-associated CpGs for previously published GWAS loci.

GWAS (Consortium)	p-value	FDR-corrected p-value	Expected	Observed
BMI (GIANT)	6.8E-01	1	12	11
T2D (DIAGRAM)	6.6E-02	7.3E-01	12	17
Fasting Glucose (MAGIC)	3.7E-01	1	12	13
Fasting Insulin (MAGIC)	3.7E-01	1	12	13
HDL (LIPIDS)	4.7E-01	1	13	13
HOMA_B (MAGIC)	4.3E-01	1	11	12
HOMA_IR (MAGIC)	9.4E-01	1	11	7
LDL (LIPIDS)	3.0E-04	3.3E-03	12	25
Total Cholesterol (LIPIDS)	2.5E-01	1	12	14
Triglycerides (LIPIDS)	8.6E-01	1	12	9
WHR adjusted for BMI (GIANT)	8.6E-03	9.5E-02	12	21

Enrichment analysis was conducted using *Meta-Analysis Gene-set Enrichment of variaNT Associations* (MAGENTA, Segrè *et al.* (2010)). p-values were determined by comparing the number of top 5% genes of the ranked list in the gene set ("observed") vs. the respective number in permuted gene sets of the same size ("expected"). The permutation test was based on 10,000 permutations.

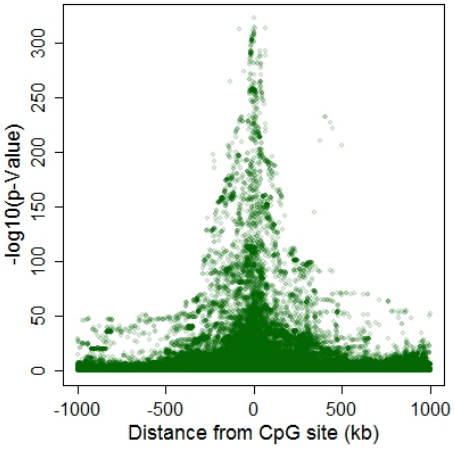

Figure 4.11: *cis*-associations of genetic variants with methylation. Genomic distance (in kb) between SNP and CpG is plotted against -$\log_{10}$(p-value) of the respective association. Only SNPs with a minor allele frequency of $\geq$ 1% were considered. See Appendix Table A.4 for additional quality criteria.

CpGs, showed significant association at $p < 5.8 \times 10^{-8}$ (with non-significant p-value of the heterogeneity test). Strength of association tended to relate to the physical proximity of a SNP to a CpG (Figure 4.11).

4.2.7 Causality and direction of the observed methylation-BMI associations

Three approaches were employed towards deciphering the causality and direction of the observed associations. First, an *ad hoc* Mendelian randomization (MR) approach was applied (see Section 3.5.2), second, longitudinal associations between DNA methylation and change in BMI were studied, and third, for selected CpGs, a formal MR experiment was conducted (see Section 3.5.2).

Methylation causal to BMI

The chosen *ad hoc* MR approach relies on the assumption that given methylation at a CpG causally affects BMI, a SNP that strongly associates with the CpG should show an association with BMI ("observed") similarly as strong as the effect predicted from the product of the effect sizes between SNP and CpG, and between CpG and BMI. To compare observed and predicted effects, observed SNP-BMI associations were retrieved from a previously published large GWAS of the GIANT consortium ($n \approx 100,000$, Speliotes *et al.* (2010)), whereas predicted effects were determined from the available data, using the *cis*-SNP with the strongest association with the respective CpG (*cis*-SNPs identified for the 175 eligable CpGs, see Section 4.2.6 above), and meta-analyzing results from different cohorts using inverse-variance weighted fixed-effects meta-analysis. To determine significance, p-values were first multiplied by the number of SNPs allocated to the CpG, followed by Bonferroni correction for the number of CpGs. Using this approach, a single CpG (cg26663590, upstream of *NFATC2IP*) showed evidence of being causal to BMI (Figure 4.12A).

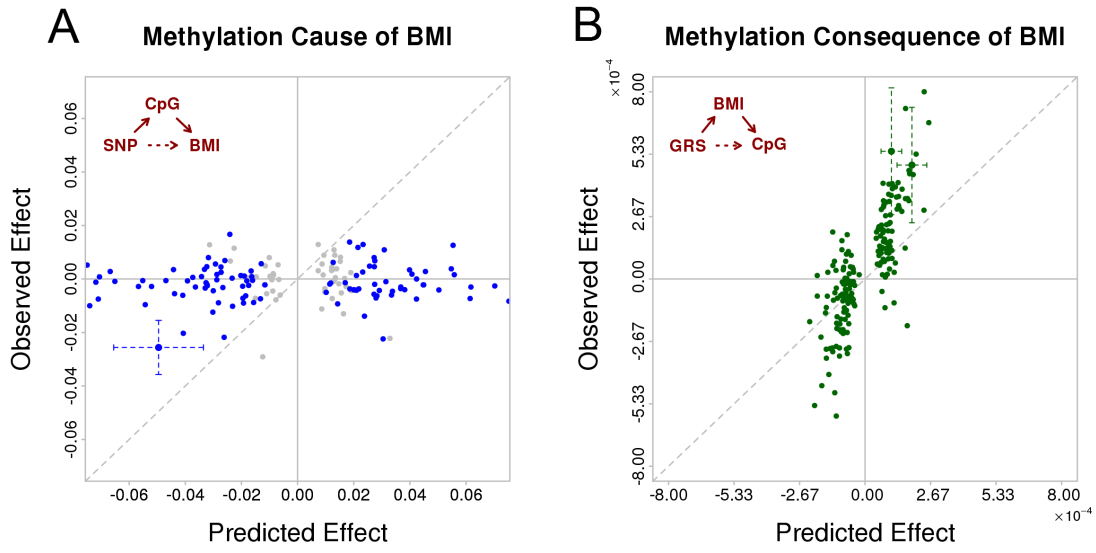

Figure 4.12: Causality of the BMI-methylation associations. *Ad hoc* MR approach. **A** CpG causal to BMI: Plot of predicted versus observed effect sizes for SNP-BMI associations, where the SNP from the *cis*-area of the respective CpG was chosen that associated most strongly with the CpG. Grey, SNP-CpG association not significant; blue, SNP-CpG association significant. For a single CpG (cg26663590), a significant SNP-BMI association was also observed; 95% confidence intervals of predicted and observed effects are shown. **B** CpG consequential to BMI: Plot of predicted versus observed effect sizes for genomics risk score (GRS)-CpG association. Two SNPs (cg00138407 and cg06500161) were significantly associated with the GRS; 95% confidence intervals of predicted and observed effects are shown.

Next, association of methylation with subsequent annual percentage change in BMI over a 7-year follow-up period was assessed in a subset of 2948 participants from EpiMigrant ($n = 1513$) and KORA F4 ($n = 1435$), using linear models adjusted for the discovery covariates and baseline BMI, followed by inverse-variance weighted fixed-effects meta-

analysis. Methylation at none of the 187 CpGs showed statistically significant evidence of causing change in BMI. However, the direction of effect was confirmed for cg26663590 (*NFATC2IP* locus, $p = 0.018$) (Table 4.8). In addition, two further CpG sites showed nominal significance in both MR and the longitudinal approach, namely cg00634542 (*SLC11A1* locus) and cg00711896 (*ZHF48* locus).

Table 4.8: Causality of the methylation-BMI associations. Shown are CpGs with significant evidence for being causal or consequential to BMI from either *ad hoc* MR approach or longitudinal analysis, or nominal significance in both. Results from formal MR are also shown, where available. Note that for the consequential direction, consistency of effect sizes between the *ad hoc* MR approach (predicted/observed GRS-CpG association) and the longitudinal analysis (ΔBMI-CpG association) is required for results to support each other, since GRS was formed such that is relates positively with BMI. A similar consistency is not required for the causal direction.

				Predicted		Observed		Disc. Longitudinal			MR	
CpG ID	**Chr.**	**Pos.**	**Nearest gene**	**Dir.**	***p*-value**	**Dir.**	***p*-value**	**Dir.**	**Dir.**	***p*-value**	**Dir.**	***p*-value**
Causal direction												
cg00634542	2	219254588	SLC11A1	+	1.7E-05	+	1.1E-02	+	+	1.6E-02	+	2.5E-01
cg26663590	16	28959310	NFATC2IP	-	1.3E-09	-	9.6E-07*	+	+	1.8E-02		
cg00711896	16	30410051	ZNF48	+	2.4E-03	+	7.7E-03	+	+	4.5E-03		
Consequential direction												
cg08648047	1	11028561	C1orf127	+	9.7E-06	+	4.1E-03	+	+	7.6E-03	+	7.2E-01
cg16815882	1	35908609	KIAA0319L	+	3.2E-05	+	2.6E-02	+	+	1.7E-03		
cg17901584	1	55353706	DHCR24	-	1.2E-08	-	8.6E-03	-	-	1.9E-02	+	7.5E-01
cg00138407	3	47386505	KLHL18	+	6.3E-07	+	8.0E-05*	+	+	1.4E-03	+	4.4E-01
cg10549088	3	64277154	PRICKLE2	+	2.5E-06	+	1.3E-02	+	+	1.7E-02	+	3.2E-01
cg16846518	3	128062608	EEFSEC	-	6.1E-05	-	3.7E-02	-	-	2.3E-03	-	1.1E-01
cg10438589	4	14531493	LINC00504	+	1.1E-05	+	1.2E-01	+	+	1.1E-04*	+	8.6E-02
cg06690548	4	139162808	SLC7A11	-	3.0E-08	-	2.0E-01	-	-	4.6E-05*		
cg27269962	7	127540997	SND1	+	3.7E-05	+	2.2E-02	+	+	1.9E-02		
cg04126866	10	85932763	C10orf99	+	1.5E-05	+	6.4E-03	+	+	1.6E-02		
cg16578636	10	92987457	PCGF5	-	1.2E-06	-	6.2E-01	-	-	7.0E-05*	-	3.8E-01
cg06603309	11	2724144	KCNQ1	-	4.2E-07	-	5.4E-01	-	-	3.5E-05*	-	4.0E-01
cg09777883	11	112093696	BCO2	+	3.2E-06	+	8.7E-03	+	+	3.6E-02	+	1.3E-01
cg26687842	13	41055491	LINC00598	+	1.2E-05	+	9.8E-03	+	+	3.8E-02	+	3.2E-01
cg25096107	14	106037781	IGHA2	-	1.1E-06	-	2.5E-02	-	-	1.2E-04*	-	9.4E-01
cg06192883	15	52554171	MYO5C	+	3.3E-08	+	4.5E-02	+	+	7.1E-04		
cg09664445	17	2612406	CLUH	+	6.2E-08	+	1.8E-02	+	+	5.8E-04	+	6.4E-02
cg11024682	17	17730094	SREBF1	+	5.7E-09	+	2.1E-02	+	+	9.6E-06*	+	3.0E-01
cg08857797	17	40927699	VPS25	+	1.0E-08	+	3.1E-02	+	+	1.1E-02	+	5.1E-01
cg27087650	19	45255796	BCL3	-	2.2E-05	-	7.8E-03	-	-	1.3E-03		
cg18217136	20	36157651	PPIAP3	+	7.9E-06	+	9.0E-03	+	+	6.3E-03	+	1.8E-01
cg24403644	20	42574624	TOX2	+	3.4E-05	+	2.4E-02	+	+	2.3E-02		
cg08309687	21	35320596	LINC00649	-	1.1E-07	-	3.8E-02	-	-	1.9E-02		
cg06500161	21	43656587	ABCG1	+	7.1E-10	+	1.1E-04*	+	+	4.2E-08*	+	6.8E-02

* Significant after Bonferroni correction. Chr., chromosome; Disc., discovery study; Dir., effect direction; MR, Mendelian randomization; Pos., genomic position.

Finally, the three mentioned CpGs with either nominal significance in both approaches or Bonferroni significance in either approach were further examined in a formal MR experiment based on an instrumental variable *two-stage least squares* (TSLS) approach (see Section 3.5.2). As a prerequisite for TSLS, absence of an independent effect of the re-

spective *cis*-SNP with BMI was tested. If the hypothesis of an association could not be significantly rejected, the respective CpG was excluded from further steps. Furthermore, to avoid weak instrument bias, single cohorts were only included in meta-analysis of TSLS results if the F statistic of the SNP-CpG association was at least 10 (see Section 3.5.2 for details). Causality was not confirmed for cg00634542 (*SLC11A1* locus, $p = 0.25$); however, when the weak instrument requirement $F \geq 10$ was relaxed to $F \geq 5$, p-value became significant ($p = 2.7 \times 10^{-3}$). cg26663590 (*NFATC2IP* locus) was excluded from MR analysis since absence of an independent association of the SNP with BMI could not be excluded ($p = 0.083$). cg00711896 (*ZHF48* locus) was excluded from TSLS due to weak instrument ($F < 10$) in all cohorts. At $F \geq 5$, the effect was not significant ($p = 0.85$).

Together, little evidence for CpGs being causal to BMI was obtained, although power for a formal MR approach seemed to be insufficient.

Methylation consequential to BMI

For the *ad hoc* MR approach, a *genomic risk score* (GRS) was defined as the sum of expected risk alleles from the SNPs previously reported to associate with BMI (Speliotes *et al.*, 2010) (less one SNP, rs7359397, significantly associated with one of the 187 CpGs independent of BMI, see Section 3.5.2 for instrument requirements) weighted with the effect sizes derived from the same publication. In accordance to the description above for the causal direction, observed GRS-CpG associations were then compared with predicted associations derived as the product of the GRS-BMI and BMI-CpG associations (with BMI inverse normal transformed to match Speliotes *et al.* (2010)). Across all CpGs, observed versus predicted GRS-CpG associations were strongly correlated (Pearson's $\rho = 0.79$, $p = 6.4 \times 10^{-42}$), suggesting that altered methylation is a consequence of BMI at the majority of the identified CpGs (Figure 4.12B). Specifically, after correction for multiple comparisons, methylation at two CpGs, cg00138407 (*KLHL18* locus, $p = 8.0 \times 10^{-5}$) and cg06500161 (*ABCG1* locus, $p = 1.1 \times 10^{-4}$), showed significant evidence for being consequential to BMI.

When association of CpGs with previous change in BMI was investigated in a subgroup of $n = 1698$ KORA S4/F4 subjects, adjusted for the discovery covariates as well as baseline BMI, and changes in behavioral factors and white blood cell proportions during follow-up, seven CpGs showed significant association (Table 4.8). An additional trend for association was observed for 76 CpGs ($p < 0.05$), corresponding to a strong enrichment of nominally significant p-values (permutation $p < 1 \times 10^{-5}$).

All significant CpGs in the two approaches showed consistency of effect direction in the respective other approach. In addition, a strong correlation of effects between both approaches was observed (Pearson's $\rho = 0.67$, $p = 3.4 \times 10^{-26}$), which was not observed for the causal direction (ρ = -0.08, $p = 0.287$).

When formal MR was applied to the 24 CpGs showing Bonferroni significance in either approach or nominal significance in both approaches, requirements for MR were met for 16 CpGs, for none of which the consequential direction was confirmed. However, when the weak instrument requirement was relaxed to $F \geq 5$, nominal significance was observed for four CpGs. These included cg06500161 (*ABCG1* locus, $p = 0.002$), cg11024682 (*SREBF1* locus, $p = 0.042$), cg09664445 (*CLUH* locus, $p = 0.048$), and cg18217136 (*PPIAP3* locus, $p = 0.024$).

Together, these results provide evidence for methylation being consequential to BMI at selected loci, while this is also suggested for the majority of loci at a lower level of statistical evidence.

4.2.8 Relation to clinical traits and incident disease

To evaluate the potential clinical relevance of these findings, associations of the 187 methylation markers with obesity-related clinical traits were studied in a subgroup of 4159 subjects of the KORA F4 (n = 1697) and EpiMigrant (n = 2462) cohorts with data on clinical traits and fasting blood samples available. 92 of the 187 CpGs were significantly associated with HDL, LDL and total cholesterol, triglycerides (TGs), systolic blood pressure, C-reactive protein (CRP), glucose, insulin, HbA1c and WHR after adjustmend for BMI (Figure 4.13), suggesting that these associations are not (fully) mediated by BMI. The strongest independent associations were observed for TGs (e.g., $p = 5.6 \times 10^{-65}$ with *ABCG1* locus), HDL cholesterol (e.g., $p = 3.8 \times 10^{-45}$ with *ABCG1* locus), CRP (e.g., $p = 6.5 \times 10^{-36}$ with *CRELD2* locus) and HbA1c (e.g., $p = 1.6 \times 10^{-12}$ with *ABCG1* locus). No independent signals were observed for weight, height and diastolic blood pressure.

Furthermore, association of the 187 CpGs with incident type 2 diabetes (T2D) was studied in 3064 subjects from EpiMigrant (with T2D defined as HbA1c > 6.5% or physician diagnosis). Using logistic regression including the discovery covariates, 61 CpGs were identified that showed a significant association with incident T2D after Bonferroni correction, with 16 remaining significant after additional adjustment for BMI.

In the available data, BMI was associated with all traits associated with CpGs, except LDL and total cholesterol. In order to get some indication of whether the identified CpGs account for a part of these associations of BMI with clinical traits, direct (remaining after adjustment for CpGs) and indirect parts of the BMI-trait association were determined using linear and logistic regression for continuous traits and incident T2D, respectively, and tested using a bootstrap procedure as described in Section 3.3.3, followed by inverse-variance weighted fixed-effects meta-analysis. For all traits including incident T2D, a significant part of the association with BMI was indirect, i.e. not independent of the 187 BMI-related CpG sites, suggesting that methylation at these CpGs is intermediate between BMI and trait in terms of either being a confounder, a mediator, or associated with another variable that confounds or mediates the association (Figure 4.14). The proportion of

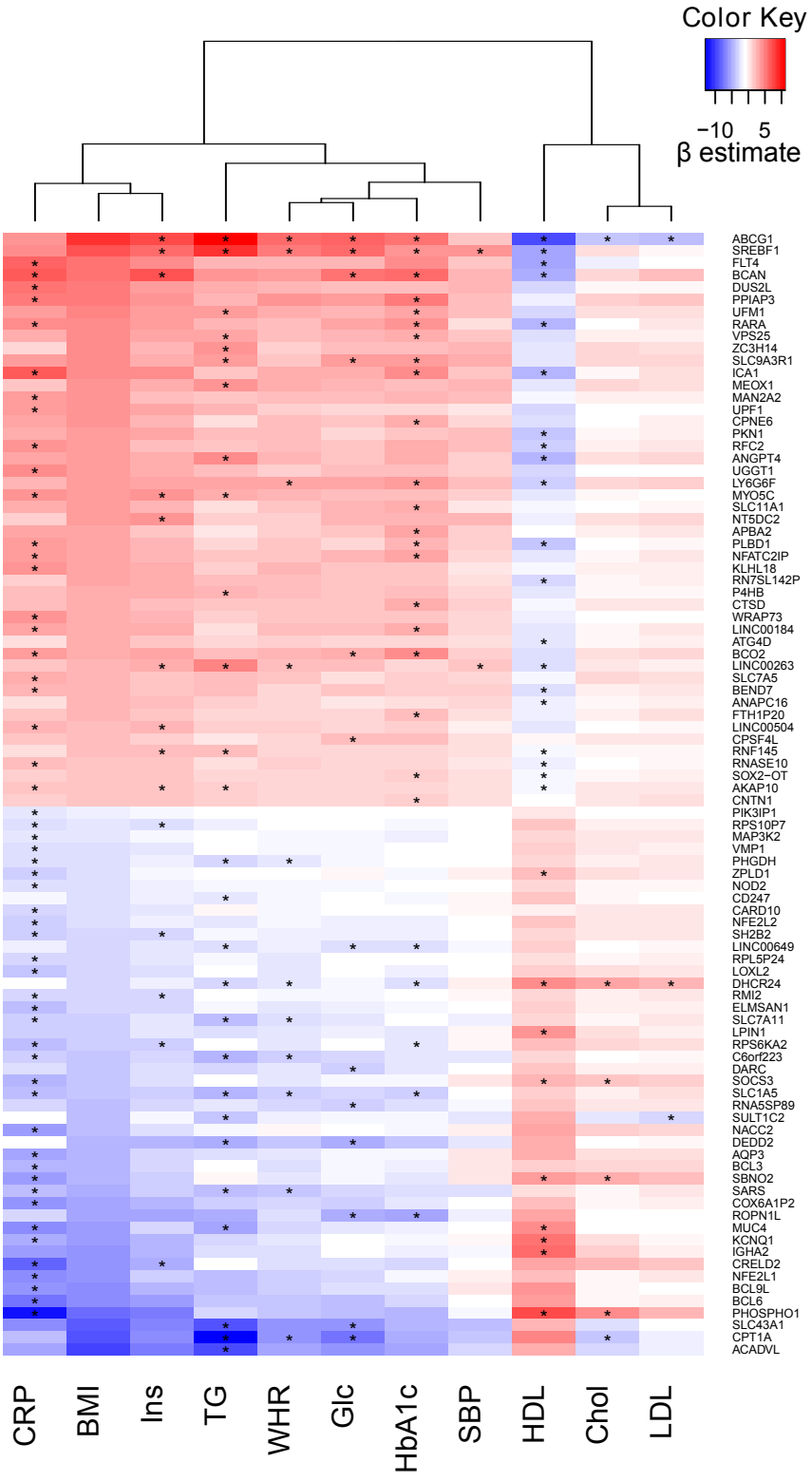

Figure 4.13: Associations between methylation at the BMI-associated CpG sites and clinical traits. Heatmap including the 92 CpGs showing significant association with at least one of the traits independent of BMI at $p < 2.1 \times 10^{-5}$ (significance denoted by black stars). Colors represent association strengths without conditioning on BMI, to allow for comparison with BMI effect. Associations were derived from linear models with clinical trait as response and CpG, (BMI) and discovery covariates as independent variables. Results were combined using inverse-variance weighted fixed-effects meta-analysis. BMI, body mass index; Chol, total cholesterol; CRP, C-reactive protein; Glc, fasting glucose; HDL, high density lipoprotein cholesterol; Ins, fasting insulin; LDL, low density lipoprotein cholesterol; SBP, systolic blood pressure; TG, triglycerides; WHR, waist-hip ratio

indirect effect by total effect ranged from 19.7% (insulin, $p_{\text{indirect}} = 5.7 \times 10^{-4}$) to 81.7% (TGs, $p_{\text{indirect}} = 2.4 \times 10^{-22}$), when all 187 CpGs were included. When the contribution of single CpGs to the BMI-trait associations was studied, 23 CpGs were shown to significantly contribute to the association of BMI with at least one of the traits (Figure 4.14). These findings raise the possibility that these methylation markers contribute to the development of metabolic and cardiovascular complications as a consequence of obesity.

4.2.9 Discussion

To explore the role of site-specific DNA methylation in the development of obesity and related metabolic disturbances, a large multi-ethnic EWAS of BMI was conducted based on more than 10,000 subjects from 13 studies. As a result, 187 methylation sites were identified that showed stable association with BMI.

BMI-related methylation is involved in lipid and glucose metabolism

Methylation at 92 of the 187 identified loci was also significantly associated with obesity-related clinical traits or incident T2D. Together with the finding that the 187 loci were enriched for biological pathways related to lipid metabolism and insulin signaling as well as for previously published lipid GWAS loci, this provides evidence for a role of the identified CpG sites in lipid and glucose metabolism. In addition, for a substantial number of CpG sites, a significant part of the BMI-trait association was diminished after adjustment for methylation. This might be explained by either a confounding or a mediating effect of methylation, or by methylation being associated with a variable that confounds or mediates the association. Mediation seems more likely at the majority of CpG sites where Mendelian randomization experiments and longitudinal analyses provide evidence for methylation being consequential rather than causal to BMI.

Few CpG sites deserve mentioning for which evidence is provided that first, methylation is consequential to BMI, and second, methylation significantly accounts for BMI-clinical trait associations. These include cg06500161 (*ABCG1* locus), which was the top marker showing a strong positive association with BMI and a negative association with *ABCG1* gene expression, and accounted for a significant part of the BMI association with TGs (27.6%), HDL cholesterol (16.9%), glucose (11.6%), HbA1c (10.9%), incident T2D (8.8%) and WHR (4.4%). *ABCG1* encodes *ATP binding cassette transporter subfamily G member 1* (ABCG1), which is involved in reverse cholesterol transport by promoting cellular cholesterol efflux from macrophages to HDL (Ye *et al.*, 2011). Cholesterol efflux capacity from macrophages is thought to play a role in atheroprotection. Accordingly, an inverse association with carotid intima-media thickness and coronary artery disease was recently reported (Khera *et al.*, 2011). In addition, genetic variants in *ABCG1* associated with reduced *ABCG1* gene expression were recently shown to increase cardiovascular disease risk (Schou *et al.*, 2012). *In vivo*, *ABCG1* knockout had a stage-dependent influence on the

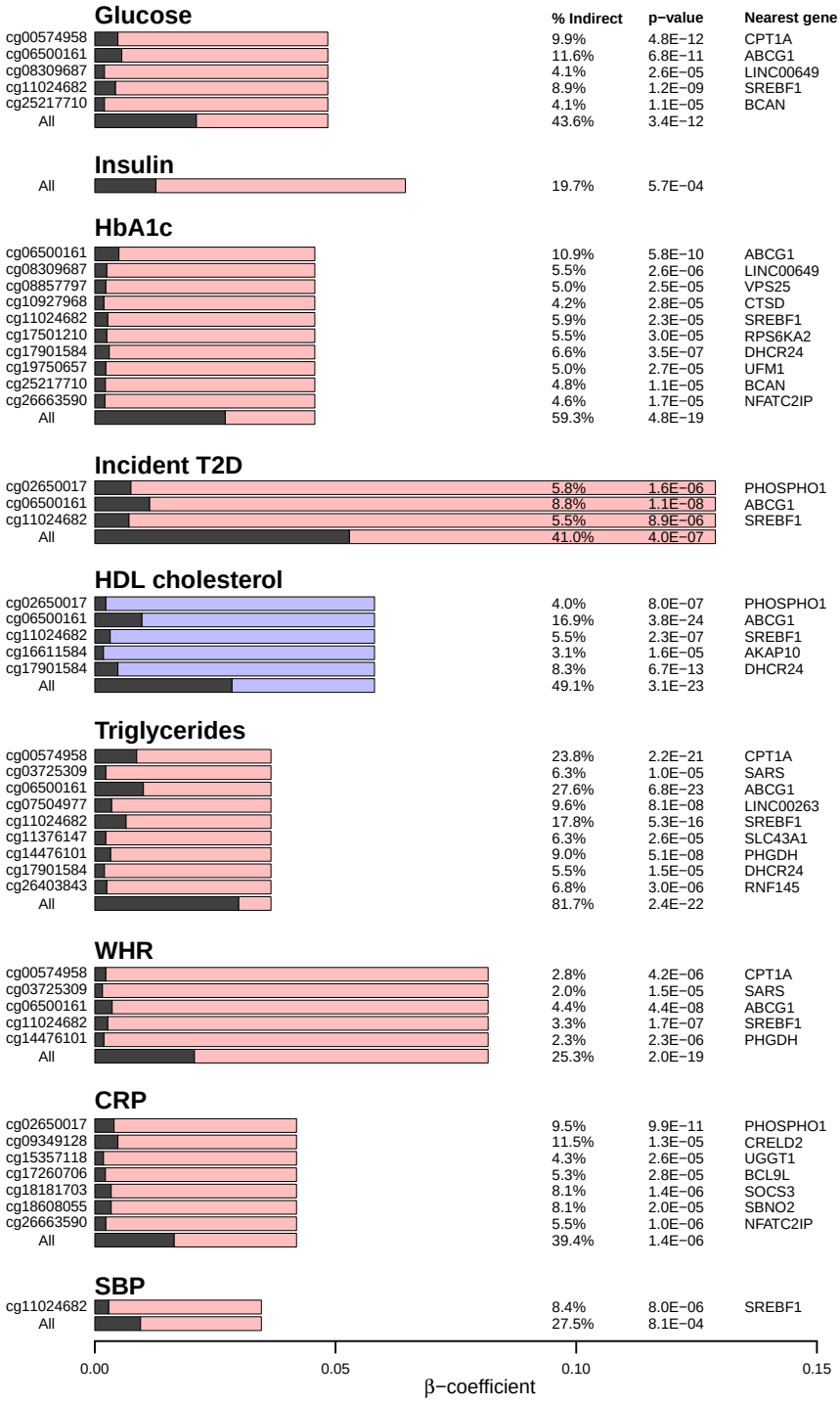

Figure 4.14: Contribution of CpG sites to BMI-trait associations. Barplots represent total BMI-trait association (red bars, positive; blue bars, negative) and "indirect" part of the association that is accounted for by single CpGs or all 187 CpGs (grey bars). Only CpG sites with a significant indirect effect (Bonferroni correction, $p < 3.0 \times 10^{-5}$) are shown. Proportion of indirect effect and bootstrap p-value is given. Total and direct effect were derived from linear models with clinical trait as response and BMI as well as discovery covariates as independent variables, without and with additional adjustment for methylation at one or several CpGs, respectively. Indirect effect was determined as difference between total and direct effect and significance tested using a bootstrap procedure (Section 3.3.3), with 10,000 bootstrap samples. For incident T2D, β coefficient was derived from logistic regression and is thus on a different scale than for the continuous traits where it was derived from linear regression. CRP, C-reactive protein; HDL, high density lipoprotein; SBP, systolic blood pressure; T2D, type 2 diabetes; WHR, waist-hip ratio.

atherosclerotic process, with a diminished lesion formation in early stages (Meurs *et al.*, 2012). Interestingly, body weight loss upon intervention has previously been reported to upregulate *ABCG1* gene expression in adipose tissue (Johansson *et al.*, 2012), and bariatric surgery increased cellular cholesterol efflux via ABCG1 (Aron-Wisnewsky *et al.*, 2011). These results support the finding that body mass affects methylation at *ABCG1* rather than the opposite way. They are also consistent with the strong correlation of blood and adipose tissue *ABCG1* methylation. In this thesis, for the first time DNA methylation is reported as mechanism of body weight effect on *ABCG1* expression. Besides a putative role of *ABCG1* in lipid metabolism, the results of the present study suggest a role in the development of T2D. The same CpG site, cg06500161, was recently found to associate with fasting insulin and homeostasis model assessment of insulin resistance (HOMA-IR) in a cross-sectional EWAS (Hidalgo *et al.*, 2014). Besides a role for cellular cholesterol efflux, evidence is emerging that ABCG1 is involved in intracellular cholesterol distribution (Tarling and Edwards, 2011). In pancreatic β-cells, cholesterol homeostasis plays a role for normal insulin secretion (Sturek *et al.*, 2010). Accordingly, *ABCG1* deficiency was associated with reduced insulin secretion and β-cell function in mice (Sturek *et al.*, 2010, Kruit *et al.*, 2012).

Another CpG site with good evidence for methylation being consequential to BMI and accounting for BMI-trait associations was cg11024682 (*SREBF1* locus). For this CpG site, a positive association with BMI (third strongest among all sites), and a negative association with *SREBF1* gene expression was observed. In addition, methylation putatively accounted for a significant part of the BMI association with TGs (17.8%), glucose (8.9%), systolic blood pressure (8.4%), HbA1c (5.9%), incident T2D (5.5%), HDL cholesterol (5.5%) as well as WHR (3.3%). *SREBF1* encodes the isoforms *Sterol regulatory element-binding transcription factor 1a* (SREBP1a) and c (SREBP1c). These are transcriptional regulators of genes involved in lipogenesis, fatty acid desaturation, cholesterol uptake and synthesis (Shimano, 2001) as well as gluconeogenesis and glycogen synthesis (Ruiz *et al.*, 2014). Methylation at *SREBF1* has been found to decrease in rats after changing high-fat, high-sucrose diet to control diet (Uriarte *et al.*, 2013). Together with evidence for substantial increase in SREBP1 protein after repeated fasting-refeeding cycles (Kochan, 2003), this suggests a role of SREBP1 in the lipogenic potential of adipose tissue after repeated dieting. Of note, the intronic microRNA 33b (miR33b) is co-transcribed with *SREBF1* (Rotllan and Fernández-Hernando, 2012) and might explain a part of the presented findings. miR33a and b are negative regulators of genes involved in cellular cholesterol efflux (including *ABCG1*), fatty acid oxidation (including *CPT1*, which showed a strong negative association with BMI in the present study), and insulin signaling (Rotllan and Fernández-Hernando, 2012).

For three further CpG sites, results suggest that methylation is consequential to BMI and that these CpG sites partly account for BMI-clinical trait associations: cg17901584

(*DHCR24* locus, negative association with BMI, negative association with *DHCR24* gene expression), cg08309687 (*LINC00649* locus, negative association with BMI) as well as cg08857797 (*VPS25* locus, positive association with BMI). *DHCR24* encodes *24-dehydrocholesterol reductase*, which is involved in cholesterol biosynthesis (Waterham *et al.*, 2001). It was among the genes with significantly decreased expression following bariatric surgery in obesity subjects with T2D, showing also association with improvement in HbA1c and fasting glucose (Berisha *et al.*, 2011). In addition, genetic variation in *DHCR24* was shown to associate with T2D in an isolated population, although without replication in larger populations (Rampersaud *et al.*, 2007). Besides its role in sterol synthesis, DHCR24 binds tumor suppressor p53 upon oxidative stress and protects it from degradation (Wu *et al.*, 2004). p53 has been shown to be involved in the development of insulin resistance in mice with excessive calorie intake through induction of proinflammatory cytokines in adipose tissue (Minamino *et al.*, 2009). Less is known about the *LINC00649* and *VPS25* loci, which mediated a part of the BMI association with glycemic traits, and no strong association with expression of nearby genes could be shown. *VPS25* showed a weak positive association with *G6PC* transcription, a gene coding *glucose-6-phosphatase*, the key enzyme of hepatic gluconeogenesis, the upregulation of which is a characteristic of T2D (Gautier-Stein *et al.*, 2012).

It remains to be determined how increased body mass might influence DNA methylation at these loci. A recent study demonstrated strong effects of the fatty acid palmitate on DNA methylation and gene expression in pancreatic islets *in vivo* (Hall *et al.*, 2014). This raises the possibility that the effect of body mass on methylation might be explained by increased non-esterified fatty acid (NEFA) concentrations (Section 1.1.4, de Ferranti and Mozaffarian (2008)). Few genes regulated by palmitate in islets Hall *et al.* (2014) were also (at least nominally) associated with BMI-related CpG sites in the present study, including *SCD*, *UBE2E3* and *GPX4*. Of note, moderate correlation of blood and pancreatic methylation of the 187 CpG sites was shown.

Methylation and the pathogenesis of obesity

Methylation at the majority of CpG sites was consequential to BMI in this study, whereas evidence for a causal effect was obtained only at three loci. Previous longitudinal studies have shown association of cord blood methylation with childhood body composition (Relton *et al.*, 2012, Godfrey *et al.*, 2011), suggesting that methylation is either causally involved in weight regulation or might otherwise be a non-causal marker. In addition, a previous cross-sectional EWAS of 96 obese and lean adolescents showed an enrichment for obesity risk genes and thereby provide evidence that a part of the identified loci might be causal rather than consequential to BMI (Xu *et al.*, 2013). The main difference to the present study is that these were studies on children and adolescents, whereas the present study was based on adults. Age shows a strong relationship with site-specific DNA methy-

lation (Bell *et al.*, 2012), and interaction of age with BMI in relation to DNA methylation has been observed (Almén *et al.*, 2014). During ageing, environmental and behavioral effects on DNA methylation might accumulate, thereby possibly reverting methylation at early markers of weight regulation, or increasing methylation variability. Consequently, the power for identifying methylation effects on BMI might be reduced. Diverging effects of the *FTO* risk locus with childhood versus adult obesity have been reported recently (Sovio *et al.*, 2011). In addition, the study by Xu *et al.* (2013) focused on extreme groups of obese and lean subjects, and on a different ethnicity, i.e. African-Americans. Of note, ethnicity did not cause a lot of heterogeneity in BMI effect in the present study based on Europeans and South Asians. Together, age differences seem to be a likely reason for the observed differences between the present investigation and earlier studies. Age can also be the reason underlying the small overlap of significant CpG sites between the present study and previous EWAS of BMI (Xu *et al.*, 2013, Wang *et al.*, 2010, Almén *et al.*, 2012). The only available large EWAS in adults found one association of BMI with methylation at *HIF3A* (Dick *et al.*, 2014), which is replicated in the discovery meta-analysis of the present study ($p = 8.8 \times 10^{-3}$).

Little is known about the three loci showing a potential causal effect on BMI. The CpG site cg26663590, located on chromosome 16 upstream of the gene *NFATC2IP*, showed the strongest evidence of being causal to BMI, and was also associated with HbA1c. The corresponding genomic locus has previously been identified in a GWAS on BMI (Speliotes *et al.*, 2010) and contains the candidate gene *SH2B1* (*SH2B adaptor protein 1*). The protein SH2B1 has a known role in energy and glucose homeostasis (Ren *et al.*, 2007). However, methylation did not show a significant association with *SH2B1* gene expression ($p = 0.112$).

The results on causal inference obtained in this study should be interpreted with care. Two different MR approaches were applied to study causality of BMI-methylation associations in a relatively large subsample of approximately 4000 subjects. Although they provide some evidence for methylation at the majority of CpG sites being consequential rather than causal to BMI, they are subject to certain limitations. The *ad hoc* MR approach did not query model assumptions (which are, admittedly, difficult and partly impossible to test), so results have to be interpreted with care, whereas the statistically more solid instrumental variable approach that involved an assessment of selected model assumptions suffered from limited power. Specifically, instrument strength (i.e., the association of the genetic variant with the putative independent variable) of the single studies was often too low to include them in the meta-analysis for reasons of weak instrument bias (Palmer *et al.*, 2012), resulting in power reduction. Importantly, meta-analyzing a large number of small studies (with weak instruments) does not avoid weak instrument bias (Burgess *et al.*, 2011). Thus, meta-analyses of larger studies are needed to obtain reliable causal estimates and possibly stronger evidence for causality.

When investigating the direction of association between BMI and methylation, one should not exclude a bidirectional effect, feedback mechanisms, or even more complex causal constructs as a possibility. Interestingly, Milagro *et al.* (2011) found methylation at one locus, *ATP10A*, to be both predictive of weight loss success during intervention, and to be affected by weight reduction, suggesting a bidirectional effect. Methylation at *ATP10A* showed a significant negative association with BMI in the present study, which is consistent with a positive association between weight loss and methylation in the study by (Milagro *et al.*, 2011). However, while they discuss ATP10A as a protein with a plausible role in body fat regulation, no association of methylation with *ATP10A* gene expression was observed in the present study.

Strengths and Limitations

Important strengths of this study include its large sample size and careful validation of the discovered associations in an independent large replication study, as well as the availability of gene expression, SNP and clinical data, including incident disease information, for a large proportion of discovery samples. Further strengths are the stringent quality control (e.g., close examination of SNPs in probes), the comprehensive study of pathway and functional enrichment, and the careful examination of causality using two MR approaches and longitudinal analyses. In addition, evidence is provided that the observed methylation signatures in whole blood might reflect methylation in metabolically relevant tissues. Together, this is the first large EWAS of BMI which, supported by elaborated downstream analyses, provides a comprehensive picture of the role of DNA methylation in adult obesity.

Several limitations of this study deserve comment. To begin with, the study is cross-sectional. Although approaches to decipher causality were applied, the difficulty to test the underlying model assumptions, and limited power and risk for bias in the case of weak genetic instruments require replication of MR findings in larger studies, complemented by experimental evidence. In addition, methylation was measured in whole blood samples, which is the only tissue available in all included cohorts. This raises two issues, namely first, the question of whether whole blood methylation is representative of methylation in metabolically relevant tissues, and second, the issue of cell type confounding. The first issue was answered by showing high correlation of blood methylation with methylation in tissues including spleen as well as omental and subcutaneous fat. Conversely, the low representation of tissue-specific gene expression in whole blood (Emilsson *et al.*, 2008) raises the question of how far other tissue-specific methylation signatures are not represented in blood. This remains to be determined. To address the second issue, proportions of six white blood cell types were estimated using the method of Houseman *et al.* (2012) and were included as covariates in the model. This might diminish a potential confounding effect of cell proportions on the association between BMI and methylation, but might not

completely abolish it. The quality of the estimated cell proportions depends on the quality of the external data underlying the estimation, and only those cell types are accounted for that were part of the external data (i.e., $CD4^+$ T cells, $CD8^+$ T cells, B cells, natural killer cells, monocytes and granulocytes). Subtypes of these cell types, e.g., Th1 and Th2 subtypes of $CD4^+$ T cells (Brand *et al.*, 2012) might also differ in their methylation profile. The development of improved methods to deal with cell type confounding in whole blood methylation and expression studies is an ongoing research focus (Houseman *et al.*, 2014, Zou *et al.*, 2014, Jaffe and Irizarry, 2014). Of note, in the study by Xu *et al.* (2013), obesity was not associated with proportions of neutrophils, eosinophils, basophils, monocytes, lymphocytes, $CD4^+$ T cells and $CD8^+$ T cells, reducing the probability of cell type confounding for these cell types. Finally, only one technique, the Infinium HumanMethylation450K BeadChip, was used to determine methylation stage in all discovery and replication cohorts. It is advisable to validate microarray results using a different technique to improve the reliability of the reported results (Wang *et al.*, 2006). A suitable technology is the Sequenom MassArray EpiTyper, which was used previously to validate results obtained with the Infinium HumanMethylation450K BeadChip (Milagro *et al.*, 2011, Zeilinger *et al.*, 2013). This technique would also allow for a denser coverage of CpGs at candidate loci.

Conclusions

In this work, the first large EWAS of BMI identified 187 methylation sites showing solid association with BMI. These sites were comprehensively studied in downstream analyses. This revealed an enrichment for functional genomic features, for biological pathways and for previously published lipid GWAS loci. A large number of the identified CpG sites were associated with gene expression at nearby genes and showed strong relation to genetic variation in *cis*. Cross-tissue analysis showed strong correlations with methylation in different tissues including omental and subcutaneous fat. MR and longitudinal analyses indicate that methylation at the majority of loci was consequential rather than causal to BMI. Finally, methylation at selected sites explained a part of the association of BMI with clinical traits and incident T2D. Together, these findings provide new evidence for methylation as a mechanisms linking obesity with its metabolic comorbidities. Furthermore, methylation at selected sites explained a part of the association of BMI with clinical traits and incident T2D, suggesting methylation as a candidate mechanism underlying the development of obesity-related comorbidities.

4.3 Metabolic signature of weight change: an integrative metabolomics and transcriptomics approach

As reviewed in Sections 1.2.3 and 1.2.4, previous cross-sectional efforts suggest a relationship between obesity and the human blood metabolome and transcriptome. In addition, weight loss upon behavioral intervention was associated with changes in the blood metabolome, suggesting that the observed obesity-related molecular signatures are at least in part reversible.

However, the effect of long-term body weight change on the human blood metabolome and transcriptome in the general population – rather than under clinical settings – is less well explored. The few studies that investigate the association with weight change in prospective cohorts are based on a small sample size and restricted to a small set of lipoprotein subclasses (Mäntyselkä *et al.*, 2012, Naganuma *et al.*, 2009). In addition, although multi-omics approaches have been fruitful in different applications in enhancing the understanding of complex molecular pathways (Zhang *et al.*, 2013, Zhou *et al.*, 2012, Acharjee *et al.*, 2011, Inouye *et al.*, 2010a, Dutta *et al.*, 2012), the potential of integrating multiple omics techniques has rarely been used in the study of weight-associated metabolic effects in humans (Oberbach *et al.*, 2011, Valcárcel *et al.*, 2014).

In light of these considerations, the present study focuses on the investigation of metabolomic and transcriptomic consequences of weight change over a 7-year follow-up period in the general population (Figure 4.15).

This section is based on the manuscript

- **Wahl S**[*], Vogt S[*], Stückler F, Krumsiek J, Bartel J, Schramm K, Carstensen M, Rathmann W, Roden M, Jourdan C, Kangas AJ, Soininen P, Ala-Korpela M, Nöthlings U, Boeing H, Theis F, Meisinger C, Waldenberger M, Suhre K, Gieger C, Kastenmüller G, Illig T, Linseisen J, Peters A, Prokisch H, Herder C, Thorand B[#], Grallert H[#]. "Metabolic signature of weight change: an integrative metabolomics and transcriptomics approach." *under review.*

4.3.1 Weighted correlation analysis reveals four metabolite and two gene expression modules related to body weight change

In the population-based KORA S4/F4 cohort, two-platform serum metabolomics and whole blood transcriptomics measurements were available from the follow-up examination F4 for 1631 and 689 participants, respectively (Table 4.9, see Section 2.1 for data retrieval and Section 3.1 for data preprocessing). Previous studies have shown that clusters of related genes may be more reproducibly associated with a phenotype or disease

[*,#] contributed equally

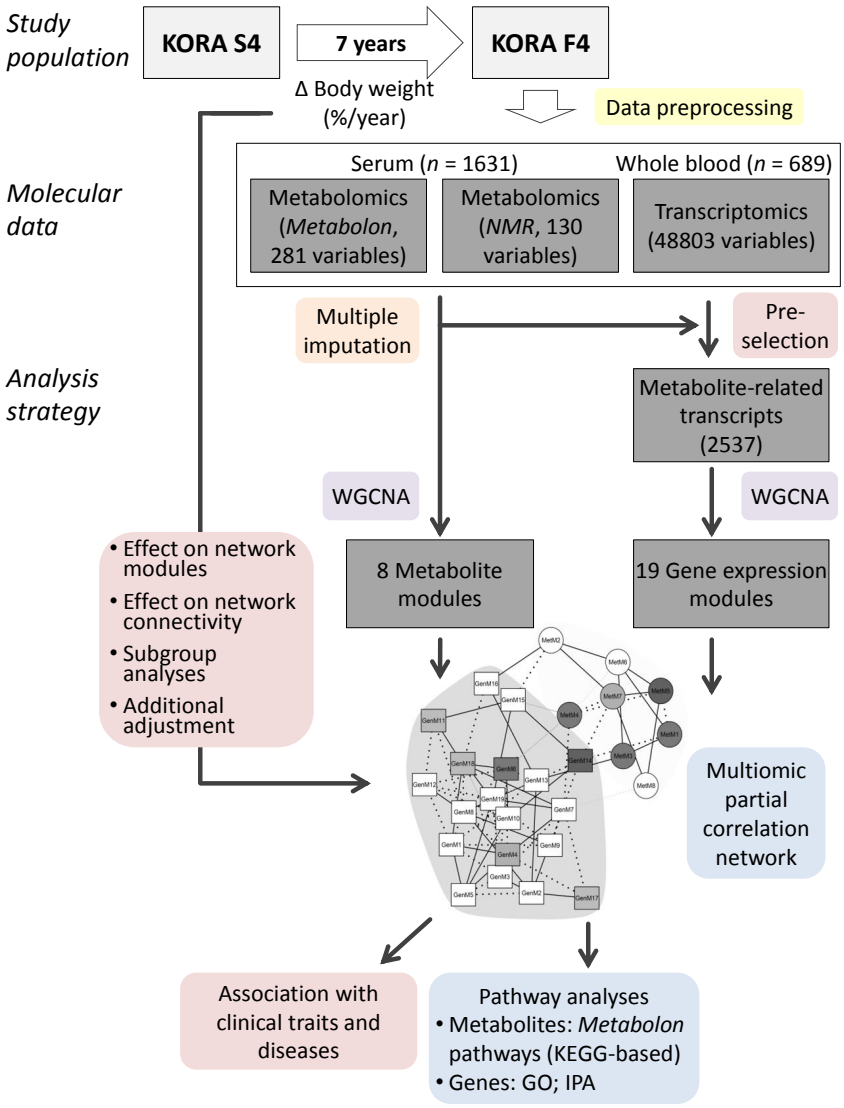

Figure 4.15: Metabolic consequences of body weight change: Study design and analysis strategy. Omics measurements are described in Section 2.1. Color coding of statistical methods: yellow, data preprocessing and quality control (Section 3.1); orange: missing data handling (Section 3.2); red, univariate data analysis (Section 3.3); violet, multivariate data analysis (Section 3.4); blue, extraction of biological knowledge (Section 3.5). GO, gene ontology; IPA, Ingenuity pathway analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; NMR, nuclear magnetic resonance; WGCNA, weighted correlation network analysis.

than single genes (Chuang *et al.*, 2007), and that testing groups of metabolites instead of single metabolites improved power in a genome-wide association study (Inouye *et al.*, 2012). Thus, the strategy in this study was to cluster metabolomics and transcriptomics data prior to testing for association with previous weight change. This was achieved through *weighted correlation network analysis* (WGCNA, see Section 3.4.1). Clustering was jointly performed on the 411 serum metabolites (281 from the Metabolon platform [M] and 130 from the NMR platform [N], after multiple imputation as described in Section

Table 4.9: Characteristics of the KORA S4/F4 study population for the metabolomics and transcriptomics study of weight change.

Variable	Metabolomics data ($n = 1631$)	Combined metabolomics & transcriptomics data ($n = 689$)
	Mean(sd)	
Body weight (kg), baseline	78.3 (14.7)	78.5 (13.2)
Body weight (kg), follow-up	79.7 (15.6)	79.3 (13.8)
Δ Body weight (%)	1.8 (6.8)	1.0 (6.5)
Δ Body weight/year (%)	0.3 (1.0)	0.1 (0.9)
BMI (kg/m^2), baseline	27.7 (4.5)	28.5 (4.3)
BMI (kg/m^2), follow-up	28.2 (4.7)	28.8 (4.5)
Age (years), baseline	54.2 (8.7)	61.8 (4.3)
Age (years), follow-up	61.2 (8.7)	68.8 (4.3)
	Relative frequency (%)	
Sex (male/female)	50.8 / 49.2	50.1 / 49.9
Weight change direction (reduction/gain)	39.3 / 60.7	45.9 / 54.1

3.2). Prior to clustering of gene transcripts, they were pre-selected based on their association with metabolite concentrations, to keep the focus on genes related to the blood metabolome (see Section 3.4.1 for methodological details). 2537 "metabolite-related transcripts" were identified that showed at least a suggestive association ($p < 10^{-5}$) with at least one metabolite.

WGCNA generated 8 metabolite modules (MetM) and 19 gene expression modules (GenM). Four of the metabolite modules were significantly associated with previous annual percentage body weight change (ΔBW), as determined through association with the respective module eigengene (ME) (see Section 3.4.1 for definition) in linear models adjusted for age, sex and baseline body weight (positive associations for MetM1, $p = 1.2 \times 10^{-24}$; MetM3, $p = 2.2 \times 10^{-4}$; MetM4, $p = 7.3 \times 10^{-17}$; negative association for MetM5, $p = 1.7 \times 10^{-14}$, all significant after Bonferroni correction for 27 modules, Figure 4.16, first column). Of the gene expression modules, two were significantly associated with ΔBW (positive association for GenM6, $p = 3.8 \times 10^{-12}$; negative association for GenM14, $p = 1.9 \times 10^{-4}$). Note that ΔBW is a variable spanning the whole weight change range, with weight loss coded as negative ΔBW values and weight gain as positive ΔBW values. Thus, effect directions displayed in Figure 4.16 have to be interpreted as the average linear effect across the weight change range, and effect directions have to be inverted to construe the effect of weight reduction. Using the example of MetM1, the positive association of ΔBW with MetM1 can be interpreted as increase in the ME with increasing weight gain, and as decrease in the ME with increasing weight loss. Stratified models were conducted as a sensitivity analysis (Figure 4.16, columns 2-11) and are the subject of Section 4.3.4 below.

To investigate the interrelatedness of the identified ΔBW-related metabolite and gene expression modules conditional on all other modules and the above-mentioned covariates, a

partial correlation network was constructed from the MEs (Figure 4.17; see Section 3.5.3 for methodological details). The six ΔBW-related modules were interrelated. The strongest positive partial correlation was observed between MetM1 and MetM3 ($p = 2.3 \times 10^{-54}$). In addition, MetM1 showed a strong negative correlation with MetM5 ($p = 1.8 \times 10^{-72}$), as well as with GenM14 ($p = 6.8 \times 10^{-29}$).

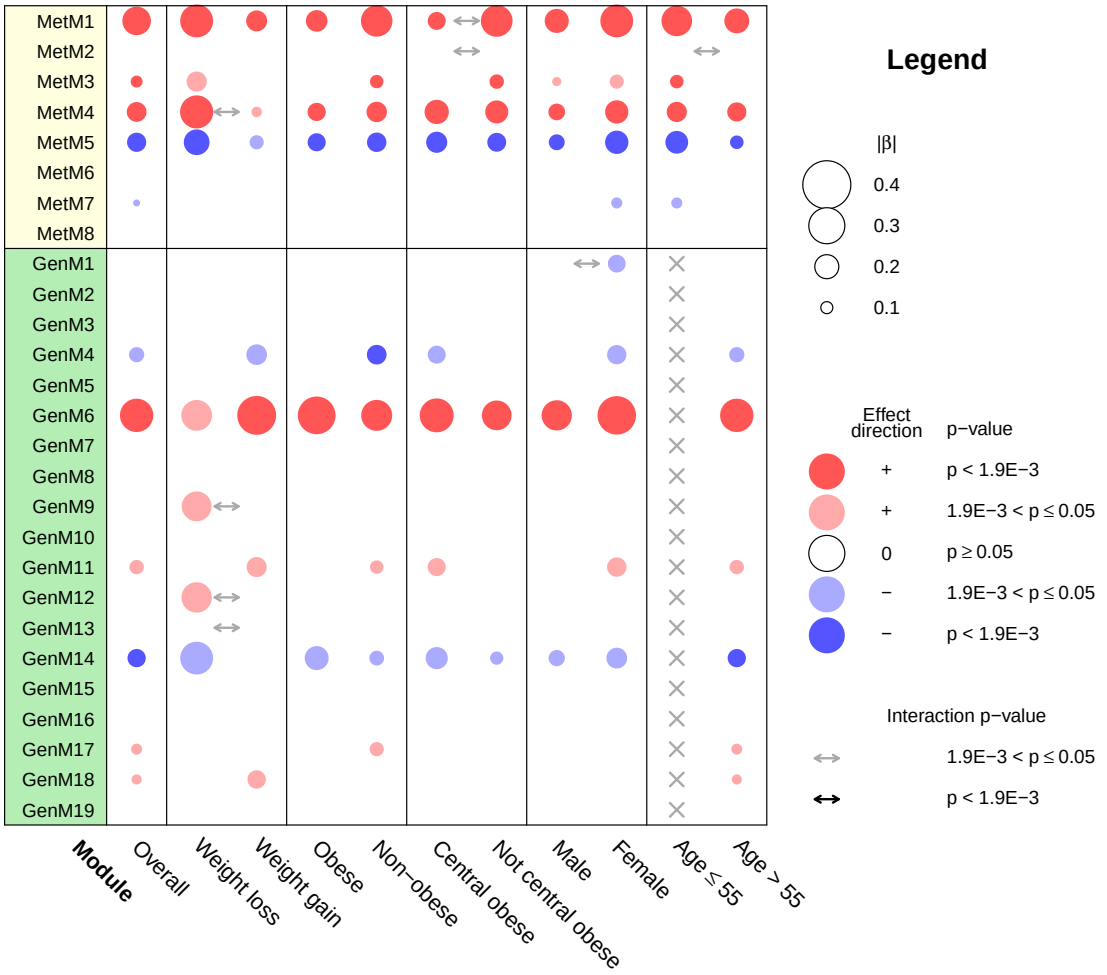

Figure 4.16: Association of annual percentage body weight change (ΔBW) with omics modules in the overall study population and in subpopulations. Bubbles represent effect strengths and significance, as described in the legend. Models were adjusted for age, sex and baseline body weight. Significance threshold $p < 1.9 \times 10^{-3}$ corresponds to Bonferroni correction for 27 modules. For subgroup analyses (columns 2-11), interaction models were fitted, to obtain main ΔBW effect in the respective subgroups, and ΔBW:subgroup interaction effect indicating difference in effect between the subgroups (see Section 3.3.1 for details). Gene expression analysis was restricted to a subgroup of 689 subjects aged >55 years at baseline. No effect estimates are available for the younger subgroup in this population (indicated as grey crosses). Central obesity was defined as waist-hip ratio (WHR) > 1 in males and > 0.85 in females. GenM, gene expression module; MetM, metabolite module.

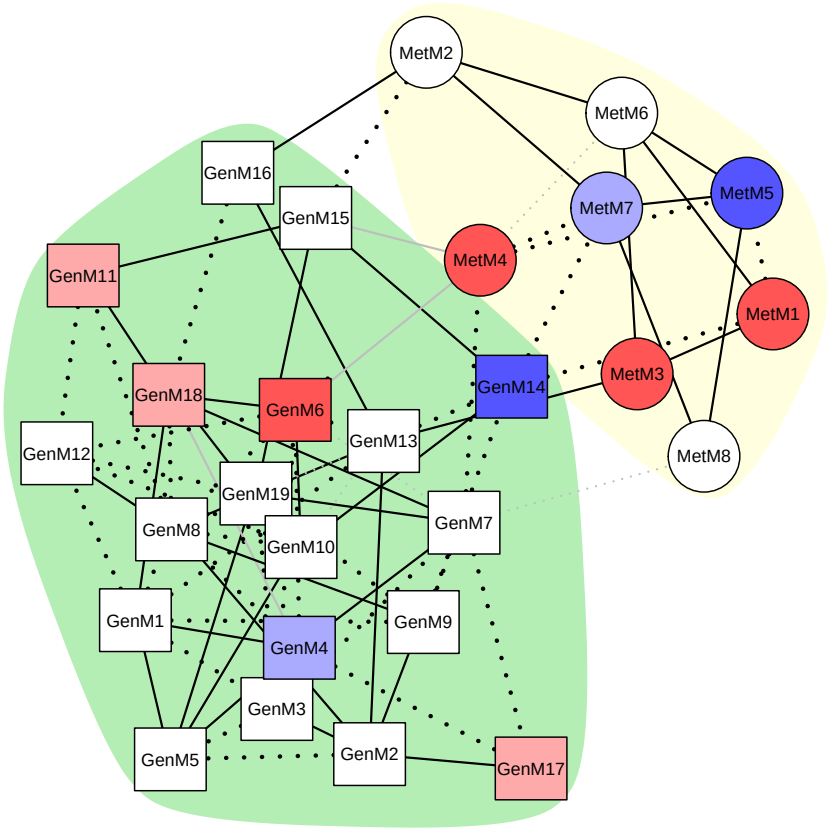

Figure 4.17: Multi-omic partial correlation network comprising the 8 metabolite and the 19 gene expression modules. Nodes represent omics modules (circle, metabolite module (MetM); rectangle, gene expression module (GenM)), colored according to their association with ΔBW (red, significant positive association; blue, significant negative association; bright color, significant $p < 1.9 \times 10^{-4}$; light color, $p < 0.05$). Edges represent partial correlations (ζ) between pairs of modules (represented by their module eigengenes (MEs)), conditional on all other presented modules and the covariates age, sex, and ΔBW (solid black line, $\zeta > 0.1$; dotted black line, $\zeta < -0.1$; solid grey line, $0.05 < \zeta < 0.1$; dotted grey line, $-0.1 < \zeta < -0.05$). Background color reflects metabolite (yellow) vs. gene expression (green) modules. See Section 3.5.3 for methodological details.

4.3.2 The four ΔBW-related metabolite modules cover major branches of metabolism

Metabolites were assigned to super- and sub-pathways in accordance with the Metabolon classification (Appendix Table A.5). The four ΔBW-related metabolite modules comprised a total of 147 metabolites. Together, these metabolites covered major branches of metabolism captured by the metabolomics platforms, including lipid metabolism, amino acids and peptides, carbohydrate metabolism, cofactors and vitamins, nucleotides and energy metabolism (Figure 4.18). This suggests a global impact of body weight change on the serum metabolome.

For each MetM, metabolites were ranked by their contribution to the module defined as the correlation with the respective ME (see Section 3.4.1). MetM1 (comprising 60

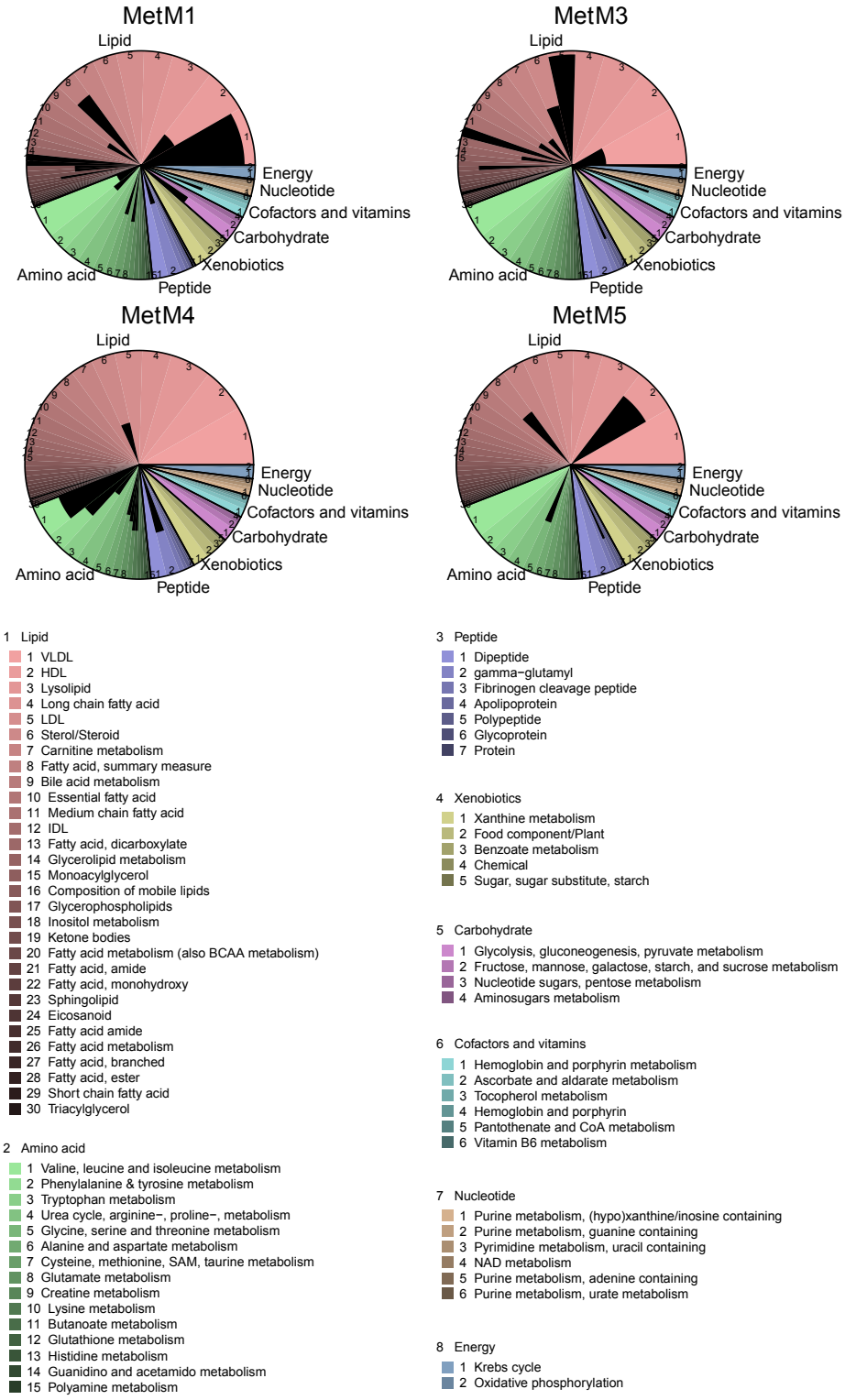

Figure 4.18: Coverage of the serum metabolome by the metabolite modules (MetM) related to annual percentage body weight change (ΔBW). Pie chart with color indicating super-/sub-pathway as described in the legend, and size of wedges representing the number of metabolites in the data set corresponding to the respective sub-pathway. Sorted by pathway size. Black wedges represent number of metabolites from the respective module significantly associated with ΔBW in the respective sub-pathway.

metabolites) was strongly determined by constituents of all very low density lipoprotein (VLDL) subclasses, total serum triglycerides (TGs), TGs in small high density lipoprotein (S-HDL) and measures of primarily saturated and monounsaturated fatty acids, which all showed a module membership strength of above 0.8 (Figure 4.19). Together with isoleucine [N], glycoprotein, glutamate, urate, lactate, phenylalanine and pyruvate, these most connected metabolites were also most strongly associated with ΔBW in single metabolite models (Figure 4.19). When a formal enrichment analysis was performed (see Section 3.5.1), MetM1 was significantly enriched for metabolites belonging to the super-pathway "Lipids" and the sub-pathways "VLDL" and "Triglycerides" (all $p < 10^{-5}$), confirming the pre-dominant role of these metabolite classes for MetM1.

MetM3 (comprising 39 metabolites) was mainly driven by constituents of low density lipoprotein (LDL) and intermediate density lipoprotein (IDL) subclasses and XS-VLDL, measures of serum cholesterol as well as apolipoprotein B (module membership strengths > 0.8, Figures 4.18 and 4.19). In addition, a significant enrichment for the super-pathway "Lipids" and the sub-pathways "LDL" and "IDL" was observed ($p < 10^{-5}$). The most contributing metabolites of MetM4 (comprising 26 metabolites) were the branched-chain amino acids (BCAAs) valine, leucine and isoleucine, and the peptide gamma-glutamylleucine, with an enrichment for the super-pathway "Amino acids" and the sub-pathway "Valine, leucine and isoleucine metabolism" ($p < 10^{-5}$). Finally, MetM5 comprised 22 metabolites and was mostly driven by constituents of L- and XL-HDL as well as apolipoprotein A1, with a significant enrichment for the super-pathway "Lipids" ($p = 1.6 \times 10^{-4}$) and the sub-pathway "HDL" ($p < 10^{-5}$).

These results demonstrate that ΔBW strongly associates with lipoprotein constituents, amino acids and peptides, as well as metabolites of energy metabolism, and that clustering helped to reveal pathways jointly and strongly associated with ΔBW. Of note, metabolites reflecting biological pathways that are less well covered by the metabolomics platforms are less likely to cluster in modules sharing association with ΔBW. These include the positive association of ΔBW with the tryptophane metabolites hydroxytryptophane [M] and kynurenine [M], which are successors of tryptophan in the serotonin and niacin biosynthesis pathways, respectively, and negative association with serotonine (5HT) [M]. They also include the positive association with the xenobiotics caffeine [M] and piperine [M], an alkaloid found in pepper, and negative association with quinate [M] and catechol sulfate [M], the positive association with bradykinine, des-arg(9) [M], the active metabolite of the vasodilating peptide hormone bradykinine, and the positive association with the metabolites N1-methyl-3-pyridon-4-carboxamide [M] and N1-methyladenosine [M] from the nucleotide super-pathway.

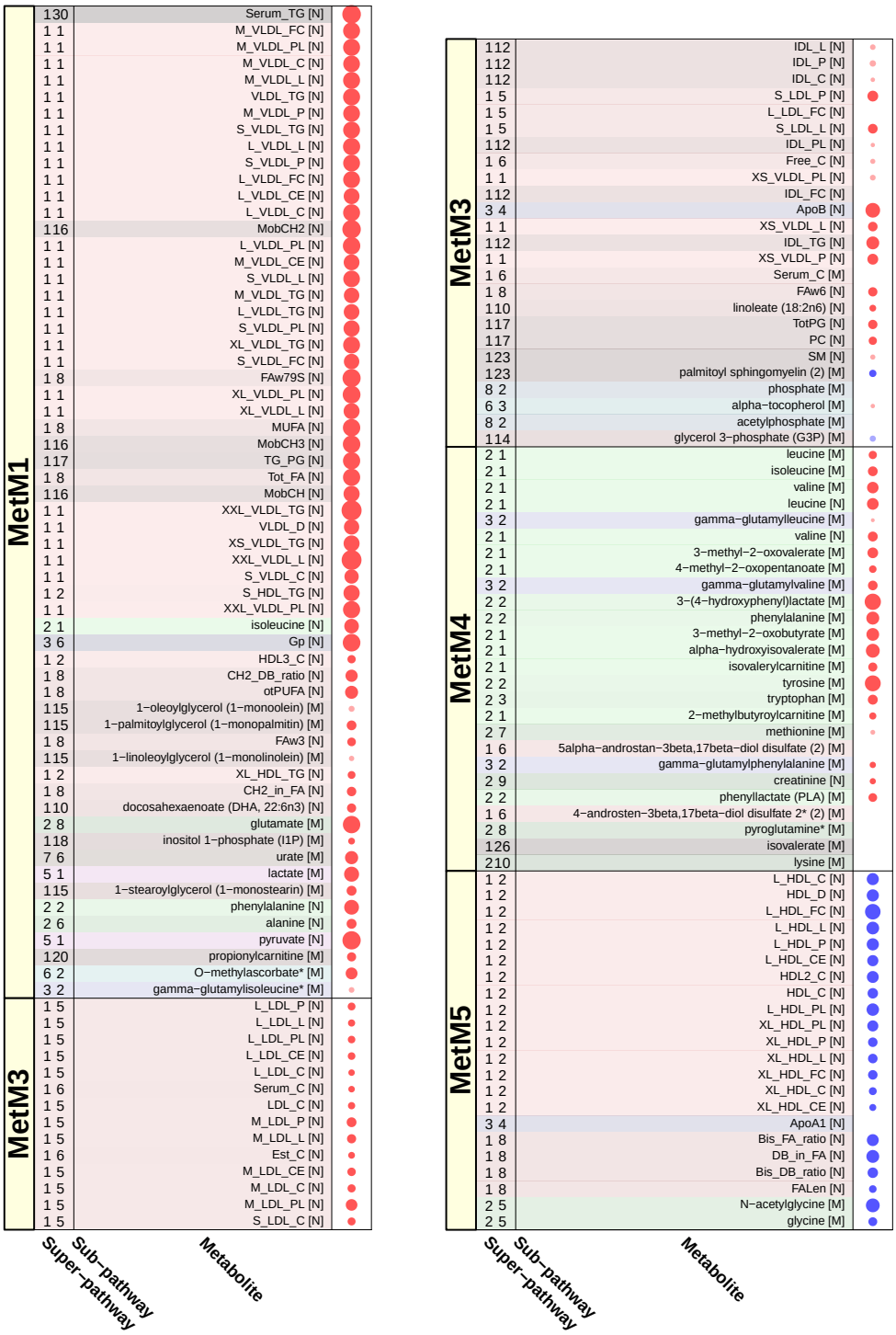

Figure 4.19: Association of annual percentage body weight change (ΔBW) with members of associated metabolite modules (MetM). Bubbles represent effect strengths and significance, see legend of Figure 4.16. Models were adjusted for age, sex and baseline body weight. For single metabolites, the significance threshold was chosen as $p < 1.2 \times 10^{-4}$ corresponding to Bonferroni correction for 441 tests. Metabolites are sorted by their module membership strength, as determined by the correlation of metabolite concentration with the module eigengene (ME) (see Section 3.4.1). Background colors correspond to super- and sub-pathway annotation, see legend of Figure 4.18.

Lipoprotein subclasses

The associations of ΔBW with lipoprotein subclasses (positive association with VLDL, LDL and S-HDL subclasses, and negative association with larger HDL particles as well as with HDL and LDL particle size; Figure 4.19) are in agreement with the observations of two smaller prospective studies that analyzed the effect of weight change over similar time periods (9 and 6.5 years, respectively) on lipoprotein subclasses (Mäntyselkä *et al.*, 2012, Naganuma *et al.*, 2009). Specifically, ΔBW was positively associated with increases in VLDL and LDL subclasses, and with decreases in L-HDL, whereas S-HDL behaved oppositely (Mäntyselkä *et al.*, 2012). ΔBW was also negatively related to LDL and HDL particle sizes (Mäntyselkä *et al.*, 2012, Naganuma *et al.*, 2009). The clustering of S-HDL-TG within the VLDL module in the present study is also in agreement to their close correlation in Inouye *et al.* (2010a), where S-HDL behaved differently from larger HDL subclasses with regard to metabolite-transcript associations.

Mechanisms by which body weight increase gives rise to the described changes may include an increased release of non-esterified fatty acids (NEFAs) from adipose tissue, triggering hepatic TG and VLDL production (Klop *et al.*, 2013) and increasing the activity of hepatic lipase (Brunzell and Hokanson, 1999). Hepatic lipase is involved in the exchange of TGs from VLDL against cholesterol esters from HDL, thereby promoting the production of small dense LDL. Together with phospholipid transfer protein (PLTP) and cholesterol ester transfer protein (CETP), which also show increased levels upon obesity (Tzotzas *et al.*, 2009), hepatic lipase is centrally involved in regulating HDL particle size. Interestingly, this was reflected in oppositional associations of genetic variants in the respective genes *LIPC, PLTP* and *CETP* with small versus large HDL subclasses (Tukiainen *et al.*, 2012).

Together, the lipoprotein signature related to positive ΔBW (i.e., weight gain) largely corresponds to an unfavorable, atherogenic lipid profile. For instance, large VLDL and small HDL particles were found to be positively, larger HDL particles to be negatively associated with coronary artery disease severity (Freedman *et al.*, 1998). A role of smaller HDL particle size in cardiovascular disease risk has also been reported by Arsenault *et al.* (2009). In a large prospective cohort of 4,594 initially healthy adults, a lipoprotein pattern characterized by decreased L-HDL, increased S-/M-LDL, and increased TGs was associated with an increased cardiovascular disease incidence after a mean follow-up of 12 years (Musunuru *et al.*, 2009). Furthermore, VLDL particle size, which was positively associated with ΔBW in this study, predicted type 2 diabetes (T2D) incidence over a 13-year follow-up of 26,836 initially healthy women (Mora *et al.*, 2010). In line with these findings, a strong positive association of MetM1 (representing VLDL subclasses) and a strong negative association of MetM5 (representing HDL subclasses) with markers of insulin resistance was observed (HbA1c: $p = 1.9 \times 10^{-5}$ and 7.0×10^{-7}; oral glucose tolerance test (OGTT) 2-hours glucose: $p = 2.1 \times 10^{-9}$ and 5.1×10^{-8}, respectively) (Figure 4.20).

Amino acid metabolism

ΔBW was strongly associated with amino acid concentrations, most prominently BCAAs, phenylalanine, tyrosine and glutamate. The increase of these amino acids in obesity is long known (Felig *et al.*, 1969), and has also been observed in more recent studies (e.g., Newgard *et al.* (2009)). The underlying mechanism might be an impaired catabolism of BCAA in adipose tissue upon obesity (Pietiläinen *et al.*, 2008). Experimental studies show that BCAAs inhibit the insulin receptor substrate via the mTOR/p70S6K/S6K pathway (Lu *et al.*, 2013). Accordingly, in the study by Newgard *et al.* (2009), addition of BCAAs to a high-fat diet in rats promoted the development of insulin resistance. Recently, BCAAs, phenylalanine and tyrosine were shown to associate with future insulin resistance (Würtz *et al.*, 2012), future T2D (Wang *et al.*, 2011), and prevalent metabolic syndrome (Wiklund *et al.*, 2014). In this study, MetM4 (representing BCAA metabolites) associated positively with markers of insulin resistance (HbA1c: $p = 7.2 \times 10^{-10}$; OGTT 2-hours glucose: $p = 9.0 \times 10^{-9}$) and metabolic syndrome prevalence ($p = 3.8 \times 10^{-5}$) (Figure 4.20).

Energy metabolism

ΔBW was positively associated with pyruvate, lactate, and alanine in this study. Concentrations of these metabolites have been shown to be elevated in obesity (Newgard *et al.*, 2009). Elevated levels of the three metabolites are markers of mitochondrial dysfunction (Haas *et al.*, 2008). Mitochondria, the cells' power plants, produce adenosine triphosphate (ATP) from carbohydrates, fats, and proteins via tricarboxylic acid cycle and β-oxidation (Rogge, 2009). In states of insufficient oxygen supply or mitochondrial dysfunction, pyruvate from glycolysis is converted to lactate via lactic acid fermentation and to alanine via transamination. Obesity is associated with decreased fatty acid β-oxidation, rendering obese individuals more dependent on the glycolytic pathway for ATP production (Rogge, 2009). At the same time, mitochondrial size is reduced, and respiratory chain activity diminished in obesity (Kelley *et al.*, 2002), explaining the accumulation of pyruvate and the formation of lactate and alanine. The results of the present study suggest that weight change, even in non-obese subjects, gives rise to alterations in mitochondrial function.

Pyruvate, lactate and alanine clustered together with VLDL in MetM1 in this study. Aluminum-induced mitochondrial dysfunction was found to promote VLDL secretion in human hepatocytes (Mailloux *et al.*, 2007), suggesting a role of mitochondrial dysfunction as a further link between ΔBW and dyslipidemia as well as cardiometabolic disease. In addition, levels of pyruvate, lactate and alanine were predictive of future glucose intolerance, independent of body mass (Würtz *et al.*, 2012). For lactate, a predictive role for future diabetes independent of traditional risk factors such as fasting glucose and insulin was reported (Juraschek *et al.*, 2013). It is also associated with carotid atherosclerosis (Shantha *et al.*, 2013) and predictive for heat failure and all-cause mortality (Matsushita *et al.*, 2013).

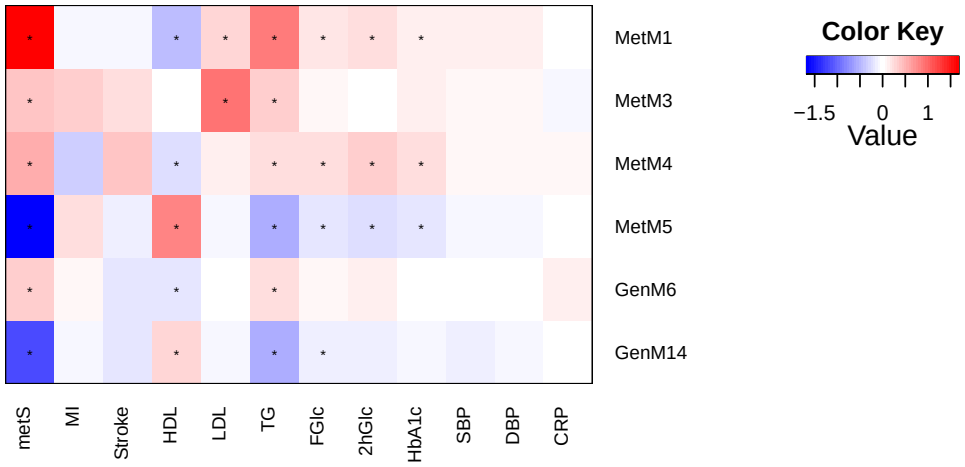

Figure 4.20: Association of the identified omics modules with clinical traits. Results are derived from linear (continuous traits, log-transformed) and logistic (binary traits) regression models adjusted for age, sex, body weight, lipid-lowering medication (overall, statins, fibrates), antihypertensive medication, antidiabetic medication as well as corticoid intake. Significant associations ($p < 6.9 \times 10^{-4}$ corresponding to Bonferroni correction for 72 tests) are denoted as black stars. 2hGlc, oral glucose tolerance test (OGTT) 2-hours glucose; CRP, C-reactive protein; DBP, diastolic blood pressure; FGlc, fasting glucose; GenM, gene expression module; HDL, high density lipoprotein cholesterol; LDL, low density lipoprotein cholesterol; MetM, metabolite module; metS, metabolic syndrome, defined according to the ATP III criteria (Adult Treatment Panel III, 2002), see legend of Table 4.10; MI, myocardial infarction; SBP, systolic blood pressure; TG, triglycerides. See Table 4.10 for descriptives of clinical traits.

Table 4.10: Description of clinical traits in the KORA F4 study.

Trait	**Mean (sd)**	**Median (range)**
HDL cholesterol (mg/dl)	56.2 (14.5)	54.0 (21.0, 123.0)
LDL cholesterol (mg/dl)	140.0 (34.6)	138.0 (44.0, 291.0)
Triglycerides (mg/dl)	132.7 (90.1)	111.0 (26.0, 1627.0)
Fasting glucose (mg/dl)	101.1 (20.0)	97.0 (67.0, 341.0)
OGTT 2-hours glucose (mg/dl)	118.6 (39.9)	111.0 (35.0, 465.0)
HbA1c (%)	5.6 (0.6)	5.5 (4.4, 12.1)
Systolic blood pressure (mmHg)	125.2 (18.6)	124.0 (57.5, 223.5)
Diastolic blood pressure (mmHg)	76.1 (9.9)	75.5 (38.5, 120.5)
C-reactive protein (mg/L)	2.9 (6.5)	1.5 (0.2, 145.0)
Trait	**Frequency**	**Relative frequency (%)**
Metabolic syndrome (ATP III) (yes/no)	385 / 1246	23.6 / 76.4
Myocardial infarction (yes/no)	60 / 1570	3.7 / 96.3
Stroke (yes/no)	45 / 1585	2.8 / 97.2

ATP III, Adult Treatment Panel III criteria (Adult Treatment Panel III, 2002), according to which metabolic syndrome prevalence is defined as three of (1) abdominal obesity (waist circumference $>$ 102 cm in males and $>$ 88 cm in females), (2) high TG ($\geq$ 150 mg/dl), (3) low HDL cholesterol ($<$ 40 mg/dl in males and $<$ 50 mg/dl in females), (4) hypertension (systolic blood pressure $\geq$ 130 mmHg or diastolic blood pressure $\geq$ 85 mmHg) and (5) high fasting glucose ($\geq$ 110 mg/dl); HDL, high density lipoprotein; LDL, low density lipoprotein; OGTT, oral glucose tolerance test.

4.3.3 ΔBW associates with the lipid-leukocyte module and a novel gene expression module

Two of the 19 modules of metabolite-related transcripts showed association with ΔBW. GenM14, which comprised 17 transcripts (Figure 4.21) and was negatively associated with ΔBW, contained all of the 11 transcripts of the "lipid-leukocyte (LL) module" previously described as a leukocyte gene expression module strongly related to blood lipids (Inouye *et al.*, 2010b) and a large number of serum metabolites including lipoprotein subclasses, lipids, glycoproteins and amino acids (Inouye *et al.*, 2010a). The authors discussed this module as being involved in basophil and mast cell-related immune response and allergy. For instance, the core gene *HDC* codes for a protein converting histidine to histamine, which is secreted by basophils and mast cells in response to IgE sensitization. Accordingly, when Ingenuity pathway analysis was applied (see Section 3.5.1), "FC Epsilon RI Signaling" ($p = 1.1 \times 10^{-4}$), "Histamin Biosynthesis" ($p = 9.1 \times 10^{-4}$) and "Airway Inflammation in Asthma" ($p = 3.7 \times 10^{-3}$) were the top three canonical pathways (Table 4.11), although results are based on only one gene per pathway. To address the question of whether the association of ΔBW with GenM14 reflects merely a shift in the proportion of basophil/mast cells upon weight change, the model was adjusted for the transcripts associated with basophil proportions in the publication by Whitney *et al.* (2003) through the top 25 principal components (explaining 84.6% of variance in the data) (see Section 3.3.2 for cell type confounding correction and Section 3.4.1 for PCA). Association of GenM14 with ΔBW remained stable ($p = 7.7 \times 10^{-4}$). Still, residual confounding cannot be completely excluded.

As mentioned above, GenM14 showed strong partial correlation with the VLDL-related MetM1. Including MetM1 as covariate in the model of GenM14 on ΔBW abolished the association (p = 0.284), suggesting that MetM1 accounts for the association between ΔBW and GenM14. In agreement with this, causal inference provided evidence that the LL module was mainly responsive to serum metabolites rather than showing a causal effect on them (Inouye *et al.*, 2010a).

The core of GenM14 (module membership strength > 0.8) comprised the LL genes *HDC, GATA2, SLC45A3, MS4A2*, and *SPRYD5*. Six further transcripts, *IL4, TRIM49L1, TEX101, EPAS1/HIF2-α, CCNA1* and *CAV2*, were co-expressed with the LL module genes, although with weaker module membership (ranging from 0.48 to 0.54), suggesting that they might share functionality with the LL module genes. Indeed, *IL4* codes for the cytokine interleukin 4 which has long been known to induce differentiation of naïve T cells to Th2 cells that play a role in allergen response. The co-expression of *IL4* within the basophil-related gene cluster confirms the identification of basophils as cells reacting to allergens by IL4 secretion (Sokol *et al.*, 2008). *EPAS1/HIF2-α* encodes a component of the hypoxia inducible transcription factor (HIF), which regulates responses to reduced oxygen and for which also a role in regulating inflammation (Imtiyaz *et al.*, 2010) and

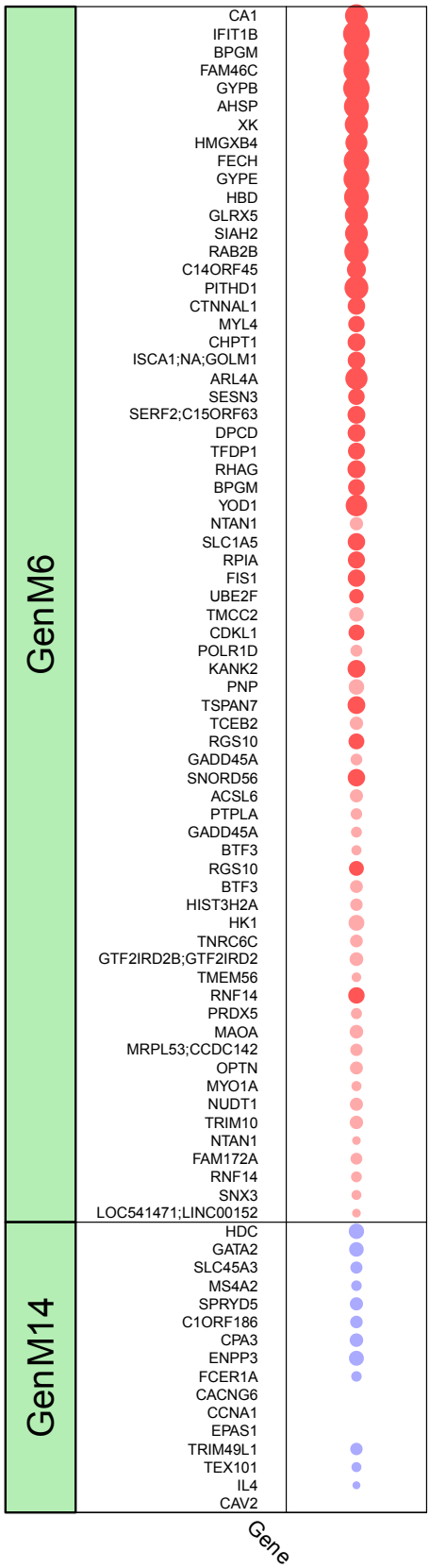

Figure 4.21: Association of annual percentage body weight change (ΔBW) with members of associated gene expression modules (GenM). Bubbles represent effect strengths and significance, see legend of Figure 4.16. Models were adjusted for age, sex and baseline body weight. For single transcripts, the significance threshold was chosen as $p < 2.0 \times 10^{-5}$ corresponding to Bonferroni correction for 2537 metabolite-related transcripts. Genes are sorted by their module membership strength, as determined by the correlation of transcript level with the module eigengene (ME) (see Section 3.4.1). Gene annotations where derived from the UCSC database.

Table 4.11: Ingenuity pathway analysis. Top 5 canonical pathways enriched in gene expression modules (GenM) 6 and 14.

Canonical pathway	Ratio	p-value	Molecules
GenM6			
Xanthine and Xanthosine Salvage	1/1	3.4E-03	PNP
Guanine and Guanosine Salvage I	1/2	6.7E-03	PNP
Adenine and Adenosine Salvage I	1/2	6.7E-03	PNP
Trehalose Degradation II (Trehalase)	1/3	1.0E-02	HK1
Heme Biosynthesis from Uroporphyrinogen-III I	1/3	1.0E-02	FECH
GenM14			
Fc Epsilon RI Signaling	3/106	1.1E-04	MS4A2,FCER1A,IL4
Histamine Biosynthesis	1/1	9.1E-04	HDC
Airway Inflammation in Asthma	1/4	3.7E-03	IL4
Tec Kinase Signaling	2/148	7.9E-03	MS4A2,FCER1A
Role of NFAT in Regulation of the Immune Response	2/160	9.1E-03	MS4A2,FCER1A

Ratio, number of genes that are member of the respective module and pathway divided by the total number of genes in the pathway. p-value is derived from Fisher's exact test (Section 3.5.1).

energy balance (Zhang *et al.*, 2011) has been reported. Also for TRIM proteins, a role within innate immune defense has been reported (Ozato *et al.*, 2008), although little is known about the specific member TRIM49L1.

Of note, the association between GenM14 and ΔBW as well as the VLDL-related MetM1 is negative, suggesting a reduced expression of these genes with weight gain and subsequent VLDL increase. Although these findings are in line with the negative association between the LL module and VLDL metabolites reported by Inouye *et al.* (2010a), they are contradictory to an analysis by Gonen *et al.* (1987), in which VLDL was found to trigger the release of histamine from human basophils. Furthermore, obesity is a risk factor for asthma, and weight gain was found to increase the risk of developing airway hyperresponsiveness (Shore, 2010). A postulated mechanism is the increase of number and function of viable basophils through action of leptin (Suzukawa *et al.*, 2011). It remains to be determined how these results fit to the observation of decreased gene expression associated with basophil/mast cell level or function upon weight gain in the present study.

Besides GenM14, a larger gene expression module (GenM6, comprising 71 transcripts) was identified as being strongly positively related to ΔBW. In contrast to GenM14, ΔBW association with GenM6 was largely stable towards adjustment for the metabolite MEs, suggesting independent effects ($p = 2.2 \times 10^{-8}$). The core of GenM6 (module membership strength > 0.8) comprised *CA1, IFIT1L, BPGM, FAM46C, GYPB, AHSP, XK, HMGXB4, FECH, GYPE, HBD* and *GLRX5*. A manual literature search revealed at least 9 of these 12 genes as red blood cell-related genes. For instance, *HBD* encodes the hemoglobin delta subunit, *AHSP* encodes α-hemoglobin stabilizing protein, BPGM regulates oxygen affinity of hemoglobin, and FECH codes for the enzyme ferrochelatase/heme synthase which is involved in heme synthesis. A formal enrichment analysis for gene ontology (GO) terms (see Section 3.5.1) revealed "hemoglobin metabolic process" and

"hemoglobin complex" among the top three enriched biological pathways (all $p = 8.4 \times 10^{-4}$, not significant after multiple testing correction, Table 4.12). In Ingenuity pathway analysis, "heme biosynthesis from uroporphyrinogen-III I" was among the top five upregulated canonical pathways ($p = 9.9 \times 10^{-3}$, not significant after multiple testing correction) (Table 4.11).

Table 4.12: Gene ontology enrichment analysis. Top 5 pathways enriched in gene expression modules (GenM) 6 and 14.

Term name	Term ID	Term Ontology	Ratio	p-value
GenM6				
bicarbonate transport	GO:0015701	BP	2/2	8.4E-04
hemoglobin metabolic process	GO:0020027	BP	2/2	8.4E-04
hemoglobin complex	GO:0005833	CC	2/2	8.4E-04
organic anion transport	GO:0015711	BP	5/30	1.5E-03
anion transport	GO:0006820	BP	5/32	2.0E-03
GenM14				
catecholamine metabolic process	GO:0006584	BP	2/5	4.6E-04
phenol-containing compound metabolic process	GO:0018958	BP	2/6	6.8E-04
negative regulation of myeloid leukocyte differentiation	GO:0002762	BP	2/7	9.5E-04
RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001077	MF	2/7	9.5E-04
RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001228	MF	2/8	1.3E-03

Ratio, number of genes that are member of the respective module and pathway divided by the total number of genes in the pathway. p-value is derived from Fisher's exact test (Section 3.5.1).

Consequently, it was hypothesized that the transcripts in GenM6 are reflective of red blood cell development, since immature red blood cells, reticulocytes, contain remnant mRNA (Goh *et al.*, 2007) which is depleted during erythrocyte maturation. An increased hematopoiesis upon diet-induced obesity in rats has been observed, putatively through action of leptin in the bone marrow (Trottier *et al.*, 2012), whereas glycosylated hemoglobin shows an inverse relationship with erythrocyte survival (Virtue *et al.*, 2004). A shift towards a larger proportion of immature red blood cells upon weight gain would be consistent with these observations. Accordingly, adjustment of the model for total red blood cell count, hemoglobin concentration, hematocrit, mean corpuscular haemoglobin (MCH), mean corpuscular haemoglobin concentration (MCHC) as well as mean cell volume of erythrocytes (MCV) did not abolish the association ($p = 2.2 \times 10^{-12}$), whereas adjustment for the first 5 principal components of the red blood cell distribution width (RDW)-related transcripts reported by Whitney *et al.* (2003) did ($p = 0.340$), whereby RDW is proportional to reticulocyte count (Roberts and El Badawi, 1985).

Neither of both ΔBW-related GenM's seemed to comprise genes with a well-established relationship to lipid metabolism, as might be expected after preselecting metabolite-related transcripts. Examplarily, the genes *LIPC, CETP* and *PLTP* discussed above within the context of lipoprotein metabolism were looked up, as well as *ABCG1* which has been discussed together with *CETP* as strongly upregulated transcript in adipose tissue in response to diet-induced weight loss (Johansson *et al.*, 2012). Whereas *ABCG1* transcripts tended to show a negative association with ΔBW (best $p = 6.7 \times 10^{-5}$ for transcript ILMN_2329927, which clustered in GenM1), transcripts of the other three genes were not related to either ΔBW or metabolites. These results could have been expected considering the tissue origin of these proteins. In line with these findings, the strong obesity-related changes in adipose tissue gene expression were weakly represented by blood cell transcriptomics in the study by Emilsson *et al.* (2008).

4.3.4 Stability of the multi-omic associations

Next, different analyses were performed to assess the stability of the multi-omic network and its relation to ΔBW. First, it was argued that metabolic effects of weight loss (negative ΔBW) and weight gain (positive ΔBW) might not be strictly opposing, and that diverging effects might remain unexplored when linear models are used. Therefore, stratified analyses (see Section 3.3.1) were performed in the group of subjects with weight loss ($n = 641$; 316 with gene expression data) and in the group of subjects with weight gain ($n = 990$; 373 with gene expression data) (Figure 4.16, second and third column). Overall, weight loss and weight gain tended to show opposing associations with the modules (Figure 4.16: same color of circles denoting association). By trend, associations of MetM1, 3, 4 and 5, and GenM14 were stronger in subjects with weight loss than with weight gain. In contrast, GenM6 showed by trend a stronger association in the group with weight gain. However, none of these differences were significant (Figure 4.16: black arrows).

In addition, the effect of ΔBW on inter- and intra-module connectivity of the network elements was investigated, since previous studies suggested sensitivity of metabolic network topology towards external factors (Inouye *et al.*, 2010a, Valcárcel *et al.*, 2014) (see Section 3.4.1 for statistical methodology). No significant differences in network connectivity were observed between the groups with weight gain and weight loss (all $p > 0.01$).

The generally opposing associations of weight loss and weight gain with the blood metabolome are in line with the studies by Mäntyselkä *et al.* (2012) and Naganuma *et al.* (2009), where the majority of ΔBW associations with lipoprotein measures were linear across the weight change range, and weight loss and weight gain showed opposite effects. Interestingly, the effect of weight loss ($\geq$ 5% across 6.5 years) versus stable weight on VLDL subclasses and L-HDL was stronger in absolute terms than the effect of weight gain ($\geq$ 5%) versus stable weight (Mäntyselkä *et al.*, 2012). These findings are in accordance with the stronger associations for MetME1 and MetME5 observed

for weight loss in this study. Although larger studies in subjects with a larger range of weight change might have more power to differentially investigate the effects of weight loss versus weight gain, the presented results suggest that differences are not large and that in general, weight loss is capable of reverting the effects of weight gain on the blood metabolome and transcriptome. Accordingly, it was shown in a randomized controlled trial that normalization of obesity led to a reversal of an unfavorable LDL subfraction pattern (Siri-Tarino *et al.*, 2009).

Further subgroup analyses were performed, assuming that weight change effect might depend on (central) obesity, on sex and on age (Figure 4.16). Again, no significant subgroup-specific effects were observed.

Body weight change over a period of 7 years might be due to several reasons, including changes in lifestyle, the occurrence of diseases, and changes in medication. For these reasons, the sensitivity of the observed associations with ΔBW towards adjustment for changes in lifestyle factors, for the occurrence of diseases, and finally for changes in medication was investigated in three separate models (Figure 4.22). None of the three models showed a change in effect sizes across the modules, indicating that the observed associations were primarily due to the change in body weight *per se* rather than the mechanisms that might have facilitated weight change. Note, however, that the majority of the variables reflecting changes in lifestyle, disease and medication were obtained from interviews and might therefore have insufficient accuracy. Also, nutrition was only obtained from the baseline timepoint, so the effect of changes could not be investigated.

Together, these results suggest that the metabolite-gene network and its relation to weight change reflect a largely stable system.

Several extensions of this study seem worthwhile. First, it would be interesting to obtain a higher resolution of body weight measurements during follow-up, as well as of metabolomics and gene expression measurements, to decipher the longitudinal sequence of metabolic changes and to study the metabolic processes related to weight cycling. In addition, extending whole blood transcriptomics to different tissues seems extremely promising, considering that blood might only weakly reflect weight-related transcriptional changes in tissues (Emilsson *et al.*, 2008), and that blood metabolites originate from different tissues. In addition, the issue of cell type confounding in whole blood transcriptomics has to be mentioned, which is discussed in the context of epigenomics in Section 4.2.9. Similar to the approach applied in Section 4.2, the adjustment for transcripts previously reported to relate to certain cell types, as it was done in this section, is subject to the quality of the initial data and analysis strategy (i.e., Whitney *et al.* (2003)), and limited to the cell types selected. Nevertheless, gene expression signatures identified in blood will be of large practical relevance since blood is most easily accessible also in a clinical setting. Also, in the context of weight change and its metabolic consequences, integrating metabolomics and blood cell transcriptomics data is relevant from the perspective that blood cells may

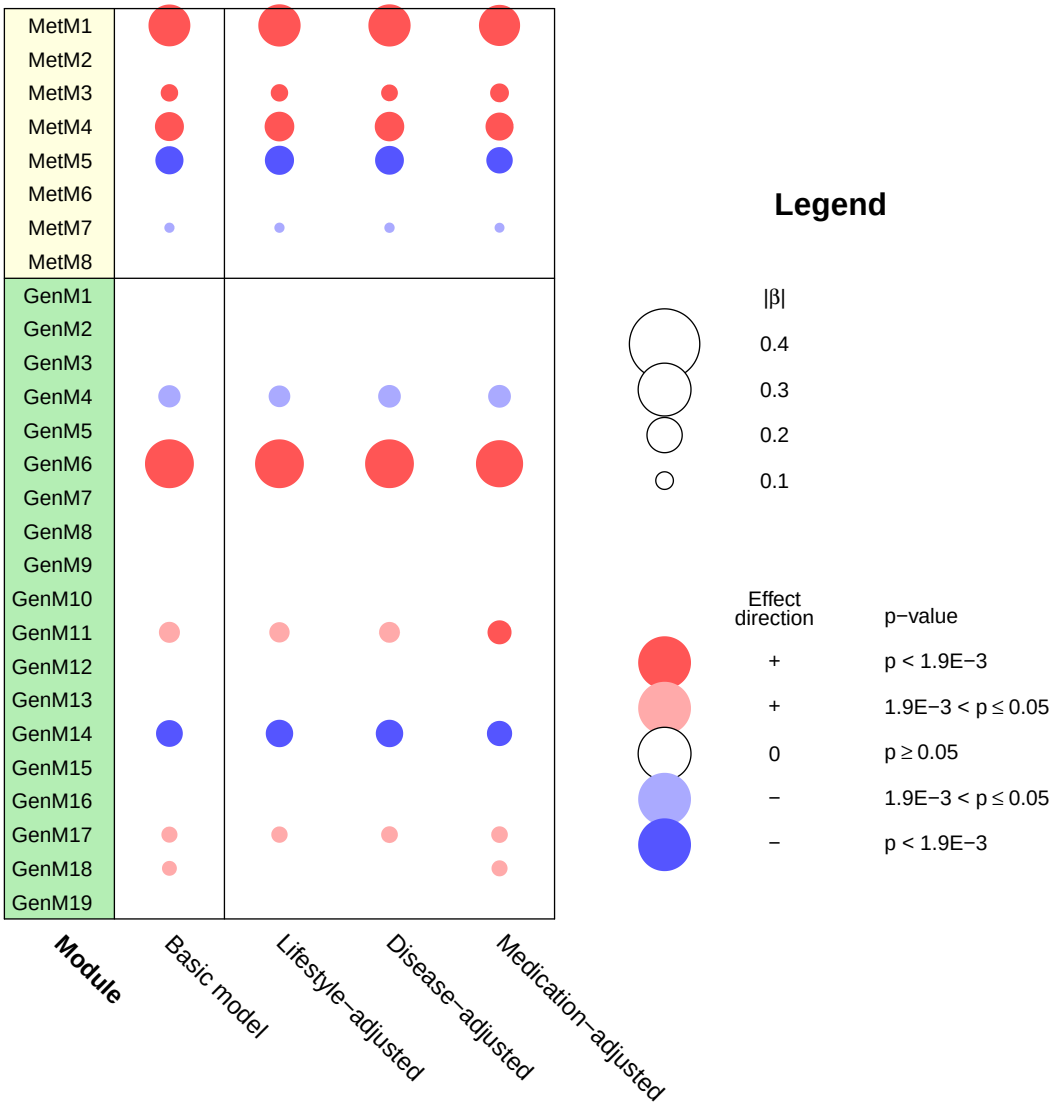

Figure 4.22: Association of annual percentage body weight change (ΔBW) with omics modules adjusting for factors driving weight change. Bubbles represent effect strengths and significance, as described in the legend. Models were adjusted for age, sex and baseline body weight. Significance threshold $p < 1.9 \times 10^{-3}$ corresponds to Bonferroni correction for 27 models. Lifestyle-adjusted: changes in physical activity, baseline nutritional score, changes in sleeping behavior, smoking and alcohol drinking were included as covariates in the model. Disease-adjusted: incident diabetes, cancer, myocardial infarction and stroke were included as covariates in the model. Medication-adjusted: change in the intake of beta-blockers, antidiabetic drugs, systemic corticoids, oral contraceptives and antidepressants were included as covariates in the models. GenM, gene expression module; MetM, metabolite module.

interact with blood substances in the etiology of atherosclerotic events (Inouye *et al.*, 2010a).

4.3.5 Conclusions

Through the integration of two-platform serum metabolomic and whole blood transcriptomic data and the formation of modules of closely connected molecules, a comprehensive characterization of metabolic effects of body weight change over a 7-year period in a large population-based cohort was obtained. Weight gain and weight loss were strongly and opposingly associated with the blood metabolome, with VLDL, LDL, large HDL subclasses, TGs, BCAAs and markers of energy metabolism as core molecules of the four metabolite modules. These associations point towards the development of dyslipidemia, disturbed amino acid metabolism as well as mitochondrial dysfunction upon weight gain. Two weight change-related gene expression modules pinpoint immune cells (mast cells, basophils) and reticulocytes as blood cell types putatively playing a role in weight change-related blood metabolism. Metabolite and gene expression modules were associated with clinical phenotypes, suggesting a role in linking excess body weight with metabolic and cardiovascular comorbidities. The findings of this study also support the hypothesis that clustering omics data prior to analyzing associations with a phenotype has increased power for identifying biologically relevant pathways (Chuang *et al.*, 2007, Inouye *et al.*, 2012). GenM14 ("LL module") was found to be associated with weight change, although none of the contributing genes showed a univariate association with weight change that would have passed significance after correction for multiple comparisons.

Together, this study provides evidence for a largely reversible effect of long-term weight gain in the general population on an integrated blood metabolomic and transcriptomic network. This improves the knowledge on molecular processes elicited by weight change and potentially linking it to comorbidities.

4.4 Metabolomic determinants of weight loss during lifestyle intervention in obese children

Childhood obesity is primarily treated with lifestyle intervention approaches based on physical activity and nutritional as well as behavior modification (Han *et al.*, 2010). The degree of overweight reduction during such intervention programs differs largely between participants, and not all children achieve sufficient overweight reduction (Reinehr *et al.*, 2004, Sabin *et al.*, 2007, Ford *et al.*, 2010).

Hence, the search for factors predicting a child's response to a lifestyle intervention is of great interest. With knowledge of such factors, lifestyle based therapeutic options could be focused on the children that are likely to benefit most, reducing the psychosocial and financial burden of unsuccessful participation (Reinehr *et al.*, 2003). In addition, a thorough understanding of the metabolic processes underlying the large inter-individual variability in weight loss is essential for the development of personalized intervention strategies.

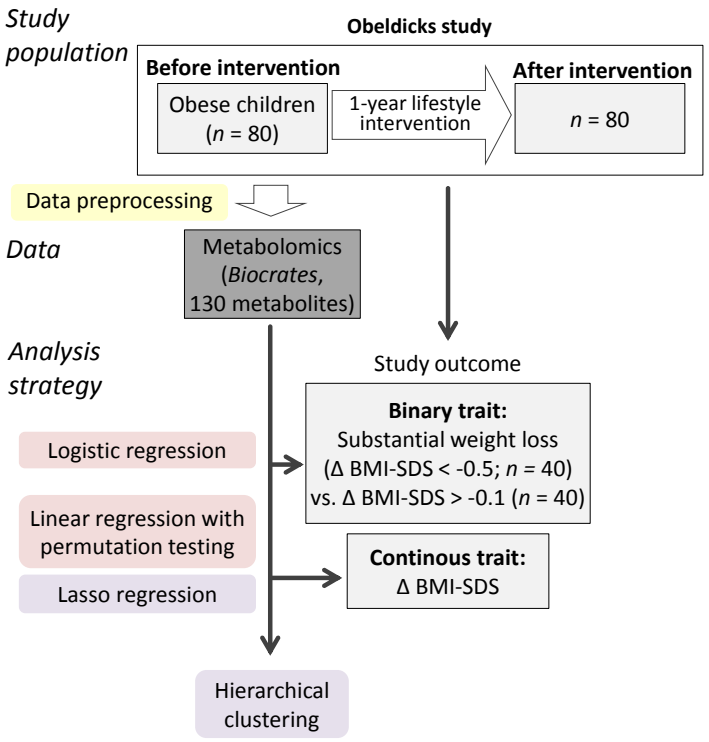

Figure 4.23: Prediction of weight loss in the Obeldicks study: Study design and analysis strategy. The Obeldicks study is described in Section 2.3. Color coding of statistical methods: yellow, data preprocessing and quality control (Section 3.1); red, univariate data analysis (Section 3.3); violet, multivariate data analysis (Section 3.4). BMI-SDS, body mass index standard deviation score; LASSO, least absolute shrinkage and selection operator.

Recently, metabolomics has become an attractive tool in exploring metabolic determinants of weight loss success (Pathmasiri *et al.*, 2012). The aim of the present study was to identify serum metabolites and anthropometric as well as clinical variables associated with weight loss of obese children during the intervention program *Obeldicks*, and to build a multivariate predictive model for BMI-SDS change during intervention.

This section was published as

- **Wahl S**, Holzapfel C, Yu Z, Breier M, Kondofersky I, Fuchs C, Singmann P, Prehn C, Adamski J, Grallert H, Illig T, Wang-Sattler R, Reinehr T (2013). "Metabolomics reveals determinants of overweight reduction during lifestyle intervention in obese children." *Metabolomics*, **9**(6), 1157-1167.

4.4.1 Study characteristics at baseline and changes upon lifestyle intervention

The overall study design and analysis strategy is visualized in Figure 4.23. By design, baseline age, sex and pubertal stage, but also weight, BMI and BMI-SDS distribution did not differ significantly between the 40 children who substantially reduced their BMI-SDS (ΔBMI-SDS $\leq$ -0.5) and the 40 children who did not (ΔBMI-SDS $>$ -0.1) (Table 4.13).

During the intervention, ΔBMI-SDS ranged from -1.49 to +0.49 and differed significantly between children with and without substantial BMI-SDS reduction, with a mean (sd) ΔBMI-SDS of -0.68 (0.27) and +0.07 (0.15), respectively ($p = 1.4 \times 10^{-14}$). Children

with substantial BMI-SDS reduction significantly improved their waist circumference (-6.0 (15.2) cm, $p = 5.8 \times 10^{-3}$) as well as their metabolic risk profile (fasting insulin: -5.3 (9.3) mU/l, $p = 2.2 \times 10^{-4}$; homeostasis model assessment of insulin resistance (HOMA-IR): -0.5 (4.9), $p = 4.8 \times 10^{-4}$; HDL: +3.9 (10.2) mg/dl,$p = 4.8 \times 10^{-2}$; triglycerides (TGs): -17.9 (34.4) mg/dl, $p = 5.3 \times 10^{-3}$; systolic blood pressure: -7.6 (19.5) mmHg, $p = 2.3 \times 10^{-3}$). In contrast, children without substantial BMI-SDS reduction mostly did not (see Supplementary Table 2 of the original publication (Wahl *et al.*, 2013a)).

Table 4.13: Baseline characteristics of the Obeldicks study population.

Variable	Children with substantial BMI-SDS reduction (n = 40)	Children without substantial BMI-SDS reduction (n = 40)	p-value
Age (years)	10.9 (2.3)	10.9 (2)	0.969
Sex (% male)	50	55	0.751
Pubertal stage (% prepubertal)	52.5	50	1.000
Weight (kg)	64.1 (16.3)	66.3 (18.8)	0.641
BMI (kg/m^2)	27.3 (3.3)	28 (4.6)	0.749
BMI-SDS	2.35 (0.43)	2.37 (0.45)	0.837
Waist circumference (cm)	83.8 (10.5)	92.4 (12.7)	0.009

Data are shown as mean (standard deviation) if not indicated otherwise. p-values were derived from Wilcoxon rank-sum test and χ^2 tests for continuous and dichotomous variables, respectively. BMI, body mass index; BMI-SDS, BMI standard deviation score.

4.4.2 Pre-intervention variables associated with BMI-SDS reduction

In total, 144 pre-intervention variables, including 130 metabolites and 14 anthropometric or clinical traits, were subjected to univariate logistic regression with the binary outcome "substantial BMI-SDS reduction", adjusted for sex, baseline age, pubertal stage and BMI-SDS. None of the variables reached significance after correction for multiple testing (Benjamini-Hochberg, see Section 3.3.4).

Assuming an increased power when replacing dichotomized by continuous ΔBMI-SDS as response variable, linear regression models were fitted to the continuous outcome ΔBMI-SDS. 18 variables showed a significant positive association with ΔBMI-SDS after correction for multiple testing (permutation p-values ranging from 5.3×10^{-3} to 1.0×10^{-4}, Figure 4.24, see Section 3.3.3 for permutation test). These variables included waist circumference, arginine and lysophosphatidylcholine (LPC) a C18:0 serum concentrations, as well as serum concentrations of 13 diacyl phosphatidylcholines (PCs) and two acyl-alkyl PCs, which were all long-chained and unsaturated. Most of these variables were also nominally associated with substantial BMI-SDS reduction (Figure 4.24). By trend, a positive association was observed for all measured diacyl PCs (see detailed association results in Supplementary Table 2 of the original publication (Wahl *et al.*, 2013a)). None of the baseline clinical traits (blood pressure, blood lipid and insulin resistance parameters) were

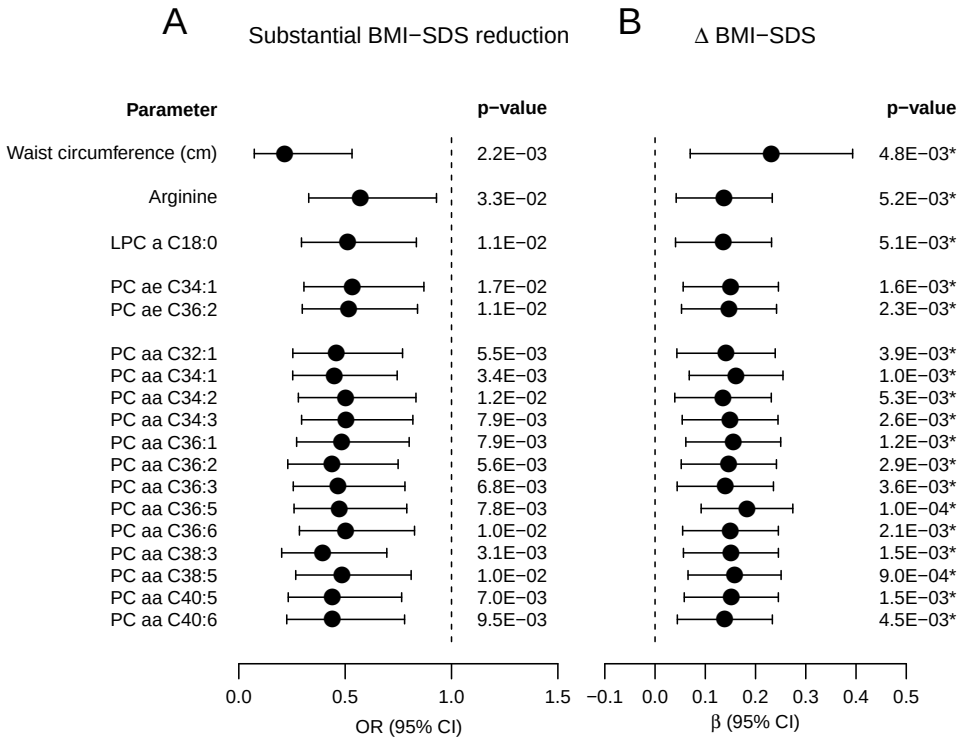

Figure 4.24: Baseline parameters associated with overweight reduction during the intervention. Associations with **A** the binary response "substantial BMI-SDS reduction" (ΔBMI-SDS $\leq$ -0.5 vs. ΔBMI-SDS > -0.1, odds ratios (OR) with 95% confidence interval (CI)) and **B** the continuous response ΔBMI-SDS (β estimates with 95% CI, permutation p-values, see Section 3.3.3) are shown for the 18 variables significantly associated with ΔBMI-SDS after correction for multiple testing. Results are based on univariate regression models adjusted for sex and baseline age, pubertal stage and BMI-SDS. The unit of variables is μmol/L, if not indicated otherwise. *Significant after correction for multiple testing. BMI-SDS, body mass index standard deviation score; Cx:y, acyl-group with chain length x and y double bonds; LPC a, lysophosphatidylcholine with acyl chain; PC aa, diacyl phosphatidylcholine; PC ae, acyl-alkyl phosphatidylcholine.

significantly associated with ΔBMI-SDS after correction for multiple testing.

4.4.3 Prediction of overweight reduction

ΔBMI-SDS was further investigated using a multivariate approach, *least absolute shrinkage and selection operator* LASSO (see Section 3.4.2), in order to identify markers that represent groups of highly correlated baseline variables playing a role in the determination of successful overweight reduction, and to assess their predictive potential. Three out of the 144 variables were selected into the predictive model, namely waist circumference, PC aa C36:5, and PC aa C32:2. Figure 4.25 shows coefficient paths and variable stability for these variables. The strongest effect and highest stability, that is, the highest selection frequency across the cross-validation (CV) folds, was observed for PC aa C36:5 ($\beta = 0.0152$, selection frequency 100%). Of note, LASSO coefficients are not comparable with the coefficients of the univariate linear regression models due to the shrinkage behavior of LASSO

(see Section 3.4.2).

In terms of prediction accuracy, the model had R^2 and Q^2 values of 0.267 and 0.116, respectively (Figure 4.26). The significance of the prediction was assessed using a permutation test with the null hypothesis stating that a Q^2 value of 0.116 would be observed by chance (see Section 3.3.3). The corresponding p-value was 4.6×10^{-3}, so this hypothesis was rejected. Thus, it could be shown that the predictive model comprising three metabolic variables explains a significant part of ΔBMI-SDS in obese children during one-year lifestyle intervention.

The three variables selected into the LASSO model were also univariately associated with ΔBMI-SDS (Figure 4.24), with the exception of PC aa C32:2, for which a univariate association was observed only by trend. The selected variables represented groups of correlated variables significantly associated with ΔBMI-SDS in the univariate regression analysis, as can be seen from the correlation and clustering results (Figure 4.27).

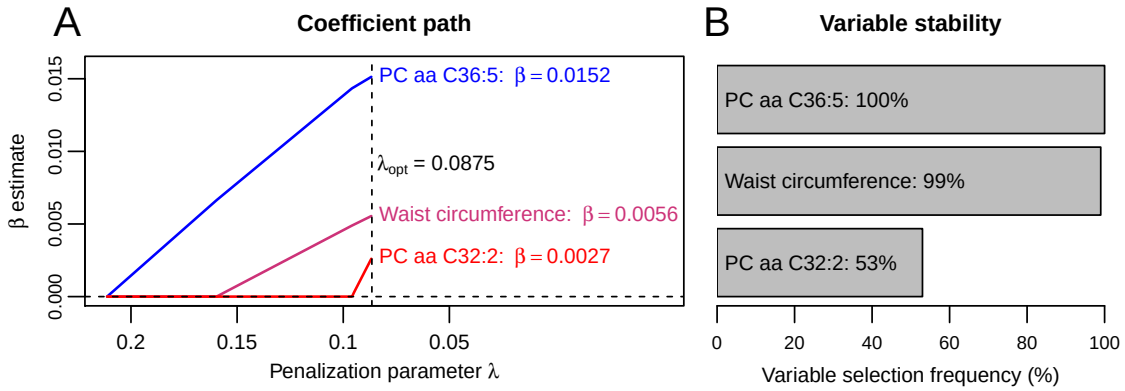

Figure 4.25: LASSO regression results. Pre-intervention variables selected as predictors for ΔBMI-SDS. **A** Coefficient paths truncated at the chosen penalization parameter $\lambda_{opt} = 0.0875$ (vertical dashed line). β estimates are plotted against a sequence of the penalization parameter λ ranging from the λ threshold beyond which no variables were retained in the model, to λ_{opt}. β estimates are displayed for λ_{opt}. **B** Variable stability, defined as the frequency with which a variable was selected by the LASSO approach across the 100 outer cross-validation (CV) loops, for the chosen variables. Cx:y, acyl-group with chain length x and y double bonds; PC aa, diacyl phosphatidylcholine.

4.4.4 Discussion

In this study, pre-intervention factors determining response to lifestyle intervention in obese children were investigated using a targeted metabolomics approach combined with clinical and anthropometric measurements, followed by univariate and multivariate statistical analysis. The factors that showed the strongest association as well as the most stable predictive potential for weight loss were serum concentrations of diacyl phosphatidylcholines (PCs), and waist circumference.

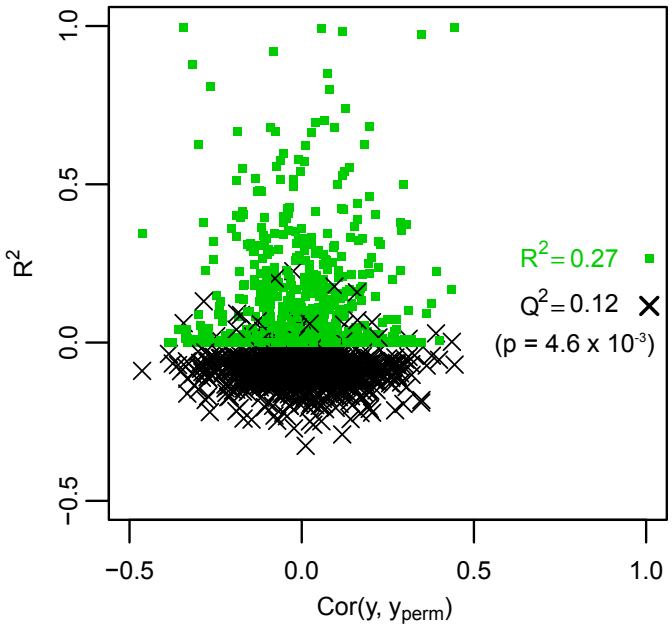

Figure 4.26: Permutation test results for the LASSO approach. Data for the first 1,000 permutations are shown. R^2 values (green squares) and Q^2 values (black crosses) are plotted against the Pearson's correlation between original and permuted outcome vector. R^2 is limited to ≥ 0, whereas Q^2 is not. At correlation = 1, R^2 and Q^2 values of the original data are plotted. Permutation-based p-value for Q^2 is given. Cor, Pearson's correlation coefficient; perm, permutation.

Phosphatidylcholines and weight loss

Children with substantial BMI-SDS reduction had lower pre-intervention serum concentrations in several PC species compared to children without substantial BMI-SDS reduction. PCs are produced in most mammalian cells via the cytidine diphosphate (CDP)-choline pathway (DeLong *et al.*, 1999). In the liver, 30% of PC synthesis occurs via the phosphatidylethanolamine methyltransferase (PEMT) pathway (Li and Vance, 2008). The enzyme PEMT methylates phosphatidylethanolamine to produce PCs, which constitutes the only endogenous pathway of choline synthesis. The PC species derived from both pathways differ in chain length and degree of saturation (DeLong *et al.*, 1999).

The long-chain unsaturated PCs C34:1, C34:3, C36:2, C36:3, C36:5, C38:5 and C40.6 were negatively associated with BMI-SDS reduction in this study and have recently been shown to be down-regulated in livers of PEMT-/- mice (Jacobs *et al.*, 2010). Also, total serum PC concentration was reduced in PEMT-/- mice. Most interestingly, PEMT-/- mice were protected from high-fat diet-induced obesity, having an increased energy expenditure and normal peripheral insulin sensitivity. These effects were prevented by choline supplementation. Thus, they are attributable to reduced choline availability upon diminished choline

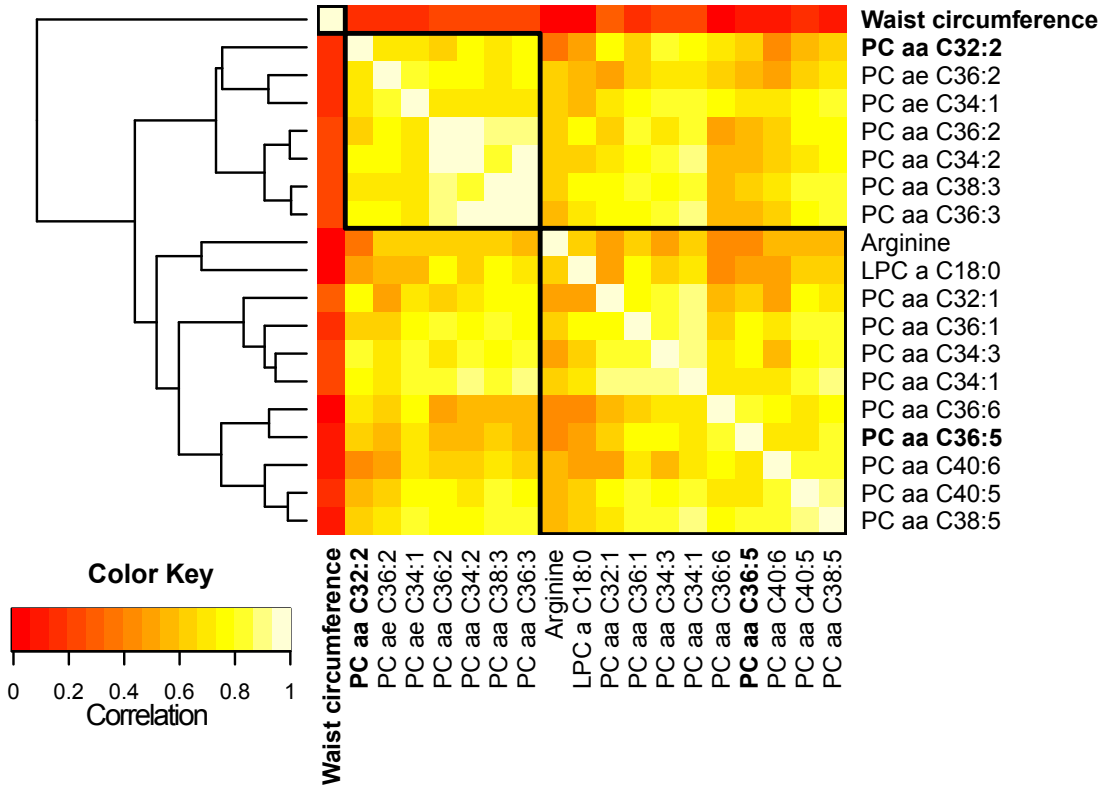

Figure 4.27: Correlation among parameters associated with overweight reduction. Heatmap of the matrix of pairwise Pearson's correlation coefficients and hierarchical clustering dendrogram are shown (see Section 3.4.1). Variables selected in the LASSO model are written in bold font. Dendrogram was cut vertically at correlation = 0.4, the resulting clusters are framed. Cx:y, acyl-group with chain length x and y double bonds; LPC a, lysophosphatidylcholine with acyl chain; PC aa, diacyl phosphatidylcholine; PC ae, acyl-alkyl phosphatidylcholine.

de novo production via PEMT, and an increased consumption of choline by increased compensatory PC production via the CDP-choline pathway (Jacobs *et al.*, 2010). A protective effect of low plasma choline levels on body mass has also been observed in a human population-based study (Konstantinova *et al.*, 2008). Low choline levels could increase energy expenditure via several mechanisms, one being the attenuation of acetylcholine signaling in the brain (Gautam *et al.*, 2006, Jacobs *et al.*, 2010).

It might therefore be hypothesized that the observed PC signature in children with substantial weight loss may reflect a reduced PEMT activity. Once these children change their nutritional habits, and thereby reduce the dietary intake of choline, they might have a greater potential to reduce their weight. This assumption is supported by a dietary intervention study in overweight adults, where a PC species that is likely PEMT-derived was negatively associated with body fat reduction (Smilowitz *et al.*, 2009).

Abdominal adipose tissue and weight loss

Waist circumference is an established marker of abdominal obesity in children (Taylor *et al.*, 2000, Schwandt *et al.*, 2008). In this study, a higher waist circumference at baseline was inversely associated with BMI-SDS reduction during intervention. This observation is consistent with the negative link between markers of abdominal fat mass and weight loss success as well as improvement in insulin sensitivity observed upon lifestyle intervention in adults (Teixeira *et al.*, 2004, Thamer *et al.*, 2007). However, the opposite association has been reported (Wabitsch *et al.*, 1992, Carmichael *et al.*, 1998).

There is biological evidence for a role of abdominal adipose tissue in weight regulation. It is well recognized that abdominal adipose tissue is an endocrine organ that contributes to the subclinical inflammation associated with obesity by secreting a range of bioactive molecules called adipokines (Wajchenberg, 2000). Of note, an increasing number of studies in both children (Fleisch *et al.*, 2007, Reinehr *et al.*, 2009b, Murer *et al.*, 2011) and adults (Verdich *et al.*, 2001, Shih *et al.*, 2006) showed higher serum levels of the adipokine leptin to be associated with weight gain or poor response to lifestyle intervention. Although leptin exerts anorexigenic functions, suppressing food intake and increasing energy expenditure, these negative associations might be explained by the presence of leptin resistance or central leptin insufficiency (Kalra, 2008, Reinehr *et al.*, 2009b).

In addition, high baseline levels of the adipokine adiponectin predicted weight gain over 4 years in adults (Hivert *et al.*, 2011) and promoter methylation of the tumor necrosis factor-α (TNF-α) gene, which positively regulated circulating TNF-α concentration, was negatively associated with weight loss success (Campión *et al.*, 2009). A further line of evidence connects abdominal obesity with resistance to weight loss during lifestyle intervention via the central action of insulin. Abdominal adipose tissue has been reported to associate with cerebral insulin resistance (Tschritter *et al.*, 2009), which was related to impaired body fat loss during lifestyle intervention (Tschritter *et al.*, 2012).

Together, these findings corroborate a complex role of abdominal fat in weight regulation and might contribute to the explanation why higher waist circumference is associated with poorer weight loss success during lifestyle intervention in the present study. Adipokine measurement was not subject of this study, so it could not be investigated whether the observed association was mediated by these factors.

Predictive potential of the LASSO model and comparison to other studies

Widely used multivariate approaches in metabolomics data analysis are partial least squares (PLS)-related methods. However, classical PLS regression has the disadvantage that variable effect strengths are not readily obtained and sparse models containing only a few important predictor variables for assessment in future studies cannot be derived easily. We therefore chose to use a LASSO regression approach, which provides, besides

measures of prediction accuracy for the whole model, measures of effect strength and variable stability for the selected variables. Using this approach, a model was obtained that comprised three pre-intervention variables that explained a significant part of ΔBMI-SDS. Although no hard cut-offs exist for R^2 and Q^2 values in this regularized regression setting, the prediction accuracy of the presented model seemed rather moderate (R^2 = 0.267, Q^2 = 0.116). A recent investigation of urinary metabolite traits predictive of substantial BMI change in a 3-week treatment camp for adolescents reported higher values of prediction accuracy (Pathmasiri *et al.*, 2012). A direct comparison is difficult since their study differed from the present study in terms of statistical methods, length and characteristics of intervention as well as metabolomics technique and investigated biofluids. Overweight change over the course of one year in an outpatient intervention program might be more strongly influenced by environmental and psychosocial factors and therefore be less predictable by the here investigated metabolic variables. Also, Pathmasiri *et al.* included post-intervention metabolite levels in their prediction model, unlike in the present study, were the aim was to obtain a model with prognostic applicability. Results of both studies require external validation in larger independent data sets. Other studies searching for metabolic predictors of weight loss success investigated single parameters and found better insulin sensitivity (i.e., lower HOMA-IR, lower fasting insulin or absence of type 2 diabetes (T2D)) (Harden *et al.*, 2007, Madsen *et al.*, 2009, Ford *et al.*, 2010) as well as lower serum TG levels (Harden *et al.*, 2007, Madsen *et al.*, 2009) as predictors of weight loss. In the present study, these parameters were not identified as significant predictors. However, HOMA-IR and serum TGs showed a borderline significant negative association with ΔBMI-SDS.

Strengths and limitations

This is one of the first studies applying a metabolomics approach to identify metabolic predictors of overweight reduction in obese children upon lifestyle intervention. In addition to the univariate identification of pre-intervention variables associated with overweight reduction, a carefully validated LASSO approach was used to build a predictive model for BMI-SDS change. As a limitation of this study, only a small group of children was investigated. Larger studies might allow for the development of sex-, age- and maturity-specific predictive models. The underlying study population did not represent a random group of obese children. Therefore, the predictive potential of the variables on which the children were matched (sex, age, and pubertal stage) could not be assessed (Sabin *et al.*, 2007, Danielsson *et al.*, 2012). Moreover, weight loss success is not only determined by compliance regarding participation at meetings, but also by implementation of the recommendations into daily life. This might be strongly influenced by environmental and psychosocial factors, which were not ascertained in this study. Furthermore, the present analysis was limited to changes in BMI-SDS as outcome. Further investigations should be aimed at identifying predictors for secondary outcomes such as changes in body fat distri-

bution and insulin sensitivity. In addition, studies investigating metabolite changes during lifestyle intervention might give additional information about the mechanisms underlying weight change.

Conclusions

The obtained results confirm a role of phosphatidylcholine metabolism in human energy regulation and success in overweight reduction as has previously been observed in animal studies. They further corroborate the connection between abdominal obesity and impaired overweight reduction. These are both important aspects for understanding the large inter-individual variation in response to lifestyle interventions, which is a prerequisite for the development of individualized intervention programs.

5 Summary and outlook

Advances in the field of high-throughput omics technologies offer the opportunity to simultaneously measure hundreds or thousands of molecules. In post-genomic obesity research, they provide valuable tools for the molecular characterization of obesity-related pathomechanisms on different system levels. In this thesis, four studies were conducted that employed combinations of omics data to address different aspects of obesity research. Four central goals of post-genomic obesity research were addressed: (1) Defining the mechanisms linking selected risk loci to obesity and type 2 diabetes (T2D), (2) explaining part of the missing heritability through the study of DNA methylation, (3) understanding the etiology and consequences of obesity and weight change using different omics data, and (4) identifying metabolomic determinants of weight loss response to lifestyle intervention. In the following, the core findings and scientific contributions of the thesis are summarized and future perspectives are given, in view of the addressed objectives.

5.1 Key findings

First, an extensive characterization and comparison of metabolomic responses to different oral and intravenous challenges was provided. This revealed in particular previously unreported changes in different phospholipid metabolites, as well as diverging metabolite changes in response to intravenous as compared to oral glucose challenge. In addition, the concept of metabolomics measurements during challenge tests was successfully applied in studying genotype-challenge interactions. Specifically, new insights into a putative role of sphingomyelin and phospholipid metabolism in *TCF7L2*-conferred T2D risk were obtained. These perturbations could only be detected through the challenge tests, demonstrating the utility of the approach in revealing early metabolic abnormalities prior to a change in conventional parameters of glucose homeostasis.

Second, the first large EWAS of BMI was conducted. It comprised more than 10,000 subjects of European and South Asian origin from 13 studies, and revealed stable methylation-BMI association for 187 loci. Downstream analyses provide solid evidence for an enrichment of these loci in regions of open chromatin, an enrichment for relevant biological pathways and for loci previously reported in lipid GWAS. They further show that a large number of the identified CpG sites were associated with gene expression at nearby genes

and with genetic variation. Three different approaches were introduced to decipher direction and causality of the observed associations, two Mendelian randomization (MR) approaches and a longitudinal regression approach. From these, evidence was obtained that change in methylation at the majority of loci was consequential to change in BMI. Furthermore, methylation at selected sites explained a part of the association of BMI with clinical traits and incident T2D, suggesting methylation as a candidate mechanism underlying the development of obesity-related comorbidities.

Third, a comprehensive investigation of the metabolomic and transcriptomic signature associated with previous body weight change over a 7-year period is provided. Applying a weighted correlation network analysis (WGCNA) approach to aggregate metabolites and transcripts to modules of closely connected molecules prior to association testing allowed the identification of modules of metabolites or transcripts jointly related to weight change. Together, these modules indicate a global effect of weight change on major branches of metabolism, including lipoprotein and lipid metabolism, amino acid metabolism, energy metabolism/mitochondrial function, basophil/mast cell function and red blood cell development. Weight gain and weight loss showed largely opposing effects on the modules, which were also cross-sectionally related to insulin resistance traits.

In the last study, combined serum metabolomic, anthropometric and clinical data were used to build a predictive model of weight loss success for obese children during the 1-year lifestyle intervention study "Obeldicks". In a careful model building and validation strategy using the regularized regression approach LASSO, a sparse model that contained waist circumference as well as two phosphatidylcholines and predicted a significant part of weight loss success was identified. These results point towards a role of abdominal fat as well as phospholipid metabolism in weight regulation, thereby contributing to the understanding of inter-individual variation in weight loss response.

Throughout the thesis, new insights were obtained through the integration of multiple omics levels. For instance, metabolomics was successfully used as a tool to more deeply characterize results from genomic approaches. The integration of epigenomic data with genomic and transcriptomic data helped to understand genetic regulation of DNA methylation and its effect on gene expression. In addition, genomics data were an essential element of MR approaches. Finally, metabolomics data from two platforms and transcriptomics data provided complementary information on the effects of body weight change.

5.2 Future perspectives

The growing field of multi-omics entails several challenges. In the human body, the different system levels act on different time scales and in different cellular compartments (Somvanshi and Venkatesh, 2014). With the exception of genomics, omics measurements constitute a snapshot of the state of a system at a specific time point in a specific tis-

sue. To be able to fully understand the dynamics of physiological processes, multiple measurements in optimally many different tissues have to be conducted. Measurements in epidemiological studies, including most of the studies that are part of this thesis, are frequently restricted to one time point and to easily accessible biofluids such as blood, which for instance only weakly reflects obesity-related transcriptional processes in adipose tissue (Emilsson *et al.*, 2008). Studies integrating omics data from different timepoints and tissues will help to understand the interplay of the system levels.

The frequent use of blood samples to study epigenomics and transcriptomics poses a further challenge. Whole blood represents a mixture of different cell types that are characterized by highly specific epigenomic and transcriptomic signatures (Houseman *et al.*, 2012, Reinius *et al.*, 2012, Zhu *et al.*, 2012). Since BMI associates with a shift in blood cell proportions (Bellows *et al.*, 2011, Trottier *et al.*, 2012), cell type proportions represent potential confounders of methylation/gene expression - BMI associations. This issue is addressed by two different statistical approaches in this thesis that both rely on previously published cell-specific methylation or expression signatures. Although these approaches diminish cell type confounding, they are limited by the quality of the external data, and with regard to the specific cell types considered (see detailed discussion in Sections 4.2.9 and 4.3.5). The development of improved methods to deal with cell type confounding in whole blood methylation and expression studies merits further efforts and is already an ongoing research focus (Houseman *et al.*, 2014, Zou *et al.*, 2014, Jaffe and Irizarry, 2014).

Another issue of future relevance is the inference of causality in systems epidemiology. With the exception of genomics, the system levels do not act in a unidirectional way (Schnabel *et al.*, 2012). For instance, the dogma of unidirectional negative regulation of gene expression by DNA methylation has been refuted (Portela and Esteller, 2010, Zilberman *et al.*, 2007). Instead, dynamic interactions and feedback loops between system levels challenge the integrated analysis. Thus, in observational studies, one cannot conclude causality from associations between omics levels, or between omics and phenotype data. Specific statistical methodology has been developed to infer causality from observational data. An example is the concept of Mendelian randomization (MR), which was applied in this thesis. In MR, genetic variants play an important role as instrumental variables (Section 3.5.2). Although the two MR approaches applied in this thesis provide evidence for causality, they rely on several model assumptions that are difficult and partly impossible to test (see discussion in Section 4.2.9), so the obtained results should be interpreted with care. In addition, the formal MR approach requires that the individual studies have substantial power, in order to prevent weak instrument bias. In the future, meta-analyses of larger studies should be performed to obtain more reliable evidence for causality. In addition, the epidemiological findings obtained in this thesis should be followed up in controlled weight change intervention studies as well as in *in vitro* and *in vivo* experiments.

An increasing number of multi-omic data resources are becoming available, and novel omics technologies – such as glycomics, i.e. the study of sugar species in a biosample (Zhang *et al.*, 2014), and metagenomics, i.e. the study of the genomes of microorganisms e.g. in the human gut – provide additional information (Norheim *et al.*, 2012). In addition, the *Encyclopedia Of DNA Elements* (ENCODE) consortium is creating a comprehensive catalogue of functional elements in the human genome sequence (Rosenbloom *et al.*, 2012). The development and adoption of statistical methods to meet the specific characteristics of novel omics data is already being worked on (e.g., Wahl *et al.* (2014)) and remains a challenging task at the interface of biology and statistics. In addition, the integrated analysis of multiple data types requires considerable computational capacities and bioinformatic tools. Databases linking omics features with biological pathways can support the interpretation of multi-omics data (Connor *et al.*, 2010, Dutta *et al.*, 2012). However, maintaining databases with good coverage and specificity for features from different omics layers will become more and more challenging as the complexity of available data increases.

A particular post-genomic concern is tackling the missing heritability of obesity. In this thesis, DNA methylation as a possible mechanisms underlying the missing heritability was explored. Animal studies suggest that environmentally aquired epigenetic modifications can be inherited to subsequent generations (Guerrero-Bosagna and Skinner, 2012), raising the possibility that they might contribute to the heritability of obesity independent of genetic variations. In the presented study, evidence for a causal role of methylation in the etiology of obesity was obtained only for few CpG sites. Further work should focus on other promising strategies of addressing the missing heritability. This includes the study of other epigenetic mechanisms such as histone modification, chromatin remodeling and RNA inference (Portela and Esteller, 2010, Rakyan *et al.*, 2011). In addition, gene-gene and gene-environment interactions deserve further investigation, as well as imprinted loci and disease-associated haplotypes, which might be explored through family studies. Furthermore, even larger GWAS efforts will have a chance to detect rare and low-penetrance genetic variants as well as previously untagged structural variants.

5.3 Conclusion

This thesis demonstrates the enormous value and potential of multi-omics strategies for post-genomic obesity research, while also acknowledging and providing solutions to the challenges arising from omics data integration. The obtained insights provide a basis for understanding the complex molecular processes underlying obesity and weight change, and linking them to metabolic derangements such as insulin resistance and dyslipidemia. Thereby, they present a promising starting point for the development of individualized treatment, early detection and prevention strategies for obesity and comorbidities such as T2D and cardiovascular diseases.

References

Abayomi K, Gelman A, Levy M (2008). "Diagnostics for multivariate imputations." *Applied Statistics*, **57**, 273–291.

Acharjee A, Kloosterman B, de Vos RCH, *et al.* (2011). "Data integration and network reconstruction with omics data using Random Forest regression in potato." *Anal Chim Acta*, **705**(1-2), 56–63.

Adult Treatment Panel III (2002). "Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report." *Circulation*, **106**(25), 3143–3421.

Almén MS, Jacobsson JA, Moschonis G, *et al.* (2012). "Genome wide analysis reveals association of a FTO gene variant with epigenetic changes." *Genomics*, **99**(3), 132–137.

Almén MS, Nilsson EK, Jacobsson JA, *et al.* (2014). "Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity." *Gene*, **548**(1), 61–67.

Ambroise C, McLachlan GJ (2002). "Selection bias in gene extraction on the basis of microarray gene-expression data." *Proc Natl Acad Sci U S A*, **99**(10), 6562–6566.

Andreasen CH, Stender-Petersen KL, Mogensen MS, *et al.* (2008). "Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation." *Diabetes*, **57**(1), 95–101.

Anway MD, Cupp AS, Uzumcu M, Skinner MK (2005). "Epigenetic transgenerational actions of endocrine disruptors and male fertility." *Science*, **308**(5727), 1466–1469.

Anzeneder L, Kircher F, Feghelm N, Fischer R, Seissler J (2011). "Kinetics of insulin secretion and glucose intolerance in adult patients with cystic fibrosis." *Horm Metab Res*, **43**(5), 355–360.

Aron-Wisnewsky J, Julia Z, Poitou C, *et al.* (2011). "Effect of bariatric surgery-induced weight loss on SR-BI-, ABCG1-, and ABCA1-mediated cellular cholesterol efflux in obese women." *J Clin Endocrinol Metab*, **96**(4), 1151–1159.

Arsenault BJ, Lemieux I, Després JP, *et al.* (2009). "HDL particle size and the risk of coronary heart disease in apparently healthy men and women: the EPIC-Norfolk prospective population study." *Atherosclerosis*, **206**(1), 276–281.

Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.* (2014). "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays." *Bioinformatics*, **30**(10), 1363–1369.

Atwood LD, Heard-Costa NL, Cupples LA, *et al.* (2002). "Genomewide linkage analysis of body mass index across 28 years of the Framingham Heart Study." *Am J Hum Genet*, **71**(5), 1044–1050.

Bell CG, Walley AJ, Froguel P (2005). "The genetics of human obesity." *Nat Rev Genet*, **6**(3), 221–234.

Bell JT, Pai AA, Pickrell JK, *et al.* (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." *Genome Biol*, **12**(1), R10.

Bell JT, Tsai PC, Yang TP, *et al.* (2012). "Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population." *PLoS Genetics*, **8**(4), e1002629.

Bellows CF, Zhang Y, Simmons PJ, Khalsa AS, Kolonin MG (2011). "Influence of BMI on level of circulating progenitor cells." *Obesity (Silver Spring)*, **19**(8), 1722–1726.

Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Stat*, **57**(1), 290–300.

Benjamini Y, Yekutieli D (2001). "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, **29**(4), 1165–1188.

Berisha SZ, Serre D, Schauer P, Kashyap SR, Smith JD (2011). "Changes in whole blood gene expression in obese subjects with type 2 diabetes following bariatric surgery: a pilot study." *PLoS One*, **6**(3), e16729.

Berndt SI, Gustafsson S, Mägi R, *et al.* (2013). "Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture." *Nat Genet*, **45**(5), 501–512.

Bibikova M, Barnes B, Tsan C, *et al.* (2011). "High density DNA methylation array with single CpG site resolution." *Genomics*, **98**(4), 288–295.

Bjornsson HT, Sigurdsson MI, Fallin MD, *et al.* (2008). "Intra-individual change in DNA methylation over time with familial clustering." *JAMA*, **299**(24), 2877–2883.

Boccard J, Veuthey JL, Rudaz S (2010). "Knowledge discovery in metabolomics: an overview of MS data handling." *J Sep Sci*, **33**(3), 290–304.

Bochud M, Rousson V (2010). "Usefulness of Mendelian randomization in observational epidemiology." *Int J Environ Res Public Health*, **7**(3), 711–728.

Bodner TE (2008). "What improves with increased missing data imputations?" *Structural Equation Modeling*, **15**(4), 651–675.

Boks MP, Erks EM, Weisenberger DJ, *et al.* (2009). "The relationship of DNA methylation with age, gender and genotype in twins and healthy controls." *PLoS One*, **4**(8), e6767.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics*, **19**(2), 185–193.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2010). "A basic introduction to fixed-effect and random-effects models for meta-analysis." *Research Synthesis Methods*, **1**(2), 97–111.

Boslem E, MacIntosh G, Preston AM, *et al.* (2011). "A lipidomic screen of palmitate-treated MIN6 β-cells links sphingolipid metabolites with endoplasmic reticulum (ER) stress and impaired protein trafficking." *Biochem J*, **435**(1), 267–276.

Bouchard C (2008). "Gene-environment interactions in the etiology of obesity: defining the fundamentals." *Obesity (Silver Spring)*, **16 Suppl 3**, S5–S10.

Bouchard C, Tremblay A, Després JP, *et al.* (1990). "The response to long-term overfeeding in identical twins." *N Engl J Med*, **322**(21), 1477–1482.

Bouchard C, Tremblay A, Després JP, *et al.* (1994). "The response to exercise with constant energy intake in identical twins." *Obes Res*, **2**(5), 400–410.

Bouchard L, Rabasa-Lhoret R, Faraj M, *et al.* (2010). "Differential epigenomic and transcriptomic responses in subcutaneous adipose tissue between low and high responders to caloric restriction." *Am J Clin Nutr*, **91**(2), 309–320.

Boulesteix AL, Strobl C, Augustin T, Daumer M (2008). "Evaluating microarray-based classifiers: an overview." *Cancer Inform*, **6**, 77–97.

Bound J, Jaeger DA, Baker RM (1995). "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American Statistical Association*, **90**, 443–450.

Bouwens M, Grootte Bromhaar M, Jansen J, Müller M, Afman LA (2010). "Postprandial dietary lipid-specific effects on human peripheral blood mononuclear cell gene expression profiles." *Am J Clin Nutr*, **91**(1), 208–217.

Bradfield JP, Taal HR, Timpson NJ, *et al.* (2012). "A genome-wide association meta-analysis identifies new childhood obesity loci." *Nat Genet*, **44**(5), 526–531.

Braga-Neto UM, Dougherty ER (2004). "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, **20**(3), 374–380.

Brand S, Kesper DA, Teich R, *et al.* (2012). "DNA methylation of TH1/TH2 cytokine genes affects sensitization and progress of experimental asthma." *J Allergy Clin Immunol*, **129**(6), 1602–10.e6.

Breier M, Wahl S, Prehn C, *et al.* (2014). "Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples." *PLoS One*, **9**(2), e89728.

Broadhurst DI, Kell DB (2006). "Statistical strategies for avoiding false discoveries in metabolomics and related experiments." *Metabolomics*, **2**(4), 171–196.

Brooker RJ (2005). *Genetics: analysis & principles*. 2nd edition. McGraw-Hill, New York.

Brunzell JD, Hokanson JE (1999). "Dyslipidemia of central obesity and insulin resistance." *Diabetes Care*, **22 Suppl 3**, C10–C13.

Burgess S, Thompson SG, CRPCHDGC (2011). "Avoiding bias from weak instruments in Mendelian randomization studies." *Int J Epidemiol*, **40**(3), 755–764.

Butte A (2002). "The use and analysis of microarray data." *Nat Rev Drug Discov*, **1**(12), 951–960.

Calinski T, Harabasz J (1974). "A dendrite method for cluster analysis." *Communications in Statistics*, **3**(1), 1–27.

Campión J, Milagro FI, Goyenechea E, Martínez JA (2009). "TNF-alpha promoter methylation as a predictive biomarker for weight-loss response." *Obesity*, **17**(6), 1293–1297.

Carmichael HE, Swinburn BA, Wilson MR (1998). "Lower fat intake as a predictor of initial and sustained weight loss in obese subjects consuming an otherwise ad libitum diet." *J Am Diet Assoc*, **98**(1), 35–39.

Chen C, Grennan K, Badner J, *et al.* (2011). "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods." *PLoS One*, **6**(2), e17238.

Chen Ya, Lemire M, Choufani S, *et al.* (2013). "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray." *Epigenetics*, **8**(2), 203–209.

Choquet H, Meyre D (2011a). "Genetics of obesity: What have we learned?" *Curr Genomics*, **12**(3), 169–179.

Choquet H, Meyre D (2011b). "Molecular basis of obesity: current status and future prospects." *Curr Genomics*, **12**(3), 154–168.

Christensen BC, Houseman EA, Marsit CJ, *et al.* (2009). "Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context." *PLoS Genetics*, **5**(8), e1000602.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007). "Network-based classification of breast cancer metastasis." *Mol Syst Biol*, **3**, 140.

Claussnitzer M, Dankel SN, Klocke B, *et al.* (2014). "Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms." *Cell*, **156**(1-2), 343–358.

Cole SR, Platt RW, Schisterman EF, *et al.* (2010). "Illustrating bias due to conditioning on a collider." *Int J Epidemiol*, **39**(2), 417–420.

Cole TJ (1990). "The LMS method for constructing normalized growth standards." *Eur J Clin Nutr*, **44**(1), 45–60.

Cole TJ, Bellizzi MC, Flegal KM, Dietz WH (2000). "Establishing a standard definition for child overweight and obesity worldwide: international survey." *BMJ*, **320**(7244), 1240–1243.

Collins LM, Schafer JL, Kam CM (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." *Psychol Methods*, **6**(4), 330–351.

Connor SC, Hansen MK, Corner A, Smith RF, Ryan TE (2010). "Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes." *Mol Biosyst*, **6**(5), 909–921.

Cook S, Weitzman M, Auinger P, Nguyen M, Dietz WH (2003). "Prevalence of a metabolic syndrome phenotype in adolescents: findings from the third National Health and Nutrition Examination Survey, 1988-1994." *Arch Pediatr Adolesc Med*, **157**(8), 821–827.

Cornelis MC, Hu FB (2013). "Systems Epidemiology: A New Direction in Nutrition and Metabolic Disease Research." *Curr Nutr Rep*, **2**(4).

Danielsson P, Svensson V, Kowalski J, *et al.* (2012). "Importance of age for 3-year continuous behavioral obesity treatment success and dropout rate." *Obes Facts*, **5**(1), 34–44.

Das UN, Rao AA (2007). "Gene expression profile in obesity and type 2 diabetes mellitus." *Lipids Health Dis*, **6**, 35.

Davies DL, Bouldin DW (1979). "A cluster separation measure." *IEEE Trans Pattern Anal Mach Intell*, **1**(2), 224–227.

de Ferranti S, Mozaffarian D (2008). "The perfect storm: obesity, adipocyte dysfunction, and metabolic consequences." *Clin Chem*, **54**(6), 945–955.

Dedeurwaerder S, Defrance M, Bizet M, *et al.* (2013). "A comprehensive overview of Infinium HumanMethylation450 data processing." *Brief Bioinform*, pp. 1–13.

Dedeurwaerder S, Defrance M, Calonne E, *et al.* (2011). "Evaluation of the Infinium Methylation 450K technology." *Epigenomics*, **3**(6), 771–784.

DeHaven CD, Evans AM, Dai H, Lawton KA (2010). "Organization of GC/MS and LC/MS metabolomics data into chemical libraries." *J Cheminform*, **2**(1), 9.

DeLong CJ, Shen YJ, Thomas MJ, Cui Z (1999). "Molecular distinction of phosphatidylcholine synthesis between the CDP-choline pathway and phosphatidylethanolamine methylation pathway." *J Biol Chem*, **274**(42), 29683–29688.

Deo RC, Hunter L, Lewis GD, *et al.* (2010). "Interpreting metabolomic profiles using unbiased pathway models." *PLoS Comput Biol*, **6**(2), e1000692.

Després JP (2006). "Is visceral obesity the cause of the metabolic syndrome?" *Ann Med*, **38**(1), 52–63.

Devlin B, Roeder K, Wasserman L (2001). "Genomic control, a new approach to genetic-based association studies." *Theor Popul Biol*, **60**(3), 155–166.

Dick KJ, Nelson CP, Tsaprouni L, *et al.* (2014). "DNA methylation and body-mass index: a genome-wide analysis." *The Lancet*, **383**(9933), 1990–1998.

Didelez V, Sheehan N (2007). "Mendelian randomization as an instrumental variable approach to causal inference." *Stat Methods Med Res*, **16**(4), 309–330.

Dina C, Meyre D, Gallina S, *et al.* (2007). "Variation in FTO contributes to childhood obesity and severe adult obesity." *Nat Genet*, **39**(6), 724–726.

Donchenko V, Zannetti A, Baldini PM (1994). "Insulin-stimulated hydrolysis of phosphatidylcholine by phospholipase C and phospholipase D in cultured rat hepatocytes." *Biochim Biophys Acta*, **1222**(3), 492–500.

D'Orazio P, Burnett RW, Fogh-Andersen N, *et al.* (2005). "Approved IFCC recommendation on reporting results for blood glucose (abbreviated)." *Clin Chem*, **51**(9), 1573–1576.

Drechsler J (2011). "Multiple imputation in practice - a case study using a complex German establishment survey." *Advanced Statistical Analysis*, **95**, 1–26.

Du P, Kibbe WA, Lin SM (2008). "lumi: a pipeline for processing Illumina microarray." *Bioinformatics*, **24**(13), 1547–1548.

Du P, Zhang X, Huang CC, *et al.* (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." *BMC Bioinformatics*, **11**, 587.

Dudoit S, Shaffer JP, Boldrick JC (2003). "Multiple hypothesis testing in microarray experiments." *Statistical Science*, **18**, 71–103.

Dupuy A, Simon RM (2007). "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting." *J Natl Cancer Inst*, **99**(2), 147–157.

Dutta T, Chai HS, Ward LE, *et al.* (2012). "Concordance of changes in metabolic pathways based on plasma metabolomics and skeletal muscle transcriptomics in type 1 diabetes." *Diabetes*, **61**(5), 1004–1016.

Duvillard L, Florentin E, Lizard G, *et al.* (2003). "Cell surface expression of LDL receptor is decreased in type 2 diabetic patients and is normalized by insulin therapy." *Diabetes Care*, **26**(5), 1540–1544.

Ebbeling CB, Pawlak DB, Ludwig DS (2002). "Childhood obesity: public-health crisis, common sense cure." *Lancet*, **360**(9331), 473–482.

Efron B, Tibshirani RJ (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC, Boca Raton.

El-Maarri O, Becker T, Junen J, *et al.* (2007). "Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males." *Human Genetics*, **122**(5), 505–514.

Emilsson V, Thorleifsson G, Zhang B, *et al.* (2008). "Genetics of gene expression and its effect on disease." *Nature*, **452**(7186), 423–428.

Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E (2009). "Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems." *Anal Chem*, **81**(16), 6656–6667.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). *Regression - Models, methods and applications*. Springer.

Fall T, Hägg S, Mägi R, *et al.* (2013). "The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis." *PLoS Med*, **10**(6), e1001474.

Faraway JJ (2002). *Practical Regression and Anova using R*.

Felig P, Marliss E, Cahill Jr G (1969). "Plasma amino acid levels and insulin secretion in obesity." *N Engl J Med*, **281**(15), 811–816.

Fisher RA (1922). "On the interpretation of χ^2 from contingency tables, and the calculation of p." *Journal of the Royal Statistical Society*, **85**(1), 87–94.

Fleisch AF, Agarwal N, Roberts MD, *et al.* (2007). "Influence of serum leptin on weight and body fat growth in children at high risk for adult obesity." *J Clin Endocrinol Metab*, **92**(3), 948–954.

Floegel A, Wientzek A, Bachlechner U, *et al.* (2014). "Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study." *Int J Obes (Lond)*.

Fontaine-Bisson B, Wolever TMS, Chiasson JL, *et al.* (2007). "Tumor necrosis factor alpha -238G>A genotype alters postprandial plasma levels of free fatty acids in obese individuals with type 2 diabetes mellitus." *Metabolism*, **56**(5), 649–655.

Ford AL, Hunt LP, Cooper A, Shield JPH (2010). "What reduction in BMI SDS is required in obese adolescents to improve body composition and cardiometabolic health?" *Arch Dis Child*, **95**(4), 256–261.

Fraga MF, Ballestar E, Paz MF, *et al.* (2005). "Epigenetic differences arise during the lifetime of monozygotic twins." *Proceedings of the National Academy of Sciences of the United States of America*, **102**(30), 10604–10609.

Franks PW, Ekelund U, Brage S, *et al.* (2007). "PPARGC1A coding variation may initiate impaired NEFA clearance during glucose challenge." *Diabetologia*, **50**(3), 569–573.

Frayling TM, Timpson NJ, Weedon MN, *et al.* (2007). "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity." *Science*, **316**(5826), 889–894.

Freedman DS, Otvos JD, Jeyarajah EJ, *et al.* (1998). "Relation of lipoprotein subclasses as measured by proton nuclear magnetic resonance spectroscopy to coronary artery disease." *Arterioscler Thromb Vasc Biol*, **18**(7), 1046–1053.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, **33**(1), 1–22.

Gardiner-Garden M, Frommer M (1987). "CpG islands in vertebrate genomes." *J Mol Biol*, **196**(2), 261–282.

Gault CR, Obeid LM, Hannun YA (2010). "An overview of sphingolipid metabolism: from synthesis to breakdown." *Adv Exp Med Biol*, **688**, 1–23.

Gautam D, Gavrilova O, Jeon J, *et al.* (2006). "Beneficial metabolic effects of M3 muscarinic acetylcholine receptor deficiency." *Cell Metab*, **4**(5), 363–375.

Gautier-Stein A, Soty M, Chilloux J, *et al.* (2012). "Glucotoxicity induces glucose-6-phosphatase catalytic unit expression by acting on the interaction of HIF-1α with CREB-binding protein." *Diabetes*, **61**(10), 2451–2460.

Geeleher P, Hartnett L, Egan LJ, *et al.* (2013). "Gene-set analysis is severely biased when applied to genome-wide methylation data." *Bioinformatics*, **29**(15), 1851–1857.

Genolini C, Falissard B (2011). "KmL: a package to cluster longitudinal data." *Comput Methods Programs Biomed*, **104**(3), e112–e121.

Genolini C, Pingault JB, Driss T, *et al.* (2013). "KmL3D: a non-parametric algorithm for clustering joint trajectories." *Comput Methods Programs Biomed*, **109**(1), 104–111.

Gerrits A, Li Y, Tesson BM, *et al.* (2009). "Expression quantitative trait loci are highly sensitive to cellular differentiation state." *PLoS Genet*, **5**(10), e1000692.

Getty-Kaushik L, Song DH, Boylan MO, Corkey BE, Wolfe MM (2006). "Glucose-dependent insulinotropic polypeptide modulates adipocyte lipolysis and reesterification." *Obesity (Silver Spring)*, **14**(7), 1124–1131.

Ghosh S, Dent R, Harper ME, *et al.* (2010). "Gene expression profiling in whole blood identifies distinct biological pathways associated with obesity." *BMC Med Genomics*, **3**, 56.

Ghosh S, Dent R, Harper ME, Stuart J, McPherson R (2011). "Blood gene expression reveal pathway differences between diet-sensitive and resistant obese subjects prior to caloric restriction." *Obesity (Silver Spring)*, **19**(2), 457–463.

Gieger C, Geistlinger L, Altmaier E, *et al.* (2008). "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum." *PLoS Genet*, **4**(11), e1000282.

Godfrey KM, Sheppard A, Gluckman PD, *et al.* (2011). "Epigenetic gene promoter methylation at birth is associated with child's later adiposity." *Diabetes*, **60**(5), 1528–1534.

Goh SH, Josleyn M, Lee YT, *et al.* (2007). "The human reticulocyte transcriptome." *Physiol Genomics*, **30**(2), 172–178.

Gonen B, O'Donnell P, Post TJ, Quinn TJ, Schulman ES (1987). "Very low density lipoproteins (VLDL) trigger the release of histamine from human basophils." *Biochim Biophys Acta*, **917**(3), 418–424.

Goodacre R, Broadhurst D, Smilde AK, *et al.* (2007). "Proposed minimum repreport standards for data analysis in metabolomics." *Metabolomics*, **3**, 231–241.

Graham JW, Olchowski AE, Gilreath TD (2007). "How many imputations are really needed? Some practical clarifications of multiple imputation theory." *Prevention Science*, **8**(3), 206–213.

Greenland S, Morgenstern H (2001). "Confounding in health research." *Annu Rev Public Health*, **22**, 189–212.

Guerrero-Bosagna C, Skinner MK (2012). "Environmentally induced epigenetic transgenerational inheritance of phenotype and disease." *Molecular and Cellular Endocrinology*, **354**(1-2), 3–8.

Guh DP, Zhang W, Bansback N, *et al.* (2009). "The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis." *BMC Public Health*, **9**, 88.

Gungor N, Saad R, Janosky J, Arslanian S (2004). "Validation of surrogate estimates of insulin sensitivity and insulin secretion in children and adolescents." *J Pediatr*, **144**(1), 47–55.

Haas RH, Parikh S, Falk MJ, *et al.* (2008). "The in-depth evaluation of suspected mitochondrial disease." *Mol Genet Metab*, **94**(1), 16–37.

Hall E, Volkov P, Dayeh T, *et al.* (2014). "Effects of palmitate on genome-wide mRNA expression and DNA methylation patterns in human pancreatic islets." *BMC Med*, **12**, 103.

Hall P, Wilson SR (1991). "Two guidelines for bootstrap hypothesis testing." *Biometrics*, **47**, 757–762.

Hamel FG, Bennett RG, Upward JL, Duckworth WC (2001). "Insulin inhibits peroxisomal fatty acid oxidation in isolated rat hepatocytes." *Endocrinology*, **142**(6), 2702–2706.

Han JC, Lawlor DA, Kimm SYS (2010). "Childhood obesity." *Lancet*, **375**(9727), 1737–1748.

Hanzu FA, Vinaixa M, Papageorgiou A, *et al.* (2014). "Obesity rather than regional fat depots marks the metabolomic pattern of adipose tissue: an untargeted metabolomic approach." *Obesity (Silver Spring)*, **22**(3), 698–704.

Harden KA, Cowan PA, Velasquez-Mieyer P, Patton SB (2007). "Effects of lifestyle intervention and metformin on weight management and markers of metabolic syndrome in obese adolescents." *J Am Acad Nurse Pract*, **19**(7), 368–377.

Hardy RJ, Thompson SG (1998). "Detecting and describing heterogeneity in meta-analysis." *Stat Med*, **17**(8), 841–856.

Harrell, Jr FE (2006). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* Springer.

Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001). "Maximum Likelihood Estimation of Optimal Scaling Factors for Expression Array Normalization."

Haslam DW, James WPT (2005). "Obesity." *Lancet*, **366**(9492), 1197–1209.

Hastie T, Tibshirani R (2004). "Efficient quadratic regularization for expression arrays." *Biostatistics*, **5**(3), 329–340.

Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning: Data mining, inference, and prediction.* 2nd edition. Springer.

He M, Su H, Gao W, *et al.* (2010). "Reversal of obesity and insulin resistance by a non-peptidic glucagon-like peptide-1 receptor agonist in diet-induced obese mice." *PLoS One*, **5**(12), e14205.

Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA (2002). "Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology." *Am J Epidemiol*, **155**(2), 176–184.

Hidalgo B, Irvin MR, Sha J, *et al.* (2014). "Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study." *Diabetes*, **63**(2), 801–807.

Higgins JPT, Thompson SG (2002). "Quantifying heterogeneity in a meta-analysis." *Stat Med*, **21**(11), 1539–1558.

Hill AB (1965). "The environment and disease: association or causation?" *Proc R Soc Med*, **58**(5), 295–300.

Hivert MF, Sun Q, Shrader P, *et al.* (2011). "Higher adiponectin levels predict greater weight gain in healthy women in the Nurses' Health Study." *Obesity (Silver Spring)*, **19**(2), 409–415.

Ho JE, Larson MG, Vasan RS, *et al.* (2013). "Metabolite profiles during oral glucose challenge." *Diabetes*, **62**(8), 2689–2698.

Holle R, Happich M, Löwel H, Wichmann HE (2005). "KORA - a research platform for population based health research." *Gesundheitswesen*, **67**, S19–S25.

Holzapfel C, Grallert H, Huth C, *et al.* (2010). "Genes and lifestyle factors in obesity: results from 12,462 subjects from MONICA/KORA." *Int J Obes (Lond)*, **34**(10), 1538–1545.

Horvath S, Dong J (2008). "Geometric interpretation of gene coexpression network analysis." *PLoS Comput Biol*, **4**(8), e1000117.

Houseman EA, Accomando WP, Koestler DC, *et al.* (2012). "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics*, **13**, 86.

Houseman EA, Molitor J, Marsit CJ (2014). "Reference-free cell mixture adjustments in analysis of DNA methylation data." *Bioinformatics*, **30**(10), 1431–1439.

Howie BN, Donnelly P, Marchini J (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLoS Genet*, **5**(6), e1000529.

Huxley R, Mendis S, Zheleznyakov E, Reddy S, Chan J (2010). "Body mass index, waist circumference and waist:hip ratio as predictors of cardiovascular risk–a review of the literature." *Eur J Clin Nutr*, **64**(1), 16–22.

Illig T, Gieger C, Zhai G, *et al.* (2010). "A genome-wide perspective of genetic variation in human metabolism." *Nat Genet*, **42**(2), 137–141.

Imtiyaz HZ, Williams EP, Hickey MM, *et al.* (2010). "Hypoxia-inducible factor 2alpha regulates macrophage function in mouse models of acute and tumor inflammation." *J Clin Invest*, **120**(8), 2699–2714.

Inouye M, Kettunen J, Soininen P, *et al.* (2010a). "Metabonomic, transcriptomic, and genomic variation of a population cohort." *Mol Syst Biol*, **6**, 441.

Inouye M, Ripatti S, Kettunen J, *et al.* (2012). "Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis." *PLoS Genet*, **8**(8), e1002907.

Inouye M, Silander K, Hamalainen E, *et al.* (2010b). "An immune response network associated with blood lipid levels." *PLoS Genet*, **6**(9), e1001113.

Ioannidis JPA (2007). "Non-replication and inconsistency in the genome-wide association setting." *Hum Hered*, **64**(4), 203–213.

Irizarry RA, Hobbs B, Collin F, *et al.* (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics*, **4**(2), 249–264.

Issaq HJ, Van QN, Waybright TJ, Muschik GM, Veenstra TD (2009). "Analytical and statistical approaches to metabolomics research." *J Sep Sci*, **32**(13), 2183–2199.

Jacobs RL, Zhao Y, Koonen DPY, *et al.* (2010). "Impaired de novo choline synthesis explains why phosphatidylethanolamine N-methyltransferase-deficient mice are protected from diet-induced obesity." *J Biol Chem*, **285**(29), 22403–22413.

Jaffe AE, Irizarry RA (2014). "Accounting for cellular heterogeneity is critical in epigenome-wide association studies." *Genome Biol*, **15**(2), R31.

Janssen I, Katzmarzyk PT, Srinivasan SR, *et al.* (2005). "Combined influence of body mass index and waist circumference on coronary artery disease risk factors among children and adolescents." *Pediatrics*, **115**(6), 1623–1630.

Janszky I, Ahlbom A, Svensson AC (2010). "The Janus face of statistical adjustment: confounders versus colliders." *Eur J Epidemiol*, **25**(6), 361–363.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000). "The large-scale organization of metabolic networks." *Nature*, **407**(6804), 651–654.

Johansson LE, Danielsson APH, Parikh H, *et al.* (2012). "Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance." *Am J Clin Nutr*, **96**(1), 196–207.

Jolliffe IT (2002). *Principal component analysis*. 2nd edition. Springer.

Juraschek SP, Selvin E, Miller ER, Brancati FL, Young JH (2013). "Plasma lactate and diabetes risk in 8045 participants of the atherosclerosis risk in communities study." *Ann Epidemiol*, **23**(12), 791–796.e4.

Kahn SE, Hull RL, Utzschneider KM (2006). "Mechanisms linking obesity to insulin resistance and type 2 diabetes." *Nature*, **444**(7121), 840–846.

Kalra SP (2008). "Central leptin insufficiency syndrome: an interactive etiology for obesity, metabolic and neural diseases and for designing new therapeutic interventions." *Peptides*, **29**(1), 127–138.

Kelley DE, He J, Menshikova EV, Ritov VB (2002). "Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes." *Diabetes*, **51**(10), 2944–2950.

Kendall M (1938). "A new measure of rank correlation." *Biometrika*, **30**(1-2), 81–89.

Khera AV, Cuchel M, de la Llera-Moya M, *et al.* (2011). "Cholesterol efflux capacity, high-density lipoprotein function, and atherosclerosis." *N Engl J Med*, **364**(2), 127–135.

Kim K, Doi A, Wen B, *et al.* (2010). "Epigenetic memory in induced pluripotent stem cells." *Nature*, **467**(7313), 285–290.

Kleemann R, van Erk M, Verschuren L, *et al.* (2010). "Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance." *PLoS One*, **5**(1), e8817.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer (New York).

Klop B, Elte JWF, Cabezas MC (2013). "Dyslipidemia in obesity: mechanisms and potential targets." *Nutrients*, **5**(4), 1218–1240.

Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I (2009). "Fewer permutations, more accurate P-values." *Bioinformatics*, **25**(12), i161–i168.

Kochan Z (2003). "Increased lipogenic potential of rat adipose tissue after repeated dieting–the role of SREBP-1 transcription factor." *Cell Mol Biol Lett*, **8**(4), 901–909.

Kolz M, Johnson T, Sanna S, *et al.* (2009). "Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations." *PLoS Genet*, **5**(6), e1000504.

Konstantinova SV, Tell GS, Vollset SE, *et al.* (2008). "Divergent associations of plasma choline and betaine with components of metabolic syndrome in middle age and elderly men and women." *J Nutr*, **138**(5), 914–920.

Kretschmer A, Möller G, Lee H, *et al.* (2014). "A common atopy-associated variant in the Th2 cytokine locus control region impacts transcriptional regulation and alters SMAD3 and SP1 binding." *Allergy*, **69**(5), 632–642.

Kromeyer-Hauschild K, Gläßer N, Zellner K (2008). "Waist circumference percentile in Jena children (Germany) 6- to 18-years of age." *Aktuel Ernaehr Med*, **33**(3), 116–122.

Kromeyer-Hauschild K, Wabitsch M, Kunze D, *et al.* (2001). "Perzentile für den Body-mass-Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben." *Monatsschr Kinderheilkd*, **149**(8), 807–818.

Krug S, Kastenmüller G, Stückler F, *et al.* (2012). "The dynamic range of the human metabolome revealed by challenges." *FASEB J*, **26**(6), 2607–2619.

Kruit JK, Wijesekara N, Westwell-Roper C, *et al.* (2012). "Loss of both ABCA1 and ABCG1 results in increased disturbances in islet sterol homeostasis, inflammation, and impaired β-cell function." *Diabetes*, **61**(3), 659–664.

Krumsiek J, Suhre K, Evans AM, *et al.* (2012a). "Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information." *PLoS Genet*, **8**(10), e1003005.

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011). "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data." *BMC Syst Biol*, **5**, 21.

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2012b). "Bayesian independent component analysis recovers pathway signatures from blood metabolomics data." *J Proteome Res*, **11**(8), 4120–4131.

Kryszczuk K, Hurley P (2010). "Estimation of the number of clusters using multiple clustering validity indices." In N Gayer, J Kittler, F Roli (eds.), *Multiple classifier systems*. Springer.

Laird NM, Ware JH (1982). "Random-effects models for longitudinal data." *Biometrics*, **38**(4), 963–974.

Langevin SM, Houseman EA, Christensen BC, *et al.* (2011). "The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood." *Epigenetics*, **6**(7), 908–919.

Langfelder P, Horvath S (2008). "WGCNA: an R package for weighted correlation network analysis." *BMC Bioinformatics*, **9**, 559.

Langfelder P, Zhang B, Horvath S (2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." *Bioinformatics*, **24**(5), 719–720.

Larrouy D, Barbe P, Valle C, *et al.* (2008). "Gene expression profiling of human skeletal muscle in response to stabilized weight loss." *Am J Clin Nutr*, **88**(1), 125–132.

Lauritzen SL (1996). *Graphical Models.* Oxford Statistical Science Series. Oxford University Press.

Lee H, Jaffe AE, Feinberg JI, *et al.* (2012). "DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth." *Int J Epidemiol*, **41**(1), 188–199.

Lee KJ, Carlin JB (2010). "Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation." *Am J Epidemiol*, **171**(5), 624–632.

Lee YH, Nair S, Rousseau E, *et al.* (2005). "Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes." *Diabetologia*, **48**(9), 1776–1783.

Li H, Xie Z, Lin J, *et al.* (2008). "Transcriptomic and metabonomic profiling of obesity-prone and obesity-resistant rats under high fat diet." *J Proteome Res*, **7**(11), 4775–4783.

Li S, Zhao JH, Luan J, *et al.* (2010). "Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study." *PLoS Med*, **7**(8), e1000332.

Li Z, Vance DE (2008). "Phosphatidylcholine and choline homeostasis." *J Lipid Res*, **49**(6), 1187–1194.

Lien LF, Haqq AM, Arlotto M, *et al.* (2009). "The STEDMAN project: biophysical, biochemical and metabolic effects of a behavioral weight loss intervention during weight loss, maintenance, and regain." *OMICS*, **13**(1), 21–35.

Little RJA, Rubin DB (2002). *Statistical analysis with missing data.* John Wiley & Sons, New Jersey.

Liu J, Morgan M, Hutchison K, Calhoun VD (2010). "A study of the influence of sex on genome wide methylation." *PLoS ONE*, **5**(4), e10028.

Liu Y, Aryee MJ, Padyukov L, *et al.* (2013). "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis." *Nat Biotechnol*, **31**(2), 142–147.

Loos RJF, Bouchard C (2003). "Obesity–is it a genetic disorder?" *J Intern Med*, **254**(5), 401–425.

Loos RJF, Lindgren CM, Li S, *et al.* (2008). "Common variants near MC4R are associated with fat mass, weight and risk of obesity." *Nat Genet*, **40**(6), 768–775.

Lopez-Miranda J, Williams C, Lairon D (2007). "Dietary, physiological, genetic and pathological influences on postprandial lipid metabolism." *Br J Nutr*, **98**(3), 458–473.

Lu J, Xie G, Jia W, Jia W (2013). "Insulin resistance and the metabolism of branched-chain amino acids." *Front Med*, **7**(1), 53–59.

Luo J, Schumacher M, Scherer A, *et al.* (2010). "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data." *Pharmacogenomics J*, **10**(4), 278–291.

Lupi R, Dotta F, Marselli L, *et al.* (2002). "Prolonged exposure to free fatty acids has cytostatic and pro-apoptotic effects on human pancreatic islets: evidence that beta-cell death is caspase mediated, partially dependent on ceramide pathway, and Bcl-2 regulated." *Diabetes*, **51**(5), 1437–1442.

Luís PBM, Ruiter JPN, Ijlst L, *et al.* (2011). "Role of isovaleryl-CoA dehydrogenase and short branched-chain acyl-CoA dehydrogenase in the metabolism of valproic acid: implications for the branched-chain amino acid oxidation pathway." *Drug Metab Dispos*, **39**(7), 1155–1160.

Lyssenko V, Lupi R, Marchetti P, *et al.* (2007). "Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes." *J Clin Invest*, **117**(8), 2155–2163.

Madsen KA, Garber AK, Mietus-Snyder ML, *et al.* (2009). "A clinic-based lifestyle intervention for pediatric obesity: efficacy and behavioral and biochemical predictors of response." *J Pediatr Endocrinol Metab*, **22**(9), 805–814.

Maes HH, Neale MC, Eaves LJ (1997). "Genetic and environmental factors in relative body weight and human adiposity." *Behav Genet*, **27**(4), 325–351.

Mailloux R, Lemire J, Appanna V (2007). "Aluminum-induced mitochondrial dysfunction leads to lipid accumulation in human hepatocytes: a link to obesity." *Cell Physiol Biochem*, **20**(5), 627–638.

Maksimovic J, Gordon L, Oshlack A (2012). "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." *Genome Biol*, **13**(6), R44.

Malpique R, Figueiredo H, Esteban Y, *et al.* (2014). "Integrative analysis reveals novel pathways mediating the interaction between adipose tissue and pancreatic islets in obesity in rats." *Diabetologia*, **57**(6), 1219–1231.

Marabita F, Almgren M, Lindholm ME, *et al.* (2013). "An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform." *Epigenetics*, **8**(3), 333–346.

Marshall A, Altman DG, Holder RL (2010). "Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study." *BMC Medical Research Methodology*, **10**, 112.

Marshall WA, Tanner JM (1969). "Variations in pattern of pubertal changes in girls." *Arch Dis Child*, **44**(235), 291–303.

Marshall WA, Tanner JM (1970). "Variations in the pattern of pubertal changes in boys." *Arch Dis Child*, **45**(239), 13–23.

Martin DI, Cropley JE, Suter CM (2011). "Epigenetics in disease: leader or follower?" *Epigenetics*, **6**(7), 843–848.

Matsushita K, Williams EK, Mongraw-Chaffin ML, *et al.* (2013). "The association of plasma lactate with incident cardiovascular outcomes: the ARIC Study." *Am J Epidemiol*, **178**(3), 401–409.

Matthews DR, Hosker JP, Rudenski AS, *et al.* (1985). "Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man." *Diabetologia*, **28**(7), 412–419.

Mayer B (ed.) (2011). *Bioinformatics for Omics Data*, volume 719 of *Methods in Molecular Biology*. Springer.

McTernan PG, Harte AL, Anderson LA, *et al.* (2002). "Insulin and rosiglitazone regulation of lipolysis and lipogenesis in human adipose tissue in vitro." *Diabetes*, **51**(5), 1493–1498.

Meisinger C, Löwel H, Heier M, Kandler U, Döring A (2007). "Association of sports activities in leisure time and incident myocardial infarction in middle-aged men and women from the general population: the MONICA/KORA Augsburg cohort study." *Eur J Cardiovasc Prev Rehabil*, **14**(6), 788–792.

Meng Q, Mäkinen VP, Luk H, Yang X (2013). "Systems Biology Approaches and Applications in Obesity, Diabetes, and Cardiovascular Diseases." *Curr Cardiovasc Risk Rep*, **7**(1), 73–83.

Meurs I, Lammers B, Zhao Y, *et al.* (2012). "The effect of ABCG1 deficiency on atherosclerotic lesion development in LDL receptor knockout mice depends on the stage of atherogenesis." *Atherosclerosis*, **221**(1), 41–47.

Mihalik SJ, Goodpaster BH, Kelley DE, *et al.* (2010). "Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity." *Obesity (Silver Spring)*, **18**(9), 1695–1700.

Milagro FI, Campión J, Cordero P, *et al.* (2011). "A dual epigenomic approach for the search of obesity biomarkers: DNA methylation in relation to diet-induced weight loss." *FASEB*, **25**(4), 1378–1389.

Miller OJ, Schnedl W, Allen J, Erlanger BF (1974). "5-Methylcytosine localised in mammalian constitutive heterochromatin." *Nature*, **251**, 636–637.

Millstein J, Zhang B, Zhu J, Schadt EE (2009). "Disentangling molecular relationships with a causal inference test." *BMC Genet*, **10**, 23.

Minamino T, Orimo M, Shimizu I, *et al.* (2009). "A crucial role for adipose tissue p53 in the regulation of insulin resistance." *Nat Med*, **15**(9), 1082–1087.

Mäntyselkä P, Kautiainen H, Saltevo J, *et al.* (2012). "Weight change and lipoprotein particle concentration and particle size: a cohort study with 6.5-year follow-up." *Atherosclerosis*, **223**(1), 239–243.

Moleres A, Campión J, Milagro FI, *et al.* (2013). "Differential DNA methylation patterns between high and low responders to a weight loss intervention in overweight or obese adolescents: the EVASYON study." *FASEB J*, **27**(6), 2504–2512.

Moore DS, McCabe GP, Duckworth WM, Sclove SL (2003). *The practice of business statistics companion*, chapter Bootstrap methods and permutation tests. W. H. Freeman.

Mora S, Otvos JD, Rosenson RS, *et al.* (2010). "Lipoprotein particle size and concentration by nuclear magnetic resonance and incident type 2 diabetes in women." *Diabetes*, **59**(5), 1153–1160.

Morris AP, Voight BF, Teslovich TM, *et al.* (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." *Nat Genet*, **44**(9), 981–990.

Mühlberger N, Behrend C, Stark R, Holle R (2003). "Datenbankgestützte Online-Erfassung von Arzneimitteln im Rahmen gesundheitswissenschaftlicher Studien - Erfahrungen mit der IDOM-Software." *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, **34**(4), 601–611.

Murer SB, Knöpfli BH, Aeberli I, *et al.* (2011). "Baseline leptin and leptin reduction predict improvements in metabolic variables and long-term fat loss in obese children and adolescents: a prospective study of an inpatient weight-loss program." *Am J Clin Nutr*, **93**(4), 695–702.

Musunuru K, Orho-Melander M, Caulfield MP, *et al.* (2009). "Ion mobility analysis of lipoprotein subfractions identifies three independent axes of cardiovascular risk." *Arterioscler Thromb Vasc Biol*, **29**(11), 1975–1980.

Naganuma R, Sakurai M, Miura K, *et al.* (2009). "Relation of long-term body weight change to change in lipoprotein particle size in Japanese men and women: the INTERMAP Toyama Study." *Atherosclerosis*, **206**(1), 282–286.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (2004). "The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents." *Pediatrics*, **114**(2 Suppl 4th Report), 555–576.

Naukkarinen J, Surakka I, Pietiläinen KH, *et al.* (2010). "Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes." *PLoS Genet*, **6**(6), e1000976.

Neel JV (1962). "Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?" *Am J Hum Genet*, **14**(4), 353–362.

Newgard CB, An J, Bain JR, *et al.* (2009). "A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance." *Cell Metab*, **9**(4), 311–326.

Ng M, Fleming T, Robinson M, *et al.* (2014). "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013." *The Lancet*.

Nägele H, Gebhardt A, Niendorf A, Kroschinski J, Zeller W (1997). "LDL receptor activity in human leukocyte subtypes: regulation by insulin." *Clin Biochem*, **30**(7), 531–538.

Norheim F, Gjelstad IMF, Hjorth M, *et al.* (2012). "Molecular nutrition research: the modern way of performing nutritional science." *Nutrients*, **4**(12), 1898–1944.

Normand SL (1999). "Meta-analysis: formulating, evaluating, combining, and reporting." *Stat Med*, **18**(3), 321–359.

Novotná R, De Vito P, Currado L, Luly P, Baldini PM (2003). "Involvement of phospholipids in the mechanism of insulin action in HEPG2 cells." *Physiol Res*, **52**(4), 447–454.

Oberbach A, Blüher M, Wirth H, *et al.* (2011). "Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes." *J Proteome Res*, **10**(10), 4769–4788.

Oberbach A, von Bergen M, Blüher S, Lehmann S, Till H (2012). "Combined serum proteomic and metabonomic profiling after laparoscopic sleeve gastrectomy in children and adolescents." *J Laparoendosc Adv Surg Tech A*, **22**(2), 184–188.

O'Connell J, Gurdasani D, Delaneau O, *et al.* (2014). "A general approach for haplotype phasing across the full spectrum of relatedness." *PLoS Genet*, **10**(4), e1004234.

Ogita K, Ai M, Tanaka A, *et al.* (2008). "Serum concentration of small dense low-density lipoprotein-cholesterol during oral glucose tolerance test and oral fat tolerance test." *Clin Chim Acta*, **387**(1-2), 36–41.

Ordovas JM, Shen J (2008). "Gene-environment interactions and susceptibility to metabolic syndrome and other chronic diseases." *J Periodontol*, **79**(8 Suppl), 1508–1513.

Oude Luttikhuis H, Baur L, Jansen H, *et al.* (2009). "Interventions for treating obesity in children." *Cochrane Database Syst Rev*, **1**, CD001872.

Owen CG, Whincup PH, Orfei L, *et al.* (2009). "Is body mass index before middle age related to coronary heart disease risk in later life? Evidence from observational studies." *Int J Obes (Lond)*, **33**(8), 866–877.

Ozato K, Shin DM, Chang TH, Morse 3rd HC (2008). "TRIM family proteins and their emerging roles in innate immunity." *Nat Rev Immunol*, **8**(11), 849–860.

Palmer TM, Lawlor DA, Harbord RM, *et al.* (2012). "Using multiple genetic variants as instrumental variables for modifiable risk factors." *Stat Methods Med Res*, **21**(3), 223–242.

Pannacciulli N, Bunt JC, Koska J, Bogardus C, Krakoff J (2006). "Higher fasting plasma concentrations of glucagon-like peptide 1 are associated with higher resting energy expenditure and fat oxidation rates in humans." *Am J Clin Nutr*, **84**(3), 556–560.

Pathmasiri W, Pratt KJ, Collier DN, *et al.* (2012). "Integrating metabolomic signatures and psychosocial parameters in responsivity to an immersion treatment model for adolescent obesity." *Metabolomics*, **8**(6), 1037–1051.

Pearson H (2007). "Meet the human metabolome." *Nature*, **446**(7131), 8.

Pellis L, van Erk MJ, van Ommen B, *et al.* (2012). "Plasma metabolomics and proteomics profiling after a postprandial challenge reveal subtle diet effects on human metabolic status." *Metabolomics*, **8**(2), 347–359.

Perez-Cornago A, Brennan L, Ibero-Baraibar I, *et al.* (2014). "Metabolomics identifies changes in fatty acid and amino acid profiles in serum of overweight older adults following a weight loss intervention." *J Physiol Biochem*, **70**(2), 593–602.

Petersen AK, Zeilinger S, Kastenmüller G, *et al.* (2014). "Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits." *Hum Mol Genet*, **23**(2), 534–545.

Petretto E, Mangion J, Dickens NJ, *et al.* (2006). "Heritability and tissue specificity of expression quantitative trait loci." *PLoS Genet*, **2**(10), e172.

Petronis A (2010). "Epigenetics as a unifying principle in the aetiology of complex traits and diseases." *Nature*, **465**, 721–727.

Philibert RA, Plume JM, Gibbons FX, Brody GH, Beach SR (2012). "The impact of recent alcohol use on genome wide DNA methylation signatures." *Frontiers in Genetics*, **3**, 54.

Pidsley R, Y Wong CC, Volta M, *et al.* (2013). "A data-driven approach to preprocessing Illumina 450K methylation array data." *BMC Genomics*, **14**, 293.

Pietiläinen KH, Naukkarinen J, Rissanen A, *et al.* (2008). "Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity." *PLoS Med*, **5**(3), e51.

Pietiläinen KH, Sysi-Aho M, Rissanen A, *et al.* (2007). "Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects–a monozygotic twin study." *PLoS One*, **2**(2), e218.

Pischon T, Boeing H, Hoffmann K, *et al.* (2008). "General and abdominal adiposity and risk of death in Europe." *N Engl J Med*, **359**(20), 2105–2120.

Ploner A (2011). *Heatplus: Heatmaps with row and/or column covariates and colored clusters*. R package version 2.1.0.

Portela A, Esteller M (2010). "Epigenetic modifications and human disease." *Nature Biotechnology*, **28**(10), 1057–1068.

Price ME, Cotton AM, Lam LL, *et al.* (2013). "Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array." *Epigenetics Chromatin*, **6**(1), 4.

R Core Team (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Radmacher MD, McShane LM, Simon R (2002). "A paradigm for class prediction using gene expression profiles." *J Comput Biol*, **9**(3), 505–511.

Raessler S, Rubin DB, Zell ER (2008). "Incomplete Data in Epidemiology and Medical Statistics." *Handbook of Statistics*, **27**, 569–601.

Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology*, **27**(1), 85–95.

Rakyan VK, Down TA, Balding DJ, Beck S (2011). "Epigenome-wide association studies for common human diseases." *Nature Reviews Genetics*, **12**(8), 529–541.

Ram O, Goren A, Amit I, *et al.* (2011). "Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells." *Cell*, **147**(7), 1628–1639.

Ramos-Roman MA, Sweetman L, Valdez MJ, Parks EJ (2012). "Postprandial changes in plasma acylcarnitine concentrations as markers of fatty acid flux in overweight and obesity." *Metabolism*, **61**(2), 202–212.

Rampersaud E, Damcott CM, Fu M, *et al.* (2007). "Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations." *Diabetes*, **56**(12), 3053–3062.

Randall JC, Winkler TW, Kutalik Z, *et al.* (2013). "Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits." *PLoS Genet*, **9**(6), e1003500.

Ranganath LR (2008). "The entero-insular axis: implications for human metabolism." *Clin Chem Lab Med*, **46**(1), 43–56.

Rathmann W, Haastert B, Icks A, *et al.* (2003). "High prevalence of undiagnosed diabetes mellitus in Southern Germany: target populations for efficient screening. The KORA survey 2000." *Diabetologia*, **46**(2), 182–189.

Ray S, Turi RH (1999). *Determination of number of clusters in k-means clustering and application in colour image segmentation*, pp. 137–143. Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99).

Reik W, Walter J (2001). "Genomic imprinting: parental influence on the genome." *Nat Rev Genet*, **2**(1), 21–32.

Reinehr T (2011). "Effectiveness of lifestyle intervention in overweight children." *Proc Nutr Soc*, **70**(4), 494–505.

Reinehr T, Andler W (2004). "Changes in the atherogenic risk factor profile according to degree of weight loss." *Arch Dis Child*, **89**(5), 419–422.

Reinehr T, Brylak K, Alexy U, Kersting M, Andler W (2003). "Predictors to success in outpatient training in obese children and adolescents." *Int J Obes Relat Metab Disord*, **27**(9), 1087–1092.

Reinehr T, de Sousa G, Toschke AM, Andler W (2006). "Long-term follow-up of cardiovascular disease risk factors in children after an obesity intervention." *Am J Clin Nutr*, **84**(3), 490–496.

Reinehr T, Hebebrand J, Friedel S, *et al.* (2009a). "Lifestyle intervention in obese children with variations in the melanocortin 4 receptor gene." *Obesity (Silver Spring)*, **17**(2), 382–389.

Reinehr T, Kiess W, Kapellen T, Andler W (2004). "Insulin sensitivity among obese children and adolescents, according to degree of weight loss." *Pediatrics*, **114**(6), 1569–1573.

Reinehr T, Kleber M, de Sousa G, Andler W (2009b). "Leptin concentrations are a predictor of overweight reduction in a lifestyle intervention." *Int J Pediatr Obes*, **4**(4), 215–223.

Reinehr T, Wolters B, Knop C, *et al.* (2014). "Changes in the serum metabolite profile in obese children with weight loss." *Eur J Nutr*.

Reinius LE, Acevedo N, Joerink M, *et al.* (2012). "Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility." *PLoS ONE*, **7**(7), e41361.

Relton CL, Groom A, St Pourcain B, *et al.* (2012). "DNA methylation patterns in cord blood DNA and body size in childhood." *PLoS One*, **7**(3), e31821.

Ren D, Zhou Y, Morris D, *et al.* (2007). "Neuronal SH2B1 is essential for controlling energy and glucose homeostasis." *J Clin Invest*, **117**(2), 397–406.

Rhee KE, Phelan S, McCaffery J (2012). "Early determinants of obesity: genetic, epigenetic, and in utero influences." *Int J Pediatr*, **2012**(463850), 1–9.

Roberts GT, El Badawi SB (1985). "Red blood cell distribution width index in some hematologic diseases." *Am J Clin Pathol*, **83**(2), 222–226.

Rogge MM (2009). "The role of impaired mitochondrial lipid oxidation in obesity." *Biol Res Nurs*, **10**(4), 356–373.

Römisch-Margl W, Prehn C, Bogumil R, *et al.* (2011). "Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics." *Metabolomics*, **8**(1), 133–142.

Rönn T, Volkov P, Davegårdh C, *et al.* (2013). "A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue." *PLoS Genetics*, **9**(6), e1003572.

Rosenbloom KR, Dreszer TR, Long JC, *et al.* (2012). "ENCODE whole-genome data in the UCSC Genome Browser: update 2012." *Nucleic Acids Res*, **40**(Database issue), D912–D917.

Rotllan N, Fernández-Hernando C (2012). "MicroRNA Regulation of Cholesterol Metabolism." *Cholesterol*, **2012**, 847849.

Rubin DB (1978). "Multiple imputations in sample surveys." In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Ruiz JR, Labayen I, Ortega FB, *et al.* (2010). "Attenuation of the effect of the FTO rs9939609 polymorphism on total and central body fat by physical activity in adolescents: the HELENA study." *Arch Pediatr Adolesc Med*, **164**(4), 328–333.

Ruiz R, Jideonwo V, Ahn M, *et al.* (2014). "Sterol regulatory element-binding protein-1 (SREBP-1) is required to regulate glycogen synthesis and gluconeogenic gene expression in mouse liver." *J Biol Chem*, **289**(9), 5510–5517.

Ruiz-Grande C, Alarcón C, Mérida E, Valverde I (1992). "Lipolytic action of glucagon-like peptides in isolated rat adipocytes." *Peptides*, **13**(1), 13–16.

Sabin MA, Ford A, Hunt L, *et al.* (2007). "Which factors are associated with a successful outcome in a weight management programme for obese children?" *J Eval Clin Pract*, **13**(3), 364–368.

Salsberry PJ, Reagan PB (2010). "Effects of heritability, shared environment, and nonshared intrauterine conditions on child and adolescent BMI." *Obesity (Silver Spring)*, **18**(9), 1775–1780.

Sancho V, Trigo MV, González N, *et al.* (2005). "Effects of glucagon-like peptide-1 and exendins on kinase activity, glucose transport and lipid metabolism in adipocytes from normal and type-2 diabetic rats." *J Mol Endocrinol*, **35**(1), 27–38.

Scherag A, Dina C, Hinney A, *et al.* (2010). "Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups." *PLoS Genet*, **6**(4), e1000916.

Schnabel RB, Baccarelli A, Lin H, Ellinor PT, Benjamin EJ (2012). "Next steps in cardiovascular disease genomic research–sequencing, epigenetics, and transcriptomics." *Clin Chem*, **58**(1), 113–126.

Schou J, Frikke-Schmidt R, Kardassis D, *et al.* (2012). "Genetic variation in ABCG1 and risk of myocardial infarction and ischemic heart disease." *Arterioscler Thromb Vasc Biol*, **32**(2), 506–515.

Schousboe K, Willemsen G, Kyvik KO, *et al.* (2003). "Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries." *Twin Res*, **6**(5), 409–421.

Schurmann C, Heim K, Schillert A, *et al.* (2012). "Analyzing Illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium." *PLoS One*, **7**(12), e50938.

Schwandt P, Kelishadi R, Haas GM (2008). "First reference curves of waist circumference for German children in comparison to international values: the PEP Family Heart Study." *World J Pediatr*, **4**(4), 259–266.

Scuteri A, Sanna S, Chen WM, *et al.* (2007). "Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits." *PLoS Genet*, **3**(7), e115.

Seghieri M, Rebelos E, Gastaldelli A, *et al.* (2013). "Direct effect of GLP-1 infusion on endogenous glucose production in humans." *Diabetologia*, **56**(1), 156–161.

Segrè AV, DIAGRAMC, MAGICi, *et al.* (2010). "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits." *PLoS Genet*, **6**(8).

Shaham O, Wei R, Wang TJ, *et al.* (2008). "Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity." *Mol Syst Biol*, **4**, 214.

Shantha GPS, Wasserman B, Astor BC, *et al.* (2013). "Association of blood lactate with carotid atherosclerosis: the Atherosclerosis Risk in Communities (ARIC) Carotid MRI Study." *Atherosclerosis*, **228**(1), 249–255.

Shenker NS, Polidoro S, van Veldhoven K, *et al.* (2013). "Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking." *Human Molecular Genetics*, **22**(5), 843–851.

Shih LY, Liou TH, Chao JCJ, *et al.* (2006). "Leptin, superoxide dismutase, and weight loss: initial leptin predicts weight loss." *Obesity (Silver Spring)*, **14**(12), 2184–2192.

Shimano H (2001). "Sterol regulatory element-binding proteins (SREBPs): transcriptional regulators of lipid synthetic genes." *Prog Lipid Res*, **40**(6), 439–452.

Shin SY, Petersen AK, Wahl S, *et al.* (2014). "Interrogating causal pathways linking genetic variants, small molecule metabolites and circulating lipids." *Genome Med*, **6**(3), 25.

Shore SA (2010). "Obesity, airway hyperresponsiveness, and inflammation." *J Appl Physiol (1985)*, **108**(3), 735–743.

Siri-Tarino PW, Williams PT, Fernstrom HS, Rawlings RS, Krauss RM (2009). "Reversal of small, dense LDL subclass phenotype by normalization of adiposity." *Obesity (Silver Spring)*, **17**(9), 1768–1775.

Skurk T, Rubio-Aliaga I, Stamfort A, Hauner H, Daniel H (2011). "New metabolic interdependencies revealed by plasma metabolite profiling after two dietary challenges." *Metabolomics*, **7**(3), 388–399.

Slieker RC, Bos SD, Goeman JJ, *et al.* (2013). "Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array." *Epigenetics Chromatin*, **6**(1), 26.

Smilowitz JT, Wiest MM, Watkins SM, *et al.* (2009). "Lipid metabolism predicts changes in body composition during energy restriction in overweight humans." *J Nutr*, **139**(2), 222–229.

Smith GD, Ebrahim S (2003). "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" *Int J Epidemiol*, **32**(1), 1–22.

Smyth GK (2005). "Limma: linear models for microarray data." In R Gentleman, V Carey, S Dudoit, R Irizarry, W Huber (eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420. Springer, New York.

Soininen P, Kangas AJ, Würtz P, *et al.* (2009). "High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism." *Analyst*, **134**(9), 1781–1785.

Sokol CL, Barton GM, Farr AG, Medzhitov R (2008). "A mechanism for the initiation of allergen-induced T helper type 2 responses." *Nat Immunol*, **9**(3), 310–318.

Somvanshi PR, Venkatesh KV (2014). "A conceptual review on systems biology in health and diseases: from biological networks to modern therapeutics." *Syst Synth Biol*, **8**(1), 99–116.

Sovio U, Mook-Kanamori DO, Warrington NM, *et al.* (2011). "Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development." *PLoS Genet*, **7**(2), e1001307.

Speakman JR (2007). "A nonadaptive scenario explaining the genetic predisposition to obesity: the "predation release" hypothesis." *Cell Metab*, **6**(1), 5–12.

Speakman JR (2013). "Functional analysis of seven genes linked to body mass index and adiposity by genome-wide association studies: a review." *Hum Hered*, **75**(2-4), 57–79.

Speliotes EK, Willer CJ, Berndt SI, *et al.* (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index." *Nat Genet*, **42**(11), 937–948.

Staiger D, Stock JH (1997). "Instrumental variables regression with weak instruments." *Econometrica*, **65**(3), 557–586.

Standaert ML, Avignon A, Yamada K, Bandyopadhyay G, Farese RV (1996a). "The phosphatidylinositol 3-kinase inhibitor, wortmannin, inhibits insulin-induced activation of phosphatidylcholine hydrolysis and associated protein kinase C translocation in rat adipocytes." *Biochem J*, **313 (Pt 3)**, 1039–1046.

Standaert ML, Bandyopadhyay G, Zhou X, Galloway L, Farese RV (1996b). "Insulin stimulates phospholipase D-dependent phosphatidylcholine hydrolysis, Rho translocation, de novo phospholipid synthesis, and diacylglycerol/protein kinase C signaling in L6 myotubes." *Endocrinology*, **137**(7), 3014–3020.

Stunkard AJ, Foch TT, Hrubec Z (1986). "A twin study of human obesity." *JAMA*, **256**(1), 51–54.

Sturek JM, Castle JD, Trace AP, *et al.* (2010). "An intracellular role for ABCG1-mediated cholesterol transport in the regulated secretory pathway of mouse pancreatic beta cells." *J Clin Invest*, **120**(7), 2575–2589.

Suhre K, Gieger C (2012). "Genetic variation in metabolic phenotypes: study designs and applications." *Nat Rev Genet*, **13**(11), 759–769.

Suhre K, Meisinger C, Döring A, *et al.* (2010). "Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting." *PLoS One*, **5**(11), e13953.

Suhre K, Shin SY, Petersen AK, *et al.* (2011). "Human metabolic individuality in biomedical and pharmaceutical research." *Nature*, **477**(7362), 54–60.

Suzukawa M, Nagase H, Ogahara I, *et al.* (2011). "Leptin enhances survival and induces migration, degranulation, and cytokine synthesis of human basophils." *J Immunol*, **186**(9), 5254–5260.

Suzuki MM, Bird A (2008). "DNA methylation landscapes: provocative insights from epigenomics." *Nature Reviews Genetics*, **9**(6), 465–476.

Svegliati-Baroni G, Saccomanno S, Rychlicki C, *et al.* (2011). "Glucagon-like peptide-1 receptor activation stimulates hepatic lipid oxidation and restores hepatic signalling alteration induced by a high-fat diet in nonalcoholic steatohepatitis." *Liver Int*, **31**(9), 1285–1297.

Szymanska E, Bouwman J, Strassburg K, *et al.* (2012). "Gender-dependent associations of metabolite profiles and body fat distribution in a healthy population with central obesity: towards metabolomics diagnostics." *OMICS*, **16**(12), 652–667.

Takamura T, Misu H, Matsuzawa-Nagata N, *et al.* (2008). "Obesity upregulates genes involved in oxidative phosphorylation in livers of diabetic patients." *Obesity (Silver Spring)*, **16**(12), 2601–2609.

Tan GD, Neville MJ, Liverani E, *et al.* (2006). "The in vivo effects of the Pro12Ala PPARgamma2 polymorphism on adipose tissue NEFA metabolism: the first use of the Oxford Biobank." *Diabetologia*, **49**(1), 158–168.

Tarling EJ, Edwards PA (2011). "ATP binding cassette transporter G1 (ABCG1) is an intracellular sterol transporter." *Proc Natl Acad Sci U S A*, **108**(49), 19719–19724.

Taylor AE, Ebrahim S, Ben-Shlomo Y, *et al.* (2010). "Comparison of the associations of body mass index and measures of central adiposity and fat mass with coronary heart disease, diabetes, and all-cause mortality: a study using data from 4 UK cohorts." *Am J Clin Nutr*, **91**(3), 547–556.

Taylor RW, Jones IE, Williams SM, Goulding A (2000). "Evaluation of waist circumference, waist-to-hip ratio, and the conicity index as screening tools for high trunk fat mass, as measured by dual-energy X-ray absorptiometry, in children aged 3-19 y." *Am J Clin Nutr*, **72**(2), 490–495.

Teixeira PJ, Going SB, Houtkooper LB, *et al.* (2004). "Pretreatment predictors of attrition and successful weight management in women." *Int J Obes Relat Metab Disord*, **28**(9), 1124–1133.

Teschendorff AE, Marabita F, Lechner M, *et al.* (2013). "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data." *Bioinformatics*, **29**(2), 189–196.

Thamer C, Machann J, Stefan N, *et al.* (2007). "High visceral fat mass and high liver fat are associated with resistance to lifestyle intervention." *Obesity (Silver Spring)*, **15**(2), 531–538.

Then C, Wahl S, Kirchhofer A, *et al.* (2013). "Plasma metabolomics reveal alterations of sphingo- and glycerophospholipid levels in non-diabetic carriers of the transcription factor 7-like 2 polymorphism rs7903146." *PLoS One*, **8**(10), e78430.

Thorleifsson G, Walters GB, Gudbjartsson DF, *et al.* (2009). "Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity." *Nat Genet*, **41**(1), 18–24.

Tibshirani R (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, **58**(1), 267–288.

Timper K, Grisouard J, Sauter NS, *et al.* (2013). "Glucose-dependent insulinotropic polypeptide induces cytokine expression, lipolysis, and insulin resistance in human adipocytes." *Am J Physiol Endocrinol Metab*, **304**(1), E1–13.

Tollefsbol T (2011). *Handbook of epigenetics: the new molecular and medical genetics.* Academic Press, London.

Toperoff G, Aran D, Kark JD, *et al.* (2012). "Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood." *Human Molecular Genetics*, **21**(2), 371–383.

Tost J (2010). "DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker." *Molecular Biotechnology*, **44**, 71–81.

Touleimat N, Tost J (2012). "Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation." *Epigenomics*, **4**(3), 325–341.

Trottier MD, Naaz A, Li Y, Fraker PJ (2012). "Enhancement of hematopoiesis and lymphopoiesis in diet-induced obese mice." *Proc Natl Acad Sci U S A*, **109**(20), 7622–7629.

Tschritter O, Preissl H, Hennige AM, *et al.* (2009). "The insulin effect on cerebrocortical theta activity is associated with serum concentrations of saturated nonesterified Fatty acids." *J Clin Endocrinol Metab*, **94**(11), 4600–4607.

Tschritter O, Preissl H, Hennige AM, *et al.* (2012). "High cerebral insulin sensitivity is associated with loss of body fat during lifestyle intervention." *Diabetologia*, **55**(1), 175–182.

Tukiainen T, Kettunen J, Kangas AJ, *et al.* (2012). "Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci." *Hum Mol Genet*, **21**(6), 1444–1455.

Tung YCL, Yeo GSH (2011). "From GWAS to biology: lessons from FTO." *Ann N Y Acad Sci*, **1220**, 162–171.

Tzotzas T, Desrumaux C, Lagrost L (2009). "Plasma phospholipid transfer protein (PLTP): review of an emerging cardiometabolic risk factor." *Obes Rev*, **10**(4), 403–411.

Uriarte G, Paternain L, Milagro FI, Martínez JA, Campion J (2013). "Shifting to a control diet after a high-fat, high-sucrose diet intake induces epigenetic changes in retroperitoneal adipocytes of Wistar rats." *J Physiol Biochem*, **69**(3), 601–611.

Valcárcel B, Ebbels TMD, Kangas AJ, *et al.* (2014). "Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity." *J R Soc Interface*, **11**(94), 20130908.

van Buuren S (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical Methods in Medical Research*, **16**(3), 219–242.

van Buuren S, Boshuizen HC, Knook DL (1999). "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine*, **18**(6), 681–694.

van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006). "Fully conditional specification in multivariate imputation." *Journal of Statistical Computation and Simulation*, **76**(12), 1049–1064.

van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate imputation by chained equations in R." *Journal of Statistical Software*, **45**(3), 1–67.

Van Gaal LF, Mertens IL, De Block CE (2006). "Mechanisms linking obesity with cardiovascular disease." *Nature*, **444**(7121), 875–880.

van Ommen B, Keijer J, Heil SG, Kaput J (2009). "Challenging homeostasis to define biomarkers for nutrition related health." *Mol Nutr Food Res*, **53**(7), 795–804.

Varma S, Simon R (2006). "Bias in error estimation when using cross-validation for model selection." *BMC Bioinformatics*, **7**, 91.

Vendrell J, El Bekay R, Peral B, *et al.* (2011). "Study of the potential association of adipose tissue GLP-1 receptor with obesity and insulin resistance." *Endocrinology*, **152**(11), 4072–4079.

Verdich C, Toubro S, Buemann B, *et al.* (2001). "Leptin levels are associated with fat oxidation and dietary-induced weight loss in obesity." *Obes Res*, **9**(8), 452–461.

Villanueva-Peñacarrillo ML, Márquez L, González N, Díaz-Miguel M, Valverde I (2001). "Effect of GLP-1 on lipid metabolism in human adipocytes." *Horm Metab Res*, **33**(2), 73–77.

Vinayavekhin N, Homan EA, Saghatelian A (2010). "Exploring disease through metabolomics." *ACS Chem Biol*, **5**(1), 91–103.

Virtue MA, Furne JK, Nuttall FQ, Levitt MD (2004). "Relationship between GHb concentration and erythrocyte survival determined from breath carbon monoxide concentration." *Diabetes Care*, **27**(4), 931–935.

von Ruesten A, Steffen A, Floegel A, *et al.* (2011). "Trend in obesity prevalence in European adult cohort populations during follow-up since 1996 and their predictions to 2015." *PLoS One*, **6**(11), e27455.

Wabitsch M, Hauner H, Böckmann A, *et al.* (1992). "The relationship between body fat distribution and weight loss in obese adolescent girls." *Int J Obes Relat Metab Disord*, **16**(11), 905–911.

Wabitsch M, Moss A, im Kindes und Jugendalter (AGA) AA (2009). "Therapie der Adipositas im Kindes- und Jugendalter - Evidenzbasierte Leitlinien der Arbeitsgemeinschaft der Adipositas im Kinder- und Jugendalter (AGA) und der beteiligten medizinischen-wissenschaftlichen Fachgesellschaften, Berufsverbände und weiterer Organisationen."

Wabitsch M, Moss A, Kromeyer-Hauschild K (2014). "Unexpected plateauing of childhood obesity rates in developed countries." *BMC Med*, **12**, 17.

Wahl S, Fenske N, Zeilinger S, *et al.* (2014). "On the potential of models for location and scale for genome-wide DNA methylation data." *BMC Bioinformatics*, **15**(1), 232.

Wahl S, Holzapfel C, Yu Z, *et al.* (2013a). "Metabolomics reveals determinants of weight loss during lifestyle intervention in obese children." *Metabolomics*, **9**(6), 1157–1167.

Wahl S, Krug S, Then C, *et al.* (2013b). "Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele." *Metabolomics*, **10**, 386–401.

Wahl S, Yu Z, Kleber M, *et al.* (2012). "Childhood obesity is associated with changes in the serum metabolite profile." *Obes Facts*, **5**(5), 660–670.

Wajchenberg BL (2000). "Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome." *Endocr Rev*, **21**(6), 697–738.

Walley AJ, Jacobson P, Falchi M, *et al.* (2012). "Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue." *Int J Obes (Lond)*, **36**(1), 137–147.

Walters SJ (2004). "Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36." *Health Qual Life Outcomes*, **2**, 26.

Wang C, Feng R, Sun D, *et al.* (2011). "Metabolic profiling of urine in young obese men using ultra performance liquid chromatography and Q-TOF mass spectrometry (UPLC/Q-TOF MS)." *J Chromatogr B Analyt Technol Biomed Life Sci*, **879**(27), 2871–2876.

Wang X, Zhu H, Snieder H, *et al.* (2010). "Obesity related methylation changes in DNA of peripheral blood leukocytes." *BMC Medicine*, **8**, 87.

Wang Y, Barbacioru C, Hyland F, *et al.* (2006). "Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays." *BMC Genomics*, **7**, 59.

Warnes GR, Bolker B, Bonebakker L, *et al.* (2014). *gplots: Various R programming tools for plotting data.* R package version 2.13.0, URL `http://CRAN.R-project.org/package=gplots`.

Waterham HR, Koster J, Romeijn GJ, *et al.* (2001). "Mutations in the 3beta-hydroxysterol Delta24-reductase gene cause desmosterolosis, an autosomal recessive disorder of cholesterol biosynthesis." *Am J Hum Genet*, **69**(4), 685–694.

Weickert MO, Loeffelholz CV, Roden M, *et al.* (2007). "A Thr94Ala mutation in human liver fatty acid-binding protein contributes to reduced hepatic glycogenolysis and blunted elevation of plasma glucose levels in lipid-exposed subjects." *Am J Physiol Endocrinol Metab*, **293**(4), E1078–E1084.

Welter D, MacArthur J, Morales J, *et al.* (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic Acids Res*, **42**(Database issue), D1001–D1006.

Westphal S, Orth M, Ambrosch A, Osmundsen K, Luley C (2000). "Postprandial chylomicrons and VLDLs in severe hypertriacylglycerolemia are lowered more effectively than are chylomicron remnants after treatment with n-3 fatty acids." *Am J Clin Nutr*, **71**(4), 914–920.

Whitney AR, Diehn M, Popper SJ, *et al.* (2003). "Individuality and variation in gene expression patterns in human blood." *Proc Natl Acad Sci U S A*, **100**(4), 1896–1901.

Wiklund PK, Pekkala S, Autio R, *et al.* (2014). "Serum metabolic profiles in overweight and obese women with and without metabolic syndrome." *Diabetol Metab Syndr*, **6**(1), 40.

Willer CJ, Speliotes EK, Loos RJF, *et al.* (2009). "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation." *Nat Genet*, **41**(1), 25–34.

Wishart DS, Jewison T, Guo AC, *et al.* (2013). "HMDB 3.0–The Human Metabolome Database in 2013." *Nucleic Acids Res*, **41**(Database issue), D801–D807.

Wolfenstetter SB, Menn P, Holle R, *et al.* (2012). "Body weight changes and outpatient medical care utilisation: Results of the MONICA/KORA cohorts S3/F3 and S4/F4." *Psychosoc Med*, **9**, Doc09.

Wong CCY, Caspi A, Williams B, *et al.* (2010). "A longitudinal study of epigenetic variation in twins." *Epigenetics*, **5**(6), 516–526.

Woods SC, D'Alessio DA (2008). "Central control of body weight and appetite." *J Clin Endocrinol Metab*, **93**(11 Suppl 1), S37–S50.

World Health Organization (2000). "Obesity: preventing and managing the global epidemic. Report of a WHO consultation." *World Health Organ Tech Rep Ser*, **894**, i–xii, 1–253.

Würtz P, Tiainen M, Mäkinen VP, *et al.* (2012). "Circulating metabolite predictors of glycemia in middle-aged men and women." *Diabetes Care*, **35**(8), 1749–1756.

Wu C, Miloslavskaya I, Demontis S, Maestro R, Galaktionov K (2004). "Regulation of cellular response to oncogenic and oxidative stress by Seladin-1." *Nature*, **432**(7017), 640–645.

Wybranska I, Malczewska-Malec M, Partyka L, *et al.* (2007). "Evaluation of genetic predisposition to insulin resistance by nutrient-induced insulin output ratio (NIOR)." *Clin Chem Lab Med*, **45**(9), 1124–1132.

Xu X, Su S, Barnes VA, *et al.* (2013). "A genome-wide methylation study on obesity: differential variability and differential methylation." *Epigenetics*, **8**(5), 522–533.

Yang J, Loos RJF, Powell JE, *et al.* (2012). "FTO genotype is associated with phenotypic variability of body mass index." *Nature*, **490**(7419), 267–272.

Yano M, Watanabe K, Yamamoto T, *et al.* (2011). "Mitochondrial dysfunction and increased reactive oxygen species impair insulin secretion in sphingomyelin synthase 1-null mice." *J Biol Chem*, **286**(5), 3992–4002.

Ye D, Lammers B, Zhao Y, *et al.* (2011). "ATP-binding cassette transporters A1 and G1, HDL metabolism, cholesterol efflux, and inflammation: important targets for the treatment of atherosclerosis." *Curr Drug Targets*, **12**(5), 647–660.

Yu Z, Kastenmüller G, He Y, *et al.* (2011). "Differences between human plasma and serum metabolite profiles." *PLoS One*, **6**(7), e21230.

Yuan Y (2011). "Multiple Imputation Using SAS Software." *Journal of Statistical Software*, **45**(6), 1–25.

Zeilinger S, Kühnel B, Klopp N, *et al.* (2013). "Tobacco smoking leads to extensive genome-wide changes in DNA methylation." *PLoS ONE*, **8**(5), e63812.

Zeller T, Wild P, Szymczak S, *et al.* (2010). "Genetics and beyond–the transcriptome of human monocytes and disease susceptibility." *PLoS One*, **5**(5), e10693.

Zhang B, Horvath S (2005). "A general framework for weighted gene co-expression network analysis." *Stat Appl Genet Mol Biol*, **4**, Article17.

Zhang FF, Cardarelli R, Carroll J, *et al.* (2011). "Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood." *Epigenetics*, **6**, 623–629.

Zhang G, He P, Tan H, *et al.* (2013). "Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer." *Clin Cancer Res*, **19**(18), 4983–4993.

Zhang Y, Jiao J, Yang P, Lu H (2014). "Mass spectrometry-based N-glycoproteomics for cancer biomarker discovery." *Clin Proteomics*, **11**(1), 18.

Zhang Y, Ranta F, Tang C, *et al.* (2009). "Sphingomyelinase dependent apoptosis following treatment of pancreatic beta-cells with amyloid peptides Abeta(1-42) or IAPP." *Apoptosis*, **14**(7), 878–889.

Zhao X, Peter A, Fritsche J, *et al.* (2009). "Changes of the plasma metabolome during an oral glucose tolerance test: is there more than glucose to look at?" *Am J Physiol Endocrinol Metab*, **296**(2), E384–E393.

Zhou M, Wang S, Zhao A, *et al.* (2012). "Transcriptomic and metabonomic profiling reveal synergistic effects of quercetin and resveratrol supplementation in high fat diet fed mice." *J Proteome Res*, **11**(10), 4961–4971.

Zhu C, Liang Ql, Hu P, Wang Ym, Luo Ga (2011). "Phospholipidomic identification of potential plasma biomarkers associated with type 2 diabetes mellitus and diabetic nephropathy." *Talanta*, **85**(4), 1711–1720.

Zhu ZZ, Hou L, Bollati V, *et al.* (2012). "Predictors of global methylation levels in blood DNA of healthy subjects: a combined analysis." *International Journal of Epidemiology*, **41**(1), 126–139.

Ziegler A, König IR (2010). *A statistical approach to genetic epidemiology: concepts and applications*. 2 edition. Wiley-VCH, Weilheim.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." *Nat Genet*, **39**(1), 61–69.

Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J (2014). "Epigenome-wide association studies without the need for cell-type composition." *Nat Methods*, **11**(3), 309–311.

A Appendix

A.1 Appendix statistical methods

A.1.1 Quantile normalization (QN)

Let $\mathbf{X}$ be a data matrix with dimension $p \times n$, where p corresponds to the number of features, and n to the number of observations. Then, QN is achieved by the following algorithm (Bolstad *et al.*, 2003):

1. Sort each column of $\mathbf{X}$ to get $\mathbf{X}_{\text{sort}}$, and save the original ordering
2. Compute the vector of row means of $\mathbf{X}_{\text{sort}}$
3. Assign this mean to each element in the respective row to get $\mathbf{X}'_{\text{sort}}$
4. Rearrange each column of $\mathbf{X}'_{\text{sort}}$ to have the same ordering as the original matrix $\mathbf{X}$.

A.1.2 Missing data handling

Multiple imputation by chained equations (MICE)

Multiple imputation by chained equations (MICE), also referred to as *sequential regression multivariate imputation* and *fully conditional specification* is based on the following algorithm (van Buuren *et al.*, 1999, Raghunathan *et al.*, 2001, van Buuren, 2007):

(1) Fill all missing values with arbitrary start values.

(2) Repeat the following steps for the incomplete variable vectors $\boldsymbol{x}_k$, $k = 1, \ldots, K$ for a pre-specified number of iterations ($b = 1, \ldots, B$):

(a) Regress $\boldsymbol{x}_k$ on all other variables $\tilde{\mathbf{X}} = \left\{ \mathbf{X}_{obs}, \boldsymbol{x}^{(b)}_{1,mis}, ..., \boldsymbol{x}^{(b)}_{k-1,mis}, \boldsymbol{x}^{(b-1)}_{k+1,mis}, ..., \boldsymbol{x}^{(b-1)}_{K,mis} \right\}$ using an appropriate regression model (e.g., linear, logistic or multinomial regression if $\boldsymbol{x}_k$ is continuous, dichotomous or categorical, respectively) to obtain estimates $\hat{\boldsymbol{\theta}}$ (e.g., $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ in the case of linear regression). Potentially, transform skewed continuous variables $\boldsymbol{x}_k$ first to ensure normality.

(b) Draw from the conditional posterior distribution of the parameters given $\tilde{\mathbf{X}}$:

$$\boldsymbol{\theta}^{(b)}_k \;\sim\; p\left(\boldsymbol{\theta}_k \,\middle|\, \tilde{\mathbf{X}}\right).$$

In Bayesian linear regression (Rubin, 1987, Yuan, 2011), this corresponds to:

$$\begin{aligned} \sigma_k^{2(b)} \Big| \tilde{\mathbf{X}} &\sim \textit{Scale-inv-}\chi^2\left(n-p, \hat{\sigma}^2\right), \text{ and} \\ \boldsymbol{\beta}_k^{(b)} \Big| \tilde{\mathbf{X}}, \sigma_k^{2(b)} &\sim \text{MVN}\left(\hat{\boldsymbol{\beta}}, \sigma_k^{2(b)}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\right), \end{aligned}$$

with the standard estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ (see Section A.1.3).

(c) Draw from the full conditional posterior distribution of the missing values given the updated parameters:

$$\boldsymbol{x}_{k,mis}^{(b)} \sim p\left(\boldsymbol{x}_{k,mis} \Big| \tilde{\mathbf{X}}, \boldsymbol{\theta}_k^{(b)}\right).$$

In Bayesian linear regression, this corresponds to:

$$\boldsymbol{x}_{k,mis}^{(b)} \Big| \tilde{\mathbf{X}}, \boldsymbol{\beta}_k^{(b)}, \sigma_k^{2(b)} \sim \text{MVN}\left(\tilde{\mathbf{X}}^T\boldsymbol{\beta}_k^{(b)}, \sigma_k^{2(b)}\right).$$

(3) Use $\boldsymbol{x}_{mis}^{(b)}$ as imputed values for the missings. A robust extension of Bayesian linear regression imputation is *predictive mean matching* (PMM), where an observed value with a similar predicted value is used as imputed value (Little and Rubin, 2002), thereby ensuring that the imputed values are within a plausible range.

(4) Repeat steps (1) to (3) M times, to obtain M imputed data sets.

Although it does not necessarily converge (Raghunathan *et al.*, 2001, van Buuren *et al.*, 2006), this procedure has proven to provide unbiased and valid inference in numerous applications (Raghunathan *et al.*, 2001, van Buuren *et al.*, 1999, 2006, van Buuren, 2007, Lee and Carlin, 2010, Marshall *et al.*, 2010, Drechsler, 2011). It is however important to monitor convergence and performance of the imputation models. Abayomi *et al.* (2008), Drechsler (2011) and van Buuren and Groothuis-Oudshoorn (2011) introduce a number of diagnostic techniques.

The algorithm can be made more flexible. For instance, it can be extended to consider variables that are defined for a subpopulation only, such as the variable "intake of oral contraceptives" which is defined for females only (van Buuren, 2007). It can also be extended to consider logical bounds and constraints, e.g. a minimum of zero for metabolite concentrations. Important to the quality of MICE is the appropriate choice of covariates in the imputation models. Generally, an "inclusive" strategy should be preferred to avoid bias (van Buuren *et al.*, 1999, Collins *et al.*, 2001). Specifically, the models should include (van Buuren and Groothuis-Oudshoorn, 2011):

(i) variables that are to be included in the subsequent statistical models (including interactions when stratified or interaction models are planned, and including the response variable), to preserve correlations among the variables.

(ii) "auxiliary" variables that are correlated with the incomplete variable in question, since they contribute information for its imputation;

(iii) "auxiliary" variables that are correlated with the missingness of the incomplete variable in question, making the MAR assumption more plausible.

If the number of such variables is too large, additional criteria might be used to restrict their number, e.g. stricter thresholds for correlation, or a required number of joint observations with the incomplete variable. Auxiliary variables for auxiliary variables might also be included to a certain level.

Combination of results from multiply imputed data sets

Having generated M imputed data sets and analyzed them with standard methods, the obtained estimates $\hat{Q}$ can be combined using combination rules by Rubin (1987). These are applicable for estimates with an approximate complete-data normal distribution, and shown here for scalar estimates $\hat{Q}$:

Let $\hat{Q}^{(m)}$ be the estimate if interest obtained from imputation m, $m = 1, \ldots, M$, and $\hat{Var}(\hat{Q}^{(m)})$ its variance. Then a combined estimate can be obtained as the average of the estimates $\hat{\theta}^{(m)}$ obtained from of the M imputed data sets:

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}^{(m)}.$$

Two variance components need to be combined: the *within-imputation variance* W

$$W = \frac{1}{M} \sum_{m=1}^{M} \hat{Var}(\hat{Q}^{(m)})$$

and the *between-imputation variance* B, which is introduced as a result of uncertainty about the missing values:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{Q}^{(m)} - \bar{Q}\right)^2.$$

Total variance T can be summarized by

$$T = W + (1 + \frac{1}{M}) \cdot B.$$

Rubin (1987) derived the following asymptotic reference distribution:

$$(Q - \bar{Q}) \cdot T^{-1/2} \sim t_{\nu_M},$$

with associated degrees of freedom

$$\nu_M = \frac{M-1}{\hat{\gamma}^2},$$

where

$$\hat{\gamma} = \frac{(1+\frac{1}{M})\cdot B}{T}$$

defines the fraction of information about Q missing due to missing data. $\hat{\gamma}$ does not necessarily correlate with the fraction of missing values in the model variables, but is also influenced by the degree of correlation of these variables with auxiliary variables in the data set. For instance, a variable with a large number of missing values might potentially be imputed with a very high accuracy if it is highly correlated with other variables in the data set, resulting in a low value of $\hat{\gamma}$. Furthermore, the *relative efficiency* (RE), defined as

$$RE = \frac{T_\infty}{T_M} \approx \frac{1}{1+\frac{\hat{\gamma}}{M}},$$

describes the proportion of T that could not have been avoided by using an infinite number of imputations (Rubin, 1987). RE increases with increasing M. It should optimally be close to 1 and can provide some indication of how many imputations should be used. Although according to Rubin (1987), 5-10 imputations should be sufficient even when the fraction of missing information is moderate, a larger number of imputations M might be advisable in the case of a large amount of missing information (Bodner, 2008, Graham *et al.*, 2007).

A.1.3 Linear regression

Model

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be an $n \times 1$ continuous response vector corresponding to n independent observations, and $\mathbf{X}$ the $n \times (p+1)$ covariate matrix. Then the *linear regression model* is defined as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2\mathbf{I}_n\right),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ represents the $(p+1) \times 1$ vector of unknown regression coefficients (including the intercept), $\boldsymbol{\eta}$ the linear predictor, and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of uncorrelated error terms with the common variance σ^2 (Fahrmeir *et al.*, 2013, Faraway, 2002). That is, one assumes that the response is a linear function of the covariates (subject to random noise, represented by $\boldsymbol{\epsilon}$), with the regression coefficients β_j, $j = 1, \ldots, p$, representing the linear association of each covariate $\boldsymbol{x}_j$ with the response vector $\boldsymbol{y}$. More precisely, β_j can be intepreted as change in $\boldsymbol{y}$ per one unit increase in $\boldsymbol{x}_j$.

$\mathbf{X}$ might contain dummies $I(\boldsymbol{x}_j = c)$ of a categorical covariate with C categories $c = 1, \ldots C$ (corresponding to $C-1$ dummies), in which case the respective entries in $\boldsymbol{\beta}$ specify

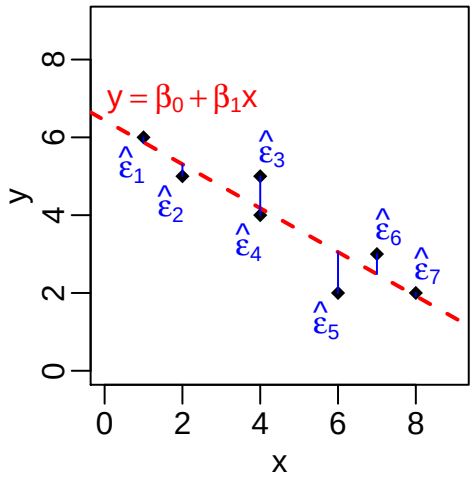

Figure A.1.1: Least squares estimation in the $p = 1$ situation.

the change in the response when comparing the respective category c with the reference category. Also, the linear predictor can be extended to contain interaction effects of two variables. That would be reasonable when one variable is assumed to modify the effect of another variable on the response. Technically, for each interaction included in the model, a column of $\mathbf{X}$ would represent the product of the two variables in question, and the corresponding entry in $\boldsymbol{\beta}$ would specify the effect modification between the variables.

Estimation

The *least squares* estimate $\boldsymbol{\beta}$ is obtained by minimizing the sum of squared errors:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \epsilon_i^2 = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}). \tag{A.1}$$

In the $p = 1$ case, this resembles finding the linear line through the "cloud of dots" which minimizes the sum of the squared distances of the dots to the line (Figure A.1.1). The solution to this minimization task is:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} \\ \hat{\boldsymbol{\sigma}^2} &= \frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{n - p - 1} \end{aligned} \tag{A.2}$$

(Fahrmeir *et al.*, 2013, Faraway, 2002).

Hypothesis testing

To allow for hypothesis tests on the effect estimates $\hat{\beta}_j$, it is commonly assumed that the error terms (approximately) follow a normal distribution: $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Given this

assumption, $\hat{\boldsymbol{\beta}}$ also follows a normal distribution (Fahrmeir *et al.*, 2013):

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}), \text{ and} \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{[j,j]}}} \sim t_{n-p-1}.$$

Thus, one can e.g. construct a test statistic t_j for the hypothesis test with the null hypothesis H_0: $\beta_j = 0$ ("no association between the response and covariate j") vs. the alternative hypothesis H_1: $\beta_j \neq 0$ ("response and covariate j are associated"):

$$t_j = \frac{\hat{\beta}_j}{\hat{\text{se}}\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{[j,j]}}}.$$

Given H_0 is true, t_j should be a draw from a t distribution with $n - p - 1$ degrees of freedom. Accordingly, a p-value can be computed as the probability of observing a test statistic at least as extreme as the observed one given H_0 is true:

$$p\text{-value} = P\left(|T| \geq |t_j|\right),$$

where T is a random t_{n-p-1} distributed variable.

A.1.4 Logistic regression

The logistic regression model is given as

$$\begin{aligned} P(y_i = 1) = \pi_i &= h(\eta_i) = h(\boldsymbol{x}_i^T\boldsymbol{\beta}) \\ g(\pi_i) = h^{-1}(\pi_i) &= \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, \end{aligned}$$

whereby $h(\cdot)$ denotes a *response function* that maps the linear predictor η_i defined on the real space into $[0, 1]$. Often, $h(\cdot)$ is chosen as the logistic function $h(\eta_i) = \frac{exp(\eta_i)}{1+exp(\eta_i)}$ (corresponding to the logit link $g(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right)$). For parameter estimation and tests, see Fahrmeir *et al.* (2013). Effect estimates obtained from logistic regression are often transformed to *odds ratios* for reasons of interpretability:

$$\begin{aligned} \text{OR}_j &= \frac{\text{Odds}(x_ij = x + 1)}{\text{Odds}(x_ij = x + 1)} = \frac{\frac{P(y_i=1|x_ij=x+1)}{1-P(y_i=1|x_ij=x+1)}}{\frac{P(y_i=1|x_ij=x)}{1-P(y_i=1|x_ij=x)}} \\ &= \frac{exp(\beta_0 + \ldots + \beta_j(x+1) + \ldots + \beta_p x_p)}{exp(\beta_0 + \ldots + \beta_j x + \ldots + \beta_p x_p)} = exp(\beta_j). \end{aligned}$$

OR_j can be interpreted as the multiplicative change in the odds (chance) of observing $y = 1$ (e.g., the disease) per one unit increase in covariate j.

A.1.5 Multiple testing procedures

The two largest classes of available procedures control either the probability of any false rejection, referred to as *family-wise error rate* (FWER), or – which is less strict – the expected proportion of false rejections among the rejections, referred to as *false discovery rate* (FDR) (Benjamini and Hochberg, 1995, Dudoit *et al.*, 2003). Using the terminology from Table A.1.1, they can be written as:

$$\begin{aligned} \text{FWER} &= P(V > 0), \text{ and} \\ \text{FDR} &= E(V/R). \end{aligned}$$

Table A.1.1: Test outcome when testing p null hypotheses (adopted from Benjamini and Hochberg (1995)).

	# H_0 **not rejected**	# H_0 **rejected**	**Total**
# true H_0	U	V	p_0
# non-true H_0	T	S	$p - p_0$
Total	$p - R$	R	p

To correct for multiple testin, a widely used approach is the *Bonferroni procedure*, according to which a null hypothesis H_j is rejected, if the corresponding p-value meets $p_j \leq \frac{\alpha}{p}$, where p denotes the number of variables and α the chosen significance level, $\alpha = 0.05$ in this thesis. Alternatively, one can compute adjusted p-values $p_j^* = p_j \cdot p$ and reject if $p_j^* < \alpha$. It can be shown that this procedure controls the FWER at α (Dudoit *et al.*, 2003):

$$\text{FWER} = P(V > 0) = P(\cup_{j=1}^{p_0} p_j \leq \frac{\alpha}{p}) \leq \sum_{j=1}^{p_0} P(p_j \leq \frac{\alpha}{p}) = \sum_{j=1}^{p_0} \frac{\alpha}{p} = \frac{p_0}{p}\alpha \leq \alpha.$$

Another frequently applied approach is the *Benjamini-Hochberg procedure* (Benjamini and Hochberg, 1995), according to which the null hypotheses $H_{(1)}, \ldots, H_{(r)}$ are rejected with r the rank of the largest p-value for which $p_{(r)} \leq \frac{r}{p}\alpha$, if existent. Adjusted p-values can be expressed as

$$p_{(j)}^* = \min_{r=j,\ldots,p} \left\{ \min \left(\frac{p}{r} p_{(r)} \right) \right\}.$$

The procedure controls the FDR at α under independence, and under certain dependence structures (positive regression dependency), which might be assumed to hold in many practical situations (Benjamini and Yekutieli, 2001).

A.1.6 Meta-analysis

Fixed-effects meta-analysis

Let $\hat{\theta}_k$, $k = 1, \ldots, K$, be estimates derived from K independent studies to be combined, for instance β coefficients from linear models. Then, the fixed-effects model postulates that a common true parameter θ underlies all studies, and that the $\hat{\theta}_k$ are realizations from a distribution with mean θ and estimation error variance σ_k^2 (Normand, 1999). Given moderately large sample sizes (central limit theorem), the $\hat{\theta}_k$ asymptotically follow

$$\hat{\theta}_k \sim N\left(\theta, \sigma_k^2\right), \; k = 1, \ldots, K.$$

The corresponding maximum likelihood estimator for θ is:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \text{ with } w_k = \frac{1}{\sigma_k^2},$$

referred to as *inverse-variance weighting* of the study estimates. Because of $\hat{\theta} \sim N\left(\theta, \frac{1}{\sum_{k=1}^{K} w_k}\right)$, standard hypothesis tests can be performed on $\hat{\theta}$.

An alternative fixed-effects method is to combine z statistics z_k from the single studies as $z = \frac{\sum_{k=1}^{K} w_k z_k}{\sqrt{\sum_{k=1}^{K} w_k^2}}$ with $w_k = \sqrt{n_k}$.

Random-effects meta-analysis

The assumption that a θ underlies every study might be unrealistic in scenarios where substantial heterogeneity is present between the studies, e.g. due to differences in phenotypic and technical study characteristics (Borenstein *et al.*, 2010). In that case, the random-effects model might be more appropriate. It assumes that the $\hat{\theta}_k$ are realizations from a distribution with study-specific means θ_k, which again are realizations from a distribution with mean θ and variance τ^2:

$$\begin{aligned} \hat{\theta}_k | \theta_k, \sigma_k^2 &\sim N\left(\theta_k, \sigma_k^2\right) \\ \theta_k | \theta, \tau^2 &\sim N\left(\theta, \tau^2\right). \end{aligned}$$

Similarly as described above, $\hat{\theta}$ can be obtained as weighted mean of the individual $\hat{\theta}_k$, this time using the weights $w_k(\tau) = \frac{1}{\sigma_k^2 + \hat{\tau}^2}$, where different estimation strategies for τ^2 have been proposed (Normand, 1999).

The decision for fixed- versus random-effects meta-analysis can be guided by the estimated heterogeneity between the studies. A commonly used test statistic is *Cochran's Q*:

$$Q = \sum_{k=1}^{K} w_k \left(\hat{\theta}_k - \hat{\theta}\right)^2 \sim \chi^2_{K-1}.$$

If the null hypothesis of homogeneity between the studies is rejected, random-effects meta-analysis might be more sensible. However, the power of tests on Q is low when K is small (Hardy and Thompson, 1998). Therefore, an alternative measure $I^2 = \frac{Q-(K-1)}{Q} \cdot 100\%$ has been proposed (Higgins and Thompson, 2002).

A.1.7 Principal component analysis

Let $\mathbf{X}$ be the $n \times p$ feature matrix with covariance matrix $\mathbf{\Sigma}$. Then, PCs $\boldsymbol{p}_j = \mathbf{X}\boldsymbol{a}_j$, $j = 1, \ldots, q \leq p$ are subsequently defined such that $\boldsymbol{a}_j = \arg\max Var(\boldsymbol{p}_j)$ under the restrictions $\boldsymbol{a}_j^T \boldsymbol{a}_j = 1$ and $\boldsymbol{a}_j^T \boldsymbol{a}_l = 0$, $l = 1, \ldots, j-1$. This optimization task can be formalized through *spectral decomposition* of the covariance matrix: $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$, where $\mathbf{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q)$ is the orthonormal $p \times q$ matrix of eigenvectors $\boldsymbol{a}_j$, and $\mathbf{\Lambda} = diag\,(\lambda_1, \ldots, \lambda_q)$ the $q \times q$ diagonal matrix of sorted eigenvalues. Then, the $n \times q$ matrix of PCs is

$$\mathbf{P} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_q) = \mathbf{X}\mathbf{A}.$$

That the resulting PCs are orthogonal and variance declines with the number of components is evident from

$$Cov(\mathbf{P}) = Cov(\mathbf{X}\mathbf{A}) = \mathbf{A}^T Cov(\mathbf{X})\mathbf{A} = \mathbf{A}^T\mathbf{\Sigma}\mathbf{A} = \mathbf{A}^T\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T\mathbf{A} = \mathbf{\Lambda}.$$

A.1.8 Cluster analysis

K-means clustering

Cluster assignment of the feature vectors $\boldsymbol{x}_j$, $j = 1, \ldots, p$, is achieved by an iterative approach (Hastie *et al.*, 2009):

(1) Start with random choice of K cluster centers $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$.

(2) Determine the distance $d(\boldsymbol{x}_j, \boldsymbol{\mu}_k)$ between each feature $\boldsymbol{x}_j$ and each cluster center $\boldsymbol{\mu}_k$ (see Section 3.4.1 for distance definitions).

(3) Assign each feature to the nearest cluster with regard to the smallest distance $d(\boldsymbol{x}_j, \boldsymbol{\mu}_k)$:

$$C(j) = \underset{1 \leq k \leq K}{\arg\min}\, d(\boldsymbol{x}_j, \boldsymbol{\mu}_k).$$

(4) Compute the new cluster means $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{r, C(r)=k} \boldsymbol{x}_r$ where $|C_k|$ represents the number of features in cluster k.

(5) Iterate steps (2) to (4) until convergence.

To choose the optimal number of clusters K, Genolini *et al.* (2013) recommend to run the algorithm for different K, and then to select the optimal number of clusters according to

some quality criterion. Different quality criteria have been proposed, all of which put the variability, or distance, between clusters in relation to that within clusters. They include the *Calinski & Harabasz* criterion:

$$\frac{tr(B)(p-K)}{tr(W)(k-1)},$$

where $tr(B)$ and $tr(W)$ represent the traces of the between- and within-cluster variance matrices (see Calinski and Harabasz (1974) for details). Furthermore, the *Ray & Turi* criterion has been proposed as the ratio of the average within-cluster distances $DW(k)$ and the minimum between-cluster distance $DB(k,l)$:

$$\frac{\frac{1}{K}\sum_{k=1}^{K} DW(k)}{\min_{k,l\in 1,\ldots,K} DB(k,l)},$$

where $DW(k) = \frac{1}{|C_k|}\sum_{r,C(r)=k} d\left(\boldsymbol{x}_{r\cdot}, \boldsymbol{\mu}_k\right)$ and $DB(k,l) = d\left(\boldsymbol{\mu}_k, \boldsymbol{\mu}_l\right)$ (Ray and Turi, 1999). Finally, Davies and Bouldin (1979) proposed to use the average proximity between any two clusters, i.e.

$$\frac{1}{K(K-1)/2}\sum_{k,l\in 1,\ldots,K}\frac{DW(k)+DW(l)}{DB(k,l)}$$

as a quality criterion. Whereas the Calinski & Harabasz criterion is to be maximized, the latter two criteria are to be minimized.

Hierarchical clustering

Hierarchical clustering (in the agglomerative mode) is achieved by the following recursive procedure (Hastie *et al.*, 2009):

(1) Assign each feature to one cluster.

(2) Determine the distance matrix $D = (d_{jl})$ of the features, where the entries resemble the pairwise dissimilarity of the features $d_{jl} = d(\boldsymbol{x}_j, \boldsymbol{x}_l)$, $j, l = 1, \ldots, p$.

(3) Define distance $D(C_r, C_s)$ between two clusters C_r and C_s. Typical definitions include

$$\begin{aligned} D(C_r, C_s) &= \max_{\boldsymbol{x}_j\in C_r, \boldsymbol{x}_l\in C_s} d(\boldsymbol{x}_j, \boldsymbol{x}_l) \ (\textit{Complete linkage}) \\ D(C_r, C_s) &= \frac{1}{|C_r||C_s|}\sum_{\boldsymbol{x}_j\in C_r}\sum_{\boldsymbol{x}_l\in C_s} d(\boldsymbol{x}_j, \boldsymbol{x}_l) \ (\textit{Average linkage}). \end{aligned}$$

Complete linkage defines cluster distance as the distance of the most distant pair of features within the clusters, whereas average linkage defines cluster distance as

the average dissimilarity observed between features of both clusters. According to Hastie *et al.* (2009), average linkage might have more desirable statistical properties.

(4) Fuse the two clusters with the smallest distance.

(5) Compute the new cluster distance matrix.

(6) Iterate steps (4) and (5) until all features are element of one cluster.

Hierarchical clustering solutions are typically visualized as trees with the leaves representing the features, branches the clusters, and merging heights the distance of the merged clusters.

A.2 Appendix Tables

Table A.1: Processing and analysis pipelines for the EWAS cohorts. Included are four discovery studies and nine replication studies included in the epigenome-wide association study (EWAS) in Section 4.2.

	Dicovery cohorts							**Replication cohorts**					
	EpiMigrant	KORA F4	EPICOR	KORA F3	Rotterdam Study 2	LifeLines Deep	Rotterdam Study	ALSPAC	Leiden Longevity	EpiMigrant Replication	TwinsUK	EGCUT Asthma	EGCUT CTG
IDAT Extraction	minfi	minfi	methylumi	minfi	Methylumi	Methylumi	Normalization script (Marc Jan Bonder)	minfi	Methylumi	minfi	Beadstudio	minfi	minfi
Background Correction	minfi	minfi	methylumi	minfi	separate colors	separate colors	separate colors	none	separate colors	minfi	none	none	none
DetP cutoff	0.01	0.01	0.01	0.01	N/A	N/A	0.01	0.01	N/A	0.01	0.05	0.01	0.01
Sample CR threshold	none	none	none	none	N/A	N/A	95	95%	N/A	none	N/A	95%	95%
Marker CR threshold	95%	95%	80%	95%	N/A	N/A	100%	90%	N/A	95%	N/A	90%	90%
Nbeads filter	3	3	3	3	N/A	N/A	3	N/A	N/A	3	3	N/A	N/A
Normalization	QN, 6 categories	QN, 6 categories	QN, 6 categories	QN, 6 categories	DASEN	DASEN	SWAN	QN, 6 categories	DASEN	QN, 6 categories	QN	SWAN	SWAN
All covariates used in Model (technical and biological)	Discovery covariates*	Discovery covariates*	Discovery covariates*	Discovery covariates*	Age, Sex	Age, Sex, Smoking	Age, Sex, Array number, Position on the Array	Age, Batch (BCD plate), measured WBC	Age, Sex, Smoking	Age, Sex, Physical Activity, Alcohol intake, Smoking, imputed WBC subsets, 20 control probe PCs	Age, Smoking, Bisulfite-converted-DNA concentration, Array number, Position on the array, family and zygosity	Age, Sex, Physical Activity, Alcohol intake, Smoking	Age, Sex, Physical Activity, Alcohol intake
Genomic Control Inflation Factor (Discovery ONLY)	1.286	1.192	1.063	0.982									

* Discovery covariates: Age, Sex, Physical activity, Alcohol intake, Smoking, Estimated WBC proportions, 20 control probe PCs.

Table A.2: Subject characteristics of the EWAS discovery cohorts. Included are the four discovery studies of the epigenome-wide association study (EWAS) in Section 4.2.

	EpiMigrant	**KORA F4**	**EPICOR**	**KORA F3**
N	2680	1709	514	484
Country	UK	Germany	Italy	Germany
Ethnicity	South Asian	European	European	European
Study Design	Prospective T2D Case/Control	Population based cohort	Myocardial Infarction Case/Control	Population based cohort
Age (yrs)	51.0 (10.1)	61.0 (8.9)	51.9 (7.5)	53.2 (9.6)
Sex (M)	67.7%	48.9%	65.8%	52.1%
Fasting glucose (mmol/L)	5.2 (0.6)	5.6 (1.1)	5.6 (1.5)*	5.8 (1.4)*
HbA1c (%)	5.5 (0.5)	5.6 (0.6)	-	5.3 (0.5)
Fasting insulin (IU/L)	12.6 (10.2)	8.9 (24.9)	10.7 (8.5)	—
% Fasting	100.0%	100.0%	62.8%	9.5%
Weight (kg)	76.2 (13.8)	79.2 (15.3)	73.9 (12.3)	78.0 (15.1)
Height (cm)	166.1 (9.2)	167.8 (9.2)	166.0 (9.0)	169.1 (9.3)
Body mass index (kg/m2)	27.6 (4.4)	28.1 (4.7)	26.8 (3.8)	27.2 (4.6)
Waist circumference (cm)	97.3 (11.2)	—	91.2 (11.4)	—
Waist-hip ratio	0.95 (0.07)	0.89 (0.09)	0.90 (0.08)	0.89 (0.09)
Alcohol (g/d)	5.6 (12.5)	15.6 (20.4)	18.2 (19.6)	16.1 (19.7)
LDL cholesterol (mmol/L)	3.4 (0.9)	3.6 (0.9)	3.9 (1.0)	3.4 (0.9)
HDL cholesterol (mmol/L)	1.3 (0.3)	1.5 (0.4)	1.5 (0.4)	1.5 (0.5)
Total cholesterol (mmol/L)	5.4 (1.0)	5.7 (1.0)	6.1 (1.2)	5.7 (1.0)
Triglycerides (mmol/L)	1.7 (1.1)	1.5 (1.1)	1.7 (1.1)	1.9 (1.4)
C-reactive protein (mg/L)	4.2 (7.2)	2.5 (5.1)	2.1 (2.7)	3.5 (4.3)
Systolic BP (mmHg)	131.6 (18.9)	124.8 (18.7)	138.0 (18.9)	128.7 (18.2)
Diastolic BP (mmHg)	81.8 (10.7)	76.1 (10.0)	85.4 (9.5)	82.8 (10.6)
HT**	33.5%	48.4%	72.0%	45.0%
Treated HT	24.4%	37.3%	18.0%	22.7%
CHD***	7.8%	—	31.0%	—
T2D****	0.0%	9.0%	0.0%	6.0%
Physically active	28.7%	57.4%	79.0%	49.8%
Smoking				
Never smoked	82.7%	43.8%	34.5%	50.2%
Ex-smoker	8.5%	42.8%	30.8%	0.0%
Current smoker	8.8%	14.5%	34.9%	49.8%

*both fasting/non-fasting subjects; **HT (Hypertension): SBP $\geq$ 140mmHg, or DBP $\geq$ 90mmHg, or who were taking anti-hypertensive or blood pressure-lowering medication for any reason (the indication for each medication was typically not recorded); ***CHD: Revascularisation by PCI or CABG; Angiographically severe coronary disease; Documented acute coronary syndrome (ACS; symptoms, ECG change and/or biochemistry); ****T2D (Type 2 Diabetes): physician diagnosis or HbA1c>6.5%.

Table A.3: Subject characteristics of the EWAS replication cohorts. Included are the nine replication studies of the epigenome-wide association study (EWAS) in Section 4.2.

	Rotterdam Study 2	**LifeLines Deep**	**Rotterdam Study**	**ALSPAC**	**Leiden Longevity**	**EpiMigrant Replication**	**TwinsUK**	**EGCUT Asthma**	**EGCUT CTG**
N	762	752	731	701	665	656		173	96
Country	The Netherlands	The Netherlands	The Netherlands	UK	The Netherlands	UK	UK	Estonia	Estonia
Ethnicity	European	European	European	European	European	South Asian	European	European	European
Study Design	Population based cohort	Population based cohort	Population based cohort	Prospective birth cohort	Population based cohort	Prospective T2D Case/Control	Population/Twins	Asthma Case/Control	young vs old
Age (yrs)	67.7 (5.9)	45.5 (13.3)	59.9 (8.2)	47.8 (4.3)	58.9 (6.6)	53.8 (10.2)	54.6 (8.9)	26.2 (7.0)	53.0 (23.7)
Sex (M)	42.6%	42.3%	45.9%	0.0%	47.3%	42.8%	0.0%	35.8%	50.0%
Fasting glucose (mmol/L)	—	5.5(0.4)*	5.6 (1.18)	5.3 (1.1)	—	5.0 (0.5)	5.1 (0.9)*	—	5.2 (0.9)
HbA1c (%)	—	4.9(0.6)	N/A		—	5.6 (0.4)	-	—	—
Fasting insulin (IU/L)	—	—	14.2 (10.8)	5.7 mu/L (5.4)	—	—	8.4 (7.4)	—	8.9 (6.4)
% Fasting	—	99.6%	100.0%	100.0%	—	100.0%	97.6%	—	—
Weight (kg)	79.8 (14.1)	77.9 (14.1)	80.7 (16.2)	—	76.1 (13.1)	72.5 (13.5)		68.7 (13.2)	77.8 (17.1)
Height (cm)	169.3 (9.2)	175.2 (8.9)	170.7 (10.0)	—	172.6 (8.8)	163.6 (9.1)	26.9 (5.0)	173.0 (10.2)	170.5 (10.8)
Body mass index (kg/m2)	27.8 (4.2)	25.4 (4.2)	27.6 (4.6)	26.6 (5.3)	25.46 (3.5)	27.0 (4.4)	79.7 (10.6)	22.8 (3.0)	26.7 (5.1)
Waist circumference (cm)	—	96.3 (10.1)	93.7 (12.9)	84.6 (12.5)	—	94.5 (11.2)	0.80 (0.10)	78.2 (10.0)	91.0 (15.2)
Waist-hip ratio	—	0.92 (0.07)	0.87 (0.08)	0.81 (0.07)	—	0.93 (0.09)	0.79 (0.06)	0.80 (0.08)	0.87 (0.10)
Alcohol (g/d)	—	—	18.2 (11.3)	*****	—	3.0 (8.6)	-	2.3 (4.7)	3.8 (9.6)
LDL cholesterol (mmol/L)	—	3.1 (0.9)	3.5 (0.98)	2.9 (0.8)	—	3.1 (0.9)	3.4 (1.0)	—	3.3 (1.0)
HDL cholesterol (mmol/L)	—	1.5 (0.4)	1.4 (0.41)	1.5 (0.4)	—	1.5 (0.4)	1.5 (0.4)	—	1.6 (0.4)
Total cholesterol (mmol/L)	—	5.1 (1.0)	5.6 (1.07)	4.9 (0.8)	—	5.2 (0.9)	5.6 (1.2)	—	5.7 (1.1)
Triglycerides (mmol/L)	—	1.1 (0.9)	1.5 (0.87)	1.0 (0.5)	—	1.4 (0.7)	1.3 (0.8)	—	1.4 (1.0)
C-reactive protein (mg/L)	—	—	2.7 (4.7)	2.1 (4.4)	—	N/A	-	—	1.9 (2.1)
Systolic BP (mmHg)	—	119.5 (13.5)	134.4 (19.8)	117.3 (11.5)	—	127.1 (18.4)	124.8 (16.4)	116.7 (12.0)	129.5 (19.0)
Diastolic BP (mmHg)	—	71.1 (9.2)	82.8 (11.3)	72.2 (10.3)	—	76.9 (10.6)	77 (10.1)	73.2 (9.4)	79.9 (10.3)
HT**	—	8.4%	53.0%	6.8%	—	26.2%	-	10.4%	50.5%
Treated HT	—	—	26.0%	—	—	—	-	1.2%	41.7%
CHD***	—	—	5.9%	—	—	0.0%	0.0%	0.0%	13.5%
T2D****	—	—	8.5%	—	—	0.0%	0.0%	0.0%	4.2%
Physically active	—	—	—	N/A	—	41.5%	N/A	74.6%	63.5%
Smoking									
Never smoked	—	47.3%	29.0%	57.7%	31.6%	90.6%	57.4%	100.0%	62.5%
Ex-smoker	—	34.3%	44.0%	34.3%	55.5%	4.5%	33.1%	0.0%	24.0%
Current smoker	—	18.4%	27.0%	8.0%	12.6%	4.9%	9.5%	0.0%	13.5%

*both fasting/non-fasting subjects; **HT (Hypertension): SBP $\geq$ 140mmHg, or DBP $\geq$ 90mmHg, or who were taking anti-hypertensive or blood pressure-lowering medication for any reason (the indication for each medication was typically not recorded); ***CHD: Revascularisation by PCI or CABG; Angiographically severe coronary disease; Documented acute coronary syndrome (ACS; symptoms, ECG change and/or biochemistry); ****T2D (Type 2 Diabetes): physician diagnosis or HbA1c>6.5%; *****23.6% of women drank alcohol 4+ times a week; 6.5% non-drinkers.

Table A.4: Genotyping and inputation details for the EWAS cohorts. Genotyping for the LOLIPOP/EpiMigrant data was performed on three different platforms.

	LOLIPOP/EpiMigrant			**KORA**	
	LOLIPOP IA610	LOLIPOP OmniX	LOLIPOP IA317	KORA F4	KORA F3
Sample Size	916	596	329	1,645	475
Call rate threshold (samples)	95%	98%	95%	97%	97%
MAF threshold	1%	1%	1%	1%	1%
Hardy-Weinberg Equilibrium threshold (P-HWE)	1.00E-06	1.00E-06	1.00E-06	5.00E-06	5.00E-06
Call rate threshold (SNPs)	98%	98%	98%	98%	98%
Info-score threshold	0.5	0.5	0.5	0.5	0.5
Platform	Illumina 610K Beadchip	Illumina Omni Express	Illumina 317K Beadchip	Affymetrix Axiom	Illumina Omni Express
Genotype calling software	Illumina Genome Studio	Illumina Genome Studio (Gencall module) and zCall	Illumina Genome Studio	Affymetrix Software	Illumina Genome Studio
Imputation panel	1000g phase1	1000g phase1	1000g phase1	1000g phase1	1000g phase1
Imputation software	IMPUTE v2.3.0	IMPUTE v2.3.0	IMPUTE v2.3.0	IMPUTE v2.3.0	IMPUTE v2.3.0
cohort-specific adjustments	PC1-5 (genetic), Coronary Heart Disease	PC1-5 (genetic)	PC1-5 (genetic)	none	none
other criteria	removed samples with gender discordance, high or low heterozygosity, cryptic relatedness, and population outliers	removed samples with gender discordance, high or low heterozygosity, cryptic relatedness, and population outliers	removed samples with gender discordance, high or low heterozygosity, cryptic relatedness, and population outliers	removed samples with gender discordance, high or low heterozygosity, cryptic relatedness, and population outliers	removed samples with gender discordance, high or low heterozygosity, cryptic relatedness, and population outliers

Table A.5: Annotation of Metabolon [M] and NMR [N] metabolites to super- and sub-pathways. According to the Metabolon annotation, based on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

Sub-pathway	Metabolite name
1 Lipid	
1 VLDL	L_VLDL_C [N], L_VLDL_CE [N], L_VLDL_FC [N], L_VLDL_L [N], L_VLDL_P [N], L_VLDL_PL [N], L_VLDL_TG [N], M_VLDL_C [N], M_VLDL_CE [N], M_VLDL_FC [N], M_VLDL_L [N], M_VLDL_P [N], M_VLDL_PL [N], M_VLDL_TG [N], S_VLDL_C [N], S_VLDL_FC [N], S_VLDL_L [N], S_VLDL_P [N], S_VLDL_PL [N], S_VLDL_TG [N], VLDL_D [N], VLDL_TG [N], XL_VLDL_L [N], XL_VLDL_P [N], XL_VLDL_PL [N], XL_VLDL_TG [N], XS_VLDL_L [N], XS_VLDL_P [N], XS_VLDL_PL [N], XS_VLDL_TG [N], XXL_VLDL_L [N], XXL_VLDL_P [N], XXL_VLDL_PL [N], XXL_VLDL_TG [N]
2 HDL	HDL_C [N], HDL_D [N], HDL2_C [N], HDL3_C [N], L_HDL_C [N], L_HDL_CE [N], L_HDL_FC [N], L_HDL_L [N], L_HDL_P [N], L_HDL_PL [N], M_HDL_C [N], M_HDL_CE [N], M_HDL_FC [N], M_HDL_L [N], M_HDL_P [N], M_HDL_PL [N], S_HDL_L [N], S_HDL_P [N], S_HDL_TG [N], XL_HDL_C [N], XL_HDL_CE [N], XL_HDL_FC [N], XL_HDL_L [N], XL_HDL_P [N], XL_HDL_PL [N], XL_HDL_TG [N]
3 Lysolipid	1-stearoylglycerophosphoinositol [M], 1-linoleoylglycerophosphoethanolamine* [M], 1-arachidonoylglycerophosphocholine* [M], 1-palmitoleoylglycerophosphocholine* [M], 1-eicosatrienoylglycerophosphocholine* [M], 1-docosahexaenoylglycerophosphocholine* [M], 1-eicosadienoylglycerophosphocholine* [M], 1-palmitoylglycerophosphocholine [M], 1-heptadecanoylglycerophosphocholine [M], 1-oleoylglycerophosphocholine [M], 1-stearoylglycerophosphocholine [M], 1-arachidonoylglycerophosphoinositol* [M], 1-stearoylglycerophosphoethanolamine [M], 1-linoleoylglycerophosphocholine [M], 1-arachidonoylglycerophosphoethanolamine* [M], 2-palmitoylglycerophosphocholine* [M], 2-oleoylglycerophosphocholine* [M], 2-stearoylglycerophosphocholine* [M], 2-linoleoylglycerophosphocholine* [M], 1-palmitoylglycerophosphoinositol* [M], 1-myristoylglycerophosphocholine [M], 1-oleoylglycerophosphoethanolamine [M], 1-palmitoylglycerophosphoethanolamine [M], 2-linoleoylglycerophosphoethanolamine* [M]
4 Long chain fatty acid	arachidonate (20:4n6) [M], margarate (17:0) [M], palmitate (16:0) [M], nonadecanoate (19:0) [M], stearate (18:0) [M], oleate (18:1n9) [M], pentadecanoate (15:0) [M], myristate (14:0) [M], dihomo-linoleate (20:2n6) [M], myristoleate (14:1n5) [M], adrenate (22:4n6) [M], palmitoleate (16:1n7) [M], eicosenoate (20:1n9 or 11) [M], 5,8-tetradecadienoate [M], stearidonate (18:4n3) [M], 10-heptadecenoate (17:1n7) [M], 10-nonadecenoate (19:1n9) [M]
5 LDL	L_LDL_C [N], L_LDL_CE [N], L_LDL_FC [N], L_LDL_L [N], L_LDL_P [N], L_LDL_PL [N], LDL_C [N], LDL_D [N], M_LDL_C [N], M_LDL_CE [N], M_LDL_L [N], M_LDL_P [N], M_LDL_PL [N], S_LDL_C [N], S_LDL_L [N], S_LDL_P [N]
6 Sterol/Steroid	Est_C [N], Free_C [N], Serum_C [M], cortisol [M], cortisone [M], androsterone sulfate [M], dehydroisoandrosterone sulfate (DHEA-S) [M], 5alpha-androstan-3beta,17beta-diol disulfate [M], epiandrosterone sulfate [M], 7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca) [M], 5alpha-androstan-3beta,17beta-diol disulfate (2) [M], 4-androsten-3beta,17beta-diol disulfate 1* (2) [M], 4-androsten-3beta,17beta-diol disulfate 2* (2) [M], Serum_C [N]
7 Carnitine metabolism	carnitine [M], palmitoylcarnitine [M], acetylcarnitine [M], hexanoylcarnitine [M], 3-dehydrocarnitine* [M], octanoylcarnitine [M], decanoylcarnitine [M], stearoylcarnitine [M], laurylcarnitine [M], oleoylcarnitine [M], nonanoylcarnitine* [M], 2-tetradecenoylcarnitine [M], cis-4-decenoylcarnitine (2) [M]
8 Fatty acid, summary measure	Bis_DB_ratio [N], Bis_FA_ratio [N], CH2_DB_ratio [N], CH2_in_FA [N], DB_in_FA [N], FALen [N], FAw3 [N], FAw6 [N], FAw79S [N], MUFA [N], otPUFA [N], Tot_FA [N]
9 Bile acid metabolism	deoxycholate [M], ursodeoxycholate [M], taurodeoxycholate [M], glycocholate [M], taurochenodeoxycholate [M], cholate [M], hyodeoxycholate [M], glycochenodeoxycholate [M], taurolithocholate 3-sulfate [M]
10 Essential fatty acid	docosahexaenoate (DHA, 22:6n3) [N], linoleate (18:2n6) [N], linoleate (18:2n6) [M], eicosapentaenoate (EPA, 20:5n3) [M], docosahexaenoate (DHA, 22:6n3) [M], docosapentaenoate (n3 DPA, 22:5n3) [M], linolenate [alpha or gamma, (18:3n3 or 6)] [M], docosapentaenoic acid (n6-DPA) [M], dihomo-linolenate (20:3n3 or n6) [M]
11 Medium chain fatty acid	caprate (10:0) [M], heptanoate (7:0) [M], laurate (12:0) [M], pelargonate (9:0) [M], undecanoate (11:0) [M], caproate (6:0) [M], caprylate (8:0) [M], 10-undecenoate (11:1n1) [M], 5-dodecenoate (12:1n7) [M]
12 IDL	IDL_C [N], IDL_FC [N], IDL_L [N], IDL_P [N], IDL_PL [N], IDL_TG [N]
13 Fatty acid, dicarboxylate	3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF) [M], dodecanedioate [M], hexadecanedioate [M], octadecanedioate [M], 2-hydroxyglutarate (2) [M]
14 Glycerolipid metabolism	glycerol [N], glycerol [M], glycerol 3-phosphate (G3P) [M], choline [M], glycerophosphorylcholine (GPC) [M]
15 Monoacylglycerol	1-palmitoylglycerol (1-monopalmitin) [M], 1-oleoylglycerol (1-monoolein) [M], 1-stearoylglycerol (1-monostearin) [M], 1-linoleoylglycerol (1-monolinolein) [M]
16 Composition of mobile lipids	MobCH [N], MobCH2 [N], MobCH3 [N]
17 Glycerophospholipids	PC [N], TG_PG [N], TotPG [N]
18 Inositol metabolism	inositol 1-phosphate (I1P) [M], myo-inositol [M], scyllo-inositol [M]

Table A.5 continued.

Sub-pathway	Metabolite name
1 Lipid	
19 Ketone bodies	acetoacetate [N], 3-hydroxybutyrate (BHBA) [N], 3-hydroxybutyrate (BHBA) [M]
20 Fatty acid metabolism (also BCAA metabolism)	butyrylcarnitine [M], propionylcarnitine [M]
21 Fatty acid, amide	linoleamide (18:2n6) [M], oleamide [M]
22 Fatty acid, monohydroxy	2-hydroxystearate [M], 2-hydroxypalmitate [M]
23 Sphingolipid	palmitoyl sphingomyelin (2) [M], SM [N]
24 Eicosanoid	12-hydroxyeicosatetraenoate (12-HETE) [M]
25 Fatty acid amide	stearamide [M]
26 Fatty acid metabolism	isovalerate [M]
27 Fatty acid, branched	15-methylpalmitate (isobar with 2-methylpalmitate) (2) [M]
28 Fatty acid, ester	n-butyl oleate [M]
29 Short chain fatty acid	valerate [M]
30 Triacylglycerol	Serum_TG [N]
2 Amino acid	
1 Valine, leucine and isoleucine metabolism	isoleucine [N], leucine [N], leucine [M], isoleucine [M], valine [M], beta-hydroxyisovalerate [M], 3-methyl-2-oxovalerate [M], 3-methyl-2-oxobutyrate [M], 2-hydroxyisobutyrate [M], 4-methyl-2-oxopentanoate [M], levulinate (4-oxovalerate) [M], 3-hydroxy-2-ethylpropionate [M], isobutyrylcarnitine [M], alpha-hydroxyisovalerate [M], isovalerylcarnitine [M], tiglylcarnitine [M], 2-methylbutyroylcarnitine [M], hydroxyisovaleroylcarnitine [M], valine [N]
2 Phenylalanine & tyrosine metabolism	phenylalanine [M], tyrosine [M], 3-methoxytyrosine [M], 3-phenylpropionate (hydrocinnamate) [M], phenyllactate (PLA) [M], 3-(4-hydroxyphenyl)lactate [M], phenol sulfate [M], phenylacetylglutamine [M], p-cresol sulfate [M], phenylalanine [N], tyrosine [N]
3 Tryptophan metabolism	tryptophan [M], serotonin (5HT) [M], kynurenine [M], indolelactate [M], indoleacetate [M], 3-indoxyl sulfate [M], indolepropionate [M], C-glycosyltryptophan* [M], hydroxytryptophane* [M], tryptophan betaine (2) [M]
4 Urea cycle, arginine-, proline-, metabolism	ornithine [M], arginine [M], urea [M], proline [M], citrulline [M], N-acetylornithine [M], homocitrulline [M], trans-4-hydroxyproline [M], dimethylarginine (SDMA + ADMA) [M], urea [N]
5 Glycine, serine and threonine metabolism	glycine [N], threonine [M], betaine [M], N-acetylglycine [M], serine [M], glycine [M], N-acetylthreonine [M]
6 Alanine and aspartate metabolism	alanine [N], N-acetylalanine [M], aspartate [M], alanine [M], asparagine [M]
7 Cysteine, methionine, SAM, taurine metabolism	methionine [M], 2-hydroxybutyrate (AHB) [M], cysteine [M], cystine [M], methylcysteine [M]
8 Glutamate metabolism	glutamine [N], glutamine [M], glutamate [M], pyroglutamine* [M]
9 Creatine metabolism	creatinine [N], creatinine [M], creatine [M]
10 Lysine metabolism	lysine [M], pipecolate [M], glutaroylcarnitine [M]
11 Butanoate metabolism	3,4-dihydroxybutyrate* [M], 2-aminobutyrate [M]
12 Glutathione metabolism	5-oxoproline [M], cysteine-glutathione disulfide [M]
13 Histidine metabolism	histidine [N], histidine [M]
14 Guanidino and acetamido metabolism	4-acetamidobutanoate [M]
15 Polyamine metabolism	N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide [M]
3 Peptide	
1 Dipeptide	glycylvaline [M], aspartylphenylalanine [M], pro-hydroxy-pro [M], leucylalanine [M], alpha-glutamyltyrosine [M], phenylalanylserine [M], phenylalanylleucine [M], leucylleucine [M], phenylalanylphenylalanine (2) [M]
2 gamma-glutamyl	gamma-glutamylglutamine [M], gamma-glutamyltyrosine [M], gamma-glutamylleucine [M], gamma-glutamylvaline [M], gamma-glutamylmethionine* [M], gamma-glutamylthreonine* [M], gamma-glutamylphenylalanine [M], gamma-glutamylisoleucine* [M]
3 Fibrinogen cleavage peptide	DSGEGDFXAEGGGVR* [M], ADSGEGDFXAEGGGVR* [M], ADpSGEGDFXAEGGGVR* [M]
4 Apolipoprotein	ApoA1 [N], ApoB [N]
5 Polypeptide	HWESASXX* [M], bradykinin, des-arg(9) [M]
6 Glycoprotein	Gp [N]
7 Protein	Alb [N]

Table A.5 continued.

Sub-pathway	Metabolite name
4 Xenobiotics	
1 Xanthine metabolism	caffeine [M], paraxanthine [M], theobromine [M], theophylline [M], 3-methylxanthine [M], 1-methylxanthine [M], 7-methylxanthine [M], 1-methylurate [M], 1,7-dimethylurate [M]
2 Food component/Plant	quinate [M], homostachydrine* [M], piperine [M], stachydrine [M], thymol sulfate [M], vanillin [M], ergothioneine (2) [M]
3 Benzoate metabolism	hippurate [M], benzoate [M], catechol sulfate [M], 4-vinylphenol sulfate [M], 4-ethylphenylsulfate [M]
4 Chemical	glycerol 2-phosphate [M]
5 Sugar, sugar substitute, starch	erythritol [M]
5 Carbohydrate	
1 Glycolysis, gluconeogenesis, pyruvate metabolism	glucose [N], lactate [N], lactate [M], pyruvate [M], glycerate [M], glucose [M], 1,5-anhydroglucitol (1,5-AG) [M], pyruvate [N]
2 Fructose, mannose, galactose, starch, and sucrose metabolism	fructose [M], mannose [M], mannitol [M], erythrose [M]
3 Nucleotide sugars, pentose metabolism	arabinose [M], arabitol [M], threitol [M]
4 Aminosugars metabolism	erythronate* [M]
6 Cofactors and vitamins	
1 Hemoglobin and porphyrin metabolism	biliverdin [M], bilirubin (Z,Z) [M], bilirubin (E,E)* [M], oxidized bilirubin* [M], bilirubin (E,Z or Z,E)* [M]
2 Ascorbate and aldarate metabolism	ascorbate (Vitamin C) [M], threonate [M], O-methylascorbate* [M]
3 Tocopherol metabolism	alpha-tocopherol [M], gamma-tocopherol [M]
4 Hemoglobin and porphyrin	heme* [M]
5 Pantothenate and CoA metabolism	pantothenate [M]
6 Vitamin B6 metabolism	pyridoxate [M]
7 Nucleotide	
1 Purine metabolism, (hypo)xanthine/inosine containing	inosine [M], hypoxanthine [M], xanthine (2) [M]
2 Purine metabolism, guanine containing	guanosine [M], 7-methylguanine [M]
3 Pyrimidine metabolism, uracil containing	uridine [M], pseudouridine [M]
4 NAD metabolism	N1-methyl-3-pyridone-4-carboxamide [M]
5 Purine metabolism, adenine containing	N1-methyladenosine [M]
6 Purine metabolism, urate metabolism	urate [M]
8 Energy	
1 Krebs cycle	acetate [N], citrate [N], malate [M], citrate [M], alpha-ketoglutarate [M], succinylcarnitine [M]
2 Oxidative phosphorylation	phosphate [M], acetylphosphate [M]

Acknowledgements

I would like to thank a number of people who supported my work as a doctoral student at the Research Unit of Molecular Epidemiology of the Helmholtz Zentrum München.

First of all, I am deeply thankful to my doctorate supervisor Prof. Dr. Thomas Illig, for his great confidence in my work, for setting up the topic of the thesis, for invaluable scientific advice, and for continuous support and availability even after he took the position in Hannover. Moreover, I wish to express my sincere gratitute to my group leader Dr. Harald Grallert, for his continuous trust and encouragement, and for ensuring pleasant working conditions. Both their support made it possible for me to study Biostatistics during my time as a doctoral student, which provided a major contribution to my scientific and personal development.

Furthermore, I am thankful to Prof. Dr. Dr. H.-Erich Wichmann, formaly head of the Institute of Epidemiolgy, Prof. Dr. Annette Peters, the head of the Institute of Epidemiology II, and Dr. Christian Gieger, the new head of the Research Unit, for providing the institutional basis for my scientific work, and for being available and supportive. Christian Gieger also receives special thanks for having facilitated research cooperations, and for valuable scientific advice.

Through different projects and cooperations, I had the pleasure to become acquainted with many inspiring scientists both inside and outside the center. In particular, I would like to say special thanks to Melanie Waldenberger, Gabi Kastenmüller and Karsten Suhre for their continuous helpfulness and excellent scientific advice, to Helmut Laumen for his contagious enthusiasm on research and his confidence in my limited statistical experience in the beginning, to the group around John Chambers at the Imperial College London, including Benjamin Lehne, Alexander Drong and Marie Loh, for a very fruitful and inspiring collaboration, to Susanne Vogt, Barbara Thorand and Thomas Reinehr for straightforward and enjoyable cooperations, to Zhonghao Yu and Rui Wang-Sattler for intensive support and valuable lessons learned in the beginnings of my work, and Tommaso Panni and Lena Riess for some delightful hours on advanced statistics. I further thank Jan Krumsiek, Janina Ried, Hans-Jörg Baurecht, Jerzy Adamski, Gabriele Möller, Cornelia Prehn, Tao Xu and a lot of other people for great help in different aspects of my work.

All members of the MEDART and MECD working groups deserve big thanks for an

enjoyable working atmosphere, for their helpfulness and open ears. I also wish to thank all staff members involved in the planning and conduction of the studies, the data management and the laboratory measurements. Great thanks also to Anja Kretschmer, Carola Marzi and Sonja Zeilinger for their invaluable feedback on my doctoral thesis.

Finally, a very big thank you to my family for their unconditional support and understanding, and to my boyfriend for accompanying me through all ups and downs of this thesis with comprehensive support and encouragement.

Eidesstattliche Versicherung

Wahl, Simone

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema

Multi-omics of obesity and weight change in the post-genomic era

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München,

Ort, Datum

Unterschrift Doktorandin/Doktorand