INTEGRATED EPIGENOMICS AND METABOLOMICS ANALYSIS IN TWINS

Yet, Idil

Awarding institution:
King's College London

INTEGRATED EPIGENOMICS AND METABOLOMICS ANALYSIS IN TWINS

IDIL YET

KING'S COLLEGE LONDON

FACULTY OF LIFE SCIENCES & MEDICINE

DIVISION OF GENETICS & MOLECULAR MEDICINE

DEPARTMENT OF TWIN RESEARCH AND GENETIC EPIDEMIOLOGY

Submitted in particular fulfilment of the

requirements for the degree of

Doctor of Philosophy

2016

ABSTRACT

Epigenetics and metabolomics are rapidly growing areas of research, in part due to recent advances in technology that have allowed for a wide coverage of the human genome.

Metabolites are small compounds present in cell and body fluids, and are involved in biochemical processes of the cell. Quantitative trait loci associated with levels of individual metabolites (mQTLs) have been identified from numerous metabolome GWAS. Here, I analysed metabolite levels in twins with the aim of identifying genetic variants that influence metabolomic traits (mQTLs) using two different metabolomics platforms, and consequently compared the results to report stable metabolites on both technologies to ultimately enable combining metabolite profiles across these two platforms.

DNA methylation is a biochemical process that is vital for mammalian development. It is present throughout the genome and is the most extensively studied epigenetic mark. Previous studies have explored the heritability of DNA methylation and have identified methylation QTLs (meQTL). Here, I identified meQTLs with the goal of assesing the evidence of genetic effects influence not only DNA methylation levels, but also variability by using MZ-twin discordance as a measure of variance.

Epigenetic mechanisms and metabolomic profiles have both been shown to play a role in gene expression and susceptibility for complex human disease. Here, I analysed the association between type 2 diabetes and metabolomic and epigenetic datasets and combined the data to identify connections between these levels of biological data at genetic variants linked to type 2 diabetes to gain more insight into the disease susceptibility and progression.

Overall, the results confirmed previous findings of strong genetic influences on metabolites and extend current knowledge about genetic effects underlying several biochemical pathways. Additionally, the results also showed genetic influences on DNA methylation, and give insights into mechanisms by which genetic impacts epigenetic processes. Lastly, the findings show that specific genetic susceptibility variants for type 2 diabetes can also impact epigenetic and metabolomics profiles, and can help improve our understanding of the disease etiology.

to Mom & Dad

“Sevgili Anneme ve Babama”

ACKNOWLEDGEMENT

I would like to thank *my Supervisor Dr. Jordana Bell* for her support and guidance throughout my study, and grand influence on me for the last four years. She was always great at sharing knowledge, experiences and helping me to put things into perspective. I am also thankful to her spending a lot of time to improve my writing.

In addition, I would like to thank *Prof. Tim Spector* for his assistance throughout my study and giving me an opportunity to work with TwinsUK cohort.

I am also indebted to my friends for their help and advice during my project, most notably Pei-Chien; Your friendship and your support in all things in academic and personal mean more to me than I can ever tell you. I would like to thank to my other friends; Abhishek, Juan, Leonie, Cristina, and Lisa. They were always ready to support me. In addition, I am thankful to the staff of *Department of Twin Research & Genetic Epidemiology* for their help.

To my family: your unwavering belief that I would succeed has been invaluable. "Sizi seviyorum."

Finally, the biggest 'thank you' must go to my husband Barbaros, who supports me in too many ways to list here. You're my inspiration. After supporting each other through PhDs, I am sure everything will seem easy.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

ABBREVATIONS

AIC: Aikake Information Criterion

BCAAs: Branched-Chain Amino Acids

BIC: Bayes Information Criterion

BMI: Body Mass Index

BMIQ: Beta Mixture Quantile Dilation

BN: Bayes Network

BP: Base Pair

CGI: CpG Islands

CpG: Cytosine-Phosphate-Guanine

DAG: Directed Acyclic Graph

ddNTP: dideoxynucleotides

DEXA: Dual-Energy X-Ray Absorptiometry

DMP: Differentially Methylated Positions

DMR: Differently Methylated Regions

DNA: Deoxyribonucleic Acid

DZ: Dizygotic

eQTL: expression QTL

EGCUT: Estonian Genome Center of the Tartu University

ERF: Erasmus Rucphen Family

ESI: Electro Spray Ionization

EWAS: Epigenome-Wide Association Studies

FDR: False Discovery Rate

FIA: Flow Injection Analysis

GC: Gas Chromatography

GWAS: Genome-Wide Association Studies

HDL: High-Density Lipoprotein

KORA: Cooperative Health Research in the Region of Augsburg

IFG: Impaired Fasting Glucose

INDEP: Independent

LCL: Lymphoblastoid Cell Lines

LD: Linkage Disequilibrium

LLS: The Leukemia & Lymphoma Society

LMER: Linear Mixed Effects Regression

MAF: Minor Allele Frequencies

MB: Metabolite

MEDIP: Methylated DNA Immunoprecipitation

meQTL: methylation QTL

mGWAS: Metabolome Genome-Wide Association Studies

MHC: Major Histocompatibility Complex

mQTL: metabolite QTL

MS: Mass Spectrometry

MT: Methylation

MWAS: Metabolome -Wide Association Studies

MZ: Monozygotic

NMR: Nuclear Magnetic Resonance

NTR: Netherlands Twin Register

PC: Principal Component

PCA: Principal Component Analysis

QIMR: Queensland Institute of Medical Research

QTL: Quantitative Trait Loci

RRBS: Reduced Representation Bisulfite sequencing

SNP: Single Nucleotide Polymorphisms

SWAN: Subset-quantile Within Array Normalization method

TwinsUK: Twins UK Registry

T1D: Type 1 Diabetes

T2D: Type 2 Diabetes

UPLC: Ultra Performance Liquid Chromatography

var meQTL: variance meQTL

WGBS: Whole Genome Bisulfite Sequencing

PUBLICATIONS RELATED TO THIS THESIS

- **Idil Yet**, Cristina Menni, So-Youn Shin, Massimo Mangino, Nicole Soranzo, Jerzy Adamski, Karsten Suhre, Tim D Spector, Gabi Kastenmüller, and Jordana T Bell (2016), Genetic influences on metabolite levels: a comparison across metabolomic platforms, PloS one, 11 (4), e0153672
- **Idil Yet**, Pei-Chien Tsai, Juan E Castillo-Fernandez, Elena Carnero-Montoro, Jordana T Bell (2016), Genetic and environmental impacts on DNA methylation levels in twins, *Epigenomics*, 8 (1), 105-117
- Harmen H.M. Draisma, Rene Pool, Michael Kobl, Rick Jansen, Ann-Kristin Petersen, Anika A.M. Vaarhorst, **Idil Yet,** Toomas Haller, Ayse Demirkan, Tonu Esko, Gu Zhu, Stefan Bohringer, Marian Beekman, Jan Bert van Klinken, Werner Ro¨misch-Margll, Cornelia Prehn, Jerzy Adamski, Anton J.M. de Craen, Elisabeth M. van Leeuwen, Najaf Amin, Harish Dharuri, Harm-Jan Westra, Lude Franke, Eco J.C. de Geus, Jouke Jan Hottenga, Gonneke Willemsen, Anjali K. Henders, Grant W. Montgomery, Dale R. Nyholt, John B. Whitfield, Brenda W. Penninx, Tim D. Spector, Andres Metspalu, P. Eline Slagboom, Ko Willems van Dijk, Peter A.C. Hoen, Konstantin Strauch, Nicholas G. Martin, Gert-Jan B. van Ommen, Thomas Illig, Jordana T. Bell, Massimo Mangino, Karsten Suhre, Mark I. McCarthy, Christian Gieger, Aaron Isaacs, Cornelia M. van Duijn & Dorret I. Boomsma (2015), Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels, *Nature Communications,* 6, 7208.
- Wei Yuan, Yudong Xia, Christopher G. Bell, **Idil Yet**, Teresa Ferreira, Kirsten J. Ward, Fei Gao, A. Katrina Loomis, Craig L. Hyde, Honglong Wu, Hanlin Lu, Yuan Liu, Kerrin S. Small, Ana Vinuela, Andrew P. Morris, Maria Berdasco, Manel Esteller, M. Julia Brosnan, Panos Deloukas, Mark I. McCarthy, Sally L. John,

Jordana T. Bell, Jun Wang & Tim D. Spector (2014), An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins, *Nature Communications,* 5, 5719.

- So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, **Idil Erte**, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, The Multiple Tissue Human Expression Resource (MuTHER) Consortium, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector & Nicole Soranzo (2014), An atlas of genetic influences on human blood metabolites, *Nature Genetics,* 46 (6), 543-50
- Cristina Menni, Eric Fauman, **Idil Erte**, John R.B. Perry, Gabi Kastenmüller, So-Youn Shin, Ann-Kristin Petersen, Craig Hyde, Maria Psatha, Kirsten J. Ward, Wei Yuan, Mike Milburn, Colin N.A. Palmer, Timothy M. Frayling, Jeff Trimmer, Jordana T. Bell, Christian Gieger, Rob P. Mohney, Mary Julia Brosnan, Karsten Suhre, Nicole Soranzo, and Tim D. Spector (2013), Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach, *Diabetes,* 62 (12), 4270-6.

CHAPTER 1

Introduction

Recent studies have explored the molecular links between metabolism and epigenetic modifications implicated in a variety of diseases (Lu and Thompson 2012). Metabolism is one of the major sources of methyl groups that are used to methylate DNA, a key epigenetic process that influences chromatin structure and gene expression, and ultimately normal development and diseases. Epigenetic mechanisms and metabolomic profiles have both been shown to play a role in gene expression and have been associated with several complex traits, including type 2 diabetes (T2D) (Katada *et al.* 2012). Here I describe both types of "-omic" profiles, that is, human metabolomics and epigenomics data, giving a brief overview of relevant literature and recent findings in each area, and lastly I summarise the aim of this thesis.

1.1 Exploring the genetic basis of complex phenotypes and disease

Complex diseases such as cardiovascular disease and T2D are an increasing global health concern. According to the World Health Organisation, cardiovascular diseases are the number one cause of death globally and almost 420 million people worldwide suffer from T2D. A better understanding of the causes of complex diseases in conjunction with an improvement of preventive medicine is one of the main aims of epidemiological studies of the disease itself as well as related risk factors. Because a genetic predisposition exists for most complex diseases, the identification of genes involved in the disease etiology has been essential. To date, genome-wide association studies (GWAS) have been one of the main approaches used to serve this purpose. In

order to gain further insight into genetic and biochemical mechanisms underlying a disease, part of this thesis expands the GWAS approach by considering metabolites and methylation as intermediate phenotypes between genes and disease.

1.2 Twin Studies and heritability

The classic twin study aims to separate the phenotypic variance into genetic and environmental components. One of the main aims of the twin design is to estimate how much of the phenotypic variance is due to genetic effects (heritability), and how much appears to be due to shared or unique environmental effects. Heritability of a trait within a population is the proportion of observed variance in the trait between individuals within a population that is due to genetic differences. Thus, the total variance of phenotype (Equation 1-1) can be calculated as sum of the variances of individual genetic and environmental effects, under the assumption that these variances can be additive and are due to independent causes (Strachan *et al.* 2011).

$$\boldsymbol{V_{Phenotype} = V_{Genetic} + V_{Environment}} \qquad \textbf{\textit{(Equation 1-1)}}$$

The genetic variation can be due to additive genetic variance (V_A) as well as non-additive genetic variance that can be related to interactions between alleles at the same locus (dominance, V_D) and/or at different loci (epistasis, V_I): $V_G = V_A + V_D + V_I$. Broad-sense heritability (H^2) takes into account the total genetic variation (V_G); Thus, H^2 can be calculated as **(**Equation 1-2);

$$\mathbf{H^2 = \frac{V_{Genetic}}{V_{Phenotype}}} \qquad \textbf{(\textit{Equation 1-2})}$$

The narrow-sense heritability heritability (h^2) takes into account only additive genetic variance (V_A) and can be calculated as (Equation 1-3);

$$\mathbf{h^2 = \frac{V_{Additive}}{V_{Phenotype}}} \qquad \textbf{(\textit{Equation 1-3})}$$

and can be estimated by comparing correlations (r) between monozygotic (MZ) and dizygotic (DZ) twins (Equation 1-4);

$$\mathbf{h^2 = 2(r_{MZ} - r_{DZ})} \qquad \textbf{(}\textit{Equation 1-4}\textbf{)}$$

The ACE model was proposed for calculating narrow-sense heritability (Neale *et al.* 1992). Three elements of the phenotypic variance (VP) are estimated in the classical twins study (Neale *et al.* 1992): the additive genetic component (A), common environment (C) and unique environment (E); which constitute the ACE components. In this framework it is also possible to study non-additive genetic effects, for example, evidence for dominant genetic effects (D) can be assessed in the ADE model.

1.3 Genome-wide association studies

Statistical tests for associations between a phenotype and genetic variants across the genome, typically single nucleotide polymorphisms (SNPs) is referred to as Genome-Wide Association Studies (GWAS). The idea underlying the GWAS approach is that a number of common SNPs are causal for a complex disease and also that these causal variants are in linkage disequilibrium (LD) with other common genetic variants, which can serve as tags of the signal. These tags can help identify causal variants, which may not even be profiled. The first GWAS was conducted in 2005 for investigating patients with age-related macular degeneration and reported two SNPs with significantly altered allele frequency compared to healthy controls (Klein *et al.* 2005). By 2016, a total of 2,423 GWAS for more than 5,000 different traits were published (Hindorff *et al.* 2009; Welter *et al.* 2014). The significant results of these GWAS and their location in the genome are displayed in Figure 1-1. Although GWAS are a very popular method to reveal novel risk variants, one drawback is the small effect size of SNPs detected to date (de Bakker *et al.* 2008).

Figure 1-1 Published GWAS by April 2016. The GWAS Catalog contains 2,423 studies and 16,617 unique SNP-trait associations ($P < 5x10^{-8}$). All traits are color-coded. Resource: The NHGRI-EBI GWAS Catalog (Welter et al. 2014).

For many traits and diseases of interest, larger sample sizes are needed to detect significant associations using the GWAS approach and this is typically achieved through meta-analysis where multiple analysts carry out the same analysis in separate cohorts and combine the results afterwards (Thompson *et al.* 2011; Zeggini and Ioannidis 2009). For example, a recent genome-wide association meta-analysis of waist and hip circumference-related traits in more than 200,000 individuals identified variants in 49 loci (33 of them novel) associated with waist to hip ratio adjusted for body mass index (BMI) and an additional 19 loci associated with related waist and hip circumference measures (Shungin *et al.* 2015).

Associations between genetic variations and various phenotypes have been reported in numerous studies, but these variants typically can only explain a limited fragment of the phenotypic diversity, leading many geneticists to raise the question of "where is the missing heritability?". The problem of the missing heritability is a widely discussed topic, for example as discussed in a recent review by Eichler. *et al.* (Eichler *et al.* 2010). Some suspected reasons for missing heritability are undetected rare mutations which are

not tagged well by common SNPs, common variants with a low penetrance, other genomic variations such as copy number variants, gene and gene-environment interactions as well as incorrect heritability estimates (Maher 2008). It has been suggested that some of the unexplained heritability might be explained by incomplete LD between the analysed SNPs and the causal variants (Yang *et al.* 2010). Moreover, larger samples, more precisely measured phenotypes, more densely genotyped SNPs as well as advanced statistical methods might help to find the missing heritability. Finally, epigenetic variation has also been proposed as an explanation for the large amount of missing heritability in complex traits (Eichler *et al.* 2010; Manolio *et al.* 2009). Despite these drawbacks GWAS have successfully identified many genetic risk variants to date for a number of complex traits and diseases.

1.4 Metabolomics

Metabolomics is the developing field of study for measuring compounds of a cell or body fluid. It is assessed that the human metabolome, which is defined as the complete set of small molecular weight molecules, covers more than 3,000 different metabolites of various biochemical classes such as sugars, amino acids, lipids or carnitines (Koal and Deigner 2010). Metabolites are the small molecular weight substances present in cells within tissues and body fluids and are involved in biochemical processes of the cell (Nicholson *et al.* 1999). Metabolomics analyses are predominantly performed on blood and urine samples, as these are easy to obtain. The measurement of metabolites reveals a view of the current state of cells. In addition to measuring levels of single metabolites, ratios between metabolite concentrations can also be used to inform on metabolic processes as well as in the search for biomarkers, for example, in a systematic screens for genetic deficiencies in newborns (Maier *et al.* 2005). Analysis of metabolites can reveal insight on functional variation in the cell and help to detect connections between different diseases (Holmes *et al.* 2008a).

Changes in the organism are magnified in the metabolome compared to the genome. Metabolomics is a promising tool in the search for biomarkers which help to detect environmental exposures, diseases, to improve the disease prognosis, to develop therapeutics or to evaluate drug toxicity (Nicholson and Lindon 2008). For example, metabolomics plays a key role in the field of cancer diagnostics, especially when early detection is difficult, such as for kidney cancer (Nagrath *et al.* 2011). The search for metabolomics biomarkers is also underway in many diseases, with moderate success so far (Barderas *et al.* 2011).

1.4.1 Metabolomics Platforms

There are two main strategies to measure metabolites, a non-targeted and a targeted approach. Whilst the non-targeted approach aims to measure all metabolites in a sample, the targeted approach focuses specifically on the quantification of selected metabolites. The most often used high-throughput methods to measure metabolites are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy (Malet-Martino and Holzgrabe 2011).

MS has to be coupled to separation techniques, such as gas chromatography (GC) or liquid chromatography (LC) and is more sensitive than NMR (Nicholson and Lindon (2008). When using GC/MS, the analyte has to be stable and sometimes requires a derivatisation step; transforming the chemical compound into a product. If a derivatisation is not possible or if the metabolites are not stable, LC/MS can be applied (Barderas *et al.* 2011). GC/MS can capture fatty acids, amino acids and sugars very well. In some cases, tandem MS (MS/MS) is applied which consists of multiple MS steps with a fragmentation step in between. The use of MS/MS facilitates the identification of the measured molecules. In NMR, the analyte does not require any treatment prior to analysis. On the other hand, MS is fast, accurate and cost- effective. Altogether, the metabolomics methods developed to date have different strengths and

weaknesses, and a single approach cannot capture the entire human metabolome - instead, a combination of different measurement techniques is essential to gain the most comprehensive insight into the metabolome.

1.4.2 Metabolome wide association studies

To date, GWAS have found associations between genotype variation and disease phenotypes, and as extension to this approach, the metabolome wide association study (MWAS) tests for systematic associations of metabolites with phenotypes and diseases (Bictash *et al.* 2010; Holmes *et al.* 2008b). Examples of MWAS include analyses of nutrition (Altmaier *et al.* 2011; Menni *et al.* 2013a), coffee consumption (Altmaier *et al.* 2009), type 2 diabetes (T2D) (Menni *et al.* 2013c), and aging and aging traits (Menni *et al.* 2013b; Yu *et al.* 2012). MWAS results can give insights into the biochemical mechanisms involved in disease susceptibility and progression, and can also be used as biomarkers of environmental exposures, phenotypes, and diseases, as previously discussed.

1.4.3 Genome-wide association studies with metabolomics

As metabolites are products of genetic as well as proteomic processes, metabolites are very closely linked to genetics in contrast to most of the other phenotypes. The examination of the genetic basis of metabolites can be conducted with metabolome GWAS (mGWAS). Numerous mGWAS have also been conducted to date (Chasman *et al.* 2009; Demirkan *et al.* 2012; Demirkan *et al.* 2015; Hicks *et al.* 2009; Illig *et al.* 2010; Kettunen *et al.* 2012; Krumsiek *et al.* 2012; Lemaitre *et al.* 2011; Nicholson *et al.* 2011; Raffler *et al.* 2013; Rhee *et al.* 2013; Ried *et al.* 2014; Rueedi *et al.* 2014; Suhre *et al.* 2011a; Suhre *et al.* 2011b; Tanaka *et al.* 2009). The results are discussed in greater detail in the introductory section to Chapter 3. There are several common themes that arise from these multiple mGWAS. The first mGWAS showed the power of considering not only single metabolites but also metabolite ratios in the analysis (Gieger *et al.*

2008). Whilst in some mGWAS all possible pair-wise metabolite ratios were analysed in a hypothesis-free approach, others focused on biologically relevant metabolite ratios. Targeted and non-targeted metabolomics are complementary in giving an improved picture of the human metabolome and mGWAS of targeted and non-targeted platforms given overlapping signals (Suhre *et al.* 2010). Moreover, as metabolomics technologies are constantly improving and are currently (in 2016) able to measure even more metabolites than previous panels, analyses of these metabolites will bring further insights into the human metabolism and disease causing mechanisms (Kastenmuller *et al.* 2015).

1.5 Epigenetics

The term "epigenetics" was first introduced by Waddington (1957) proposing a new field that combined developmental biology and genetics comprising all processes in unfolding of the genetic program for development. During the last decade, epigenetics was redefined as heritable cellular modifications that are not caused by changes in the DNA sequence (Holliday 1994; Richards 2006). This comprises a diverse range of mechanisms of which DNA methylation and histone modifications are studied the most at present. Epigenetic mechanisms can regulate gene expression and have impact on phenotypes, linking epigenetic mechanisms to development and disease susceptibility (Holliday and Pugh 1975).

1.5.1 DNA methylation

Compared to other epigenetic mechanism, DNA methylation has been the most widely studied to date due to the availability of various techniques. It is considered as one of the most stable epigenetic mechanism. Several factors are thought to influence DNA methylation profiles: genetic variation, other epigenetic modifications, stochastic changes, and environmental factors that arise during life (Lander *et al.* 2001). In mammals, DNA methylation occurs mainly at cytosine-phosphate-guanine (CpG)

dinucleotides throughout the genome (Cosgrove and Wolberger 2005). A large proportion of CpGs typically fall into short regions of roughly 1 kb with a high CpG frequency, referred to as CpG islands (CGIs). CGI shores are regions that have lower CpG density located at CGI borders (Lander *et al.* 2001; Venter *et al.* 2001). In particular during early stages of development, most of the human genes related to development were found to have a CGI in their promoter that tended to be unmethylated across different tissue types (Antequera and Bird 1993; Larsen *et al.* 1992). DNA methylation at CGI shores was observed to be up to 13 fold higher than levels at CGIs across different cells and tissues, and highly variable (Doi *et al.* 2009). CGI shores are enriched for functional signals, such as tissue-specificity in DNA methylation profiles (Doi *et al.* 2009; Irizarry *et al.* 2009), methylation changes during reprogramming (Doi *et al.* 2009), and gene expression changes linked to disease (Irizarry *et al.* 2009).

DNA methylation is an essential epigenetic mechanism that plays important roles during development, in the regulation of transcription, genomic imprinting, X-chromosome inactivation, and maintenance of chromosomal and genome stability (Bird 2002; Li *et al.* 1992; Reik 2007). DNA methylation also plays a regulatory role in gene expression, where methylation in the gene promoter is typically linked with absence or low levels of expression of the gene (Ball *et al.* 2009; Gutierrez-Arcelus *et al.* 2013; Holliday and Pugh 1975). However, DNA methylation can be both positively and negatively correlated with gene expression, depending on the position of the CpG site (Gutierrez-Arcelus *et al.* 2013) within the gene promoter and body.

1.5.2 Platforms for detecting DNA methylation

DNA methylation is one of the widely studied epigenetic mechanisms and numerous techniques have been developed for its detection. Presently, the most widely methods are based on bisulfite conversion followed by array hybridization or sequencing (Laird 2010). One such example is the Illumina Infinium HumanMethylation27k BeadChip

(Illumina 27k) (Bibikova *et al.* 2009; Steemers and Gunderson 2007), which assays DNA methylation levels at approximately 27,000 CpG sites in prometer-specific regions. A newer version of this array is the Illumina Infinium HumanMethylation450 BeadChip (Illumina 450k), which covers a wider range of approximately 450,000 CpG sites across the genome. Recently, Infinium MethylationEPIC BeadChip (Illumina EPIC) is launched as an updated version of the Illumina Infinium HumanMethylation450 BeadChip, featuring 850,000 CpGs in enhancer regions, gene bodies, promoters and CpG islands (Moran *et al.* 2016). Additionally, bisulfite conversion approaches followed by sequencing are an alternative to array-based methods. These include Whole Genome Bisulfite Sequencing (WGBS) (Cokus *et al.* 2008; Lister *et al.* 2008; Lister *et al.* 2009) and Reduced Representation Bisulfite sequencing (RRBS) (Meissner *et al.* 2008). WGBS is currently seen as the gold standard, however, it is relatively costly for large-scale experimental analysis and specific regions of the genome can be difficult to sequence with WGBS. Targeted enrichment bisulfite sequencing methods are also currently being developed where only selected genomic regions undergo bisulfite sequencing, and these typically span 2-5 million CpG sites. Another approach to detecting DNA methylation can be based on pull-down approaches, such as methylated DNA immuno-precipitation (MeDIP) (Weber *et al.* 2005; Weber *et al.* 2007). MeDIP-sequencing enables a genome-wide DNA methylome characterization and provides access to dense CpG regions of the genome and repetitive elements with potential regulatory effects (Ward *et al.* 2013). Sequencing-based methods eliminate some of the limitations of the array compositions on the other hand, selecting fragment-size, sequencing bias and performance can all impact coverage limitations and errors in measurement (Laird 2010).

1.5.3 DNA methylation heritability

There is evidence that genetics can impact DNA methylation profiles at a proportion of CpG-sites across the genome, despite developmental reprogramming of DNA methylation profiles. Reprogramming of the methylome occurs during the germ cell stage and pre-implantation during development (Reik *et al.* 2001; Reik and Walter 2001a). In the first stage, highly methylated germ cells lose much of their methylation memory in a first major wave of de-methylation, and in the fertilization stage, the germ cells undergo a second de-methylation phase when most of the methylation is erased and followed by *de novo* methylation (Reik *et al.* 2001; Reik and Walter 2001b). Primary reports in support of genetic effects on DNA methylation come from familial clustering of epigenetic variation reported at numerous loci (Bird 2002; Reik *et al.* 2001). One longitudinal study reported familial clustering of DNA methylation patterns over time in more than 200 individuals from two separate cohorts (Bjornsson *et al.* 2008). They observed that DNA methylation changes in individuals of older age were more similar within families, suggesting an effect of genotype on methylation patterns.

The majority of studies that have explored DNA methylation heritability to date have been based on the twin design (Bell and Spector 2011), with some exceptions (McRae *et al.* 2014). Numerous twin studies have been conducted to understand the regulation of DNA methylation (Bell *et al.* 2012; Bell and Saffery 2012; Gervin *et al.* 2011; Gordon *et al.* 2012; Grundberg *et al.* 2013; Heijmans *et al.* 2007; Javierre *et al.* 2010; Kaminsky *et al.* 2009; Kuratomi *et al.* 2008; Wong *et al.* 2010) and these studies have observed the effect of genetic and environmental factors on DNA methylation at particular genes (Heijmans *et al.* 2007; Wong *et al.* 2010) or throughout the genome (Javierre *et al.* 2010; Kaminsky *et al.* 2009; Kuratomi *et al.* 2008). The results are discussed in greater detail in the introductory section to Chapter 4. Overall, many studies have studied DNA methylation profiles in MZ and DZ twins, reporting

similarity between MZ twins compared to DZ twins, and implying that genetic effects contribute to DNA methylation levels in particular regions of genome. Although the heritability of individual CpGs can range between 0% and 100%, the average reported methylation heritability at all profiled CpGs across the genome is low to moderate.

1.5.4 Genetics of DNA methylation: Methylation Quantitative Trait Loci

To explore genetic impacts on DNA methylation levels further, a number of studies have examined the association between genetic variation at particular loci and DNA methylation patterns across the genome. Genetic loci, at which such associations are identified, are referred to as methylation quantitative trait loci (meQTLs). Evidence for meQTLs has been explored on a genome-wide scale using high-throughput DNA methylation analyses, identifying local (*cis*) and distal (*trans*) associations of genetic variants with methylation levels in multiple samples, across a number of cells, tissues, and ages (Banovich *et al.* 2014; Bell *et al.* 2011; Bell *et al.* 2012; Drong *et al.* 2013; Fraser *et al.* 2012; Gamazon *et al.* 2013; Gibbs *et al.* 2010; Grundberg *et al.* 2013; Gutierrez-Arcelus *et al.* 2013; Shi *et al.* 2014; Smith *et al.* 2014; van Eijk *et al.* 2012; Wagner *et al.* 2014; Zhang *et al.* 2010). The results are discussed in greater detail in the introductory section to Chapter 4. Additionally, a number of meQTL studies have also explored the overlap and direction of association of meQTLs and expression QTLs (eQTLs), reporting variability in the extent of overlap ranging between 4.8-25% (Banovich *et al.* 2014; Gibbs *et al.* 2010). Most recently, Banovich *et al.* reported almost 14,000 meQTLs in lymphoblastoid cell lines (LCLs). Interestingly, almost half of the overlapping meQTLs and eQTLs in LCLs showed positive correlation between methylation and gene expression levels. As discussed in section 1.5.1 both positive and negative correlations between methylation and gene expression levels have been reported and the relationship between methylation and gene expression depends in part on the genomic context of the CpG-site. The Banovich *et al.* meQTLs and eQTLs in

LCLs that showed positive correlation between methylation and gene expression levels were located at CpG-sites that were more distant from TSS of genes, suggesting that DNA methylation in more distal regulatory elements may be more likely to have an activating effect on expression (Banovich *et al.* 2014).

Overall, meQTLs were identified in numerous tissues and cell types and might increase our knowledge of the genetic component of gene regulation (Bell *et al.* 2011). The meQTL results identified to date suggests that genetic variation can have an effect on the methylome with implications for tissue specificity, tissue shared effects, and shared impacts across multiple gene regulatory processes.

1.5.5 Epigenome-wide association studies

To date most studies of human diseases have focused on genetic and environmental risk factors. More recent work has also underlined a role for epigenetic processes underlying disease susceptibility, where epigenetic mechanisms may mediate some of the effect of genetic and environmental risk factors towards disease. Epigenome-wide association studies (EWAS) aim to systematically investigate the association of epigenetic changes with disease, where changes may either occur prior to disease or as a consequence of the phenotype (Rakyan *et al.* 2011a). Two of the most common EWAS study designs are the case-control and disease-discordant twin design. The aim of disease-discordant twin analyses is to identify non-genetic, that is potential environmentally driven or stochastic, epigenetic changes present in the case but not the control twin. These have been applied to a large number of traits, with some interesting findings. The case-control design identified thousands of differently methylated regions (DMR) in rheumatoid arthritis (Liu *et al.* 2013) and schizophrenia (Kinoshita *et al.* 2013). Recent EWAS in disease-discordant twins have been presented for bipolar disorder (Dempster *et al.* 2011), systemic lupus erythematosus (Javierre *et al.* 2010), T1D (Rakyan *et al.* 2011b), and T2D (Nitert *et al.* 2012). Overall, their findings not only identify disease-

associated DNA methylation markers, but also suggest that epigenetic changes may be important clinical indicators of disease.

1.5.6 Environmental Epigenetics

Environmental epigenetics is a rapidly growing area of research, focusing mainly on the association between DNA methylation changes and environmental exposures. It is now becoming clear that the dynamic changes in DNA methylation patterns are partway due to environmental exposures (Szyf 2013). To date, numerous environmental factors have been discovered that influence DNA methylation. For example, direct tobacco smoking and maternal smoking have strong effects on methylation changes and smoking-related differential methylation sites have been replicated in multiple populations and across tissues (Besingi and Johansson 2014; Breitling *et al.* 2011; Buro-Auriemma *et al.* 2013; Dogan *et al.* 2014; Elliott *et al.* 2014; Guida *et al.* 2015; Harlid *et al.* 2014; Joubert *et al.* 2012; Markunas *et al.* 2014; Monick *et al.* 2012; Philibert *et al.* 2013; Shenker *et al.* 2013; Sun *et al.* 2013; Suter *et al.* 2011; Szyf 2013; Wan *et al.* 2012; Zeilinger *et al.* 2013; Zhang *et al.* 2014) . Together these studies have associated DNA methylation changes to exposure to smoking, whether during early development or during adult life. Additionally, other environmental exposures, such as sun exposure, dietary and nutrition intake, season of birth, alcohol consumption, and physical activities have also been connected to methylation levels using genome-wide approaches (Amarasekera *et al.* 2014; Breton *et al.* 2014; Dominguez-Salas *et al.* 2014; Ivorra *et al.* 2015; Lee *et al.* 2015; Philibert *et al.* 2012; Richmond *et al.* 2015; Rönn *et al.* 2013; Thapar *et al.* 2012; Voisin *et al.* 2015; Zhang *et al.* 2013; Zhao *et al.* 2013).

1.6 Aims of the thesis

In the first part of my study, my research question was to assess if GWAS of blood metabolites as functional intermediate phenotypes can give results that help to understand the role of genetic variants in dissecting human metabolic and disease

pathways. I explored human metabolomic profiles and evaluated the human genetic component of metabolite levels. I analysed metabolite levels in twins and identified many genetic variants that influence metabolomic traits. I also compared findings across metabolomics platforms to find stable and robust metabolites that may be combined or used for replication in future studies.

In the second part of my study, my research question was to test if genetic effects can influence both DNA methylation levels and DNA methylation variability. I investigated DNA methylation profiles in twin using whole blood and adipose tissue, and tested genetic influences on DNA methylation profiles.

The third part of my study aimed to link metabolomic and epigenetic datasets to T2D. I aimed to address four specific hypotheses in this section: (1) Metabolomic profiles that are characteristic of T2D associate with epigenetic variants. I performed an association study of metabolic profiles in T2D and tested whether the T2D-associated metabolites also associate with DNA methylation changes genome-wide. (2) Epigenetic variants are associated with T2D, and these may also be associated with metabolic profiles. I performed an EWAS of DNA methylation changes in T2D, comparing DNA methylation levels to T2D, to identify differentially methylated positions in T2D (T2D-DMPs). I then tested whether the T2D-DMPs also associate with metabolic profiles. (3) T2D genetic susceptibility effects are mediated via intermediate phenotypes, such as epigenetic changes or metabolic profiles. I compared the list of 81 T2D GWAS signals that have been published to date against the genetic variants that contribute to metabolomic and epigenetic profiles identified from the first two parts of my thesis. (4) Integrating genetic, epigenetic, and metabolic profiles associated with T2D can help to understand biological mechanisms underlying T2D. I fit Bayesian networks to the peak T2D-GWAS, T2D-metabolite, and T2D-DMPs results and pair-wise associations, to gain more insight into T2D susceptibility and progression.

In Summary, I initially explored the genetic basis of metabolomic and epigenetic datasets on their own (Chapter 3 and Chapter 4). I subsequently combined these results to better understand the relationship between the genetic basis of epigenomics and metabolomics in the context of genetic variants associated with T2D (Chapter 5).

CHAPTER 2

Materials and Methods

This chapter provides an overview of the data used in this thesis, starting with an overview of the cohort and genetic data, and then focusing on the metabolomics and methylation datasets that I have analysed. I provide a brief description of Metabolomics platforms and Illumina methylation arrays, and quality control procedures that I adopted through the thesis.

2.1 TwinsUK cohort

TwinsUK is the UK's largest cohort of adult twins. The registry started in 1992 and contains about 13,000 same-sex twin volunteers from all over the United Kingdom (Moayyeri *et al.* 2012). Twins from this cohort were shown to be comparable to singletons in terms of disease-related and lifestyle characteristics (Andrew *et al.* 2001).

Twins participate in regular clinical visits, during which questionnaire data are collected, a series of phenotypic tests are performed, and biological samples are also collected. Phenotype test examples include body mass index (BMI), Dual-energy X-ray absorptiometry (DEXA) scans, hearing tests, and vision tests. Biological samples examples include blood, urine, and tissue biopsies. Participation in the registry is voluntary and informed consent is obtained for all research projects.

2.2 Genotype data in TwinsUK

The genotyping and imputation steps for the TwinsUK cohort have been described in detail previously (Illig *et al.* 2010; Suhre *et al.* 2011b). Briefly, genoyping of the TwinsUK cohort was performed using a combination of Illumina arrays

(HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo). Normalized intensity data and genotype calling on the basis of the Illluminus algorithm was pooled. No calls were assigned if the most likely call had a posterior probability less than 0.95. SNPs with Hardy–Weinberg ($P<1x10^{-6}$) and with minor allele frequencies (MAF) <1% were excluded. First, the sparser HumanHap300 dataset was imputed to the HumanHap610Q using phased TwinsUK HumanHap610Q haplotypes as a reference. Next, for genotype imputation to HapMap - the combined panel was imputed using reference haplotypes from the HapMap2 project (rel 22, combined CEU+YRI+ASN panels). These analyses were performed previously by other researchers in the department for all twins with available genotype data in the cohort.

Imputation was also performed to 1000 genomes. Here, imputation was done using the denser haplotype maps from the 1000 Genomes Project (Abecasis *et al.* 2010) using the 1000 Genomes Project multi-population panel (March 2012 release for TwinsUK). "Pre-phasing" of the GWAS data was performed using IMPUTE2 without a reference panel and then fast imputation from 1000 Genome phase1 dataset was performed on the resulting haplotypes. These analyses were performed previously by other researchers in the department for all twins with available genotypes in the cohort. In this thesis, I used 7 genotype datasets (Table 2-1).

Table 2-1 Summary of the Genotype datasets

Dataset	Platform	Subjects	Chapters
1,2,3	HapMap	6055,1235,1001	3
3,4,5,6	1000G	330,83,789,459	4
7	1000G	807	5

2.3 Heritability

Heritability was calculated using the ACE model as described in the Introduction (Chapter 1). OpenMX (Boker *et al.* 2011) was used for all heritability calculations.

Heritability calculations were performed on several datasets throughout this thesis (Table 2-2). I performed the majority of heritability calculations, with the exception of the heritability results in Chapter 4, which were performed by a PhD student in the epigenomics group, Juan Edgar Castillo-Fernandez. We used the ACE model for twins assuming that the environment has a similar effect on both MZ twins and DZ twins, and then any finding of a higher correlation within MZ pairs compared to DZ pairs with respect to a particular trait indicates a genetic effect (Figure 2-1). This is driven from the fact that MZ twins are genetically identical, while DZ twins share on average just 50% of their segregated genetic variation. It follows that (A) is 1.0 for MZ pairs and 0.5 for DZ pairs. Since by assumption, MZ and DZ twins share the same common environment (C), the correlation between their latent shared environmental factors is 1 for both MZ and DZ twins.

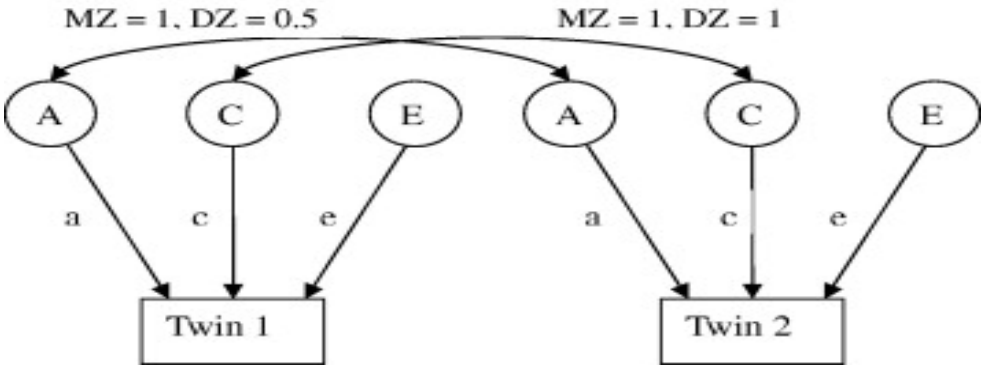

Figure 2-1 ACE models for twins

Table 2-2 Summary of the datasets used for heritability analysis in this thesis

Dataset	Data	Subjects	Chapters
1,2	Metabolon, Biocrates	1001,1001	3
3,4	Illumina 450k	660,166	4
5	Metabolon	2204	5

2.4 GWAS methods

Genome-wide association scans (GWAS) were carried out using directly genotyped and imputed SNPs using an additive linear regression model for all traits considered in this thesis. Importantly, the p-value threshold for significance is corrected for multiple testing issues. One of the simplest approaches to correct for multiple testing is the Bonferroni correction. The Bonferroni correction adjusts the alpha value from $\alpha=0.05$ to $\alpha=(0.05/k)$ where k is the number of statistical tests conducted. For a typical GWAS using 1 million SNPs, statistical significance of a SNP association would be set at 5×10^{-8}. The R packages or software programs that I used in this thesis included MERLIN (Abecasis *et al.* 2002), GEMMA (Zhou and Stephens 2012), GenABEL/ProbABEL (Aulchenko *et al.* 2007), lme4 (Bates *et al.* 2015), PLINK (Purcell *et al.* 2007), and Matrix eQTL (Shabalin 2012). Detailed information is given in each Chapter on these methods.

2.5 Metabolomics

2.5.1 Metabolomics Data Platforms

I had access to TwinsUK metabolomics data generated by MS from Metabolon Inc. (http://www.metabolon.com/) and Biocrates AG (http://www.biocrates.com/) in the TwinsUK cohort. Metabolon uses a non-targeted approach for identifying metabolites, while Biocrates uses a targeted approach. In this thesis, I used 7 metabolomics datasets (Table 2-3).

Table 2-3 Summary of the metabolomics datasets

Dataset	Platform	Subjects	Chapters
1,2	Metabolon	6055,1001	3
3,4	Biocrates	1235,1001	3
5,6,7	Metabolon	2204,36,807	5

In the following sections I review the methods used by the two metabolomics platforms, the non-targeted Metabolon (2.4.1.1) and the targeted Biocrates (2.4.1.2) platforms to detect metabolites. Both of these platforms are MS-based approaches, but they incorporate different methods to detect and quantify metabolites. A summary of the workflow in the two platforms is shown in Figure 2-2.

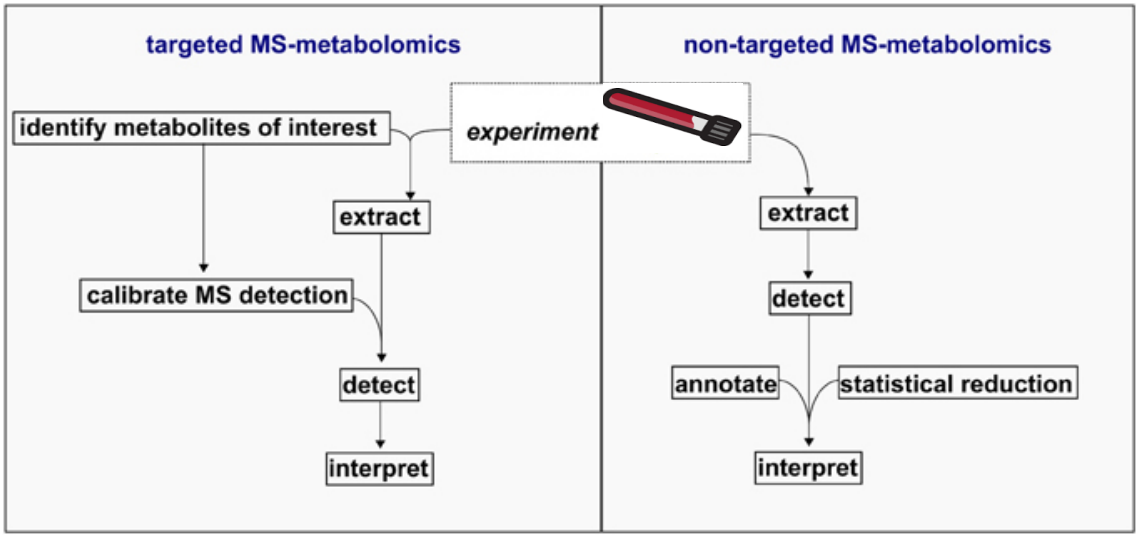

Figure 2-2 Targeted and non-targeted metabolomics workflow.Figure adapted from (Heuberger et al. 2014). Left panel shows the Biocrates workflow, which is efficient at identifying metabolites of interest. Right panel shows the Metabolon approach, aiming to identify a wider range of metabolites, not limited to a preselected library.

2.5.1.1 Metabolon

In the Metabolon data platform, chromatography is coupled with MS (Evans *et al.* 2009; Raffler *et al.* 2013; Suhre *et al.* 2011b). The Metabolon platform incorporates UPLC/MS and GC/MS. The generated spectral data are compared against an in-house library, which includes retention time and reference spectra from mass scan and fragmentation of molecules. Chromatography separates the metabolites beforehand by collision and internal standards are added only for quality control. Samples then go through Electro Spray Ionization (ESI), and charged samples go through spectrometry. Masses of fragments and ions are checked at each peak (using the area under the peak) without relating them to any standards.

One of the challenges of complex samples is the need for separation techniques like chromatography. In the Metabolon platform both GC-MS and UPLC-MS integrate chromatographic separation with MS to identify relative concentrations of a large number of small molecules in metabolomics (Lawton *et al.* 2008). GC-MS is a method that combines the features of gas chromatography and mass spectrometry to identify different substances within a test sample. GC-MS has previously been applied to identify unknown samples and drugs (Lindon *et al.* 2007) and has also been used to measure compounds in urine and tissue samples (Wilson et al., 2005). GC-MS involves two modes of ionization using electro ESI-MS/MS. This enables detection of the molecular mass by (1) electron impact ionization and (2) chemical ionization. Electron ionization is successful at picking diagnostic fragments that provide structural information for identifying metabolites (Fiehn *et al.* 2000). Chemical ionization is much more effective for providing ion information especially when identifying the molecular mass of unknowns.

UPLC-MS is an analytical chemistry technique that extends the physical separation capabilities of liquid chromatography with mass spectrometry. It is a powerful technique used for many applications and has very high sensitivity and selectivity (Ardrey, & Robert, 2003). If the GC and UPLC platforms are compared, then GC is a high-resolution separation technique for metabolite profiling that requires extensive sample pre-treatment. On the other hand, UPLC requires minimal sample preparation (Wilson *et al.* 2005). Both of the separation techniques are necessary for separating different classes of substances.

The set of 503 Metabolon metabolites profiled in the datasets used in this thesis consists of several classes of molecules: amino acids, acylcarnitines, sphingomyelins, glycerophospholipids, carbohydrates, vitamins, lipids, nucleotides, peptides, xenobiotics and steroids. Additionally, the Metabolon platform also reports unknown metabolites .

Metabolite Measurements

Serum and plasma samples were treated with methanol, shaken for 2 minutes, followed by centrifugation. The resulting extract was divided into three parts: one for analysis by UPLC-MS/MS (positive model, where the MS analysis based on positive ions), one for analysis by UPLC-MS/MS (negative model, where the MS analysis based on negative ions), and one for analysis by GC-MS. Three types of controls were analysed together with the experimental samples: samples generated from a pool of human plasma (Metabolon, Inc.) served as technical replicates throughout the data set; extracted water samples served as process blanks; and a cocktail of standards spiked into every analysed sample allowed instrument performance monitoring. Experimental samples and controls were randomized across the platform run.

The UPLC-MS/MS platform utilized a mass spectrometer, which included ESI. The instrumentation was set to monitor for positive ions in acidic extracts or negative ions in basic extracts through independent injections. Extracts were loaded onto columns with water and 95% methanol containing 0.1% formic acid or 6.5mM ammonium bicarbonate.

Samples analysed by GC-MS were dried under vacuum desiccation for a minimum of 18h prior to being derivatized (that is, transforming a chemical compound into a product) under dried nitrogen. Derivatized samples were separated on a 5% phenyldimethyl silicone column with helium as carrier gas and a temperature ramp from 60° to 340° C within a 17-min period. All samples were analysed resolving power with electron impact ionization and a 50-750 atomic mass unit scan range (Metabolon Inc.).

Metabolites were identified by automated comparison of the ion features in the experimental samples to a reference library of chemical standard entries that included retention time, molecular weight, and in-source fragments (Suhre *et al.* 2011b). Identification of structurally named chemical entities was based on comparison to a

mass spectroscopy library of >2,400 purified standards, and this procedure is part of the Metabolon profiling process. An additional 5,300 mass spectral entries have been created for structurally unnamed biochemicals. These compounds have the potential to be identified by future acquisition and further analysis. Concentrations of all analysed metabolites are reported as relative concentrations (Metabolon Inc.).

2.5.1.2 Biocrates

Targeted metabolomics was designed in the 1990s by one of the founders of Biocrates, Dr. Roscher (Biocrates Inc.). Targeted metabolomics aims to identify and quantify known and biochemically annotated metabolites. The Biocrates method is a quantitative screen of known small molecule metabolites detected with multiple reaction monitoring, which is a highly sensitive and selective method for the targeted quantitation of protein/peptide abundances in complex biological samples. Additionally, neutral loss and precursor ion scans are used for screening. In a neutral loss scan, the first mass analyzer scans all the masses. On the other hand, in precursor ion scans the product ion is selected in the second mass analyzer, and the precursor masses are scanned in the first mass analyzer. All of these processes are part of main scan experiments in MS (Lindon *et al.* 2007). Metabolites are then quantified by comparison to structurally similar molecules labelled with stable isotopes added to the samples in defined concentrations as internal standards. In the Biocrates targeted platform metabolites are measured using targeted LC-MS, which also uses ESI-MS/MS for the ionization source at the pre-separation stage. The interpretation of targeted metabolomics data are typically much more straightforward, the obtained metabolite concentrations are provided as absolute levels, and this platform is well suited for high-throughput and routine applications for cohort investigations (Sonntag *et al.* 2011).

Measurements

Serum samples (100 μl) were prepared for quantification using the AbsoluteIDQ kit (BIOCRATES AG). Sample analyses were performed on the API 4000 Q TRAP LC/MS System with an autosampler (Illig *et al.* 2010). Briefly, Biocrates uses Flow Injection Analysis (FIA) tandem MS (Illig *et al.* 2010). Platform internal standards, which are molecules with heavy isotopes, were added to the samples. These standards serve as references for calculating all metabolite concentrations. This dissolvent is used directly in tandem MS. The first step in the mass spectrometry is ionization. Ionization charges the dissolvent so that metabolites can be measured during MS more effectively. Here, spectrometry searches for the loss of specific masses when metabolites are fragmented, and these are compared to the known masses of the internal standards. The resulting peaks are analysed for whether they match specific targeted metabolites and atomized algorithms calculate absolute concentration values (in micromolar units (μuM)). Atomization of the analysed sample refers to the transformation of solid matter into atomic vapour and ionization of the atoms. These atoms are then sorted and counted with the help of mass spectrometry and used as a reference when identifying new outputs (Lindon *et al.* 2007). One issue of consideration for both targeted and non-targeted MS platforms is ion suppression. In a complex mixture, metabolites can influence each other's ionising ability. Ion count is taken to directly reflect the amount of the metabolite, therefore ionisation can impact quantification. This issue impacts both platforms, but in the targeted version quantification is improved by the use of standards.

The Biocrates metabolomics datasets used in this thesis contained 163 targeted metabolites: 41 acylcarnitines (Cx:y), hydroxylacylcarnitines [C(OH)x:y] and dicarboxylacylcarnitines (Cx:y-DC); 14 amino acids; 1 sugar; 15 sphingomyelins (SMx:y) and sphingomyelin-derivatives [SM(OH)x:y]; and 92 glycerophospholipids (PC). Glycerophospholipids are differentiated with respect to the presence of ester (a)

and ether (e) bonds in the glycerol moiety, where two letters (aa = diacyl, ae = acyl-alkyl) denote that two glycerol positions are bound to a fatty acid residue, while a single letter (a = acyl) indicates the presence of a single fatty acid residue. Lipid side chain composition is abbreviated as Cx:y, where x denotes the number of carbons in the side chain and y the number of double bonds. Further descriptions of the 163 Biocrates metabolites have previously been published (Menni *et al.* 2013a; Mittelstrass *et al.* 2011; Römisch-Margl *et al.* 2012).

In summary, Metabolon and Biocrates are two of the most commonly used metabolomics high-throughput techniques that are currently used in large cohort studies. Both are based on MS, applying either a targeted approach in the Biocrates platform or a non-targeted approach in the Metabolon platform. MS diagnostics of targeted chemical compounds are more cost effective than other current approaches, such as NMR. MS also offers a wider analytical panel and improved diagnostic quality of compounds (Biocrates Inc.). Unfortunately, targeted MS can only identify known metabolites and that causes a drastically small amount of metabolites than non-targeted MS methods.

2.6 DNA methylation

2.6.1 DNA Methylation Data Platforms

DNA methylation profiles from whole blood and adipose samples from individuals in the TwinsUK cohort were generated using Illumina Infinium HumanMethylation BeadChip arrays. In this thesis I used 8 methylation datasets (Table 2-4).

Table 2-4 Summary of the methylation datasets

Dataset	Illumina Platform	Tissue	Subjects	Chapters
1,2	450k, 27k	Blood, Blood	57,57	2
3,4,5,6	450k	Blood, Blood, Blood, Adipose	660,789,459,166	4
7,8	450k	Blood	864,807	5

2.6.2 Comparison of the Illumina 27k and Illumina 450k technology

DNA samples were interrogated utilising the Illumina 27k and Illumina 450k arrays. These platforms detect the methylation status of 27,578 CpG sites vs. 485,000 sites, respectively by microarray genotyping of bisulfite treated DNA, respectively.

2.6.2.1 Illumina Infinium Human Methylation27 BeadChip

In Illumina 27k, two beads, each containing probes of length of 50 base pairs (bp), are used to assess DNA methylation levels at each CpG site. The two bead types are the unmethylated bead type, designed to match the unmethylated version of the CpG site, and methylated bead type, which matches the methylated CpG site. The bisulfite converted DNA sample is first separated into single strands and then hybridized to the Illumina 27k array, which contains the bead probes. Following hybridization of the DNA to the bead probes, fluorescently labelled dideoxynucleotides (ddNTPs) will be incorporated at each bead probe, if the probe sequence matches the DNA, and the array is then scanned for bead intensities, the design of the type I probe and detection process is shown in the Figure 2-3.

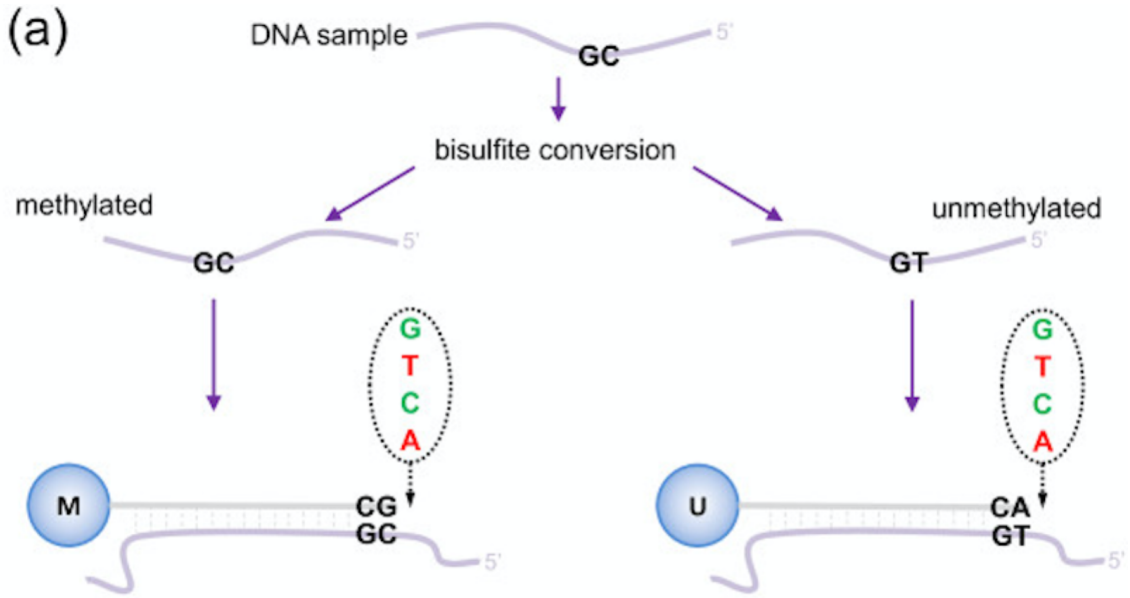

Figure 2-3 Design of the type I probes in Illumina 27k.Figure adapted from (Maksimovic et al. 2012)

Briefly, Figure 2-3 shows the type I probe design, which uses fluorescence from two different probes, unmethylated probes and methylated probes, to assess the level of methylation at a CpG-site (Maksimovic *et al.* 2012). If a CpG is methylated in the sample, the cytosine (C) base will remain unconverted after bisulfite conversion, and the genomic DNA will bind to the 'methylated' probe, which at the 3' end can bind a C. On the other hand, if the CpG is unmethylated in the sample, the C base will be converted to a thymine (T) following bisulfite conversion, and then the genomic DNA will bind to the 'unmethylated' probe, which at the 3' end can bind a T. Binding at either probe is followed by single base extension that results in the addition of a fluorescently labeled nucleotide, which is then read by the scanner to detect and quantify binding.

The degree of methylation is measured as a quantitative trait called beta values. Illumina 27k use betas to quantify methylation levels, and the beta value is calculated as the ratio of the intensity of methylated beads over intensity of the sum of the intensity of both methylated and unmethylated beads (Equation 2-1). At a single CpG-site the range of beta values is between 0 (unmethylated) and 1 (methylated).

$$\boldsymbol{beta} = \frac{\boldsymbol{methylated\ signal}}{\boldsymbol{methylated\ signal + unmethylated\ signal + 100}} \qquad \textit{(Equation 2-1)}$$

A BeadChip includes 12 samples and covers 27,578 CpG dinucleotides in the promoter of almost 15,000 genes.

2.6.2.2 Illumina Infinium Human Methylation450 BeadChip

In Illumina 450k, in addition to type I probes from Illumina 27k (28% of the Illumina 450k probes), there are also type II probes (72% of the Illumina 450k probes). In type II probes there is only one bead to detect the methylation levels, which is different than the two bead design for type 1 probes. In type II probes the colour of the incorporated

fluorescently labelled ddNTP (green or red) determines the methylation status of the CpG site. The design of the type II probe and detection process is shown in Figure 2-4.

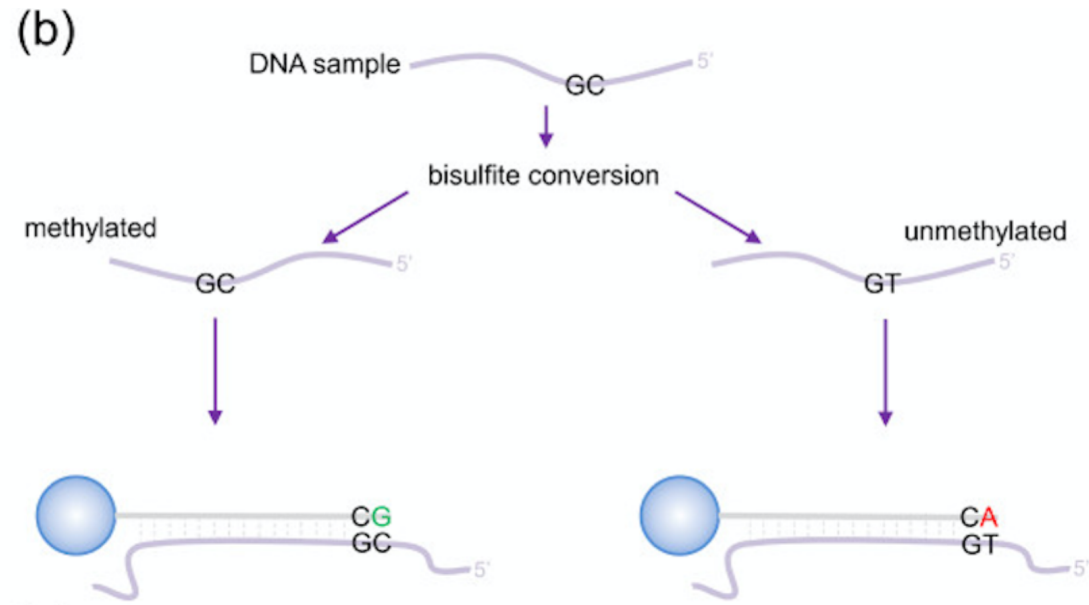

Figure 2-4 Design of the type II probes in Illumina 450k. Figure adapted from (Maksimovic et al. 2012)

Briefly, Figure 2-4 shows the type II probe design that uses only a single probe per CpG. The 3' end of each type II probe is linked to the base directly upstream of the C of the CpG (Maksimovic *et al.* 2012). Methylation state is detected by single base extension at the position of the C of the CpG, which results in the addition of a labeled G or A nucleotide, matching the 'methylated' C or 'unmethylated' T, respectively.

An Illumina 450k BeadChip includes 12 samples and covers 485,836 CpG dinucleotides. Illumina 450k uses the same beta quantification as Illumina 27k. On the other hand, the density distribution of the methylation beta values differs according to probe type, as can be seen in Figure 2-5, which represents the density plot of all methylation values from a single female individual in the cohort.

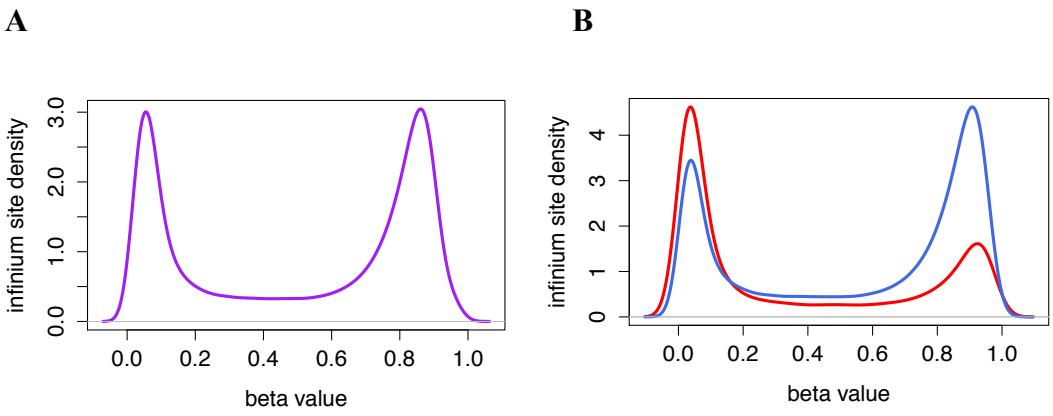

Figure 2-5 Methylation beta distribution of one subject Figure depicts A) Overall Illumina 450k and B) Signals according to probe type, with type I (red) and type II (blue) probes.

It has been suggested that the two probe types distributions should be made comparable, prior to between subject normalization, because without such adjustment, there would be a bias on type 1 probes to have higher rankings than type II probes (Teschendorff *et al.* 2013). Quantile normalization is a typically used normalization approach in gene expression datasets (Bolstad *et al.* 2003) and can be applied in other "-omics" data. Additionally, for Illumina 450k array data several other recent methods have also been developed and these are available in R packages, such as Subset-quantile Within Array Normalization method (SWAN) (Maksimovic *et al.* 2012), wateRmelon (Pidsley *et al.* 2013) and Beta Mixture Quantile Dilation (BMIQ) (Teschendorff *et al.* 2013). I used BMIQ for the normalization of the raw betas since this method transforms type II probes to fit the distribution of type I probes. After categorizing both types of probes into methylated, hemi-methylated and unmethylated, type I probes are used as a base for fitting type II probes into quantiles using the inverse of the cumulative beta distributions in each category.

2.7 Quality control data procedures for methylation and metabolomics

Many approaches have been proposed in quality control analysis of metabolomics platform data and DNA methylation data from Illumina arrays (Bock 2012; Morris *et al.* 2014; Suhre *et al.* 2011b). The approach used in this thesis for both datasets included

the following steps: 1) identification of outliers, 2) identification of batch effects, missing values and covariates that affect methylation or metabolite levels in the sample, and 3) application of data normalization and adjustment for covariates to reduce these effects. In this section I explain the steps that I undertake in quality control procedures for both the metabolomics and methylation data in the thesis, using an example dataset of 57 individuals with available Illumina 27k and Illumina 450k data. In these data I aimed to compare methylation data from the two platforms. From both platforms I selected the overlapping probes (23,678) present in both arrays for 57 individuals (Figure 2-6).

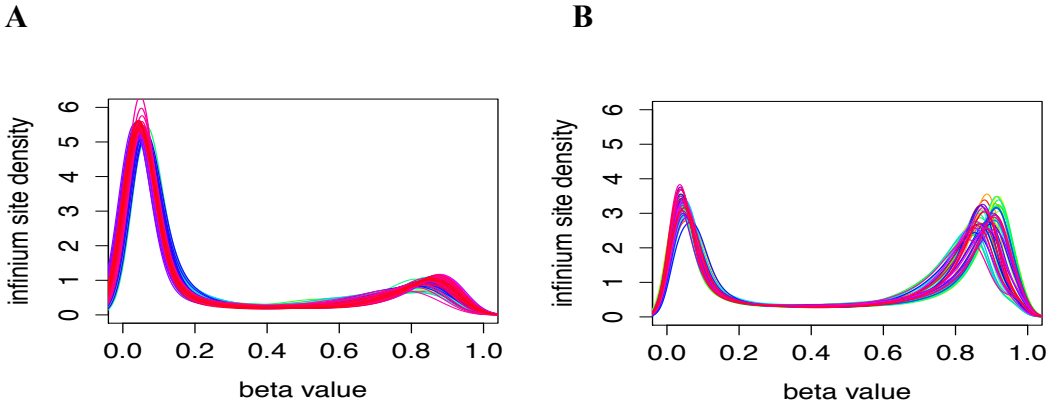

Figure 2-6 DNA methylation distribution in 57 individuals Figure depicts A) Illumina 27k and B) Illumina 450k array. Each line represents the density distribution in one individual.

2.7.1 Identification of outliers

I first tried to identify outliers visually both at the level of the methylation probe (or metabolite if using a metabolomics dataset) and at the level of the individual. Additionally, high missing rates were also considered as exclusion criteria for both individual and probe level data. Boxplots and density plots were used to explore these different patterns.

In the metabolomics data, the effect of experimental batches (i.e. run-days (1-27 (batch 1)), (28-49, 50-71 (batch 2)), (72-97, 98-122 and 123-147 (batch 3)) was explored and before this a data normalization step was also applied to adjust for variation due to run-

day tuning differences (procedure applied by Metabolon directly). Additionally, data points that were more than 4 standard deviations from the mean of each metabolite concentration were excluded.

Initial quality control of the methylation data was performed by a member of our group (Dr. Pei-Chien Tsai). She identified several outliers (subjects) in the larger Illumina 450k dataset. To explore these impacts in the dataset of 57 individuals with both Illumina 27k and Illumina 450k I assessed the distribution and missing value rates for each individual using a number of plots and summary statistics as described above, and did not exclude any individuals based on these profiles. First, I calculated the overall correlation between the 57 individuals from both arrays, which shows strong positive correlation r= 0.97 (Figure 2-7). For each individual I compared the DNA methylation levels obtained from the Illumina 27k to the methylation levels from the Illumina 450k, across altogether 23,678 probes that were present on both arrays.

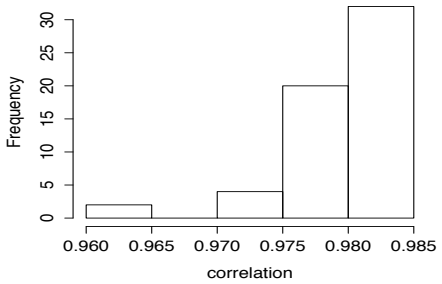

Figure 2-7 Pearson pairwise correlation in DNA methylation profiles across 57 individuals using 23,678 probes generated between Illumina 27k and Illumina 450k arrays.

Another approach to look for outliers at the individual-level was to visualize the correlation in methylation patterns between the individuals in a two-dimensional plot using gradient colours in a heatmap (Figure 2-8). In general, the 57 individuals in this dataset were moderately correlated (r= 0.89) with each other in both arrays, with slightly lower correlations (r=0.73) on the Illumina 27k data (Figure 2-8A) when compared to the Illumina 450k correlations (Figure 2-8B). Heatmaps can also be used to identify any patterns of substructure in the sample as some structures can be seen in

Figure 2-8B and these can also be used to visually identify outliers for potential exclusion in downstream analysis.

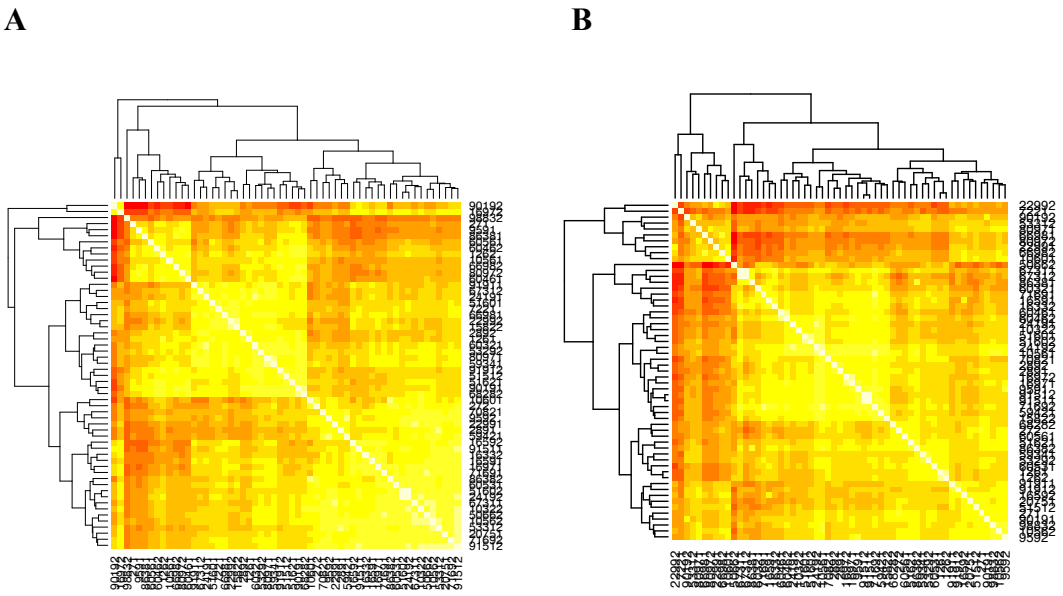

Figure 2-8 Heatmap of the Pearson pairwise correlation in DNA methylation profiles across 57 individualsindividuals using 23,678 probes generated on the A) Illumina 27k and B) Illumina 450k arrays. Greater correlation between individuals was represented as a light colour (yellow), and lower correlation between arrays was represented in red.

In summary, I used a number of approaches to detect outliers in both the methylation and metabolomics datasets on the level of the subject and probe/metabolite. I excluded altogether a small number of datapoints based on the outlier identification approaches described above, because outliers may represent potential experimental error and I wanted to minimize these effects in the datasets for downstream analyses.

2.7.2 Principal Component Analysis

I used Principal component analysis (PCA) to identify potential batch effects (i.e age, sex and plate). PCA refers to a transformation of a number of variables into a smaller number of uncorrelated variables or Principal Components (PCs). PCA allows us to identify overall patterns in the data by exploring PCs. The two main purposes of using this analysis was to first, reduce the dimensionality of the dataset and second, use the summary variables (PCs) to identify possible confounders by comparing the PCs to potential variables that may introduce patterns or noise in the data, for example, batch

effects. By definition, the first PC will always be the one that captures most of the variation in the data. In this dataset PC1 of the Illumina 27k can explain 24% of the variance of the data and the first three PCs explained in total 44% of the variance (Figure 2-9A) PC1 of the Illumina 450k can explain 30% of the variance of the data and the first three PCs explained in total 46% of the variance (Figure 2-9B).

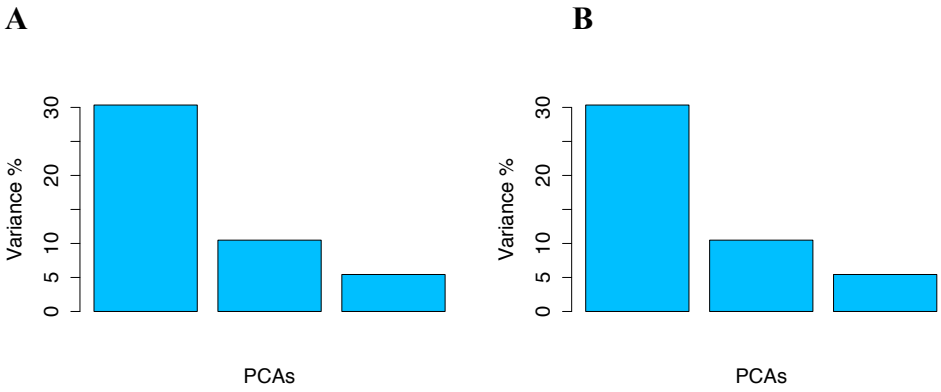

Figure 2-9 Variance distributed by the first 3 principal components on DNA methylation profiles across 57 individuals in A) Illumina 27k showing cumulative variance of 44% B) 3 Illumina 450k cumulative variance of 46%

Previous comparison of PCs in the larger Illumina 27k dataset (dataset 1) has identified significant association of the first 3 PCs with the following covariates: batch, age, plate, order of plate (Bell *et al.* 2012). For the 57 individuals, I compared the PC loadings for the first 3 PCs against known covariates in the Illumina 27k dataset of 57 individuals, and I observed significant associations between the PC and these known covariates, suggesting that these covariates are major sources of variability in the Illumina 27k data (Table 2-5).

Table 2-5 Illumina 27k PCs nominally associated (P = 0.05) with known covariates

	PC1	PC2	PC3
Proportion of variance	0.24	0.15	0.05
Cumulative Proportion	0.24	0.39	0.44
	Age Plate Order of plate	Age Plate Order of Plate	Plate BMI

Similarly, I compared the PC loadings for the first 3 PCs in the Illumina 450k dataset of 57 individuals against known covariates, and I observed significant association between the PC and the known covariates (Table 2-6).

Table 2-6 Illumina 450k PCs nominally associated (P = 0.05) with known covariates

	PC1	PC2	PC3
Proportion of variance	0.31	0.10	0.05
Cumulative Proportion	0.31	0.41	0.46
	Age Plate Order of plate	Plate BMI	Plate

I compared the first 2 PCs from both Illumina 27k and Illumina 450k but found low correlation (Figure 2-10), suggesting that the PCs capture technical variation specific to each array rather than shared technical or not biological variation. However, it is also possible that different PCs in each dataset capture different covariate effects (for example, BMI is associated with PC3 in Illumina 27k and with PC2 in Illumina 450k), but I did not investigate this further as the main aim here was give an example of a quality control procedure that I used to identify batch effects in large scale datasets.

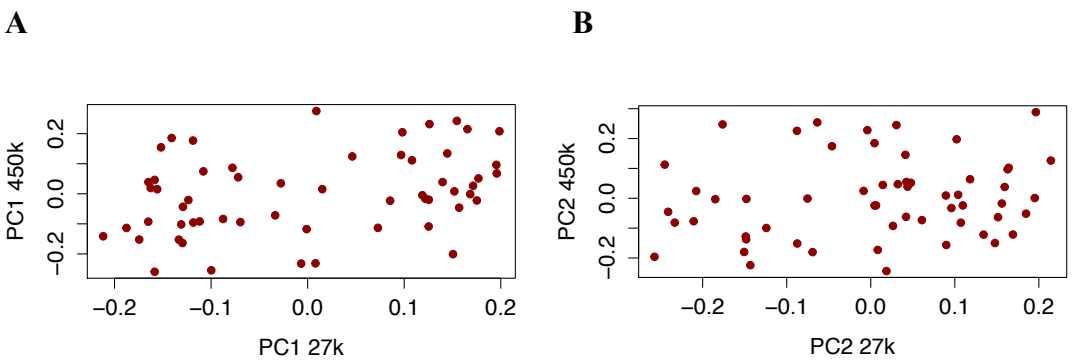

Figure 2-10 Comparison of PCs from both datasets for A) PC1 (r=0.34) B) PC2 (r=0.19)

2.7.3 Correlation between Illumina 27k and Illumina 450k

These identified covariates were then corrected for using a linear mixed effects regression (LMER) in lme4 package in R (Bates *et al.* 2015). Residuals were calculated

from the full regression model, where normalized methylation levels were fitted as the outcome and the predictors consisted of age, BMI, plate and blood cell count estimations as fixed effect terms and random effect terms family and zygosity.

I next calculated the correlation between overlapping probes, where one probe was from the Illumina 27k and the other from the Illumina 450k. Out of 23,678 overlapping probes only 2,132 (11%) on the Illumina 27k array remain as type I probe on the Illumina 450k array. The rest of the overlapping 21,546 probes have switched from type I probes on the Illumina 27k array to type II probes on the Illumina 450k array, which may introduce technical variation in the methylation signal.

Looking at correlations at the level of each probe, only positive correlations were observed (Figure 2-11), both for type I probes in both arrays (r=0.42) (Figure 2-11A) and for probes that switch type (type I from Illumina 27k and type II from Illumina 450k) (r=0.45) (Figure 2-11B). The correlations observed at the probe level are overall positive (r=0.44), but great variability in the level of correlation is observed (from 0.1 to 1), and overall the mean correlation per probe is relatively moderate.

On the other hand, across-array correlation on the level of the individual was much higher. As shown in Figure 2-11 performing pairwise correlations per individual over 23,678 probes that are measured on both arrays, results in very high positive pair-wise correlation coefficients (r=0.97), as expected because these data are generated in the same samples (same time point). In conclusion using the Illumina 27k as a replication in Illumina 450k studies, or vice versa, should be handled carefully considering the wide range of correlation coefficients detected over the 23,678 overlapping probes. In cases where across-probe correlations are modest or relatively low (r<0.3), conclusions will be difficult to interpret.

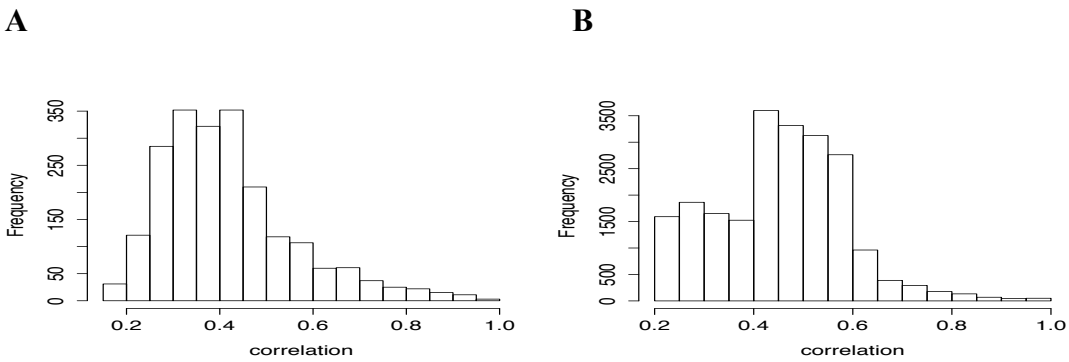

Figure 2-11 Correlation between probes that are present in both Illumina 27k and Illumina 450k in the dataset of 57 individuals for A) 2,132 probes in type I B) 21,546 probes in type II

2.7.4 Further DNA methylation probe quality control

To assess the potential of cross hybridization of the Illumina array 50bp probe sequences to multiple locations in the genome, I double checked the alignment of the Illumina 450k probes against the human genome using MAQ (v0.7.1) (Li *et al.* 2008). I allowed for probes mapping to multiple locations within 2 mismatches, and found 17,764 probes out of 485,836 probes in Illumina 450k that mapped to multiple locations within 2 mismatches (the same number of probes we identified using both hg18 and hg19). I therefore excluded these probes in all subsequent analyses.

Additionally, another factor that may impact hybridization is the presence of genetic variants in the probe sequence. Several previous studies have explored these effects and in this thesis I considered the results of one of these. Naeem *et al.* categorized both all probes with SNPs in the 50bp probe sequences and probes with SNPs located on the actual CpG-site, small insertions and deletions (INDELs), repetitive DNA, and regions with reduced genomic complexity (Naeem *et al.* 2014). The authors found that the second type of probes impact methylation levels. Since Chapter 4 in my thesis explicitly considers genetic impacts on DNA methylation level, I excluded probes previously identified by Naeem *et al.* to contain genetic variants on CpG sites. However, I included such probes in the context of MZ twin analyses in Type 2 diabetes in Chapter 5. This is because a postdoctoral fellow in the group (Dr.Pei-Chien Tsai) compared whether

methylation levels seem to be influenced by the SNP on the probe, by comparing methylation levels at probes without SNPs and probes with SNPs (and different number and location of SNPs), but did not see much difference, suggesting that the hybridization is not strongly influenced by the SNPs in the probe in the majority of cases. Therefore, although I included these probes containing SNPs in Chapter 5, I report results both including and removing these probes.

CHAPTER 3

Metabolites

3.1 Introduction

The aim of this chapter is to identify metabolite QTLs (mQTLs) from mGWAS. GWAS of blood metabolites, as functional intermediate phenotypes, give greater power to understand the role of genetic variants in dissecting human metabolic and disease pathways.

This chapter is divided into 3 sections based on the separate projects. The first two sections (3.3 and 3.4) report multi-cohort collaborative mGWAS that I contributed towards. These were each performed on large-scale datasets profiled on one of two separate metabolite platforms, Metabolon and Biocrates, and descriptive statistics for these are reported in Table 3-1. In the final section (3.5), I then extend this work to my own research focus aiming to perform and compare mGWAS results from the two metabolite platforms, where data were obtained in the same subset of samples.

Part of this work, specifically the large-scale collaborative mGWAS analyses in sections one and two, has been published (Draisma *et al.* 2015; Shin *et al.* 2014) and a manuscript based on the third section is recently published (Yet *et al., 2016*).

Table 3-1 Descriptive statistics for Metabolon and Biocrates TwinsUK datasets

Data	Biocrates	Metabolon
Individuals	1235	6056
Sex (F/M)	813 / 422	5622 / 434
Mean age (SD)	57.95 (11.13)	53.43 (13.99)
MZ/DZ	764/471	3040/3025
Metabolite sensitivity	Targeted	Non-targeted
Sample source	Serum	Serum and plasma

3.2 Metabolite GWAS

Metabolomic profiling is a powerful approach to characterize human metabolism and help understand common disease risk. mGWAS results have been reported for metabolomics datasets profiled using multiple platforms in different tissues and samples (Chasman *et al.* 2009; Demirkan *et al.* 2012; Demirkan *et al.* 2015; Hicks *et al.* 2009; Illig *et al.* 2010; Kettunen *et al.* 2012; Krumsiek *et al.* 2012; Lemaitre *et al.* 2011; Nicholson *et al.* 2011; Raffler *et al.* 2013; Rhee *et al.* 2013; Ried *et al.* 2014; Rueedi *et al.* 2014; Suhre *et al.* 2011a; Suhre *et al.* 2011b; Tanaka *et al.* 2009).The first mGWAS was performed by Gieger *et al.* using the MS platform (Gieger *et al.* 2008). The authors analysed more than 350 metabolites measured in almost in 300 serum samples. The metabolite data set covered lipids, amino acids, acylcarnitines and sugars. As an initial analysis, a mGWAS was conducted for each of the single metabolites, then following up with mGWAS of metabolite ratios. It has been suggested that ratios increase statistical power due to cancelling the common experimental error for a metabolite pair in the ratio. Metabolite ratios can also serve as substitutions for enzymatic reaction rates for closely biologically connected metabolites; consequently associations at genes encoding enzymes are typically stronger for metabolite ratios than for single metabolite concentrations. Gieger *et al* reported several associations and two of the main results, associations in the *FADS* gene cluster (fatty acid desaturase) and the *LIPC* locus (hepatic lipase) were discovered from ratio mGWAS analysis. After this initial study, multiple mGWAS have been performed on metabolite data profiled using the MS approach. Lipid based mGWAS have been conducted in targeted MS, and these have focused on phospholipids and sphingolipids (Demirkan *et al.* 2012; Hicks *et al.* 2009; Illig *et al.* 2010), or different polyunsaturated fatty acids (Lemaitre *et al.* 2011; Tanaka *et al.* 2009) and lipoprotein subfractions (Chasman *et al.* 2009). Altogether these studies have identified genetic variants at more than 50 loci associated with 200 metabolites

and 25,000 ratios. Suhre et al. focused on detecting known metabolites using MS and reported associations at 37 novel loci in 1,768 serum samples with replication in 1,052 twins (Suhre *et al.* 2011b). Krumsiek et al. (Krumsiek *et al.* 2012) focused on detecting unknown metabolites with non-targeted MS in the same sample of 1,768 serums used by Suhre et al. and identified association at 34 genetic loci. A recent study by Rhee *et al.* reported associations at 31 loci in 2,076 plasma samples using targeted MS (Rhee *et al.* 2013). Another recent study focused on both targeted and non-targeted MS and identified variants at 12 new loci from 2,652 serum samples assaying 344 metabolites (Ried *et al.* 2014). Overall, MS studies identified associations at more than 100 loci in almost 10,000 individuals assaying more than 500 metabolites.

There have also been a number mGWAS performed on metabolite datasets profiled using the NMR approach. Kettunen *et al.* identified mGWAS variants at 31 loci in 8,330 serum samples profiled using NMR (Kettunen *et al.* 2012). Another two NMR studies explored mQTLs in 1,757 and 2,893 urine samples and reported associations at 7 and 5 loci, respectively (Raffler *et al.* 2013; Suhre *et al.* 2011a). Additionally, an NMR study in 142 samples identified genetic associations at 3 loci in urine and 1 locus in plasma (Nicholson *et al.* 2011). A subsequent NMR study reported genetic associations at 11 loci in 1,436 urine samples (Rueedi *et al.* 2014). Most recently, associations at 8 loci were identified in 2,118 serum samples using NMR (Demirkan *et al.* 2015). Overall, NMR studies identified more than 60 loci in almost 10,000 individuals assaying more than 1,000 metabolites.

To date mQTLs have been identified in several tissues and samples using numerous metabolite detection platforms. In this chapter I present results from three additional mQTL studies. The findings can help improve our knowledge of inherited part of human metabolic individuality, and also give some potential insight into biological mechanisms involved disease.

3.3 Metabolon

This was a bi-cohort mGWAS collaboration with TwinsUK and KORA samples (Shin *et al.* 2014). The analysis was led by Dr. Shin, a postdoctoral fellow in Dr Soranzo's research group at the Wellcome Trust Sanger Institute. Based on altogether 7,824 adult individuals, this is the most comprehensive assessment of genetic loci in human metabolism to date. I worked on the quality control of metabolites from the TwinsUK dataset for 6,055 individuals, which I will report briefly below. I also then describe the key findings from the collaborative Metabolon mGWAS study, which led me to pursue section 3.5. A subset of the data in this section (1,052 individuals and 280 known metabolites) were previously reported (Suhre *et al.* 2011b) as a meta-analysis dataset for metabolite analyses within the KORA cohort for mGWAS.

3.3.1 Methods

3.3.1.1 Metabolomics Measurements

The TwinsUK dataset generated on the non-targeted MS platform Metabolon has also previously been described (Illig *et al.* 2010; Suhre *et al.* 2011b). The methodology describing the Metabolon procedure for measuring relative metabolite concentrations is described in detail in Chapter 2, section 2.4.1.1.

3.3.1.2 Quality Control and Statistical Analysis

The Metabolon metabolomics dataset first underwent several quality control checks as described in Chapter 2, sections 2.4.1.1 and 2.6. The merged final dataset consisted of 503 metabolites in 6,056 TwinsUK plasma and serum samples. Altogether 486 metabolites overlapped between the two cohorts. First, missing data for each sample and for each metabolite were investigated, and one TwinsUK sample with high missing rate (83%) was excluded but no metabolite was excluded because of data missingness. (Figure 3-1) shows the correlation between missingness numbers within metabolites in

three batches. Pearson correlation coefficients between missingness rates for metabolites in different batches show that batch2 and batch3 had higher correlation (r=0.87) than the correlation with batch1 (r=0.54, r=0.53). This may be due to the fact that batch1 samples are serum, and batch2 and batch3 are measured in plasma.

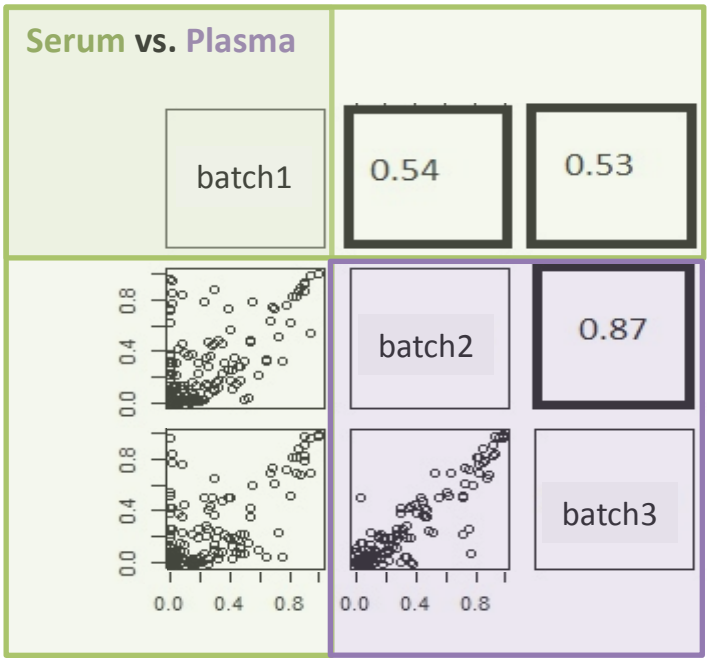

Figure 3-1 Correlation of missingness of metabolites between batches of the TwinsUK Metabolon data.

For the remaining 6,055 TwinsUK samples, the correlation between metabolite missingness rates and experimental batches was assessed. The missingness rate was shown to be correlated with experimental batches, which is likely due to different calibration of the machines on different days. Therefore, experimental batch effect was added as a covariate in the association analysis after PCA analysis. After adjustment for covariates such as age, sex and batch results in a linear model, the distribution of metabolites appeared to be in general normally distributed (Figure 3-2).

A log transformation with base 10 was applied to all the metabolites, following previous work (Suhre *et al.* 2011b). Finally, data points that were more than 4 standard deviations from the mean of each metabolite concentration were excluded (10 metabolite).

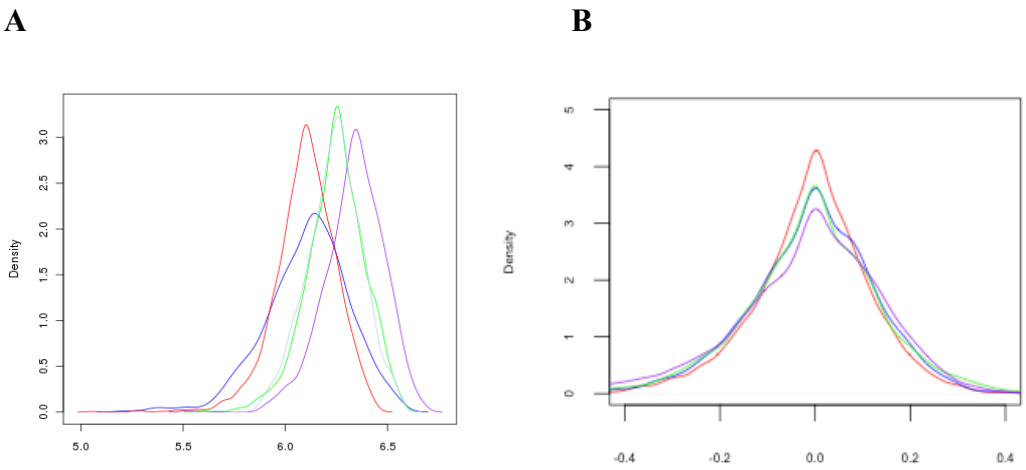

Figure 3-2 Adjustment for covariates. A) Distributions of 4 random selected different metabolites with different colour per metabolite B) Distributions of residuals of 4 random selected different metabolites after adjusting for batch effects

3.3.1.3 GWAS for metabolites

Primary association testing was carried out at each SNP (in the HapMap2–based imputed genotype data set (see Chapter 2)) for all 503 metabolite concentrations present in the TwinsUK data set after quality control steps. Linear regression models were used, assuming an additive genetic model. Age, sex and batch effect were included as covariates in the model. Associations were carried out using the Merlin software by Dr. Shin (Abecasis *et al.* 2002). Briefly, Merlin is based on the variance-component regression model and provides two family-based association tests. Merlin takes pedigree information from direct input of family pedigrees and reported twin status (monozygotic or dizygotic). The TwinsUK mGWAS results were then meta-analyzed with mGWAS results from KORA by Dr.Shin using METAL (Willer *et al.* 2010)

3.3.2 Results

We reported genome-wide significant associations at 145 metabolic loci and their connection with more than 400 metabolites in human blood from meta-analysis of TwinsUK and KORA. These results are presented at Bonferroni-corrected thresholds of $P = 1.03 \times 10^{-10}$ ($5 \times 10^{-8}/486$) and $P = 5.08 \times 10^{-13}$ ($5 \times 10–^{8}/98{,}346$) for metabolites

or metabolite ratios, respectively. The observed effect sizes ranged from -0.07 to 0.3 and nominal P-values ranged from 10^{-10} to 10^{-860} (Figure 3-3).

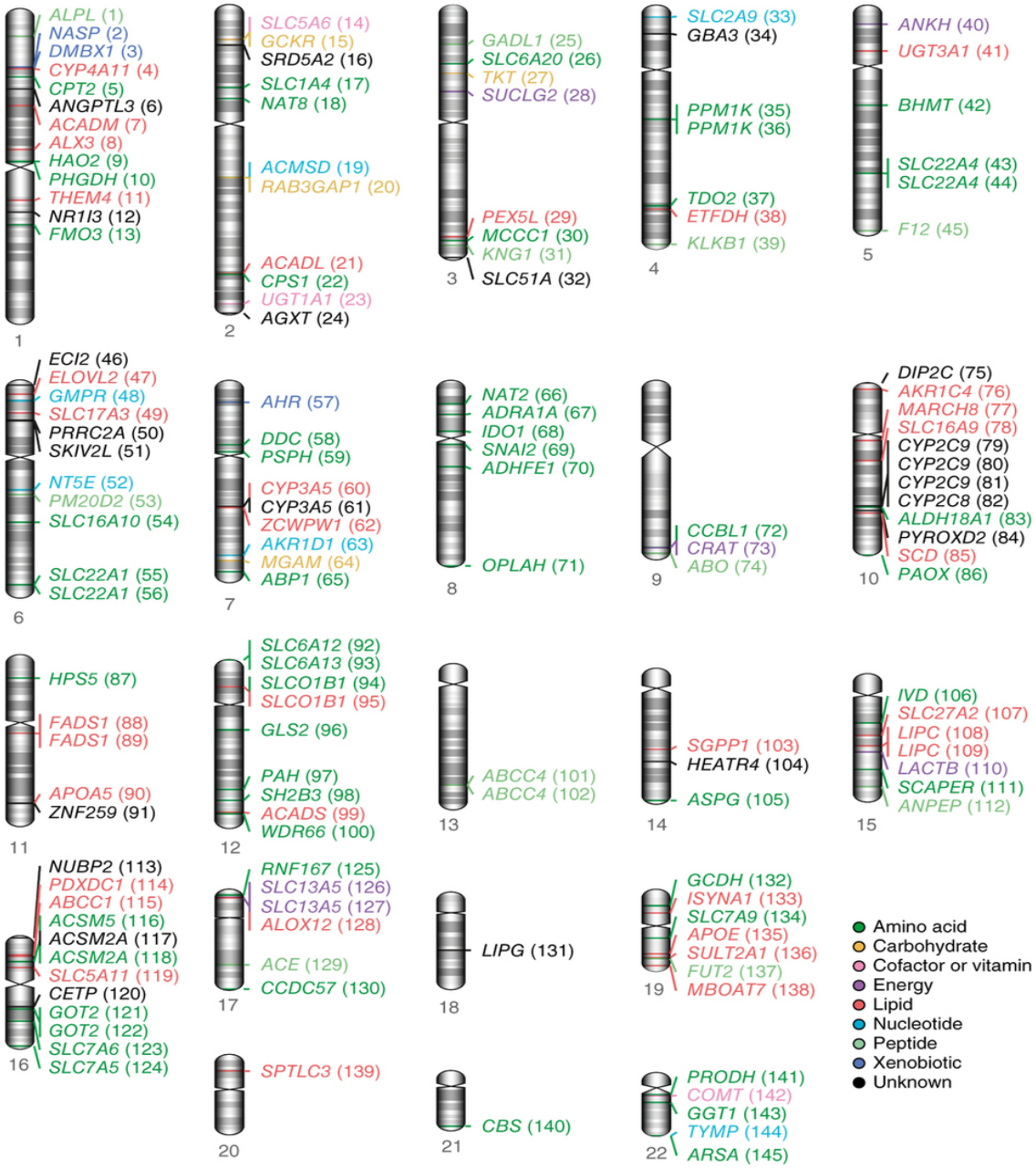

Figure 3-3 Chromosomal locations of the 145 loci identified in this study. Locus label colours (amino acid: green, carbohydrate: orange, vitamin: pink, energy: purple, lipid: red, nucleotide; cyan, peptide: light green, xenobiotic: dark blue and unknown: black) are indicative of the metabolite pathway class of the most strongly associated metabolite at each locus (adapted from (Shin et al. 2014))

3.4 Biocrates

This was a multi-cohort collaboration between TwinsUK and 6 other cohorts (TwinsUK, KORA, EGCUT, LLS, QIMR, ERF and NTR) for the ENGAGE

consortium on the Biocrates platform lead by Dr. Harmen Draisma and Dr. René Pool (Draisma et al. 2015). I worked on the quality control, normalization, and GWAS of Biocrates metabolites from TwinsUK for 1,235 individuals. I will report briefly the key findings, which lead me towards Chapter 3.5. A subset of these data (422 individuals and 163 metabolites) was previously described (Illig *et al.* 2010) as a replication dataset for metabolite analyses within the KORA cohort for GWAS.

3.4.1 Methods

3.4.1.1 Metabolomics Measurements

The TwinsUK dataset generated on the targeted MS platform Biocrates has previously been described (Illig *et al.* 2010; Mittelstrass *et al.* 2011; Römisch-Margl *et al.* 2012). The methodology describing the Biocrates procedure for measuring relative metabolite concentrations is described in detail in Chapter 2, section 2.4.2.1. Further descriptions of the 163 Biocrates metabolites have previously been published (Menni *et al.* 2013a; Mittelstrass *et al.* 2011; Römisch-Margl *et al.* 2012).

3.4.1.2 Quality Control and Statistical Analysis

The Biocrates metabolomics dataset for 163 metabolites first underwent several quality control checks as described in Chapter 2.

After much discussion, a unified analysis plan was agreed on and followed by all cohorts in the collaboration (Figure 3-4). First, I calculated the coefficient of variance (CV) as a measure of the precision and repeatability of the data as:

$$\boldsymbol{CV = \frac{sd}{mean}} \qquad \textit{(Equation 3-1)}$$

for all metabolites and plates. For the CV calculation and thresholds, one reference blood sample was measured five times on each plate across all plates. I excluded all metabolites with a mean CV (of all plates) higher than 25%, as at this threshold the data were considered unreliable and were flagged as outliers because higher values of CV

are generally observed for metabolites that are closer to the detection limit at very low concentrations.

In addition to this criterion, I also excluded all metabolites with a missing value rate larger than 5%. This resulted in the removal of 8 metabolites from the 163 reported metabolites. I then checked for outlier samples and data points having a value greater than mean ± 5sd of all measurements per metabolite and excluded 4 of these. Then, I performed imputation for all missing values using the R-package mice (van Buuren and Groothuis-Oudshoorn 2011), which applies a linear regression approach. I recorded the mean over all imputation iterations. Finally, I performed a transformation of metabolite concentrations (using natural logarithm). I performed the PCA on the level of the individuals using all metabolites and visually checked the first 5 PCs. PC1 by itself can explain 15.4% variance of the data. In total 45.1% of the variance can be explained with the cumulative sum of the 5 PCs. I compared the PC loadings for the first 5 PCs against known covariates (age, sex, and BMI), and observed significant association between the first 2 PC loadings and both age and BMI. In the downstream analyses, according to the consortium agreements the followings covariates were included: age, sex and BMI.

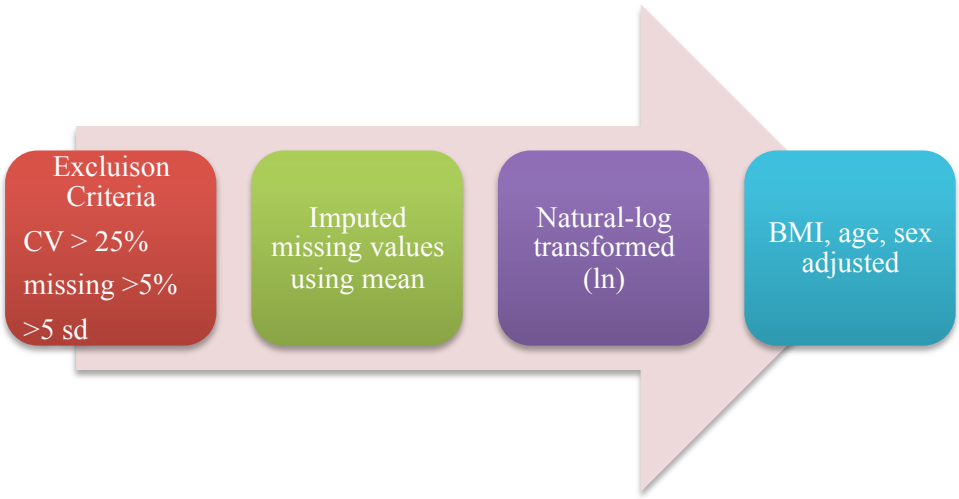

Figure 3-4 Quality Control stages for Biocrates for TwinsUK

3.4.1.3 GWAS for metabolites

I carried out primary association testing at each SNP (in the HapMap 2–based imputed genotype data set (see Chapter 2.2)) for all 151 metabolite concentrations present in the TwinsUK data set after quality control steps. Linear regression models (assuming an additive genetic model) were used. Age, sex and BMI were included as covariates. I performed genetic association tests using GenABEL/ProbABEL (Aulchenko *et al.* 2007; Aulchenko *et al.* 2010). Briefly, GenABEL/ProbABEL allow for genetic association analysis using regression modelling in correlated data, such as twin pairs and involve a two-step design. First, GenABEL fits a polygenic model of the trait incorporating the kinship matrix, and second, ProbABEL performs the genetic association (using the mmscore option). I then selected only the top results from the GenABEL/ProbABEL GWAS, and repeated the association analyses at these using an alternative approach, LMER with lme4 package in R for validation (Bates *et al.* 2015). The selection was applied due to the computational burden of the LMER approach. In the full linear mixed effects model, I regressed metabolite levels on genotype, age, sex and BMI as fixed effects, and zygosity and family structure as random effects. As these data only included batch 1 (serum) samples, batch was not included as a covariate in these analyses, and plate order was also not included as it was not significantly associated with the metabolite principal components as described in section 3.5.1.4 below. This full model was compared to the null model (without genotype) using an ANOVA F-statistic to assess model fit and significance of the genetic association.

3.4.2 Results

I performed mGWAS analysis of 2.5 million SNPs with 151 Biocrates metabolites in 1,235 TwinsUK individuals. I observed 436 genome-wide significant associations at a genome-wide threshold of $P = 3.3 \times 10^{-10}$ (Bonferroni corrected $P = 5 \times 10^{-8}/151$).

Linear mixed effects models were used for verification of the 6 most associated mQTLs (Table 3-2), showing consistent results in the same direction of association.

Table 3-2 Top results from mGWAS analysis for TwinsUK

SNP	LOCUS	CHR	TWINSUK GenABEL P	TWINSUK lme4 P	METABOLITE
rs7094971	*SLC16A9*	10	$3.5x10^{-10}$	$3.6x10^{-10}$	CO
rs2014355	*ACADS*	12	$4.7x10^{-40}$	$2\ x10^{-46}$	C4
rs211718	*ACADM*	1	$4.x10^{-11}$	$2.6x10^{-12}$	C6(C4:1-DC)
rs2286963	*ACADL*	2	$1.2x10^{-35}$	$3.3\ x10^{-38}$	C9
rs2216405	*CPS1*	2	$1.2x10^{-18}$	$9.1x10^{-20}$	Glycine
rs174547	*FADS1*	11	$1.2x10^{-16}$	$1.2x10^{-16}$	PC:aa:C36:4

The TwinsUK mGWAS results were then meta-analyzed with mGWAS results from the other 6 cohorts by two independent analysts from ERF and NTR cohorts using two different software packages (METAL (Willer *et al.* 2010) and GWAMA ((Magi and Morris 2010))). The final GWAS meta-analysis across 7 cohorts, consisting of 7,478 individuals of European descent, was performed for 129 serum metabolites (Draisma et al. 2015). The meta-analysis results identified altogether 4,086 significant associations ($P = 1.09 \times 10^{-9}$), and these involved 54 independent SNPs and 85 metabolites (Figure 3-5).

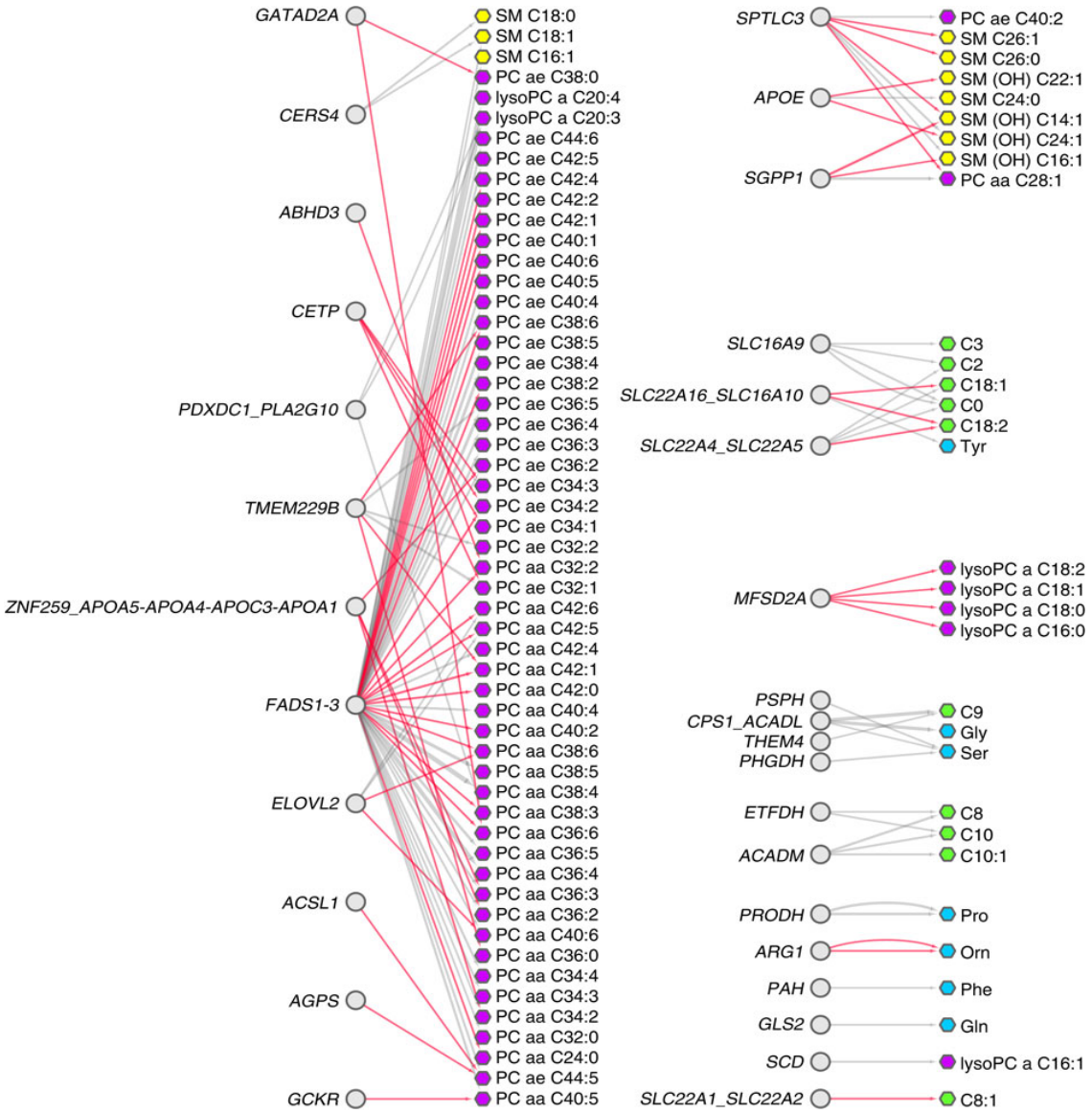

Figure 3-5 Biocrates mGWAS meta-analysis results across 7 cohorts of European descent. Grey circles are loci associated with at least one metabolite. Biochemical classes of the metabolites are indicated by colours: green, acylcarnitines; blue, amino acids; purple, glycerophospholipids; yellow, sphingolipids. Arrows point from each locus to the associated metabolite(s); arrow widths scale linearly with -log10(association P). Grey arrows denote known associations; red arrows denote newly discovered associations on meta-analysis. (figure adapted from (Draisma et al. 2015)).

3.5 Comparison between Metabolon and Biocrates

Together, the two mGWAS studies in the TwinsUK cohort described in the above sections (Chapter 3.1, Chapter 3.2) identified multiple human genetic variants that influence metabolic profiles, and many of these results replicated in additional independent cohorts and appeared as significant findings from mGWAS meta-analysis results. On the other hand, high-throughput metabolomics is typically performed using

either targeted or non-targeted mass-spectrometry platforms, neither of which captures the entire human metabolism. Here, I compared TwinsUK genetic analyses across the two frequently used metabolomic platforms, Biocrates and Metabolon, with the aim of identifying stable metabolites on both technologies to ultimately enable combining metabolite profiles across these two platforms.

Several previous studies have explored metabolomics datasets across multiple platforms (Adamski and Suhre 2013; Büscher *et al.* 2009; Mandal *et al.* 2012; Nicholson *et al.* 2011; Psychogios *et al.* 2011; Suhre *et al.* 2010). For example, (Suhre *et al.* 2010) used multiple metabolomics platforms in a case-control study of T2D. They profiled 100 individuals using three different metabolomics platforms to assess the potential of using metabolomic data in diabetes research by identifying metabolites that associate with diabetes. The study showed good agreement between known biomarkers of diabetes, especially sugar metabolites that could be replicated by multiple metabolomic platforms. Another study (Psychogios *et al.* 2011) aimed to characterize the human serum metabolome by combining targeted and non-targeted NMR, GC-MS and LC-MS methods to identify a comprehensive set of metabolites commonly detected and quantified in human serum samples. They reported good agreement between the measurement concentrations of NMR and GC-MS. However, these studies did not extensively compare genetic association findings for metabolite profiles from the same individuals to assess whether associations from datasets across mass spectrometry platforms overlap.

In this study I analysed serum samples from 1,001 twins who were profiled both using targeted (Biocrates, n=163 metabolites) and non-targeted (Metabolon, n=499 metabolites) mass spectrometry platforms. These 1,001 individuals were also included in the analyses described in sections 3.3 and 3.4. Genome-wide association scans with the high-throughput metabolic profiles (mGWAS) and their ratios (only for 43

overlapping metabolites) were performed for each dataset and the mGWAS results were compared. Although the methods for quantifying metabolites are distinct, I observed overlapping results for several metabolites measured by both platforms suggesting that these are stable and robust metabolites that may be combined or used for replication in future studies.

3.5.1 Methods

3.5.1.1 Study Population and Sample Collection

The 1,001 participants in this study were selected from the TwinsUK cohort (Moayyeri *et al.* 2013a). The sample consisted of 79 MZ twin pairs, 215 DZ twin pairs, and 413 unrelated individuals. TwinsUK blood serum samples for Metabolon and Biocrates platform were obtained after at least 6 hour of fasting and were inverted three times, followed by 40 min resting at 4 °C to obtain complete coagulation. The samples were then centrifuged for 10 min at 2,000g. Serum was removed from the centrifuged tubes as the top yellow translucent layer of liquid. Four aliquots of 1.5 ml were placed into skirted micro-centrifuge tubes and then stored in a −45 °C until sampling.

3.5.1.2 Metabolomics Measurements

The same serum samples from 1,001 individuals in this study were profiled on two separate MS platforms, Biocrates and Metabolon. Biocrates kits were applied to quantify a targeted set of 163 stable metabolites, while Metabolon uses a non-targeted approach for measuring 499 metabolites, as previously described in Chapter 2 and sections 3.3 and 3.4 above.

3.5.1.3 Genotyping and Imputation

Genotyping and quality control assessment of the TwinsUK dataset was previously described in Chapter 2. Here, I used genotype data in 1,001 individuals 2.5 million

directly genotyped and imputed SNPs, with imputation to HapMap2 project (rel 22, combined CEU+YRI+ASN panels) (see Chapter 2 for further details).

3.5.1.4 Quality Control and Statistical Analysis

The Biocrates and Metabolon metabolomics datasets in the subset of 1,001 individuals first underwent several quality control checks as described in Chapter 2. Both datasets were investigated for missingness. Metabolites or individuals with missing values greater than 15% were excluded from further analysis. Additionally, outliers at more than 4 standard deviations from the mean of each metabolite were excluded. In total, 11 metabolites were removed from the Metabolon dataset (out of 499 total) and 3 metabolites were removed from Biocrates dataset (out of 163 total). PCA was performed on the metabolomics profiles in both datasets. I compared the first five PCs with covariates (sex, age, BMI, day) to assess which variables should be included in mGWAS analyses. Sex, age and BMI were nominally associated with at least one PC and as a result included as covariates in the downstream mGWAS analyses.

Altogether, there were 488 (Metabolon) and 160 (Biocrates) metabolites that passed quality control checks, and of these 43 metabolites overlapped (Figure 3-6), that is, were assigned to be the same molecule by both platforms. In 8 cases, specifically for the lyso-phosphatidylcholines, the two platforms actually measure not the same but similar molecules, which differ by having a different position of the chain in the two detection technologies. In my analysis, first, Pearson correlation was computed between the 43 metabolite profiles across platforms to assess the similarities in metabolite measurements. These correlation analyses were extended to compare all metabolites across the two platforms. Additionally, ratios for the matching metabolite pairs (43) within each platform were also included in downstream mGWAS analysis. Both Biocrates and Metabolon datasets used here were log transformed (base 10).

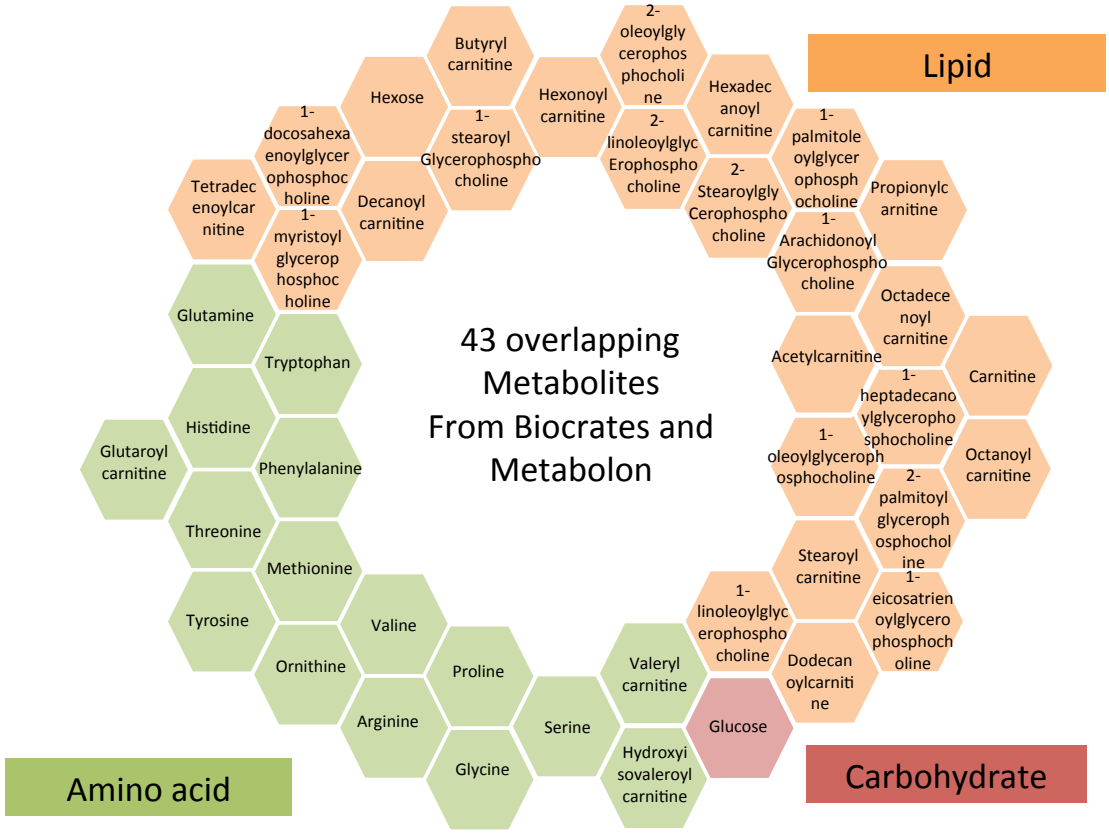

Figure 3-6 Forty-three overlapping metabolites from both platforms separated into 3 pathways (amino acid: green, lipid: orange, carbohydrate: red).

Since the 1,001 individuals included twins, I was able to calculate estimates of twin-based heritability in the metabolite profiles. Heritability was computed for the 43 overlapping metabolites by comparing metabolite profiles in MZ and DZ twin pairs using the ACE model in OpenMx software (Boker *et al.* 2011), as described in Chapter 2. The goal of these analyses was to establish the influence of genetic effects on metabolite profiles, and relate the results to mQTL findings.

To assess evidence for mQTLs, I performed a mGWAS in GEMMA (Zhou and Stephens 2012), which implements a genome-wide efficient mixed model association algorithm specifically suitable for the analysis of related individuals, and provides exact P values from linear mixed models. GEMMA tests for association between each metabolite and each SNP, using one of three commonly used test statistics (the Wald test, the likelihood ratio or score). Here I reported all three but considered the Wald test

when setting the significance threshold. I used Bonferroni correction to account for multiple testing, resulting in genome-wide significance thresholds of $P = 3x10^{-10}$ for Biocrates and $P = 1x10^{-10}$ for Metabolon. Finally, metabolite ratios were also included in the mGWAS analysis, because of the success in mGWAS results for metabolite ratios from previous studies (Gieger *et al.* 2008; Raffler *et al.* 2013; Suhre *et al.* 2010). Ratios were only calculated for the 43 metabolites (Figure 3-6) within each platform. The p-gain statistic (Petersen *et al.* 2012) was used for quantifying the decrease in p-value for the association with the ratio, compared to the p-values of the two corresponding independent metabolite concentrations (Gieger *et al.* 2008). P-gain is calculated as the minimum p-value of the metabolite, which is one of the pair in the ratio (Ma, Mb) and divided by the p-value of the ratio:

p-gain = (minimum (Ma, Mb)/ratio(Ma/Mb)) ***(Equation 3-2)***

The critical value for p-gain is $X/(2x\alpha)$ for type I error rate of α and applying a correction for X tests. When previous studies have analyzed ratios (Illig *et al.* 2010; Suhre *et al.* 2011b), they suggested multiple testing correction should be applied assuming a type I error rate of $\alpha = 0.05$, this leads to a critical p-gain threshold of 903/(2x0.05)=9,030, which implies a Bonferroni correction for 903 tests in this study.

3.5.2 Results

3.5.2.1 Platform comparison: Metabolites profiles

After quality control assessment, there were 488 (Metabolon) and 160 (Biocrates) metabolites available for analysis in serum samples from 1,001 individuals. Of these, 43 were designated as overlapping molecules or very similar molecules by both platforms (APPENDIX A Table S3-1). Comparisons of the 43 metabolites showed a mean correlation coefficient (r) of 0.44 with a maximum correlation for octanoylcarnitine (r=0.92), minimum correlation for 1-docosahexaenoylglycerophosphocholine (r=0), and

weak correlations ($0<r<0.2$) for 7 metabolites (APPENDIX A Table S3-1), which included lipids and an amino-acid. Using hierarchical clustering of the correlation matrix, I observed that metabolites tend to cluster first within platform, and then within type of the metabolite (Figure 3-7). One clear exception is hexose (Biocrates), which clusters with glucose in the Metabolon cluster, as expected. A second exception is C0 (Biocrates), which clusters near proline, valine, tyrosine, and proionylcarnitine in the Metabolon cluster.

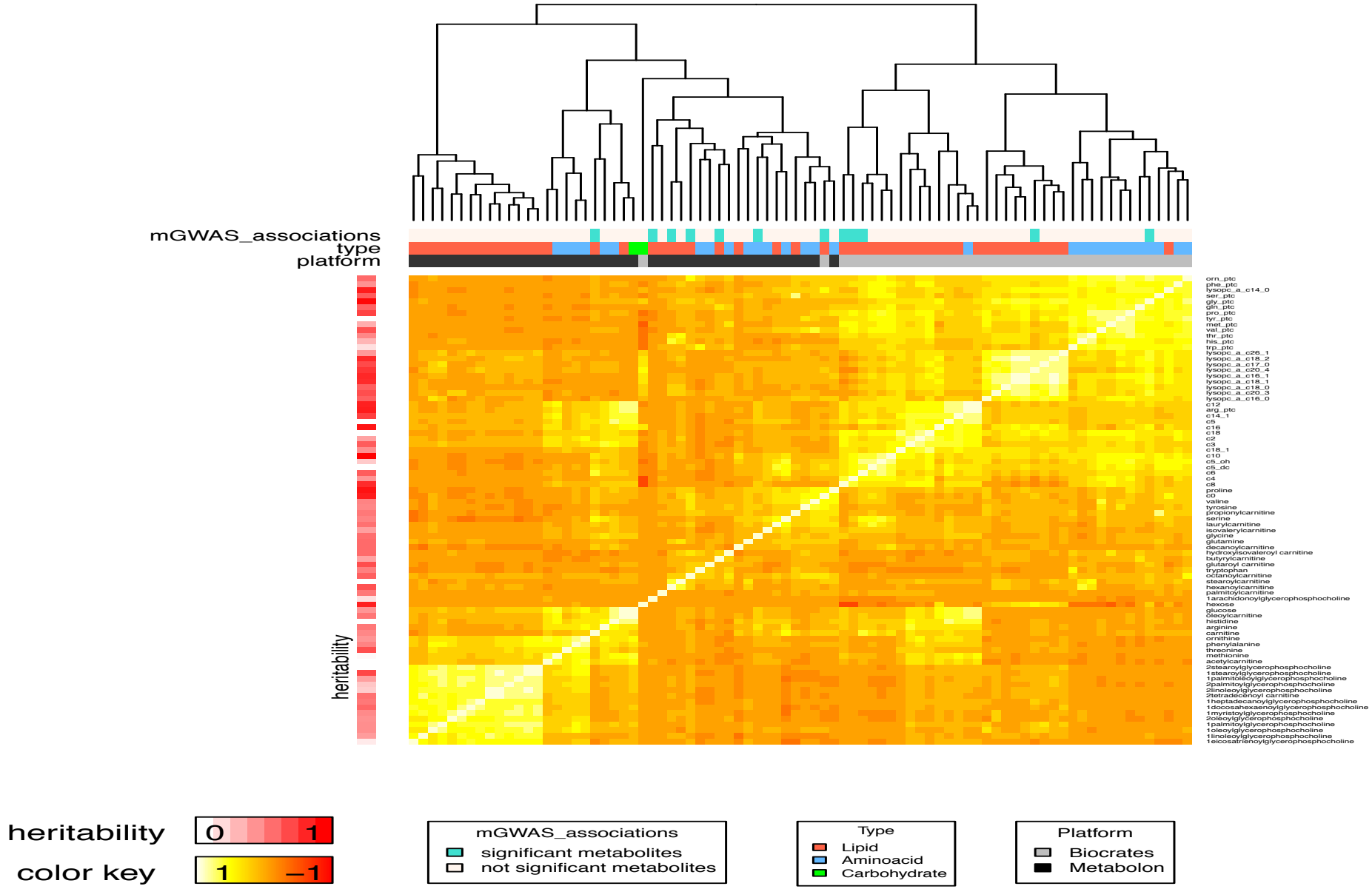

Figure 3-7 Hierarchical cluster of the correlation across 43 overlapping metabolites from both platforms. Upper colour bars represent the significant metabolites, metabolite pathway/type, and platform specification. The left colour bar represents the heritability of the metabolite from red (high) to white (low).

3.5.2.2 Heritability

I calculated twin-based heritability estimates of the metabolite profiles, focusing on the 43 overlapping metabolites (APPENDIX A Table S3-1). Of the 43 metabolites, 37 (Biocrates) and 34 (Metabolon) were at least moderately heritable in twins (h^2>0.2). There were 29 metabolites with evidence for heritability on both platforms (h^2 ranging from 0.29 to 0.72, APPENDIX A Table S3-1). Of these, the 9 most heritable profiles were observed for 6 lipids (h^2: 0.4 to 0.72) and 3 amino acids (h^2: 0.42 to 0.7), indicating that these are stable profiles and highly likely to be under genetic influence.

3.5.2.3 mGWAS results: overlapping and complementary mQTLs

In total, 488 and 160 metabolites were tested separately on the Metabolon and Biocrates platforms in two mGWAS analyses. All significant association results are reported at a stringent Bonferroni cut-off: $P = 1\times10^{-10}$ ($5\times10^{-8}/488$) for Metabolon and $P = 3\times10^{-10}$ ($5\times10^{-8}/160$) for Biocrates (APPENDIX A Table S3-2).

In total, 61 genome-wide significant metabolite associations were identified at 26 independent loci: 42 metabolites were associated with 25 loci on the Metabolon platform, and 19 metabolites were associated with 8 loci on the Biocrates platform (Table 3-3, APPENDIX A Table S3-2). Of the 26 independent loci, genome-wide significant metabolite associations at 7 loci were identified on both platforms. There were 19 loci that had associations only with metabolites from one platform (18 loci in Metabolon and 1 locus in Biocrates). All results reported as significant with single metabolites or ratios here were also reported at a relaxed significance threshold ($P = 1x10^{-5}$) in the TwinsUK results from sections 3.3 and 3.4 despite the use of different GWAS programs (MERLIN and GENABEL).

Table 3-3 Significant mGWAS results

	Loci[a]	All associated metabolites	Associated metabolites from set of 43 overlapping metabolites[b]
Metabolon (M)	25	42	6
Biocrates (B)	8	19	7
Overlap	7	22(13M + 9B)	6
Total	26	61 (35M+12B+7M&B+7B&M)	13

[a]Unique loci

[b]Metabolites with significant mGWAS results from the set of 43 matching metabolites only. In all cases the reciprocal platform mGWAS result surpassed nominal significance with the same direction of association.

3.5.2.4 Overlapping mQTLs: genetic associations identified on both platforms

Associations at 7 independent loci were identified in both platforms, namely with SNPs in the regions of the *ACADS, ACADM, ACADL, FADS1, SGPP1, SLC16A9* and *CPS1* genes (Table 3-4). The 7 loci associate with 22 metabolites in total: 9 metabolites from Biocrates and 13 metabolites from Metabolon.

Table 3-4 mGWAS results at 7 loci associated with metabolites in both platforms

Gene	Chr	Peak SNP	Biocrates ($P = 3\times10^{-10}$)	Metabolon ($P = 1\times10^{-10}$)
	1	rs211718		*X-11421 (3.8×10^{-8})
ACADM	1	rs4949874	C6 (4.1×10^{-11})	Hexanoylcarnite (1.62×10^{-13})
	1	rs2172507	*C8 (2.4×10^{-8})	Octanoylcarnitine (4.8×10^{-11})
ACADL	2	rs7601356	C9 (9.7×10^{-38})	
		rs12612970		X-13431 (3.5×10^{-25})
CPS1	2	rs4673553	Glycine (5.27×10^{-17})	Glycine (7.1×10^{-27})
				X-08988 (1.6×10^{-11})
SLC16A9	10	rs1171614	C0 ($4.6\text{x}10^{-12}$)	
	10	rs1171617		Carnitine (2.3×10^{-13})
FADS1	11	rs174546	*PC ae C42:5	*1-linoleoylglycerophosphoethanolamine
	11	rs174547	(1.9×10^{-8})	(1.19×10^{-8})
	11	rs174547	lysoPC aa C20:4	*1-arachidonoylglycerophosphocholine
			(2×10^{-14})	(2.9×10^{-10})
				*arachidonate (20:4n6) (5.5×10^{-10})
ACADS	12	rs2066938	c4 (2.9×10^{-44})	Butyrylcarnitine (1.75×10^{-114})
SGPP1	14	rs7157785	*PC aa C28:1	1-stearoylglycerol (2.77×10^{-14})
			(3.8×10^{-8})	*X-10510 (1.24×10^{-9})

*Shown at a relaxed genome-wide cut-off ($5\text{x}10^{-8}$)

Of the 22 associated metabolites, 6 metabolites associated with 5 loci were named for the overlapping metabolite compound on both platforms. These included C6 (Biocrates,

P = 4.1×10^{-11}) = Hexanoylcarnite (Metabolon, P = 1.6×10^{-13}), C8 (Biocrates, P = 2.4×10^{-8}) = Octanoylcarnitine (Metabolon, P = 4.8×10^{-11}), Glycine (Biocrates, P = 5.27×10^{-17}) = Glycine (Metabolon, P = 7.1×10^{-27}), C0 (Biocrates, P = 4.6×10^{-12}) = Carnitine (Metabolon, P = 2.3×10^{-13}), C4 (Biocrates, P = 2.9×10^{-44}) = Butyrylcarnitine (Metabolon, P = 1.75×10^{-114}), and lysoPC aa c20:4 (Biocrates, P = 2×10^{-14}) = 1-arachidonoylglycerophosphocholine (Metabolon, P = 2.9×10^{-10}), as designated by Biocrates and Metabolon, respectively. For three of the 5 loci with matching named metabolites, there were also associations with other metabolites, which do not necessarily match across platforms (Table 3-4).

In one case genetic variants in locus *ACADL* were associated with both a Biocrates metabolite C9 (P = 9.7×10^{-38}) and an unknown Metabolon metabolite (X-13431 (P = 3.5×10^{-25})), which were recently shown to be identical molecules (Krumsiek *et al.* 2012).

In one case, metabolite associations with genetic variants at the *SGPP1* locus did not match exactly in name for PC aa C28:1 (Biocrates) and 1-stearoylglycerol (Metabolon) (Table 3-4). However, both of these are lipid metabolites, which could share the C18:0 fatty acid chain.

3.5.2.5 Complementary mQTLs: genetic associations identified in only one platform

There were 19 loci that had associations only with metabolites from one platform (18 loci in Metabolon and 1 locus in Biocrates) and all were associated with metabolites that were not measured in the other platform (APPENDIX A Table S3-2).

The 18 Metabolon-specific mGWAS results included associations with 29 metabolites. Of these 29 metabolites, 17 were unknowns, 4 were lipids and 3 were amino acids and these were not included in Biocrates, considering that Biocrates consists mostly of lipids

and amino acids. The 5 remaining metabolites were 2 drugs, a carbohydrate, a nucleotide, and a peptide.

There was only 1 locus (*DYNC1H1*) where genetic variants showed significant mGWAS results on the Biocrates platform only with 4 metabolites, and in all 4 cases these were with lipids that Metabolon does not measure.

3.5.2.6 mGWAS of metabolite ratios and p-gain

In the second stage of the analysis I focused on mGWAS of 903 (43x42/2) ratios for the overlapping 43 metabolites, and results were reported at a threshold of $P = 5x10^{-11}$ (5e-8/903) (APPENDIX A Table S3-3). In the 903 mGWAS there were 104 significant ratios on Metabolon, and 167 significant ratios on Biocrates, and 101 of these ratios overlapped, that is, 101 (same) ratios had significant mGWAS results on both platforms (APPENDIX A Table S3-4).

I next wanted to assess whether ratios had more evidence for mGWAS signal, compared to their corresponding single metabolite mGWAS results, which might indicate that these metabolite pairs are also linked to each other in the biological process. I calculated the p-gain statistic for the 101 significant overlapping ratios from both platforms (APPENDIX A Table S3-4). In total 25 ratios from Biocrates and 14 ratios from Metabolon showed a significance (p-gain>9030) when using metabolites as pairs in ratios, which also indicates that ratios may contain more information than the two corresponding metabolite concentrations alone. These 25 and 14 metabolites fall in 4 loci, of which 3 overlap. The 4 loci were identified; *ACADM, CPS1, FADS1* and *ACADS* (APPENDIX A Table S3-3). Of these, the first locus had 1 matching mQTLs, the second locus (*CPS1*) is only identified by Biocrates, the third locus had 9 matching mQTLs, and the fourth locus had 2 matching mQTLs (APPENDIX A Table S3-3). All of the mGWAS ratio loci identified here overlapped with the previously reported 7 overlapping mGWAS main effects loci (Figure 3-8).

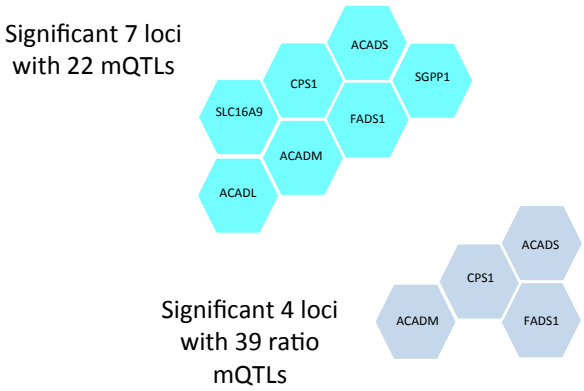

Figure 3-8 Seven loci reported with mQTLs and four loci reported with ratio mQTLs

3.6 Discussion and Conclusion

This Chapter describes the results of two large-scale collaborative mGWAS studies and a bi-platform metabolite comparison using mGWAS with the objective of identifying metabolites measured on more than one platform where signals overlap and may be combined in future studies, for example for replication analysis.

The key results in Chapter 3.3 identified genetic impacts on over a hundred individual metabolites in the largest mGWAS study to date, and included many novel findings. These findings may lead to a better understanding of inherited variation in blood metabolic diversity and propose potential new possibilities for drug development and for understanding disease.

The key results in Chapter 3.4 also identified genetic effects on many individual metabolites, and included novel findings (5 novel loci associated with serum metabolites). Given the targeted nature of the metabolomics platform used in this section (Biocrates, focusing on lipids) these results may give new insights specifically to cardiovascular and metabolic disease.

The key results in Chapter 3.5 identified 7 loci showing genetic associations with metabolites on both platforms. Moderate correlation and heritability estimates were obtained for these metabolites and these results were predominantly consistent with

recent reported mGWAS (Illig *et al.* 2010; Suhre *et al.* 2011b), some of which are based on results from extended cohorts that include the samples used in the current analysis. The findings support the hypothesis demonstrating the complementary nature of both MS platforms where a combination of targeted and non-targeted MS can identify a wider range of metabolites than applying each platform separately.

Positive correlation was observed when comparing metabolomic profiles at the 43 overlapping metabolites measured on both platforms (mean r=0.44). Of the 43 metabolites that overlapped, 37 and 34 metabolites were moderately heritable in data from Biocrates and Metabolon respectively (h^2>0.2). Overall, there is a moderate correlation between matching metabolites and many of these profiles showed evidence for heritability. These results suggested that performing mGWAS analysis might be a suitable approach to identify more specific metabolite overlaps and potential shared pathways.

The major mGWAS finding was that 7 unique loci showed genome-wide significant association with metabolites on both platforms. Genetic variants at 5 of these loci were associated with metabolites that were named for the overlapping compound across platforms, and in 2 locus associations were only obtained for non-matching metabolites and unknown metabolites from both platforms. Of the metabolites associated with the 7 loci, 5 metabolites (Biocrates C8, C6, C0, C4, and Glycine) had at least moderate heritability (h^2>0.26) and correlation (r>0.38) on both platforms, showing that these profiles were stable and reproducible across platforms. 1 matching metabolite, lysoPC a C20:4 [Biocrates] / 1—arachidonoylglycerophosphocholine [Metabolon], showed low heritability in one platform (0.09 in Metabolon and 0.59 in Biocrates platform) and showed relatively low correlation (r=.29) across platforms, but was still identified to associate with the same locus from both platforms at genome-wide significance. This observation may be due to the difference in the measured compounds between the two

platforms: while Metabolon specifically quantifies the lysoPC with the 20:4 fatty acid chain at *sn1* position of the glycerol backbone (lysoPC(20:4/0:0), Biocrates does not distinguish between the lysoPCs with fatty acid chains at *sn1* and *sn2* positions and only quantifies the sum concentration of the two forms (lysoPC(20:4/0:0 and lysoPC(0:0/20:4). Moreover, the quality of measurement differs for various lipids between the targeted Biocrates and the non-targeted Metabolon platform, which might also cause lower correlation between the corresponding matching metabolites. Notably, despite those differences inherent in the platforms both profiles give a robust signal of genetic association for *FADS1*.

Further comparison of the GWAS results across platforms shows that genetic variants at 5 of the 7 loci (*ACADM*, *CPS1*, *SLC16A9*, *FADS1*, *ACADS*) were associated with metabolites that were named for the overlapping compound. However, genetic variants at the *ACADL* and *SGPP1* loci only associate with non-overlapping metabolites or unknown metabolites from the Metabolon platform. One potential use of these results is to inform the function of unknown metabolites on the Metabolon platform or identify metabolites that belong to the same or related biological pathways. For example, variants in 1 locus (*ACADL*) associated with the C9 Biocrates metabolite and also with the unknown X-13431 Metabolon metabolite, which were recently reported to be the same molecule (Krumsiek *et al.* 2012). When I explored the results for similar association patterns, I observed that Metabolon metabolites X-10510 and 1-stearoylglycerol shared mQTL findings within the same locus (*SGGP1*) as the Biocrates metabolite PC aa C28:1. These results suggest a link between the molecules, where the more specific Metabolon lipid chain length can hint that the PC aa C28:1 association is possibly driven by the involvement of a 18:0 lipid chain. Alternatively, the *SGGP1* genetic variant (rs7157785) has also been associated with sphingomyelin 14:0 in a separate study (Zhou and Stephens 2012). Our platform does not include this

metabolite, but X-10510 may be also related to this sphingolipid pathway. This assumption is further supported by high partial correlation between X-10510 and other Metabolon sphingolipid molecules and genetic associations to a second sphingolipid related gene in Shin *et al.* (Shin *et al.* 2014).

Additionally, I explored the results for the 43 overlapping metabolites on both platforms to check if inconsistencies across platform signals were observed. Expectedly, the mean correlation between the 43 matching metabolites (r=0.44) is higher than the mean correlation with all metabolites between the two platforms (r=0.17). Exceptions include correlations of Biocrates metabolites with Metabolon metabolites of yet unknown chemical identity. In these cases, the high correlation could indicate matching metabolites or biochemically related metabolites and might thus again assist in the identification of unknown metabolites. There were 15 of the 43 metabolites that were highly correlated with reasonable h^2 estimates on both platforms, but no matching mQTLs were identified.

Eight metabolites of the 43 were very weakly correlated, but had moderate h^2. Four lyso-phosphatidylcholines metabolites (lysoPC a C16:0, lysoPC a C18:0, lysoPC a C18:1, lysoPC a C18:2) from the Biocrates platforms had overlapping metabolites on the Metabolon platform, but neither contained matching mQTLs nor showed high heritability or correlation. I conclude that in this instance the two platforms are likely measuring distinct signals that cannot be combined or this may be due to a relatively lower quality of measurement for these lipids on the Metabolon platform, since it is not specifically aimed at targeting lipids like Biocrates. In another example, the carnitine (C0) Biocrates metabolite showed moderate correlation (r = 0.39) with the carnitine Metabolon metabolite, and both had evidence for heritability and mapped to the same genetic variant in *SLC16A9*. These findings confirm that the carnitine signal is stable across platforms despite the observation that the carnitine Biocrates metabolite is also

well correlated with other Metabolon metabolites (proline, valine, tyrosine, and propionylcarnitine), as observed in the hierarchical cluster analysis.

Focusing on the mGWAS results that were only identified in one platform, then variants in 18 loci associated with Metabolon profiles alone, while variants in only 1 locus associated with Biocrates profiles. Overall, the two platforms are designed to focus on different metabolites, and these significant findings can enlighten on platform-specific metabolites. Eventually, combining metabolomics profiles across platforms is more informative than single-platform analysis because these platforms are complementary. In contrast to other "-omics", it is not possible to assay the entire metabolome with one platform due to large differences in the physiochemical properties of the different metabolites (e.g. lipophilic and hydrophilic metabolites).

I identified 4 loci significantly associated with metabolite ratios and these are a subset of the previously reported 7 loci associated with main effects of metabolites on both platforms (APPENDIX A Table S3-3, Figure 3-8). P-gain is a well defined measure that can be used to identify statistically significant metabolite ratios in association studies. The p-gain results for the overlapping ratios from both platforms suggest for the 25 (Biocrates) and 14 (Metabolon) cases ratios may contain more information than the two corresponding metabolite concentrations alone. (APPENDIX A Table S3-4). One possible interpretation of the associations with the 4 ratio loci may be that a SNP affects a biochemical reaction where both molecules are linked to a substrate or both are linked to a product, and the effect of the genetic variant is to deplete one of these molecules.

In summary, I identified genetic associations at 7 loci with metabolite profiles from both the Biocrates and Metabolon platforms and 4 loci out of these 7 loci were also significantly associated with metabolite ratios. The results contain information about potential shared metabolic pathways, as well as distinct metabolite profiles, and can clarify unknown metabolites. They can also guide further research on the genetic

determination of metabolites for new studies. These findings also inform the reliability of both platforms and demonstrate the complementary nature of both targeted and non-targeted MS platforms implying that future studies should combine datasets across platforms where possible, especially for replication of metabolite hits when datasets are profiled on different platforms. I identified metabolite signals that show consistent genetic associations and therefore appear stable and robust across multiple platforms, suggesting that these metabolomic profiles can be combined across platforms. These findings are informative for future studies of comparative and integrative metabolomics analyses in human samples.

CHAPTER 4

DNA Methylation

4.1 Introduction

The aim of this chapter is to investigate the influence of genetic effects on DNA methylation. Several factors are thought to influence DNA methylation profiles: genetic variation, other epigenetic mechanisms, environmental factors, and stochastic changes that accumulate during life (Bell and Spector 2011). Previous studies have explored the heritability of DNA methylation across the genome and identified meQTLs in several tissues (Banovich *et al.* 2014; Bell *et al.* 2011; Bell *et al.* 2012; Drong *et al.* 2013; Fraser *et al.* 2012; Gamazon *et al.* 2013; Gibbs *et al.* 2010; Grundberg *et al.* 2013; Gutierrez-Arcelus *et al.* 2013; Shi *et al.* 2014; Smith *et al.* 2014; van Eijk *et al.* 2012; Wagner *et al.* 2014; Zhang *et al.* 2010), and here I aim to extend this work by exploring genetic effects on DNA methylation in twins. From part of this work, I prepared a review on genetic and environmental impacts on DNA methylation levels in twins (Yet *et al.* accepted for publication).

4.1.1 DNA methylation Heritability

Variation in DNA methylation between individuals has been shown in many studies (Boks *et al.* 2009; Flanagan *et al.* 2006; Kaminsky *et al.* 2009). Heritability studies have explored the degree of DNA methylation variation and the extent to which this variation can be explained by genetic influences. In one of the initial genome-wide DNA methylation studies in twins, Kaminsky *et al.,* profiled methylation in white blood cells and buccal epithelial cells from 194 twins using a 12K CpG island microarray (Kaminsky *et al.* 2009). Consistent with evidence for heritability, the study reported that

DZ twins showed greater epigenetic differences in buccal cells compared to MZ twins. Gervin *et al*., reported low heritability in 89 twin pairs in whole blood when analysed bisulfite sequencing on 1760 CpG-sites within the major histocompatibility complex (MHC), but methylation differences within MZ pairs were lower than those observed for DZ pairs (Gervin *et al.* 2011). Another study using Illumina 27k array in whole blood samples from 172 twins demonstrated a range of heritability estimates per CpG-site with a genome wide average of 0.18 (Bell *et al.* 2011). Another study using Illumina 27k DNA methylation profiling determined the epigenetic variation present in three tissues from 22 MZ and 12 DZ newborn twin pairs and estimated the contribution of genetic and common and unique environment to DNA methylation profiles (Gordon *et al.* 2012). Twin pairs showed low mean heritability across the genome in all tissues (h^2= 0.12), but high heritabilities were detected for some of the individual CpG sites. Grundberg *et al*. profiled Illumina 450K adipose methylome data from 648 adult twins (Grundberg *et al.* 2013). DNA methylation heritability estimates were at a genome-wide average of 19%, however, the mean heritability estimate was significantly higher (h^2 = 0.34) when only including variable probes. Overall, many studies have explored DNA methylation profiles in twins, reporting higher similarity between MZ twins compared to DZ twins, and proposing that genetic effects contribute to DNA methylation levels in some regions of genome. The average reported methylation heritability at CpGs across the genome is low to moderate (12%-19%), while the heritability of one CpG can show a very wide range.

4.1.2 Genetics of DNA methylation: meQTLs

A number of studies have assessed the association between genetic variation at particular loci and DNA methylation patterns across the genome to discover genetic impacts on DNA methylation levels. The genetic loci at which associations are identified are referred to as methylation quantitative trait loci (meQTLs). Evidence for

meQTLs has been explored on a genome-wide scale using high-throughput DNA methylation analyses, identifying local (*cis*) and distal (*trans*) genetic variants associated with methylation levels in multiple samples, across a number of cells, tissues, and ages (Banovich *et al.* 2014; Bell *et al.* 2011; Bell *et al.* 2012; Drong *et al.* 2013; Fraser *et al.* 2012; Gamazon *et al.* 2013; Gibbs *et al.* 2010; Grundberg *et al.* 2013; Gutierrez-Arcelus *et al.* 2013; Shi *et al.* 2014; Smith *et al.* 2014; van Eijk *et al.* 2012; Wagner *et al.* 2014; Zhang *et al.* 2010).

Two of the initial genome-wide studies explored evidence for meQTLs in the human brain tissue (Gibbs *et al.* 2010; Zhang *et al.* 2010). Gibbs *et al*. assessed DNA methylation on the Illumina 27k and reported that up to 5% of CpG-sites had a meQTL and the bulk of signals were observed in *cis*, that is, within 1 Mb window of the CpG-site in tissue samples from four brain regions from 150 individuals (Gibbs *et al.* 2010). Additionally, SNPs that were identified as a *cis* meQTL were also observed to affect gene expression, or were expression QTLs (eQTLs) across tissues (Gibbs *et al.* 2010). Zhang *et al.* also used the Illumina 27k and reported 736 *cis* meQTLs in brain tissue (Zhang *et al.* 2010). Shared genetic control of both DNA methylation and gene expression is detected where 13% of the meQTLs also regulated expression of the gene closest to the DNA methylation site (Zhang *et al.* 2010). Overall, both studies observed over 800 *cis* meQTLs in human brain tissues. Additionally, Gamazon *et al.* (Gamazon *et al.* 2013) identified that Bipolar Disorder (BD) GWAS SNPs were *cis* meQTLs for 132 CpGs from the same brain tissue samples that was reported in previous study (Zhang *et al.* 2010).

Many studies have observed meQTL effects in other samples including, whole blood, blood cell subtypes, and blood-derived cells (Banovich *et al.* 2014; Bell *et al.* 2011; Bell *et al.* 2012; Fraser *et al.* 2012; Gutierrez-Arcelus *et al.* 2013; Smith *et al.* 2014; van Eijk *et al.* 2012). Presence of meQTLs in lymphoblastoid cell lines (LCLs) from 77 HapMap

Yoruba individuals were explored using the Illumina 27k array (Bell *et al.* 2011). The authors observed *cis* meQTLs affecting 180 CpG-sites in 173 genes, and only a small amount of *trans* meQTLs, and detected an enrichment of meQTL SNPs for eQTL effects. A following study assessed DNA methylation using Illumina 27k in 180 LCLs from two different populations using the HapMap Yoruba and CEPH samples (Fraser *et al.* 2012). Population specific patterns of DNA methylation were observed at almost half of the genes. A later study re-examined DNA methylation profiles in 64 LCLs derived from the HapMap Yoruba individuals using Illumina 450k array (Banovich *et al.* 2014). The authors observed many more (13,915) *cis* meQTLs and showed that QTLs underlying other regulatory genomic processes, such as gene expression, chromatin accessibility, and histone modifications highly overlap with meQTLs. Gutierrez-Arcelus *et al.* used Illumina 450k array in LCLs, T cells and fibroblasts derived from 204 umbilical cords from healthy newborns (Gutierrez-Arcelus *et al.* 2013). The authors reported over 20,000 methylation QTLs and expression-methylation associations, with greater level of tissue differentiation at methylation sites with meQTLs, compared to non-meQTL CpG-sites. Another study in whole blood samples from 148 individuals on the Illumina 27k array reported 575 *cis* meQTLs, which also had effects on gene expression levels (van Eijk *et al.* 2012), and were predominantly located outside of CGIs. Another study in whole blood explored DNA methylation profiles in Illumina 27k from 172 twin samples and reported 1,537 CpG sites with *cis* meQTLs that were predominantly not related to age-related differential methylation signals (Bell *et al.* 2012).

Methylation QTL effects have also been observed in other cell types and tissues, including adipose tissue (Drong *et al.* 2013; Grundberg *et al.* 2013), lung tissue (Shi *et al.* 2014) and fibroblasts (Wagner *et al.* 2014). In adipose tissue, Drong et al. explored 38 unrelated individuals for DNA methylation levels using differential methylation

hybridization and gene expression microarray (Drong *et al.* 2013). Methylation levels of 149 regions were reported having at least one SNP in *cis* (meQTL), but no overlap with eQTLs was explored (Drong *et al.* 2013). Grundberg *et al*. reported almost 100,000 *cis* meQTLs in the adipose tissue profiling 648 twins using Illumina 450k array, and found that 6% of the loci played a role in regulating both DNA methylation and gene expression (Grundberg *et al.* 2013). A recent meQTL study in lung tissue observed 34,304 *cis* meQTLs and 585 *trans* meQTLs across 210 individuals, with validation of the signals in breast and kidney tissues, and described genomic profiles of CpG-sites under *cis* and *trans* genetic control (Shi *et al.* 2014). In fibroblasts, Illumina 450k DNA methylation profiles in 62 unrelated individuals revealed evidence for association with genetic variants in *cis* at 1,676 CpG-sites (Wagner *et al.* 2014). Overall, meQTLs were identified in several tissues and cell types and can help improve our knowledge of the genetic component of gene regulation (Bell *et al.* 2011).

Variance QTL

Over last decade, research has concentrated on unravelling how genetic variation contributes to phenotypic variability in population. Many factors can influence the variance of a complex phenotypic trait, and these can include genetic effects, environmental effects, epigenetic effects, gene-environmental interaction, gene-gene interaction, and stochastic factors. A common approach to explore the impact of genetic variants on the trait is based on assessing differences in the genotypic means of the trait, aiming typically to identify SNPs that are associated with complex phenotypes. However, these methods do not consider the possibility that QTLs may contribute to the amount of variability of the phenotype. As a result, much interest has now concentrated on exploring QTLs that are associated with the variability of a phenotype (var QTLs). The mechanisms of var QTL action on the trait remain largely unknown, but these may be explained in part by interactions between QTL with other genetic or environmental

factors (Hill and Mulder 2010), that is, the QTLs involved in epistatic or gene-environment interactions may manifest as var QTLs.

So far, var QTLs have been identified across different organisms including plants (Ordas *et al.* 2008; Shen *et al.* 2012), chickens (Ronnegard and Valdar 2011), and humans. In humans, five var GWAS studies have been published so far. In 2010, Pare *et al.* identified two var QTLs impacting on BMI or smoking (Pare *et al.* 2010). This was followed by other studies identifying var QTLs for BMI (Yang *et al.* 2012), high-density lipoprotein (HDL) cholesterol (Surakka *et al.* 2012), and expression (Brown *et al.* 2014; Hulse and Cai 2013). Hulse and Cai first reported *cis* and *trans* var eQTLs in the human genome (Hulse and Cai 2013) using gene expression levels in LCLs (n=210 individuals). This study identified for the first time 218 distinct genes that were associated with 379 *cis*-acting var eQTLs and even more trans-acting var eQTLs (500 representative genes and 13,000 SNPs) in the human genome. The majority of studies to date exploring var QTL effects on trait variability in humans have focused on unrelated individuals, and have used Bartlett's (Yang *et al.* 2012) or Levene's test (Pare *et al.* 2010; Struchalin *et al.* 2010) of variance or have proposed a specific statistical models such as the double generalized linear model (DGLM) (Ronnegard and Valdar 2011) and squared residual value linear modelling (SVLM) (Struchalin *et al.* 2012) to formally test for association between genotype and trait variance in independent groups of individuals of different genotypes.

Another study design to consider in this scenario is MZ twins, and specifically the discordance in MZ twins as a measure of trait variability. It has been proposed that MZ twins are powerful study design to explore these effects compared to population-based approaches (Visscher and Posthuma 2010). Surakka *et al.* were the first to use the MZ study design genome-wide in a meta-analysis of 8 twin cohorts to exploring var QTLs associated with lipid levels (Surakka *et al.* 2012). They reported a locus associated with

HDL variability. One more recent study explored var eQTL using RNA-sequence data from 765 LCLs from the TwinsUK cohort and they identified a set of 508 variance associated SNPs (Brown *et al.* 2014) Recently, it has been proposed that var QTLs that change the variability of a phenotype could be mediated by DNA methylation (Feinberg and Irizarry 2010). This proposed model in which DNA methylation variation is a mediator between genetic contribution and phenotypic variability has been investigated in rheumatiod arthritis (Liu *et al.* 2013). An EWAS was performed to determine differentially methylated positions (DMPs) associated with rheumatoid arthritis followed by a GWAS for each of the DMPs. Using a causal interference test 9 unique DMPs were determined as mediating the genetic component to rheumatiod arthritis disease risk. Interestingly, an association between genotype and methylation variability was found in 5 out of 9 determined DMPs, although in all cases the associated var QTL was also a QTL for the level of DNA methylation. However, the finding of 5 QTLs that are associated with DNA methylation variation and consequently with increased rheumatiod arthritis disease risk, fits within the hypothesis that var QTLs that change the variability of a phenotype could be mediated by DNA methylation.

In this chapter I investigated epigenetic profiles in monozygotic twins with aim of identifying QTLs and var QTLs for DNA methylation in whole blood and adipose tissue. I tested the evidence that genetic variants influence not only DNA methylation levels, but also DNA methylation variability. I assessed the evidence that genetic effects influence DNA methylation variability by using MZ-twin discordance as a measure of variance, aiming to identify meQTLs for DNA methylation variance (var meQTLs). This Chapter is divided into methods and 5 main results sections including blood meQTLs, blood variance meQTLs, validation of variance meQTLs, gene-environment interaction follow-up analysis, and meQTLs across tissues.

4.2 Methods

4.2.1 Datasets & QC

We obtained DNA methylation Illumina 450k profiles at 485,000 CpG dinucleotides in whole blood samples from 330 MZ female pairs, and in adipose tissue for 83 MZ female pairs with an overlapping number of 49 MZ pairs. In addition, the overall sample of 789 individuals (these included the 330 MZ pairs) were used to follow up QTL results with gene-environment interaction analysis and 459 unrelated individuals (from the 789) were used in the validation of var meQTLs. All samples were from TwinsUK as previously described in Chapter 2 (Moayyeri *et al.* 2013b; Spector and Williams 2006).Methylation profiling was performed on the bisulfite-converted DNA samples with the Illumina 450k array and the resulting datasets first underwent several quality control checks as described in Chapter 2.

4.2.2 Heritability

A PhD student in the epigenetics group in the department, Mr Juan Castillo-Fernandez calculated the heritability in blood for 442,307 probes using the Open Mx software (Boker *et al.* 2011). In a sample of 330 MZ pairs and 25 DZ pairs using the ACE model, the heritability is moderate (h^2 >40%) for 18% of the CpG sites.

4.2.3 Genotyping and Genotype imputation

TwinsUK imputed genotypes were obtained for the 1000 genomes reference set (1000G), as described in Chapter 2. For the genetic variants used in this, I excluded SNPs with Hardy–Weinberg P $<1x10^{-4}$ and MAF <5% (in the 330 and 83 twin pairs) and those with IMPUTE info value <0.8.

4.2.4 Estimating genetic impacts on trait variance

The dispersion of a dataset is measured most commonly by the variance, standard deviation, and interquartile range. Measures of dispersion such as variances are difficult to estimate accurately, for example compared to means. Therefore the detection of var

QTLs, which measure genetic impacts on trait variability, will require more observations to attain good power to detect effects, compared to QTLs that measure genetic impacts on trait means (Lee and Nelder 2006; Visscher and Posthuma 2010). In addition, the variance is a measurement of dispersion around the mean and therefore is affected by the mean effects. Over the past few years, various statistical methods have been developed to address the challenge of identifying var QTLs (Ronnegard and Valdar 2012). These methods are typically models that fit variance and mean using a two-stage approach. Three studies evaluated var QTLs using DGLM that provide a framework for modelling variance and the mean simultaneously using the dglm package in R (Hulse and Cai 2013; Ronnegard and Valdar 2011; Wang *et al.* 2014). On the other hand, Struchalin *et al.* developed an R package, VariABEL, which uses squared residual value linear modelling (SVLM) and is a two-step process (Struchalin *et al.* 2012). It first regresses the trait for a SNP effect and other covariates and uses the squared residuals in a second regression with SNP as the predictor. Another widely used methods are Bartlett's test and Levene's test for assessing the equality of variance across independent groups (eg unrelated individuals of different genotypes).

One approach to increase the power to detect var QTLs is to use a monozygotic (MZ) twin-pair design. MZ twin-pairs are considered genetically identical and a phenotypic discordant MZ twin-pair design thus provides a powerful natural setting to assess the contribution of environmental or stochastic factors to the phenotype. Using a phenotype discordant MZ twin-pair design one could identify QTLs that determine the extent of discordance. This phenotypic discordance is one measure of trait variability, and it may be due to environmental or stochastic differences within the MZ co-twins. Because MZ twins are nearly genetically identical, the phenotype discordant MZ twin-pair design would control for possible epistatic effects. Several studies have considered MZ twin discordance as a measure of trait variance, and a couple have further used this measure

to identify var QTL effects. Surakka *et al.* calculated both absolute differences and means for twins then transformed the values and adjusted for batch effects as well as by mean effect using a linear regression model (Surakka *et al.* 2012). The second study by Brown *et al.* identified var eQTLs across tissues using MZ twins (Brown *et al.* 2014). In their approach the maximum expression of one of the twins in the pair was regressed on the minimum expression in the twin pair and genotype of the twin pair. This explored whether the association between maximum and minimum expression was conditional on genotype.

My approach for estimating var meQTL was similar to that used by Surakka *et al.*. I fitted the absolute methylation discordance of the MZ pair on batch effects and the mean methylation level of the pair, and I then extracted the residuals and took the square of the residuals as a measure of variance. First, I calculated the methylation mean and absolute differences in methylation values for each MZ pair at each CpG site. Then I regressed the absolute difference on the mean and kept the residuals from the regression and squared them (Equation 4-1).

$$Y = residuals(absolute\ difference(twin1 – twin2) \sim mean(twin1, twin2))^2 \qquad \textit{(Equation 4-1)}$$

In the next step, I used the square of the residuals (Y in Equation 4-1) as a phenotype in the additive genetic model in the variance meQTL analysis, described in section 4.2.3 below. I had one measure of variance (Y) for each MZ twin pair. The var meQTL analysis was on the unit of the MZ pair for 330 and 83 independent MZ pairs in blood and adipose samples, respectively.

4.2.5 Genetic association testing

The primary associations between DNA methylation levels and genotype were tested in a sample of 330 healthy MZ female pairs. Because the aim was to identify genetic

effects influence on both DNA methylation levels and DNA methylation variability, I used a two-step approach for both.

First, the normalized methylation betas were fitted with in lme4 package in R (Bates *et al.* 2015) as the outcome on the level of the individual, and the predictors consisted of smoking, BMI, age, plate and blood cell count estimations as fixed effects, and family and zygosity as random effects, using R. Residuals from this model were extracted and normalized to N(0,1), and then used in the second step. In the second step, I applied two different models:

1) I calculated the methylation mean for the MZ pair to use as a phenotype for meQTL analysis, and

2) I calculated the methylation absolute discordance and mean for the MZ pair, fitted Equation 4.1 and took the square of the resulting residuals per MZ pair as a phenotype for the variance meQTL analysis.

I then normalized the phenotype values in steps 1) and 2) above to N(0,1) before the QTL analysis. QTL analysis was performed using association tests in the R package, matrix eQTL (Shabalin 2012). I ran an additive genetic model in matrix eQTL, that is:

$$Methylation = \alpha + \gamma \cdot genotype\_additive \qquad \textit{(Equation 4-2)}$$

where methylation is the normalized value from 1) or 2) above, alpha(α) is intercept and gamma(γ) is the additive genetic effect. I specifically tested for the significance of γ and reported the t-statistic from the association analysis.

4.2.6 Multiple Testing

To estimate the significance level taking into account multiple testing (5 million SNPs and 442,307 DNA methylation profiles), I used a permutation-based FDR approach for both *cis* and *trans* threshold, because a Bonferroni correction is very stringent as neither

the genetic nor epigenetic data were independent. I ran two different permutations separately for *cis* and *trans* results.

In each permutation, I kept the structure of the epigenetic data and permuted the twin pair labels for the genetic data. I then permuted the datasets 10 times. I calculated FDR for each p-value threshold. That is, for a given p-value threshold I counted the number of observations (SNP-CpG associations) in the real results that surpassed this p-value threshold (No). I then also counted the number of observations (SNP-CpG associations) in each permutation that surpassed this p-value threshold, and took the average number across the 10 permutations avg (Ne). I then calculated FDR as (Ne/No). An FDR = 5% corresponds to a *cis* QTL nominal $P = 4x10^{-5}$ and *trans* QTL nominal $P = 8x10^{-9}$.

4.3 Results

In this chapter, I assessed evidence for genetic impacts on DNA methylation levels and variances in MZ twins. I first investigated whether there are any missing values in probes, then removed 17,764 probes in Illumina 450k that mapped to multiple locations within 2 mismatches (the same number of probes were identified using both hg18 and hg19) using MAQ ((Li *et al.* 2008)). Then, I investigated the genetic impacts on 442,307 DNA methylation CpG-sites genome-wide. Finally, I investigated whether any of the probes have missing CpG sites contained SNPs in the probe sequence, which may affect cross-hybridization and created spurious meQTL signals. Based on a comprehensive assessment reported by Naeem *et al.* (Naeem *et al.* 2014), I later excluded all probes that were known to contain SNPs.

4.3.1 meQTL results in 330 MZ twins in whole blood

At an FDR of 5% ($P = 4x10^{-5}$), I detected *cis* meQTLs for 53,813 autosomal CpG probes (11% of 442,307 autosomal probes), mapping to 11,693 genes. I defined the *cis* interval as 100 kb upstream and downstream from the targeted CpG site (200 kb total). A 200kb window size was chosen because previous *cis* meQTL studies have used this

window size (Shi *et al.* 2014), and because results from a simulation study suggest that the multiple testing burden becomes too great when expanding the search window beyond 200kb such that the number of detected SNP-CpG pair results declines after 200 kb (Luijk *et al.* 2015).

A *cis* meQTL was defined as a SNP that was significantly associated with DNA methylation level at a CpG, and was located within the *cis* interval from that CpG-site. Figure 4-1 shows the QQplot for the *cis* meQTL signals at the CpG site with the most significant meQTL overall (probe cg23097878 with most associated rs7945565 P = 4.90×10^{-112}) and the boxplot presents the methylation levels by genotype at this *cis* meQTL. Table 4-1 reports the top 10 associations genome-wide.

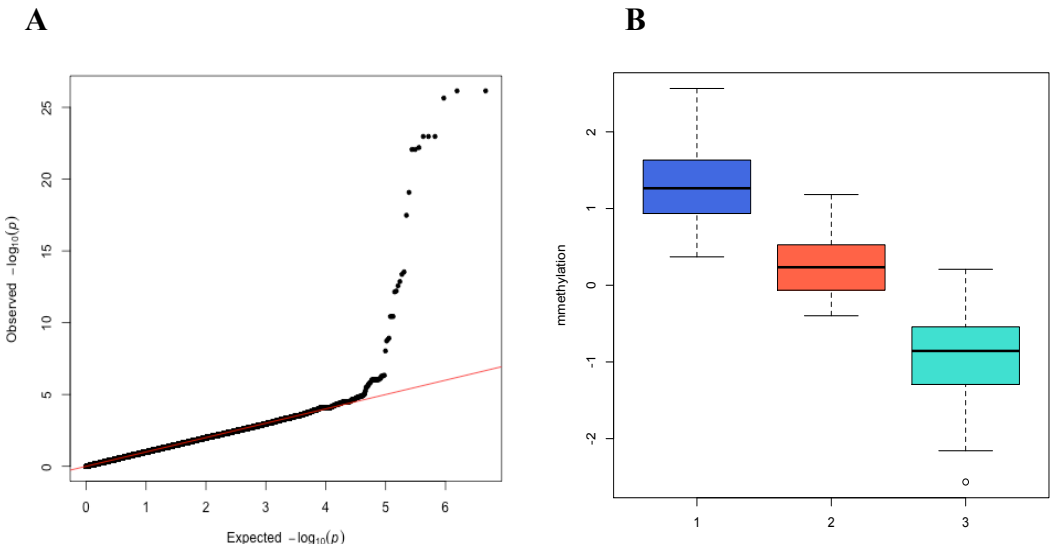

Figure 4-1 QQplot of the top significant cis meQTLprobe, cg23097878 (A), and barplot showing the mean methylation levels across three genotype categories for the most associated SNP for this probe (rs7945565, P = 4.90×10^{-112}) (B).

Table 4-1 The 10 top-ranked CpG-sites with cis meQTLs

SNP	Probe	Beta	t-stat	P	Chr	Gene
rs7945565	cg23097878	1.27	34.77	4.90×10^{-112}	11	*CRY2*
rs12218872	cg17191567	-1.24	-34.15	4.97×10^{-110}	10	-
rs2438251	cg02082929	1.24	33.98	1.79×10^{-109}	2	*GCC2*
rs4074915	cg21028142	-1.25	-33.81	6.42×10^{-109}	17	*NPLOC4*
rs690461	cg24786174	1.27	33.15	9.84×10^{-107}	18	*ZNF516*
rs4980503	cg04850017	1.25	32.62	5.48×10^{-105}	11	*RCOR2*
rs6677965	cg01427815	1.23	32.49	1.48×10^{-104}	1	*UCK2*
rs804227	cg07863524	1.25	32.34	4.65×10^{-104}	17	*OR3A4*
rs7174099	cg17847044	-1.27	-32.29	6.87×10^{-104}	15	*MAPKBP1*
rs132717	cg11420782	1.26	32.19	1.57×10^{-103}	22	*APOL4*

Next, at an FDR at 5% ($P = 8 \times 10^{-9}$) I detected *trans* meQTLs at 15,392 CpG probes (3% of 442,307 autosomal probes), mapping to 6,380 genes. The most significant trans meQTL is in probe cg02082929 with most associated rs826698, $P = 1.77 \times 10^{-105}$) in GCC2 gene. I next considered the characteristics of the genetic variants that were identified as meQTLs, focusing on the most-associated SNP per CpG-site. The genetic variants associated with methylation levels in *cis* were overrepresented in regions close to the methylation site (±100 kb) and specifically at regions very close to the DNA methylation site (1kb) (Figure 4-2).

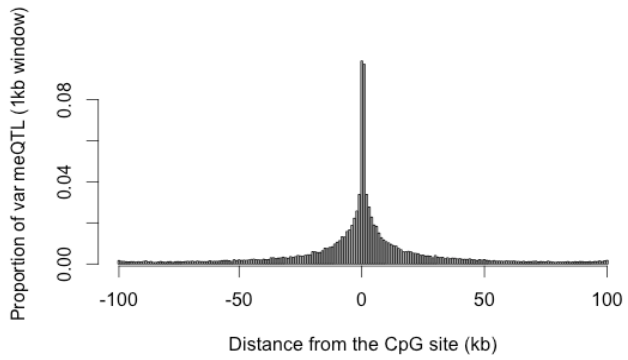

Figure 4-2 Genomic location of the most significantly associated SNP per CpG-site for cis meQTLs

The MAF distribution for the meQTLs is presented in Figure 4-3, and shows no bias towards rare or common MAFs (0% to 50%) in this meQTL set.

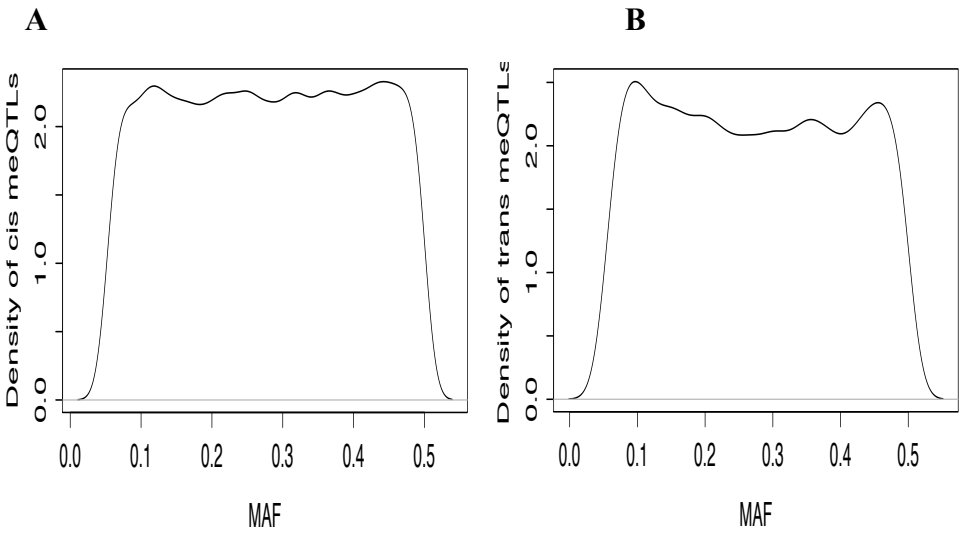

Figure 4-3 Distribution of MAF at the most significantly associated SNP per CpG-site for A) cis meQTLs, and B) trans meQTLs.

4.3.2 Variance meQTL results in 330 MZ twins in whole blood

At an FDR of 5% (P = 4x10-5), I detected cis var meQTLs (2% of 442,307 autosomal probes) for 8,106 CpG probes, mapping to 3,539 genes. I detected 3,694 CpG probes with trans var meQTLs (FDR = 5%, P =8x10-9), mapping to 2,411 genes. The sequence variants associated with the methylation traits were overrepresented in regions close to the methylation site (Figure 4-4). The MAF distribution for top hits is presented in Figure 4-5, and most of the hits from both *cis* and *trans* results fall in between MAF of 10% and 20% and targeting least common alleles occurring in the sample. Figure 4-6 shows the QQplot for the most significant meQTL probe and the boxplot represents the methylation levels across genotype categories at the most associated *cis* var meQTL SNP. Additionally, the overlap between *cis* meQTL and *cis* var meQTL is 7,220 (89%) and for *trans* results it is 987 (27%) showing that almost all of the *cis* var meQTLs are also *cis* meQTLs and quarter of *trans* variance signals are also *trans* meQTL signals. Table 4-2 presents the top ten associations that have been identified as cis var meQTL.

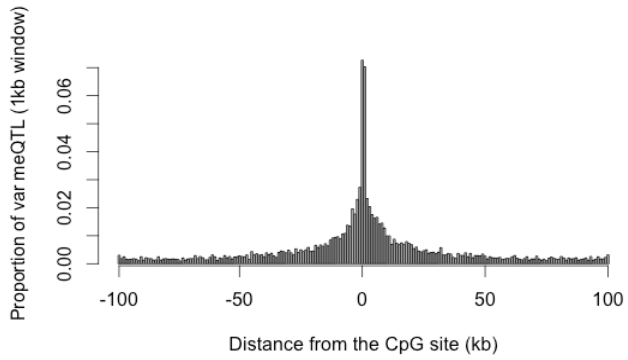

Figure 4-4 Genomic location of most associated cis var meQTL SNPs

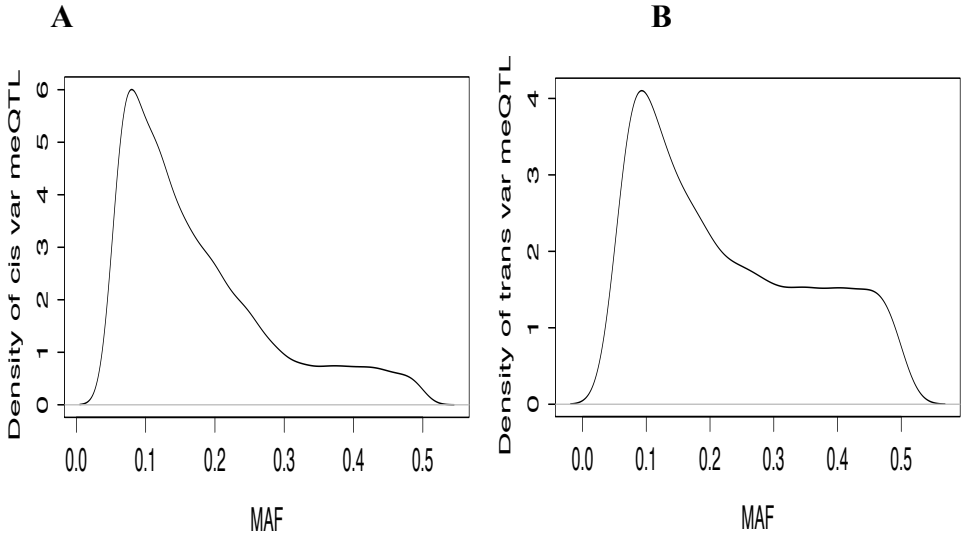

Figure 4-5 Distribution of MAF at the most-associated SNPs for var meQTLs,in A) cis var meQTLs, and B) trans var meQTLs. All cis var meQTLs have a genotype count of at least 5 in each genotype class.

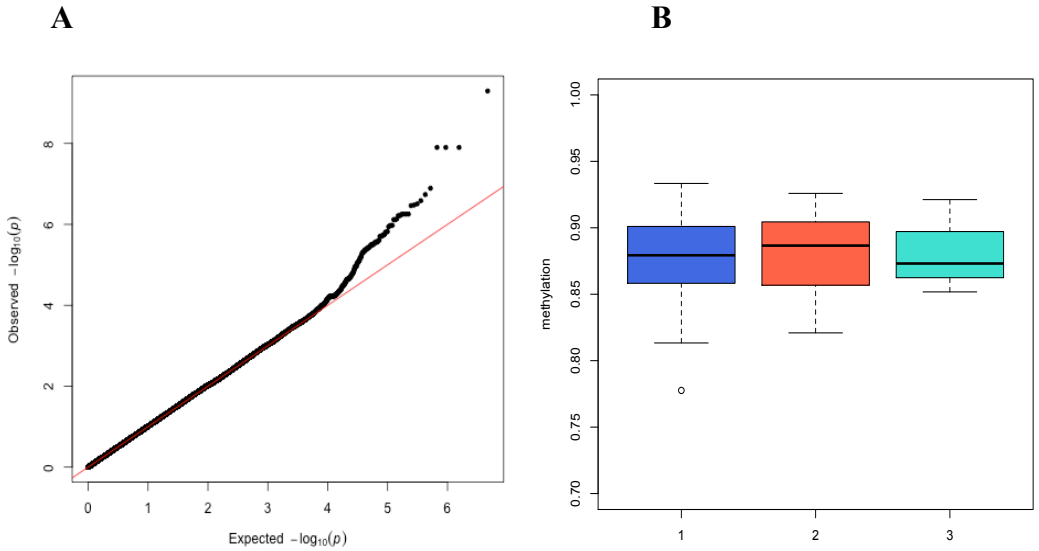

Figure 4-6 QQplot of top significant cis var meQTL (rs6100252, cg08091561, P = 3.8x10^{-42}) (A), and barplot showing the squared residual methylation levels across three genotype categories (B).

Table 4-2 Top 10 probes identified in cis var meQTL

SNP	Probe	Beta	t-stat	P	Chr	Gene
rs6100252	cg08091561	1.21	15.76	$4.88x10^{-42}$	20	*GNAS*
rs6100252	cg06200857	1.17	14.86	$1.46x10^{-38}$	20	*GNAS*
rs79171026	cg12594803	1.42	14.70	$6.19x10^{-38}$	3	*DLG1*
chr17:4899230:D	cg19427746	1.21	14.41	$8.00x10^{-37}$	17	*CAMTA2*
rs61766781	cg15603964	1.29	14.26	$3.22x10^{-36}$	1	*UBR4*
rs62195287	cg17674726	1.36	14.17	$6.64x10^{-36}$	2	*ITM2C*
chr8:74716807:D	cg12682382	1.26	14.05	$1.91x10^{-35}$	8	*UBE2W*
rs1952512	cg02988727	1.60	13.89	$8.20x10^{-35}$	14	*ZNF219*
rs3012041	cg12829045	-1.24	-13.82	$1.45x10^{-34}$	10	-
rs7736222	cg11100481	1.42	13.78	$2.11x10^{-34}$	5	-

4.3.3 Do var meQTLs capture gene-environment interactions?

Variance meQTL may capture interaction effects due to gene-gene or gene-environment interactions. As MZ twins are nearly genetically identical, this design is better suited to try to uncover gene-environment interactions that underlie var meQTLs. Therefore, including environment covariates should be undertaken with caution in var meQTL GWAS, because these may obliterate gene-environment interactions underlying the var meQTL signals. Here, I explored the idea that the var meQTLs identified in the previous section may capture some gene-environment interactions, specifically for smoking. Smoking has one of the strongest impacts on whole blood DNA methylation levels to date, where multiple studies have identified hundreds of smoking-associated DNA methylation signals, and many of these have large effects that have been replicated by multiple studies (Besingi and Johansson 2014; Breitling *et al.* 2011; Dogan *et al.* 2014; Shenker *et al.* 2013; Zeilinger *et al.* 2013). To explore whether the identified var meQTLs may capture some smoking-genetic interactions on DNA methylation levels, here I repeated the var meQTL analyses, but not controlling for smoking. I termed these new var meQTLs as 'no-smoking adjusted var meQTLs'. I then assessed if there was evidence for gene-smoking interactions at the SNPs and methylation probes that formed the 'no-smoking specific var meQTLs'. 'No-smoking

specific var meQTLs' were defined as 'no-smoking adjusted var meQTLs' that were not within the 'smoking adjusted var meQTLs' identified in section 4.3.2 above. I used the same 330 MZ pairs to characterize 'no-smoking adjusted var meQTLs' by using the above-mentioned methods but without adjusting for smoking. I then tested for evidence of gene-smoking interactions on DNA methylation levels at the CpG-sites that had 'no-smoking specific var meQTLs', and these interactions tests were performed in the largest available sample of 789 individuals, which included the sample of 330 MZ pairs.

No-smoking specific var meQTL results

At an FDR at 5% ($P = 4x10^{-5}$), I detected no-smoking adjusted cis var meQTLs for 8,119 CpG probes (2% of 442,307 autosomal probes), mapping to 3,558 genes. I detected 3,698 CpG probes with no-smoking adjusted trans var meQTLs (FDR = 5%, $P = 8x10^{-9}$) mapping to 2,114 genes. For the identified cis signals, the distributions for distance to probe and MAF both show similar results to those observed for cis var meQTL (Figure 4-4, and Figure 4-5).

I wanted to explore if correcting for smoking has a major effect on the total number of var meQTLs obtained, but the results show that this is not true. There were 8,119 *cis* and 3,698 *trans* results in the as 'no-smoking adjusted var meQTLs', and there were 8,106 *cis* and 3,694 *trans* results in the as 'smoking adjusted var meQTLs'. I then estimated the number of no-smoking specific var meQTLs as 581 *cis* var meQTLs and 1,268 *trans* var meQTLs and these were the results that were only observed in the var meQTL analysis where I did not adjust for smoking (Table 4-3).

Table 4-3 No-smoking specific var meQTLs: CpGs identified as var meQTLs only when not controlling for smoking.

Whole blood results (442,307)	***cis* ($P = 4x10^{-5}$)**	***trans* ($P = 8x10^{-9}$)**
No-smoking specific var meQTLs	581	1,268

Table 4-4 reports the top ten probes identified from the no-smoking specific var meQTL analyses.

Table 4-4 Top 10 probes identified in no-smoking specific var meQTL

SNP	Probe	Beta	t-stat	P	Chr	Gene
rs6100252	cg08091561	1.20	15.59	$2.25x10^{-41}$	20	*GNAS*
rs6026560	cg06200857	1.17	14.83	$1.98x10^{-38}$	20	*GNAS*
rs79171026	cg12594803	1.41	14.65	$9.38x10^{-38}$	3	*DLG1*
chr17:4899230:D	cg19427746	1.22	14.54	$2.66x10^{-37}$	17	*CAMTA2*
rs61766781	cg15603964	1.30	14.34	$1.47x10^{-36}$	1	*UBR4*
rs62195287	cg17674726	1.37	14.27	$2.72x10^{-36}$	2	*ITM2C*
chr8:74716807:D	cg12682382	1.26	14.17	$6.82x10^{-36}$	8	*UBE2W*
rs1952512	cg02988727	1.60	13.89	$8.32x10^{-35}$	14	*ZNF219*
rs3012041	cg12829045	-1.24	-13.83	$1.32x10^{-34}$	10	-
rs6510081	cg26353507	-1.45	-13.83	$1.34x10^{-34}$	19	*ZNF773*

Gene-smoking interactions on DNA methylation

I then focused on the 581 *cis* var meQTLs and 1,268 *trans* var meQTLs that were only identified as no-smoking specific var meQTL. I explored if these QTLs were also identified as gene-smoking interactions. For gene-smoking interaction analysis I used the ModelLINEAR_CROSS model within matrix eQTL (Shabalin 2012) with smoking as a covariate, that is,

$$Methy = \alpha + \beta \cdot smoking + \gamma \cdot genotype\_additive + \delta \cdot genotype\_additive \cdot smoking \quad \textit{(Equation 4-3)}$$

where α is intercept, β is the smoking effect, γ is the additive genetic effect and δ is the interaction between the additive genetic effect and smoking. I specifically tested for the significance of δ (sigma), reporting the t-statistics from the association analysis.

I performed these gene-smoking interaction analyses genome-wide in the dataset of 789 twins, which included the 330 MZ pairs. At a relaxed significance threshold of $P = 1x10^{-3}$, I detected evidence for *cis* gene-smoking interactions at 11,573 CpG probes (4% of 442,307 autosomal probes). At a relaxed significance threshold of $P = 1x10^{-6}$, I

detected evidence for *trans* gene-smoking interaction sat 99,259 CpG probes (2% of 442,307 autosomal probes). I then focused specifically on the SNP-CpG pairs that formed the no-smoking specific var meQTL results and assessed whether they were also observed as gene-smoking interactions at a relaxed significance threshold.

Table 4-5 summarize the results of the overlap between the no-smoking specific var meQTL and the gene-smoking interaction results at a relaxed significance threshold. Altogether there were 29 no-smoking specific *cis* var meQTLs that also had evidence for gene-smoking interactions, and 20 of them mapped to genes. There were 474 no-smoking specific *trans* var meQTLs that also had evidence for gene-smoking interactions, and 380 of these fall into genes.

I tested the observed overlap compared with the overlap expected under the null hypothesis using a resampling approach. For the results in cis, I first selected 581 probes at random out of all 442,307 probes. I then selected 11,573 probes at random out of all 442,307 probes. I then counted the number of overlapping probes in the two selected sets. I repeated this procedure 1 million times and created an empirical distribution of the number of overlapped items, and I observed an empirical p-value for 29 overlapping elements of P = 0.04 for the results in cis. For the results in trans, I first selected 1,268 probes at random out of all 442,307 probes. I then selected 99,259 probes at random out of all 442,307 probes. I then counted the number of overlapping probes in the two selected sets.

I repeated this selection 1 million times and created an empirical distribution of the number of overlapping items, and I observed an empirical p-value for 474 elements of P = 0.03 for the results in trans. I conclude that there are more overlapping items than expected by chance.

Table 4-5 Number of probes identified in no-smoking specific var meQTL and gene-smoking interactions analysis in whole blood

Whole blood results (442,307)	***cis* ($4x10^{-5}$)**	***trans* ($8x10^{-9}$)**
No-smoking specific var meQTL (a)	581	1,268
Gene-smoking interaction (b)	11,573 ($P = 1x10^{-3}$)	99,259 ($P = 1 x10^{-6}$)
Overlap between (a) and (b)	29	474

Table 4-6 shows the ten top-ranked gene-smoking interaction results that were also no-smoking specific *cis* var meQTLs.

Table 4-6 Ten top-ranked gene-smoking interaction results that were also no-smoking specific cis var meQTLs.

SNP	**Probe**	**Beta**	**t-stat**	**P**	**CHR**	**Gene**
rs2461266	cg07775417	-0.33	-4.40	$1.26x10^{-05}$	4	*AGPAT9*
rs13170638	cg18703511	0.44	4.31	$1.88x10^{-05}$	5	-
rs77326891	cg04742977	0.65	4.08	$5.06x10^{-05}$	14	*TTC7B*
rs11696169	cg15535174	-0.48	-4.03	$6.24x10^{-05}$	20	*C20orf117*
chr6:26328364:D	cg00631329	-0.46	-4.00	$7.06x10^{-05}$	6	-
rs4896242	cg14553824	0.31	3.87	$11.60 x10^{-05}$	6	*IL22RA2*
rs34381573	cg19135247	0.41	3.87	$11.91 x10^{-05}$	21	*SIK1*
rs144770262	cg13849647	0.57	3.82	$14.25 x10^{-05}$	9	*DCAF12*
rs6836337	cg15440376	-0.59	-3.82	$14.35 x10^{-05}$	4	-
rs34365416	cg01821684	-0.42	-3.79	$16.17 x10^{-05}$	7	*RAC1*

The top *cis* result was obtained for a methylation CpG site in the AGPAT9 (1-acylglycerol-3-phosphate O-acyltransferases), which is a protein coding gene associated with ovarian and colorectal cancer (Agarwal 2012; Currie *et al.* 2013; Wang *et al.* 2012). For *trans* meQTLs, the second top result was obtained for Calcium-activated chloride channel ANO1 which promotes breast cancer progression by activating EGFR and CAMK signalling (Britschgi *et al.* 2013).

4.3.4 Validation of var meQTLs in 459 unrelated individuals

Validation analysis was performed for var meQTLs in 459 unrelated individuals selected from the 789 individuals (these included some of the individuals in the original 330 MZ sample). I used two different variance analyses for this var meQTL analysis. First, I adjusted methylation for all covariates as discussed in 4.2.3 and additionally for the mean of each trait (or CpG site), and then I quantile normalized the residuals to a standard normal distribution. I then regressed the squared transformed residuals, which are a measure of variance (Visscher and Posthuma 2010), on the genotype indicator variable of each SNP to test for association of the SNP with methylation variability. I only validated the reported results for the smoking-adjusted var meQTL (8,106 *cis* and 3,694 *trans* signals). Second, I applied the Bartlett's test to these signals, which is a test for variance heterogeneity. I reported these results from Bartlett test if they reached nominal significance (P = 0.05). All of the *cis/trans* var meQTL results were validated at nominal significance using either measure for testing variability. As an example, Table 4-7 shows the ten top-ranked var meQTL results were validated with using two alternative variance methods with 459 unrelated individuals.

Table 4-7 Top 10 var meQTLs were validated with 459 unrelated individuals

SNP	Probe	Chr	Gene	P (regression)	P(Bartlett)
rs6100252	cg08091561	20	*GNAS*	$4.58x10^{-12}$	$8.26x10^{-05}$
rs6100252	cg06200857	20	*GNAS*	$2.49x10^{-03}$	0.037
rs79171026	cg12594803	3	*DLG1*	$5.19x10^{-04}$	0.049
chr17:4899230:D	cg19427746	17	*CAMTA2*	$6.17x10^{-07}$	0.0059
rs61766781	cg15603964	1	*UBR4*	$3.22x10^{-06}$	0.0094
rs62195287	cg17674726	2	*ITM2C*	$8.37x10^{-05}$	0.049
chr8:74716807:D	cg12682382	8	*UBE2W*	$3.19x10^{-03}$	$1.96 x10^{-06}$
rs1952512	cg02988727	14	*ZNF219*	$9.63x10^{-07}$	0.0073
rs3012041	cg12829045	10	-	$1.47x10^{-09}$	0.0016
rs7736222	cg11100481	5	-	$6.87x10^{-03}$	$1.48 x10^{-05}$

4.3.5 Tissue Specificity of genetic impacts on DNA methylation

To explore the tissue specificity of genetic effects on DNA methylation, I repeated the meQTL and var meQTL analysis in 83 MZ pairs profiled using Illumina 450k in

adipose tissue. I used a permutation based FDR approach as described above in section 4.2.6 to establish significance thresholds accounting for multiple testing. I selected FDR 5% to be able to make comparison between whole blood results and adipose tissue results at the same threshold.

At an FDR of 5% ($P = 1x10^{-3}$) I detected cis meQTLs for 40,601 CpG probes (9.5% of 427,767 autosomal probes), and trans meQTLs at 24,091 CpG probes (FDR 5%, $P = 1x10^{-7}$). Moreover, I detected cis var meQTLs for 15,936 CpG probes (3.72% of 427,767 autosomal probes, $P = 1x10^{-3}$) and trans var meQTLs at 24,779 CpG probes ($P = 1x10^{-7}$). The adipose tissue results are summarized in Table 4-8. Additionally, the overlap between cis meQTL and cis var meQTL is 4,240 (27%) and for trans meQTL and trans var meQTL it is 2,343 (9.7%) showing there are multiple QTLs showing both effects.

Table 4-8 Number of probes identified in QTL analysis in adipose tissue with FDR 5%

Adipose tissue FDR(5%)	***cis* ($1x10^{-3}$)**	***trans* ($1x10^{-7}$)**
meQTL	40,601	24,091
var meQTL	15,936	24,779

Comparison between adipose tissue and whole blood

Half of the *cis* meQTL signals from adipose tissue are also identified in whole blood and almost a quarter of the *trans* meQTL from whole blood are also identified in adipose tissue (Table 4-9). The overlap between adipose tissue and whole blood is 15% for *cis* var meQTL (1,210 probes) and 9% for *trans* var meQTL (500 probes) when validated with FDR = 5% (Table 4-9).

Table 4-9 Overlap of probes identified in QTL analysis in adipose tissue versus whole blood results

Adipose tissue vs. Whole blood *	***cis* (FDR 5%)**	***trans* (FDR 5%)**
Adipose meQTL	40,601	24,091
Whole blood meQTL	53,813	15,392
Overlap meQTL	**20,049 (49%)**	**2,444 (15%)**
Adipose var meQTL	15,936	24,779
Whole blood var meQTL	8,106	3,694
Overlap var meQTL	**1,210 (15%)**	**343 (9%)**

*Percentage of overlap is estimated out of the smallest denominator (whole blood or adipose)

Overall, I identified 53,813 *cis* meQTLs and 15,392 *trans* meQTLs in whole blood and 40,601 *cis* meQTLs and 24,091 *trans* meQTLs in adipose tissue. The overlap of more than 20,000 *cis* meQTLs between these tissues highlights the extent of tissue shared effects, as well as showing the importance of genetic effects in tissue differentiation. Moreover, I identified 8,106 *cis* var meQTLs and 3,694 *trans* var meQTLs in whole blood and 15,396 *cis* var meQTLs and 24,779 *trans* var meQTLs in adipose tissue. The overlap of around 15% *cis var*-meQTLs highlights that genetic impacts on methylation variability tend to be tissue specific, with some tissue shared effects.

4.4 Discussion and Conclusion

I performed QTL analysis of methylation levels and variances at over 400,000 CpG sites across the genome profiled using the Illumina 450k in both whole blood and adipose tissue. In blood, I identified more than 50,000 *cis* and more than 15,000 *trans* meQTL signals. These numbers are consistent with recent meQTLs reported in different tissues using the Illumina 450k array (Grundberg *et al.* 2013; Shi *et al.* 2014). I then

identified 8,106 *cis* var meQTLs and almost 90% of the *cis* var meQTLs are also identified as *cis* meQTL. I also identified 3,694 *trans* var meQTLs and more than quarter of them were also identified as *trans* meQTLs.

Additionally, 83 MZ pairs were used for exploring tissue-shared effects in adipose tissue, and the results showed that more than half of the probes identified had *cis* meQTLs in both blood and adipose tissues. Altogether 49 MZ pairs had both adipose and blood data available. Ideally, a fully overlapping sample across blood and adipose tissue would be the most appropriate dataset to assess tissue shared and tissue specific results. However, as these data were not available I used a partially overlapping adipose sample, as this was the largest available adipose sample with methylation data that I had access to. Due to the partial overlap in blood and adipose samples, 'true' tissue-shared meQTLs results should be consistent in both dataset, as the allele frequency and effect size distribution should be similar across these partially overlapping samples.

In contrast, the tissue-shared effects were much attenuated for *cis* var meQTLs, for which approximately 15% of probes share var meQTLs across tissues. Additionally, *trans* meQTL and var meQTL results show a greater extent of tissue-specifity compared to results in *cis*. Overall, this indicates that numerous *cis* meQTLs are likely to be detected in multiple tissue types, but the majority of trans meQTLs are tissue-specific. However, the tissue-shared results were difficult to interpret due to the large difference in sample size for MZ twins for adipose tissue and whole blood. This is why I used permutation based FDR calculations at the level of 5% for comparison of both samples. Overall, there is evidence for some shared genetic control on DNA methylation across different tissues.

The presence of var meQTL can be induced by gene–environment interactions, as well as epistasis or haplotype effects. Given that my sample was based on a twin cohort, which includes monozygotic (MZ) twin pairs, I used another measure of variability

within the dataset. Because MZ twins share the same genetic background, the impact of epistatic and haplotype effects driving var meQTL is likely to be very low in our sample. Therefore, var meQTL could capture gene environment interactions and MZ twins would be a good model in which to assess this. I tested this running two different var meQTL analyses, one adjusting for smoking and one without. The 581 *cis* and 1,268 *trans* signals that were only identified as no-smoking specific var meQTL were explored further in gene-smoking interaction analysis. From these, 29 *cis* and 474 *trans* signals identified as having gene–smoking effect. The top no-smoking specific *cis* var meQTL acted on DNA methylation variance in the *AGPAT9* gene, which has been associated with colorectal and ovarian cancer before (Agarwal 2012; Currie *et al.* 2013; Wang *et al.* 2012). Additionally, this gene was identified as a smoking-responsive gene in RNA-Seq analyses (Hackett *et al.* 2012) in which responsiveness to smoking was quantified with an index representing the % of smoking-responsive genes abnormally expressed, with smokers grouped into responders based on the proportion of smoking-responsive genes up- or down-regulated in each smoker. Although, the identified number of no-smoking specific *cis* var meQTL with evidence for gene-smoking interaction is relatively small, the observed findings still suggest that gene environment interaction plays a role in driving the significant no-smoking specific var meQTLs, and that these no-smoking specific var meQTLs can identify gene by environment interaction effects on DNA methylation.

One issue with the variance analyses above is that genotype classes with a smaller sample size, such as the rare genotype category, will have inflated standard deviations. For this reason I performed a validation test of var meQTLs with a bigger sample of 459 unrelated individuals, using another method to assess variance meQTLs - Bartlett's test for variance heterogeneity in unrelated samples. The findings showed that all of the var meQTLs in the MZ analyses validated as var meQTLs at nominal significance in the

unrelated sample. The MZ twin-pair design is a powerful approach to study genetic control of phenotypic variability as it requires a smaller sample size than its equivalent in an unrelated study and identified specifically the extent of variability due to environmental factors.

I clearly see that the distribution of var meQTLs is enriched for rare SNPs, compared to cis meQTLs, and this is likely in part due to the statistical artefact that higher variance is observed in smaller genotype groups. I included two follow-up analyses to assess the extent to which an increase of variance in small samples is likely driving these effects. The first, was a validation analysis of the *cis* var-meQTLs in all of the available unrelated subjects using Bartlett's test in section 4.3.4. The sample size of unrelated dataset is larger than the 330 MZ pairs and all of the cis var-meQTLs remained nominally significant in the unrelated dataset. The second analysis was to ensure that for all *cis* var-meQTLs every genotype group had a minimum of 5 samples. Therefore, although it is likely that the cis var-meQTLs capture to some extent increase variability of small genotype groups, the result cannot be entirely explained by this effect.

Overall, the results from this chapter have identified thousands of genetic effects on DNA methylation levels and for the first time, genetic impacts on DNA methylation variability. These findings give insights into the biological processes regulating epigenetic mechanisms in humans.

CHAPTER 5

Metabolomic and epigenetic signatures of type 2 diabetes

5.1 Introduction

The main aim of this chapter is to understand the biological mechanisms involved in T2D susceptibility and progression, by linking DNA methylation and metabolomics to T2D status in twins. Part of this work has been recently published (Menni *et al.* 2013c; Yuan *et al.* 2014).

T2D is a chronic disease, caused by failure of beta cell function and insulin resistance. Multiple GWAS have been conducted to explore genetic influences on T2D, identifying approximately 81 susceptibility variants to date (Billings and Florez 2010; DIAGRAM *et al.* 2014; Guan *et al.* 2008; McCarthy 2003; Morris *et al.* 2012; Sanghera and Blackett 2012; Zeggini *et al.* 2008). Metabolic profiles have also been linked to this disease (Menni *et al.* 2013c; Shin *et al.* 2014; Suhre *et al.* 2010; Wang *et al.* 2011). In addition, longitudinal studies have identified a link between metabolite levels and insulin resistance and T2D (Suhre *et al.* 2010; Wang *et al.* 2011). Finally, T2D-related differentially methylated regions (DMRs) have also been identified (Yuan *et al.* 2014).

In this chapter I explored four specific research questions. First, the aim was to assess whether metabolic profiles that are characteristic of T2D also associate with certain epigenetic variants. I performed an association study of metabolic profiles in T2D and I tested whether the T2D-associated metabolites also associated with DNA methylation changes genome-wide. This work is in line with a recent EWAS study from the KORA cohort, exploring the link between metabolomics and epigenetics in the first metabolite-

EWAS (Petersen *et al.* 2014). The second question that I addressed was to try to uncover if epigenetic variants are associated with T2D, and whether these variants may also be associated with metabolic profiles. I performed an epigenome-wide association study of DNA methylation changes in T2D, that is, comparing DNA methylation levels to T2D, to identify differentially methylated positions in T2D (T2D-DMPs). I then tested whether these T2D-DMPs also associate with metabolic profiles. The third research question was to assess if T2D genetic susceptibility effects can be mediated via intermediate phenotypes, such as epigenetic changes or metabolic profiles. I compared the list of 81 T2D GWAS signals that have been published to date against the genetic variants that contribute to metabolomic and epigenetic profiles identified in Chapters 3 and 4 of my thesis. Finally, in the fourth component I combined the results from the three previous analyses described here, that is, integrating genetic, epigenetic, and metabolic profiles associated with T2D to try to understand biological mechanisms underlying T2D. The fourth research aim was to attempt to infer causality in T2D by fitting Bayesian networks to the peak T2D-GWAS, T2D-metabolite, and T2D-DMP results and their pair-wise associations, to gain more insight into T2D susceptibility and progression.

5.2 Methods and Results

5.2.1 Metabolic profiles that are characteristic of T2D associate with epigenetic variants.

Collaborative study

In my first year, I contributed towards a T2D cross-sectional case-control study of metabolite levels, where we identified 42 metabolites associated with T2D (Menni *et al.* 2013c) (Figure 5-1).

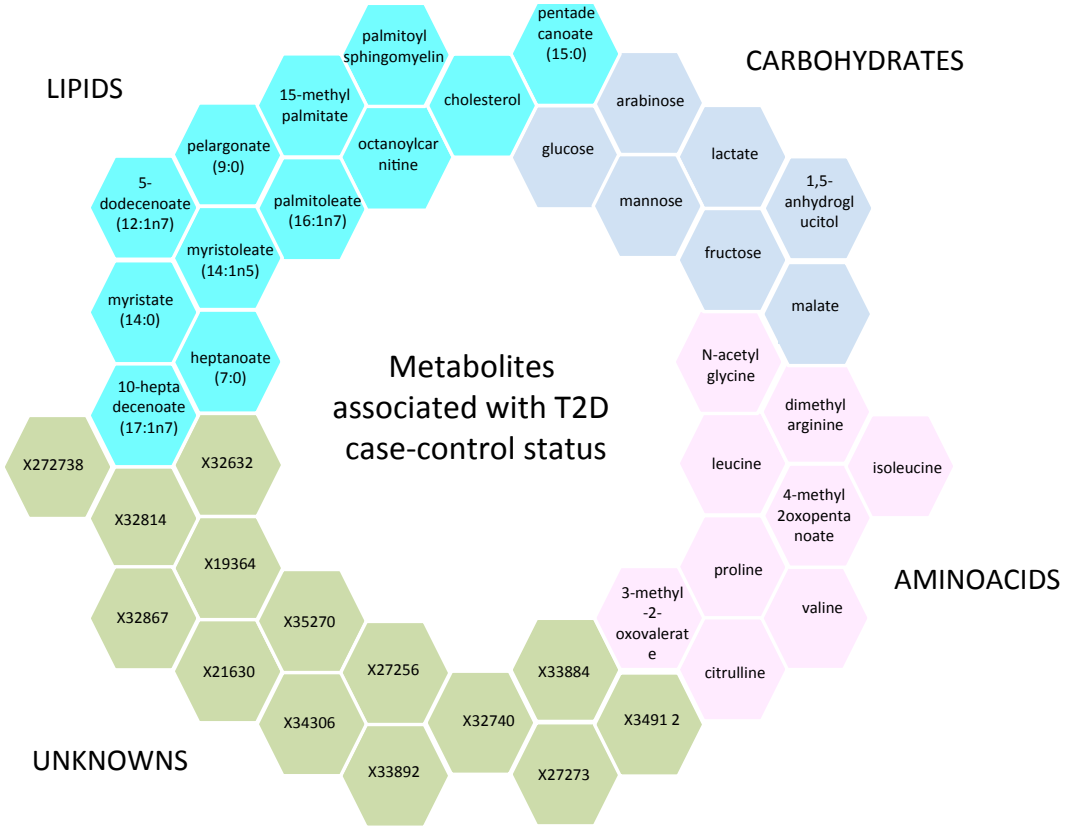

Figure 5-1 Description of 42 metabolites associated with T2D case-control status (adapted from Menni et al. 2013).

Dr Cristina Menni, a research fellow in the department, led this MWAS work and I contributed to the study. Briefly, the dataset consisted of 2,204 female subjects (115 T2D, 192 impaired fasting glucose (IFG), and 1,897 controls) who had metabolomics profiles available for 447 Metabolon metabolites after quality control. Using 1,297 monozygotic and 1,200 dizygotic twin pairs, I estimated heritability for each metabolite identified in association analysis using OpenMx (Boker *et al.* 2011). The calculated heritabilities ranged from 0% to 65%. For each T2D-control and IFG-control contrast, we fitted mixed effect regression adjusting for age, BMI, batch, and included a random effect for family relatedness, and then ran a stepwise linear regression including all the significant metabolites to look for metabolites independently associated with T2D and IFG respectively. At a Bonferroni-corrected cut off of P = $1x10^{-4}$ (0.05/447), 42 of the 447 metabolites showed significant differences among T2D case and control subjects,

and independently associated with T2D, after adjusting for all other metabolites in multivariate analyses. Furthermore, 117 metabolites were associated at nominal significance (P =0.05) and these 117 were used in downstream analyses in section 5.2.3 below. As depicted in Figure 5-1, the 42 peak metabolites fall into three principal classes: 12 are lipids (primarily medium and long-chain free fatty acids), 7 are carbohydrates, 9 are branched-chain amino acids (BCAAs) or derivatives, and 14 are unknown. Additionally, 14 metabolites from the 447 metabolites tested showed significant differences among IFG case and control subjects with a Bonferroni-corrected cut off of $1x10^{-4}$ (0.05/447), and of these 8 overlap with the results from T2D case and control analysis. I then validated the top 48 hits using in the lme4 package in R (Bates *et al.* 2015). The model can take into account relatedness between individuals and the sample contained MZ and DZ twins. The MWAS covariates included fixed-effect terms (age and BMI at the time of sampling, metabolite batch) and random effect terms (family structure and zygosity i.e DZ, MZ and singleton). In the full regression model metabolite were fitted as the outcome, and the predictors consisted of age and BMI at the time of sampling, fasting glucose levels, batch and random effect terms. This full model compared to the null model (without fasting glucose levels) using an ANOVA F-statistic to compare model fit.). Diabetes is considered to be primarily a disorder of glucose and was the strongest predictive biomarker for T2D ($P = 1.12x10^{-14}$). On the other hand, apart from carbohydrates, we also reported lipids and amino acids that associate with T2D and IFG.

Extended work for my thesis

For my thesis, I then extended this work by performing additional analyses. I tested whether the 42 T2D-associated metabolites also associated with DNA methylation changes genome-wide in 42 EWAS using the Illumina 450k methylation dataset in 807 individuals, which included 32 T2D cases and 775 controls.

The Metabolon metabolomics and DNA methylation dataset first underwent several quality control checks as described in Chapter 2. I visually checked plots to remove the outliers and subjects with missing values from both datasets. In the methylation dataset I removed probes mapping to multiple locations as described in Chapter 2. Raw DNA methylation values normalized using BMIQ (Teschendorff *et al.* 2013) and beta values on each probe were normalized to N(0,1) then fitted in linear mixed effect models. A log transformation with base 10 was applied to the 42 metabolites. The EWAS covariates included fixed-effect terms and random effect terms in LMER in lme4 package in R (Bates *et al.* 2015). In the full regression model normalized methylation levels were fitted as the outcome, and the predictors consisted of metabolites (standard normalized), smoke, sex, age, BMI, plate and blood cell count estimations and random effect terms. This full model compared to the null model (without metabolites) using an ANOVA F-statistic to compare model fit. A methylation probe was defined as a T2D-metabolite-DMP (Differentially Methylation Position) if it passed the False Discovery Rate (FDR) of 5%. FDR was calculated for each probe using the QVALUE R package (Storey 2015).

Altogether, I performed 42 EWAS, one for each of the 42 T2D-associated metabolites identified in the collaborative study described above (Menni *et al.* 2013c). The merged final dataset included a total of 42 metabolites and 474,979 methylation probes in 807 individuals. At the threshold of FDR 5% ($P = 1x10^{-6}$) I identified 121 (T2D-associated metabolite)-DMPs (APPENDIX B Table S5-1), and I present the 10 top-ranked metabolite-DMPs in Table 5-1. If I've had removed the SNP on probes as suggested by Naeem et al. ((Naeem *et al.* 2014)) than it would be 94 (T2D-associated metabolite)-DMPs.

Table 5-1 Top 10 metabolite-DMPs in blood

Probe	Chr	Location	Gene	Beta	P	Metabolite	Super_pathway
cg12612277	16	74455542	*CLEC18B*	0.18	$1.15x10^{-7}$	mannose	Carbohydrate
cg11113753	2	44065383	*ABCG5*	0.19	$1.70 x10^{-7}$	X - 11550	NA
cg03333776	1	52455148	*RAB3B*	0.17	$3.79 x10^{-7}$	15-methylpalmitate	Lipid
cg23867721	18	77631069	*KCNG2*	0.16	$3.81 x10^{-7}$	X - 11550	NA
cg05925577	1	121375906	-	0.22	$6.07 x10^{-7}$	malate	Energy
cg21831937	15	57519802	*TCF12*	-0.22	$6.96 x10^{-7}$	lactate	Carbohydrate
cg04483701	16	86253703	-	0.15	$7.03E x10^{-7}$	X - 12442	NA
cg21383495	17	54673193	-	-0.17	$7.19 x10^{-7}$	15-methylpalmitate	Lipid
cg08526784	16	1811246	*MAPK8IP3*	-0.36	$7.21 x10^{-7}$	lactate	Carbohydrate
cg03666973	2	27008764	*CENPA*	-0.29	$7.30 x10^{-7}$	arabinose	Carbohydrate

5.2.2 Epigenetic variants are associated with T2D, and these also associate with metabolic profiles.

Collaborative study

I recently contributed to an epigenetic study of T2D, led by Dr. Yuan a research fellow in the department. In this work, the aim was to identify differentially methylation regions associated with T2D (T2D-DMRs) in T2D-discordant genetically identical twins (Yuan *et al.* 2014). Whole blood samples from 27 MZ twin pair were profiled for DNA methylation using MeDIP-seq. MeDIP-seq data were generated at almost 30 million paired-end reads of length 50 bp per individual, and mapped to the human genome. He quantified levels of DNA methylation in overlapping bins of size 500 bp using MEDIPS (Yuan *et al.* 2014). He identified 31DMRs at a FDR level of 10%, using a linear mixed effects model. The strongest signal is in the promoter of the *MALT1* gene (FDR=5%), which is a signalling protein with a role in the development and function of B and T cells, as well as energy and insulin pathways.

My role was to perform a metabolite analysis in this sample of T2D-discordant MZ twins and compare the metabolite profiles to the peak T2D-DMRs. Among the 27 MZ twin pairs, 18 MZ twin pairs had plasma and/or serum metabolites profiles. For the 36 MZ individuals (18 MZ twin pairs) who had metabolite data in this sample, the mean

difference of metabolite levels within twin pair (affected–unaffected twin) was calculated. Wilcoxon signed-rank test was used to evaluate whether the mean difference for each metabolite was significantly (P<0.05) associated with the DNA methylation twin pair difference at *MALT1*. The peak T2D DNA methylation signal in the *MALT1* gene was nominally significantly associated with 7 metabolites (X–06246, N-(2-furoyl) glycine, X–12450, glycerol 2-phosphate, 4-acetamidophenol, X–11818 and taurocholate) in my results. One of the metabolites, taurocholate, was of interest because it has been associated with increasing L cell and insulin secretion as well as a decrease in blood glucose and food intake in obese type 2 diabetic volunteers (Adrian *et al.* 2012). Additionally, one of the 42 metabolites associated with methylation in the *MALT1* gene, which was identified in a T2D-EWAS below (Yuan *et al.* 2014). The probe in *MALT1* gene (cg23450680) was associated with 15-methylpalmitate (isobar with 2-methylpalmitate) at a relaxed p-value (P<0.004). I included these results for further analysis in section 5.2.3.

Extended work for my thesis

The epigenetic study led by Dr. Yuan was based on MEDIP-seq DNA methylation profiles. Although this technology provides good genome-wide coverage, the resolution of the DNA methylation signal is only at the level of 150-300bp. Conversely, the Illumina 450k array provides DNA methylation at single-base-pair resolution, but only at ~485,000 CpG-sites genome-wide. Therefore, I extended the epigenetic analyses of T2D by using the Illumina 450k dataset and a case-control design in a total of 864 individuals, which included 45 T2D cases and 819 controls. 864 include 807 individuals used in 5.2.1 part of extended work for my thesis.

The DNA methylation dataset first underwent several quality control checks as described in Chapter 2. Briefly, I visually checked plots to remove the outliers and subjects with missing values from dataset, and removed probes mapping to multiple

locations. Raw DNA methylation values were normalized using BMIQ (Teschendorff *et al.* 2013) , and beta values at each probe were normalized to N(0,1). T2D case-control status was defined according to available blood fasting glucose levels. Subjects were classified into two groups based on fasting glucose levels at time of initial sampling and at subsequent visits; T2D case subjects (fasting glucose ≥7 mmol/L or physician's letter confirming diagnosis) and T2D control subjects (3.9 mmol/L< fasting glucose <5 mmol/L). The merged final dataset included a total of 474,979 probes profiled in 45 T2D cases and 819 controls.

The normalized methylation betas were fitted with a linear mixed effect model as the outcome, and the predictors consisted of BMI, plate and blood cell count estimations, family and zygosity, as previously described. Residuals from this model were extracted and normalized to N(0,1), and then included in the second linear model. The residuals were fitted as the outcome, and the predictors consisted of T2D, smoking, sex, and age. This full model was compared to a null model (without T2D status) using an ANOVA F-statistic to compare model fit. A probe was defined as a T2D-DMP if it passed FDR of 5%. FDR was calculated for each probe using the QVALUE R package (Storey 2015).

At an FDR of 5% ($1x10^{-5}$), I identified 9 T2D-DMPs in the 864 individuals (Table 5-2, Figure 5-2). On the other hand, at nominal significance (P =0.05) there were 27,285 CpG-sites associated with T2D and I included these 27,285 results for further analysis in section 5.2.3 below. If I've had removed the SNP on probes as suggested by Naeem et al. ((Naeem *et al.* 2014)) than it would be 22,713 (T2D-associated metabolite)-DMPs.

Table 5-2 9 probes found in T2D –DMPs in blood tissue

Probe	Chr	Location	Gene	Beta	t-stat	P
cg25677697	19	50912349	*POLD1*	0.05	5.04	5.57×10^{-7}
cg26093898	17	80137377	*CCDC57*	0.05	4.92	1.03×10^{-6}
cg21733020	17	2599052	*KIAA0664*	0.05	4.88	1.29×10^{-6}
cg04607246	1	155720323	*GON4L; MSTO2P*	0.05	4.72	2.70×10^{-6}
cg18191664	8	37594173	*ERLIN2*	0.08	4.66	3.70×10^{-6}
cg13573626	14	105858487	*PACS2*	0.04	4.59	5.01×10^{-6}
cg13093042	1	207975500	*MIR29C*	0.04	4.52	7.16×10^{-6}
cg01758022	8	141557171	*EIF2C2*	0.04	4.47	8.82×10^{-6}
cg01678580	16	4674018	*MGRN1*	0.03	4.46	9.35×10^{-6}

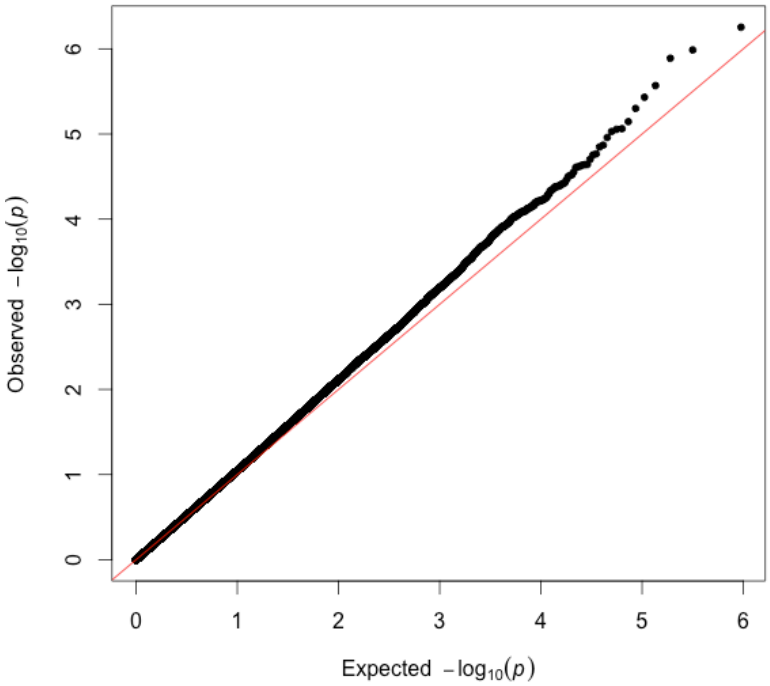

Figure 5-2 QQplot of the T2D EWAS in 45 cases and 819 controls

*Inflation factor (λ) for the QQplot calculated as 1.3

5.2.3 Bayesian Network Analysis

Bayesian Networks (BNs) are graphical probabilistic models that are able to represent joint probability distributions compactly in a factorized way (Koller and Friedman 2009; Pearl 1988). A BN consists of a graphical structure and a set of parameters. The graphical structure of a BN is a directed acyclic graph (DAG) that consists of nodes

representing variables and directed edges representing the relations between those variables. If a directed edge connects two variables, as in *A* ➔ *B, A* is called the parent variable and *B* is called the child variable. The DAG structure encodes a set of conditional independence assumptions between the variables: a variable is independent of its non-descendants given its parents. The DAG structure is also suitable for representing causal relations of a domain as the directed edges are often used to represent the causal relations between the variables (Pearl 2000). The parameters of a BN represent the conditional probability distributions between the variables that are directly connected by an edge.

A BN model can be built in several different ways. First, we can define the graphical structure, and estimate the parameters (i.e. conditional probability distributions) from the data by using maximum likelihood approach or Bayesian methods. Second, we can estimate both the graphical structure and the parameters from data by using score-based or constraint-based algorithms. Third, we can manually define both the structure and the parameters by using expert information. The performance of alternative BN models can be assessed with various scoring methods such as, log-likelihood, the Aikake Information Criterion (AIC) (Akaike 1976) or the Bayesian Information Criterion (BIC) (Schwarz 1978). In this study, I used the first approach: I build three BN structure for the alternative causal relations, I estimated the parameters of these structures from the data and finally I examined the compatibility of these structures with the data by using the AIC Score.

Aim

My aim was to attempt to infer causality in T2D by fitting BN to the peak T2D-GWAS, T2D-metabolite, and T2D-DMP results and their pair-wise associations. I compared the peak T2D-associated metabolite and methylation results from sections 5.2.1 and 5.2.2 in this Chapter, with the methylation and metabolome QTL results from Chapters 3 and 4

specifically at 81 published T2D-GWAS loci. Pair-wise associations that surpassed nominal significance were then used to fit in the BN and infer direction of association.

Methods

I first compared the list of 81 T2D GWAS signals that have been published to date (Billings and Florez 2010; Guan *et al.* 2008; McCarthy 2003; Sanghera and Blackett 2012) against the genetic variants that contribute to metabolomic and epigenetic profiles identified from my thesis (Chapter 3 and Chapter 4). If T2D genetic variants are identified as metabolite QTLs or methylation QTLs, this may give us some insight into the intermediate molecular mechanisms involved in T2D susceptibility. I also compared metabolite QTLs to methylation QTLs, because if any metabolite QTLs are also methylation QTLs this may point to potential shared mechanisms of genetic control of both processes. I then only considered the QTLs that had impacts on the metabolites or epigenetic probes that were at least nominally associated (P = 0.05) with T2D, from the analyses shown in the earlier two sections in this chapter (Chapter 5.2.1 and Chapter 5.2.2).

In total there were 81 genes (T2D-GWAS signals from the literature), 117 metabolites (T2D-MWAS nominally significant results from 5.2.1) and 27,285 probes (T2D-EWAS nominally significant results from 5.2.2) that were included in the comparisons (Figure 5-3). I performed all pairwise comparisons and only selected nominally-significant pairwise findings for downstream analysis. That is, I considered:

1) 81 T2D-GWAS SNPs as metabolomics QTLs for 117 T2D-associated metabolites, selecting only results that surpassed P = 0.05, and

2) 81 T2D–GWAS as methylation QTLs for 27,285 T2D-associated CpG-sites, selecting only results that surpassed P = 0.05, and

3) The correlation between the 117 T2D-associated metabolites and 27,285 T2D-associated CpG-sites, selecting only results that surpassed P = 0.05.

Using these pair-wise selection thresholds, I found 240 overlapping three-way associations with a unique set of 36 genetic variants, 71 methylation probes and 41 metabolites (Figure 5-3, APPENDIX B Table S5-2). I included these results in the BN analysis. If I've had removed the SNP on probes as suggested by Naeem et al. ((Naeem *et al.* 2014)) than I would've only include 56 methylation probes into BN analysis.

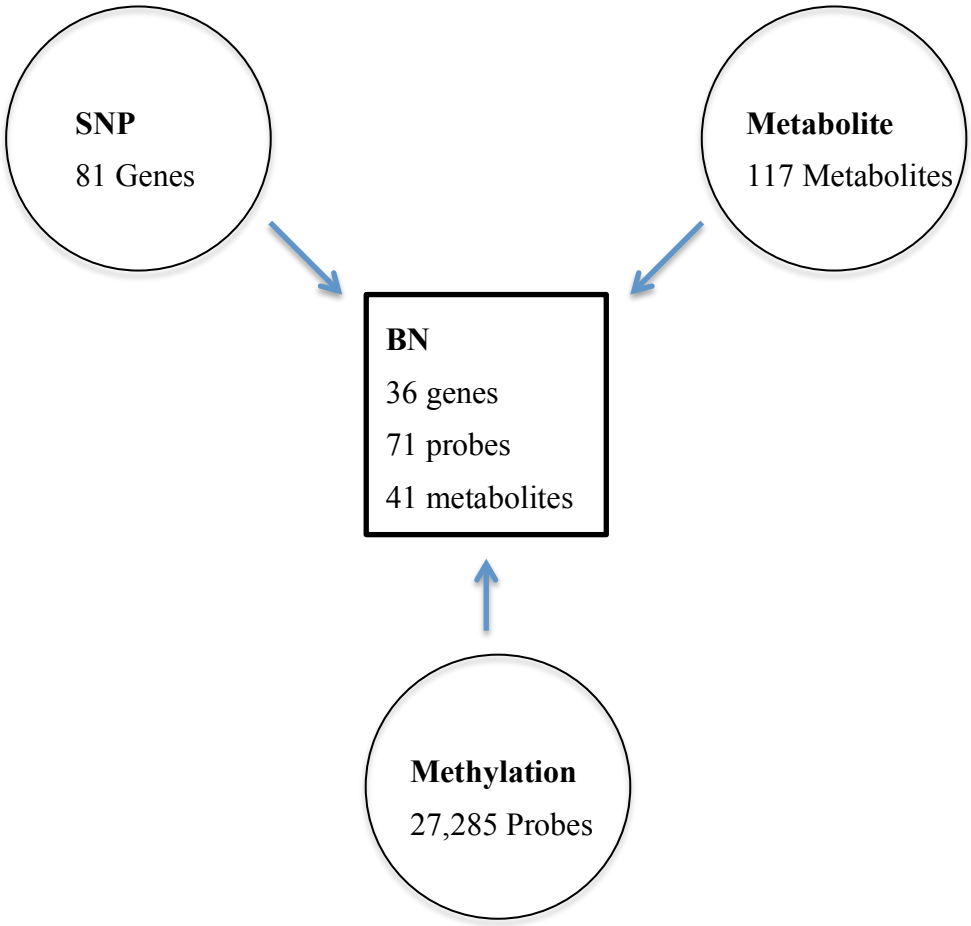

Figure 5-3 Input for BN analysis

Fitting Bayesian Networks

To infer causality of the association at the peak T2D-GWAS, T2D-metabolite, and T2D-DMP results, I used the 240 overlapping three-way association and fit a BN network to these data to explore the shared genetic impacts on T2D-association

methylation and metabolomic signatures to learn more about biological processes involved in T2D susceptibility and progression**.**

To test the networks I selected three datasets (genetic, methylation, and metabolomics – where all individuals had available T2D case-control status) that were normalized to calculate the relative frequencies of the inferred best network. Altogether, the merged normalized dataset for these BN analyses contained 807 TwinsUK individuals, which included 32 T2D cases and 775 controls with biological profiles available for the 31 genetic variants (SNP), 71 methylation profiles (MT), and 41 metabolites (MB).

I built three BN structures representing the alternative hypotheses of the causal relations between SNP, MB and MT. The first BN structure assumes that there is a causal relation from SNP to both MT and MB, and thus MT and MB are independent of each other given SNP (INDEP) (Figure 5-4a). The second BN structure assumes there is a causal relation from SNP to MB, and from MB to MT (SMbMt) (Figure 5-4B). The third BN structure assumes there is a causal relation from SNP to MT, and from MT to MB (SMtMb) (Figure 5-4C). The parameters of these networks were estimated by using the maximum likelihood approach. Afterwards, I examined the compatibility of these structures with the data by using the AIC score. I used the Akaike Information Criterion (AIC) score (AIC = 2k-2ln(L), where k is the number of parameters and L is the maximum likelihood to compare our networks. To compare the goodness of fit of one network to another, I used the relative likelihood of one network against the other following previous work (Bryois *et al.* 2014; Gutierrez-Arcelus *et al.* 2013). If I have two networks, N1 and N2 and AIC(N1)≤AIC(N2), then the relative likelihood of N2 with respect to N1 is defined as: exp((AIC(N1)–AIC(N2))/2). I kept only networks where the best model was at least ten times more likely than the second best model. The bnlearn package in R was used to build and calculate the BN models (Scutari 2009).

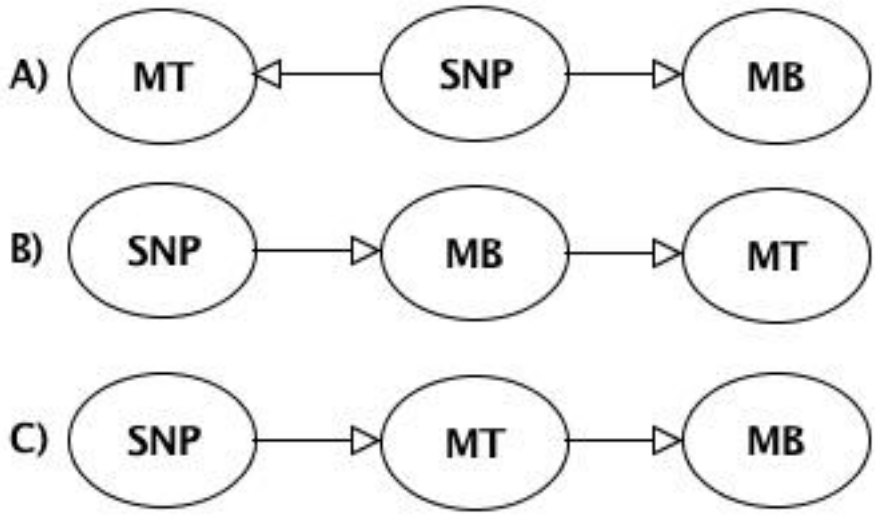

Figure 5-4 BN structures of A) INDEP B) SMbMt C) SMtMb

I used BN to test which of the three possible models best suit each set of variables. After fitting all the models at the 240 three-way associations, there were 10 models where the best model was at least ten times more likely than the second best model. Of these 10 models, 3 followed the SMtMb model and 7 followed SMbMt (Table 5-3). As an example of a SMtMb model result, I found that SNP rs9936385 in the *FTO* gene could alter DNA methylation levels locally within *FTO* and then impact on levels of the metabolite palmitoleate.

Table 5-3 10 models reported by Bayes Network

Gene	Probe	Metabolite	SNP	Score1	Score2	Score3	Best Model	RL
WFS1	cg00829753	X - 11423	rs4458523	-2584.9	-2578.7	-2584.1	2	15.47
WFS1	cg22247194	citrulline	rs4458523	-2971.9	-2966.9	-2975.9	2	12.69
ADCY5	cg14844401	X - 10506	rs11717195	-3292.3	-3286.1	-3291.9	2	18.46
ADCY5	cg14844401	malate	rs11717195	-3292.0	-3286.0	-3292.2	2	21.36
ADCY5	cg14844401	pentadecanoate (15:0)	rs11717195	-3292.9	-3287.2	-3292.4	2	13.93
FTO	cg12495954	palmitoleate (16:1n7)	rs9936385	-2984.6	-2986.9	-2977.4	3	37.12
FTO	cg12495954	X - 13215	rs9936385	-2983.9	-2987.6	-2978.7	3	13.70
KCNQ1	cg19030519	malate	rs231362	-3386.4	-3381.0	-3385.7	2	10.36
HNF4A	cg08407434	X - 13496	rs4812829	-2992.7	-2998.1	-2988.0	3	10.50
ANK1	cg26172342	heptanoate (7:0)	rs516946	-3207.5	-3202.5	-3209.8	2	12.49

Score: score from the AIC test for the 3 different models 1)INDEP 2) SMbMt 3) SMtMb

RL: Relative likelihood

On the other hand, in the SMbMt model, I observed that SNP rs231362 in the *KCNQ1* gene, could affect malate, which is a carbohydrate, and then modify DNA methylation levels. *KCNQ1* was reported for predicting and mediating impaired insulin secretion (Jonsson *et al.* 2009).

If I've had removed the SNP on probes as suggested by Naeem et al. ((Naeem *et al.* 2014)) than only 8 models would've been significant. The 2 probes in *WFS1* gene needed to be removed (cg00829753 and cg22247194).

There was no convincing evidence that the T2D-GWAS variants had independent effects on DNA methylation and metabolomics profiles at the 240 three-way associations.

5.3 Discussion and Conclusion

In this chapter I first explored the association of metabolomics and epigenetic data and one of the most common diseases, T2D. Multiple signals were identified to associate with T2D at both the metabolomics and epigenetic levels. The availability of different levels of "-omics" data in the TwinsUK cohort then allowed me to explore the relationship between T2D-associated "-omics" profiles, specifically for genetic variation, methylation and metabolomics.

One of the findings from the epigenome-wide analyses was a peak T2D-associated signal in the *MALT1* gene. The *MALT1* DMR was originally identified from a MeDIP-seq analysis (Yuan *et al.* 2014), but I was able to validate the T2D-associated signal using a different DNA methylation technology (Illumina 450k) in a case control study, but at a more relaxed significance threshold. My analyses comparing T2D-DMPs and metabolite profiles, showed significant associations of *MALT1* DNA methylation profiles with several metabolites, of which taurocholate was of most relevance to T2D. The *MALT1* T2D-DMP does not have a methylation QTL and the *MALT1* region is not

within the list of 81 genome-wide significant T2D GWAS signals. The metabolites associated with T2D and *MALT1* also do not have metabolite QTLs. Therefore, the *MALT1*- taurocholate were not further explored in the BN context. The lack of genetic signal underlying the *MALT1* result is not surprising because the *MALT1* findings were obtained in a sample of T2D-discordant genetically identical twins. It is more likely that the *MALT1*-metabolite associations are either environmentally driven, or consequence of T2D.

In the main section of this chapter I explored T2D-associated "-omic" profiles for genetic variation, methylation and metabolomics, and specifically their pair-wise and three-way associations. After restricting results to at least nominally significant associations, I obtained a list of 240 three-way associations to explore further in a BN context. I detected 10 cases of three-way associations where either genetically driven DNA methylation levels impact metabolomic profiles, or genetically driven metabolomics traits impact DNA methylation levels.

One interesting finding was obtained for a SNP in the *FTO* gene, associated with DNA methylation in the *FTO* gene, and the metabolite palmitoleate. The fatty acid, palmitoleate (16:1 n−7), which is the second most abundant monounsaturated fatty acid, is influenced by endogenous synthesis, which appears to be tissue and depot specific. It is reported as a 'lipokine', which is a lipid-controlling hormone (Cao *et al.* 2008). From my results it appears that *FTO* genetic impacts on palmitoleate are mediated via DNA methylation in the *FTO* gene. Palmitoleate travels to the muscles and liver, where it improves cell sensitivity to insulin and blocks fat accumulation in the liver (Cao *et al.* 2008). The first report linking the *FTO* (fat mass and obesity-associated) gene and obesity came from a genome-wide association study linking *FTO* variants with T2D in a European population (Frayling *et al.* 2007). However, the connection between *FTO* and diabetes was lost after correcting for body-

mass index, suggesting that *FTO* -mediated susceptibility to T2D was driven through a relationship between *FTO* and obesity. My result complements and extends this work, because the SMtMb directional effect suggests that at least some of the genetic impacts at *FTO* likely act *via* DNA methylation (which in turn may influence the expression of *FTO* or nearby genes, such as *IRX3* ((Ragvin *et al.* 2010). These effects impact palmitoleate, which is particularly relevant to metabolic disease, such as T2D.

One potential limitation of the integrative genetic/epigenetic/metabolomics T2D study is that mQTLs were explored using HapMap imputations, while I used 1000G imputations for meQTLs. However, I also identified meQTLs using PLINK (Purcell *et al.* 2007) in HapMap imputed data in an earlier version of the analyses in Chapter 4. The overlap between meQTLs in the HapMap imputed and meQTLs in the 1000G imputed results was 52%. One difference between the meQTL analyses with HapMap imputation and 1000g imputation was that when I used 1000g imputations in matrix eQTL I separately determined *cis* and *trans* results, that is, one probe can have both *cis* and *trans* meQTLs. On the other hand, in PLINK I only reported the best SNP per probe so one probe can only have *cis* or *trans* effects, which is an underrepresentation of the total number of meQTLs. Given this limitation, I used the 1000G meQTLs results throughout Chapter 5 section 5.2.3.

Overall, these results suggest that the effects of DNA methylation can be both active on metabolomic profiles, or passive, by being a consequence for metabolomic profiles. This study shows the potential of integrating other "-omics" data for common complex diseases. Nevertheless, further studies, including longitudinal studies, are needed to be conducted to explore the causal relationships between "-omics" data and ultimately T2D affection status.

CHAPTER 6

Conclusions & Future Perspectives

Numerous GWAS have been conducted to discover genes involved in human disease. The genes detected in GWAS, however, must be further investigated to better understand biochemical processes underlying the association. Metabolomics is the evolving field of measuring organic compounds of a cell or body fluid and analysis of metabolites offer to gain further insight into the function of genes. As metabolites are products of genetic processes, they are considered to enlighten biological processes and metabolic functions and are thus highly informative about biology. In this thesis the integration of metabolomics data in the GWAS approach are applied to multiple examples. Furthermore, genetic studies have also shown that the genetic contribution cannot fully explain many diseases and phenotypic traits. In order to explore the missing heritability, more studies now focus on epigenetic modifications. My thesis covered broad aspects of metabolite GWAS and methylation GWAS and below is a brief discussion of what I consider to be the key findings from each chapter.

The first part of my thesis described an application of the metabolomics GWAS approach to three projects, two large collaborations and one novel project. In this section I gained knowledge about different biochemical mechanisms. The possibility to discover novel loci that underlie metabolomic traits is one of the aims of the metabolomics GWAS approach. In total, 145 loci were detected when analyzing almost 8,000 samples in the non-targeted platform from Metabolon and 31 loci were detected when analyzing almost 7,500 samples in the targeted platform from Biocrates. Additionally, I explored the overlap in the two of the platforms in 1,001 individuals

with the aim of identifying stable metabolites on both technologies to ultimately enable combining metabolite profiles across these two platforms. Comparison of 43 metabolites named for the same compound on both platforms indicated strong positive correlations, with few exceptions. Genome-wide association scans with high-throughput metabolic profiles (mGWAS) were performed for each dataset and identified genetic variants at 7 loci that are significantly associated with 16 unique metabolites on both platforms. The 16 metabolites showed consistent genetic associations and appear to be robustly measured across platforms. These included both metabolites named for the same compound across platforms as well as unique metabolites, of which 2 (Biocrates_C9/ Metabolon_X-13431 and Biocrates_lysoPC a C28:1/ Metabolon_1-stearoylglycerol) are likely to represent the same or related biological entities. The results thereby demonstrate the complementary nature of both platforms and can be informative for future studies of comparative and integrative metabolomics analyses in samples profiled on different platforms.

One of the limitations of the mGWAS approach was the computational as well as data storage burden, especially if all pair-wise metabolite ratios were analyzed. But this limitation is also an advantage since introducing all possible metabolites and their ratios into analyses has proven to be successful even if it increases the multiple testing burden. Furthermore, the p-gain was used as an objective measure of the increase in information when considering metabolite ratios over single metabolites, and these analyses gave us further insights into metabolic pathways.

Altogether, the mGWAS results from the three studies have advanced this research area in the following directions: understanding of human metabolism by providing new insight into the role of inherited variation on the metabolome, and potentially providing new prospects for understanding the etiology of diseases and pharmaceutical research.

In the second part of my thesis, I performed GWAS analyses of epigenetic data, identifying genetic variants that influence DNA methylation levels (meQTLs) and variances (var meQTLs). The meQTL results identified to date indicate that genetic variation can have a major effect on the methylome with implications for tissue specificity, tissue shared effects, and shared impacts across multiple gene regulatory processes. I identified 53,813 *cis* meQTLs and 15,392 *trans* meQTLs in whole blood and 40,601 *cis* meQTLs and 24,091 *trans* meQTLs in adipose tissue. The overlap of more than 20,000 *cis* meQTLs between these tissues highlights the extent of tissue shared effects, as well as showing the importance of genetic effects in tissue differentiation. Additionally, the results also point out that *cis* meQTLs are very large in number and tend to be very localized (±1kb). Moreover, this is one of the first studies that profiles *trans* meQTL, showing relatively strong evidence for associations in *trans.* Overall, 18% of the methylome seems to be under genetic control when focusing on the mostly gene-rich regions assayed by Illumina 450k in whole blood.

I also explored the evidence that genetic effects may influence not only DNA methylation levels, but also variances. I explored this by using MZ-twin discordance as a measure of variance. Identifying var QTLs will give more insight into the genetic structure of complex phenotypes and gene-environment interactions. However, the frequency of var QTLs and their impact on phenotypic variability remains unknown, as a relatively unexplored area of research. I identified 8,106 *cis* var meQTLs and 3,694 *trans* var meQTLs in whole blood and 15,396 *cis* var meQTLs and 24,779 *trans* var meQTLs in adipose tissue. The overlap of around 15% *cis var*-meQTLs highlights that genetic impacts on methylation variability tend to be tissue specific, with some tissue shared effects. Strong evidence for *trans* is observed as as well as strong *cis* signals located near the probe, specifically in the 2 kb immediately surrounding the probe.

These results also overlap with eQTL results from other studies, although such shared QTLs represent a small proportion of meQTLs and eQTLs overall. The presence of genetic variants that impact both methylation and gene expression, suggest a causal mechanism by which one genetic variant might affect both processes.

One of the limitations of the methylation GWAS approach was the computational as well as data storage burden, especially when calculating the permutation based FDR, which needed to be limited to 10 permutations although it was genome-wide. In conclusion, the results confirm and extend previous findings of genetic influences on DNA methylation and give insights into mechanisms by which genetic impacts DNA methylation levels and variability.

In the final part of my thesis, I explored metabolomics and epigenetic profiles in the context of T2D. I first analysed the association of these omic data with T2D, identifying both metabolomics and epigenetic signals that were significantly associated with the disease. Identifying metabolomic and epigenetic markers for T2D, both for potential prognosis and understanding pathways involved in progression is essential and further insight to these associations will improve value of existing predictive T2D biomarkers, and thus increase the possibilities to delay, or prevent, T2D in individuals at high risk for the disease.

I then integrated genotype, metabolomics and methylation data to explore causal relationships between the identified associations. Here, I specifically focused on genetic variants that have previously been strongly associated with T2D. I explored three different models using BN, where the SNP affects metabolite and methylation independently from each other, the SNP affects metabolite and then methylation, and the SNP affects methylation then metabolite. I observed that more than half of the significant association follow model two. However, one of the interesting key finding was from model three, that involves a SNP and a methylation probe in the *FTO* gene

and a lipid-controlling hormone metabolite. This model suggests that some of the genetic impacts at *FTO* likely act *via* DNA methylation and these effects influence the levels of the metabolite palmitoleate, which is linked to T2D. This finding is interesting because these variants were associated with T2D before in separate "-omics" analysis but the integrative analysis points to a directional effect in the associations.

My recommendation for future analyses is to incorporate multiple "-omics" technologies together, for example studying combined gene expression and methylation profiles to understand gene regulation mechanisms, as well as their connection to proteomics and metabolomics where available. Using omics data integration might address fundamental biological questions that would increase our understanding of systems as a whole. Moreover, longitudinal studies are also needed to explore causal or a consequential association between the methylation, gene expression, metabolites and phenotype.

In summary, my findings show that genetic variants have major impacts on metabolite profiles in the human body, and this can give insight into biological processes, as well genetic contribution to metabolic processes. Additionally, genetic variants influence epigenetic profiles, and this might give information about the underlying biological processes that regulate human epigenetic variations. Lastly, I explored genetic variants linked to T2D by integrating these data with epigenetic and metabolomics profile to try to identify molecular processes involved in genetic susceptibility to metabolic disease.

APPENDICES

APPENDIX A Supplementary Tables for Chapter 3

Table S3-1. Metabolon and Biocrates platform comparison at 43 overlapping metabolites including correlation, heritability and peak mGWAS results

BIOCRATES	METABOLON	A	C	E	Top SNPs	Chr	p-value	Gene	A	C	E	Top SNPs	Chr	p-value	Gene	cor (r)
C8	octanoylcarnitine	0.60	2E-13	0.40	rs2172507	1	2.45E-08	ACADM	0.45	0.01	0.54	rs4949874	1	3.37E-11	ACADM	0.91
C10	decanoylcarnitine	0.72	3E-15	0.28	rs6810358	3	4.22E-08	NA	0.42	0.00	0.58	rs17650138	1	1.67E-06	NA	0.89
Proline	proline	0.53	2E-01	0.31	rs2011669	3	5.16E-07	NMNAT3	0.67	0.00	0.33	rs17172978	7	4.36E-06	NA	0.85
C6	hexanoylcarnitine	0.46	1E-12	0.54	rs4949874	1	4.14022E-11	ACADM	0.49	0.00	0.51	rs4949874	1	1.62E-13	ACADM	0.85
C4	butyrylcarnitine	0.30	4E-01	0.30	rs2066938	12	2.94439E-44	ACADS	0.26	0.50	0.24	rs2066938	12	1.77E-115	ACADS	0.74
C3	propionylcarnitine	0.46	2E-01	0.30	rs11010004	10	6.73E-10	C10orf112	0.38	0.09	0.52	rs1791780	18	1.52E-06	TAF4B	0.71
Glycine	glycine	0.70	1E-02	0.28	rs4673553	2	5.27E-17	CPS1	0.41	0.00	0.59	rs4673553	2	7.12E-27	CPS1	0.71
C2	acetylcarnitine	0.26	2E-01	0.53	rs4734517	8	2.14E-08	NA	0.00	0.64	0.36	rs160851	5	1.18E-06	NA	0.69
Valine	valine	0.49	2E-02	0.50	rs10508588	10	1.90E-09	LOC100128641	0.35	0.00	0.65	rs7785534	7	2.81E-06	NA	0.68
C5	isovalerylcarnitine	0.00	4E-01	0.65	rs181028	7	3.72E-08	NA	0.27	0.08	0.65	rs181028	7	1.19E-08	NA	0.63
Tyrosine	tyrosine	0.00	5E-01	0.54	rs2227217	6	5.49E-07	NA	0.34	0.00	0.66	rs7785534	7	2.13E-06	NA	0.62
C12	laurylcarnitine	0.63	5E-15	0.37	rs10162284	13	2.39E-06	NA	0.40	0.04	0.56	rs17100308	1	1.59E-07	AK5	0.61
Threonine	threonine	0.34	3E-01	0.37	rs7998783	13	1.21E-06	ITGBL1	0.50	0.00	0.50	rs11758855	6	7.11E-07	NA	0.6

Methionine	methionine	0.23	2E-01	0.56	rs10516614	4	1.88E-07	NA	0.02	0.27	0.71	rs1890573	9	1.39E-06	NA	0.58
Phenylalanine	phenylalanine	0.31	4E-01	0.28	rs10516614	4	1.46E-08	NA	0.35	0.00	0.65	rs12478243	2	1.08E-07	NA	0.54
Tryptophan	tryptophan	0.09	5E-01	0.45	rs17005448	4	5.55E-09	NA	0.36	0.00	0.64	rs318982	11	4.44E-07	HNT	0.53
H1	glucose	0.63	2E-01	0.15	rs11132844	4	1.70E-09	SPOCK3	0.30	0.00	0.70	rs11579657	1	1.16E-08	NA	0.52
Arginine	arginine	0.64	8E-11	0.36	rs17171540	7	4.86E-08	POU6F2	0.38	0.00	0.62	rs1714708	8	5.29E-08	CSMD1	0.5
Serine	serine	0.46	3E-01	0.29	rs10131545	14	3.27E-08	DYNC1H1	0.35	0.02	0.62	rs477992	1	1.42E-08	PHGDH	0.48
C14:1	2-tetradecenoyl carnitine	0.54	1E-14	0.46	rs757423	5	5.53E-07	SPOCK1	0.40	0.04	0.56	rs535114	3	2.36E-06	NA	0.45
C18:1	oleoylcarnitine	0.32	2E-01	0.43	rs1323570	13	3.92E-07	LMO7	0.39	0.02	0.59	rs201267	6	6.51E-07	OFCC1	0.44
lysoPC a C16:1	1-palmitoleoylglycerophosphocholine	0.61	1E-02	0.37	rs12405027	1	3.61E-07	NA	0.27	0.09	0.64	rs970048	6	2.08E-06	NA	0.44
C0	carnitine	0.64	5E-02	0.31	rs1171614	10	4.67E-12	SLC16A9	0.34	0.00	0.66	rs1171617	10	2.37E-13	SLC16A9	0.38
Histidine	histidine	0.21	5E-01	0.33	rs17005448	4	2.83E-08	NA	0.00	0.87	0.13	rs7439210	4	6.22E-10	SLC2A9	0.37
C5-DC	glutaroyl carnitine	0.00	4E-01	0.61	rs6734430	2	1.41E-06	NA	0.50	0.00	0.50	rs4949874	1	6.87E-13	NA	0.37
lysoPC a C14:0	1-myristoylglycerophosphocholine	0.64	2E-13	0.36	rs6821395	4	1.31E-06	NA	0.33	0.05	0.62	rs1455659	4	2.20E-06	NA	0.37
Glutamine	glutamine	0.44	1E-01	0.46	rs2332327	14	5.13E-08	NFATC4	0.42	0.00	0.58	rs774211	12	3.43E-08	RBMS2	0.33
lysoPC a C20:4	1-arachidonoylglycerophosphocholine	0.57	1E-01	0.33	rs174547	11	2.00E-14	FADS1	0.09	0.00	0.91	rs174547	11	2.98E-10	FADS1	0.29
C5-OH	hydroxyisovaleroyl carnitine	0.15	1E-01	0.73	rs17053040	5	1.80E-06	NA	0.42	0.32	0.26	rs10521155	17	1.78E-06	STX8	0.28
lysoPC a C18:2©	2-linoleoylglycerophosphocholine	0.61	5E-12	0.39	rs7529794	1	9.54E-07	NA	0.14	0.03	0.83	rs6871464	5	1.33E-06	NA	0.27
lysoPC a C18:2(D)	1-linoleoylglycerophosphocholine	0.61	5E-12	0.39	rs7529794	1	9.54E-07	NA	0.28	0.00	0.72	rs9829101	3	4.97E-06	NA	0.26
lysoPC a C20:3	1-eicosatrienoylglycerophosphocholine	0.50	1E-01	0.36	rs12872445	13	2.90E-06	NA	0.06	0.01	0.93	rs9829101	3	3.72E-07	NA	0.26
lysoPC a C18:1(A)	1-oleoylglycerophosphocholine	0.60	2E-12	0.40	rs10993045	9	6.62E-07	PTPDC1	0.32	0.00	0.68	rs6845407	4	7.82E-07	NA	0.26
lysoPC a C18:1(B)	2-oleoylglycerophosphocholine	0.60	2E-12	0.40	rs10993045	9	6.62E-07	PTPDC1	0.31	0.04	0.65	rs6545474	2	2.78E-06	NA	0.26

C16	palmitoylcarnitine	0.66	2E-02	0.32	rs757423	5	2.54E-06	SPOCK1	0.38	0.00	0.62	rs2350581	8	1.18E-06	NA	0.24
C18	stearoylcarnitine	0.00	5E-01	0.47	rs267573	9	4.48E-06	NA	0.00	0.52	0.48	rs4952462	2	3.21E-06	NA	0.22
Ornithine	ornithine	0.42	2E-01	0.34	rs11677129	2	1.01E-07	GALNT13	0.30	0.10	0.59	rs9924984	16	7.65E-08	NA	0.19
lysoPC a C17:0	1-heptadecanoylglycerophosphocholine	0.51	5E-02	0.44	rs2598089	7	1.54E-07	NA	0.38	0.15	0.47	rs6580376	5	5.35E-07	NA	0.1
lysoPC a C16:0(G)	2-palmitoylglycerophosphocholine	0.45	2E-01	0.38	rs17005448	4	3.15E-07	NA	0.16	0.09	0.75	rs11886877	2	8.71E-07	IL1R2	0.08
lysoPC a C18:0(E)	1-stearoylglycerophosphocholine	0.45	2E-01	0.40	rs11010603	10	4.12E-06	LOC100128641	0.52	0.48	0.00	rs974916	2	7.01E-08	FLJ42562	0.08
lysoPC a C16:0(H)	1-palmitoylglycerophosphocholine	0.45	2E-01	0.38	rs17005448	4	3.15E-07	NA	0.32	0.00	0.68	rs355829	2	1.02E-08	COBLL1	0.05
lysoPC a C18:0(F)	2-stearoylglycerophosphocholine	0.45	2E-01	0.40	rs11010603	10	4.12E-06	LOC100128641	0.00	0.06	0.94	rs974916	2	9.75E-07	FLJ42562	0.05
lysoPC a C26:1	1-docosahexaenoylglycerophosphocholine	0.29	3E-01	0.45	rs4798428	18	5.89E-06	L3MBTL4	0.43	0.01	0.56	rs7975402	12	5.08E-06	PRICKLE1	0

Table S3-2. mGWAS results for Biocrates and Metabolon platforms

				BIOCRATES					METABOLON				
GENE	**CHR**	**RS**	**PS**	**N_MISS**	**BETA**	**SE**	**p_WALD**	**METABOLITE**	**N_MISS**	**BETA**	**SE**	**p_WALD**	**METABOLITE**
CYP4B1	1	rs4646493	47,053,889						0	4.28E-01	6.27E-02	9.05E-12	10-undecenoate (11:1n1)
ACADM	1	rs211718	75,879,263						42	-2.80E-01	5.08E-02	3.83E-08	X- 11421
ACADM	1	rs4949874	75,934,477	8	-4.96E-02	7.52E-03	4.14E-11	c6	8	-3.47E-01	4.71E-02	1.62E-13	hexanoylcarnitine
ACADM	1	rs4949874	75,934,477						8	-3.15E-01	4.74E-02	3.37E-11	octanoylcarnitine
ACADM	1	rs2172507	76,103,908	1	-4.68E-02	8.39E-03	2.45E-08	c8					
GCKR	2	rs1260326	27,584,444						0	-3.11E-01	4.30E-02	4.59E-13	mannose
NAT8	2	rs10169714	73,662,281						3	-8.65E-01	4.01E-02	4.18E-103	N-acetylornithine
NAT8	2	rs7558944	73,664,417						4	-5.06E-01	5.43E-02	1.12E-20	X- 11787
NAT8	2	rs13410232	73,725,197						27	-4.92E-01	5.72E-02	7.79E-18	X- 12510
NAT8	2	rs13410232	73,725,197						15	4.45E-01	6.94E-02	1.44E-10	X- 12093
CREG2	2	rs6751877	101,342,584						2	-7.86E-01	7.79E-02	5.72E-24	N-(2-furoyl)glycine
ACADL	2	rs12612970	210,715,532						9	5.08E-01	4.90E-02	3.51E-25	X- 13431

ACADL	2	rs7601356	210,764,902	10	1.16E-01	9.06E-03	9.70E-38	c9					
CPS1	2	rs4673553	211,316,624	0	5.51E-02	6.58E-03	5.27E-17	gly	0	4.37E-01	4.07E-02	7.12E-27	glycine
CPS1	2	rs4673553	211,316,624						0	3.48E-01	5.18E-02	1.68E-11	X- 08988
UGT1A	2	rs887829	234,333,309						5	4.32E-01	5.59E-02	1.02E-14	X- 11530
UGT1A	2	rs887829	234,333,309						6	4.32E-01	5.59E-02	1.05E-14	X- 11793
UGT1A	2	rs4148325	234,338,048						7	4.58E-01	4.86E-02	4.84E-21	bilirubin (Z,Z)
UGT1A	2	rs4148325	234,338,048						7	3.74E-01	5.11E-02	2.38E-13	bilirubin (E,E)
SERPINI2	3	rs9864094	168,642,525						2	1.36E+00	1.25E-01	1.56E-27	X- 12435
SLC2A9	4	rs737267	9,543,842						0	-4.23E-01	4.62E-02	5.67E-20	urate
GBA3	4	rs3099557	22,433,119						0	4.66E-01	6.13E-02	2.96E-14	X- 11818
SLC22A5	5	rs274552	131,755,245						33	-4.07E-01	6.40E-02	1.93E-10	X- 11255
F12	5	rs2731672	176,775,080						0	-3.57E-01	4.98E-02	8.04E-13	X- 11792
SLC22A2	6	rs316020	160,589,071						5	-1.02E+00	8.13E-02	5.82E-36	X- 12798
CYP3A5	7	rs11974702	99,001,887						13	-7.78E-01	7.69E-02	4.53E-24	androsterone sulfate
CYP3A5	7	rs11974702	99,001,887						13	-6.45E-01	7.89E-02	2.91E-16	epiandrosterone sulfate
CYP3A5	7	rs1859690	99,065,108						0	-9.07E-01	1.06E-01	8.52E-18	X- 12063
OPLAH	8	rs11780874	145,221,292						2	-4.30E-01	6.20E-02	3.88E-12	5-oxoproline
SLC16A9	10	rs1171617	61,137,188						0	-3.78E-01	5.16E-02	2.37E-13	carnitine
SLC16A9	10	rs1171614	61,139,544	6	-4.18E-02	6.04E-03	4.67E-12	c0					
PYROXD2	10	rs2147896	100,138,166						0	-1.04E+00	3.51E-02	4.26E-191	X- 12092
FADS1	11	rs174546	61,326,406	0	-3.46E-02	6.17E-03	1.95E-08	pc_ae_c42_5	0	-3.13E-01	5.49E-02	1.19E-08	1-linoleoylglycerophosphoethanolamine
FADS1	11	rs174547	61,327,359	1	-6.58E-02	8.60E-03	2.00E-14	lysopc_a_c20_4	1	-3.93E-01	6.23E-02	2.98E-10	1-arachidonoylglycerophosphocholine
FADS1	11	rs174547	61,327,359						1	-3.24E-01	5.23E-02	5.59E-10	arachidonate (20:4n6)
FADS1	11	rs174556	61,337,211	0	-5.15E-02	7.66E-03	1.73E-11	pc_aa_c38_5					
FADS1	11	rs1535	61,354,548	0	-5.47E-02	7.49E-03	2.83E-13	pc_ae_c36_5					
FADS1	11	rs1535	61,354,548	0	-4.46E-02	6.79E-03	5.16E-11	pc_ae_c38_5					
FADS1	11	rs174576	61,360,086	16	-6.82E-02	7.42E-03	3.62E-20	pc_aa_c38_4					
FADS1	11	rs174576	61,360,086	16	-5.37E-02	7.22E-03	1.06E-13	pc_aa_c36_4					

FADS1	11	rs174576	61,360,086	16	-4.79E-02	6.82E-03	2.25E-12	pc_ae_c38_4					
SLCO1B1	12	rs4149056	21,222,816						8	6.99E-01	5.19E-02	2.93E-41	X- 11529
SLCO1B1	12	rs4149081	21,269,288						3	4.98E-01	6.75E-02	1.60E-13	X- 11538
SLCO1B1	12	rs2199680	21,306,763						14	4.76E-01	7.31E-02	7.07E-11	1-eicosadienoylglycerophosphocholine
ACADS	12	rs2066938	119,644,998	0	1.20E-01	8.61E-03	2.94E-44	c4	0	1.01E+00	4.42E-02	1.77E-115	butyrylcarnitine
SGPP1	14	rs7157785	63,305,309	4	4.57E-02	8.31E-03	3.81E-08	pc_aa_c28_1	4	4.51E-01	5.92E-02	2.77E-14	1-stearoylglycerol (1-monostearin)
SGPP1	14	rs7157785	63,305,309						4	3.81E-01	6.26E-02	1.24E-09	X- 10510
DYNC1H1	14	rs10131545	101,529,478	6	-9.89E-02	1.48E-02	2.25E-11	sm__oh__c16_1					
DYNC1H1	14	rs10131545	101,529,478	6	-9.91E-02	1.49E-02	3.30E-11	pc_ae_c36_2					
DYNC1H1	14	rs10131545	101,529,478	6	-8.87E-02	1.38E-02	1.28E-10	sm__oh__c22_1					
DYNC1H1	14	rs10131545	101,529,478	6	-9.26E-02	1.47E-02	3.16E-10	pc_ae_c34_2					
ACE	17	rs4329	58,917,190						0	-4.15E-01	5.29E-02	4.36E-15	aspartylphenylalanine
SULT2A1	19	rs2547231	53,076,869						0	-5.88E-01	5.90E-02	2.09E-23	X- 11440
SULT2A1	19	rs2547231	53,076,869						0	-5.14E-01	5.89E-02	2.58E-18	X- 11244
COMT	22	rs165722	18,329,013						28	4.33E-01	3.54E-02	2.61E-34	X- 11593

Table S3-3. mGWAS Ratio Results for Biocrates and Metabolon platforms

				BIOCRATES							METABOLON						
GENE	**CHR**	**RS**	**PS**	**N_MISS**	**BETA**	**SE**	**p_WALD**	**p_LRT**	**p_SCORE**	**METABOLITE RATIO**	**N_MISS**	**BETA**	**SE**	**p_WALD**	**p_LRT**	**p_SCORE**	**METABOLITE RATIO**
ACADM	1	rs4949874	75934477								8	-6.95E-02	7.29E-03	1.49E-21	8.68E-21	5.94E-20	acetylcarnitine/hexanoylcarnitine
ACADM	1	rs7534754	75957896	15	-6.40E-02	7.26E-03	1.19E-18	4.40E-18	1.88E-17	c2/c6							
CPS1	2	rs16844839	211289490	43	-7.68E-02	7.71E-03	2.11E-23	2.99E-22	4.39E-21	gly/phe							
CPS1	2	rs4673553	211316624	0	-4.38E-02	4.13E-03	3.30E-26	7.81E-25	1.86E-23	gly/ser							
CPS1	2	rs4673553	211316624	0	-5.90E-02	5.28E-03	4.67E-29	1.68E-27	6.19E-26	gly/gln							
CPS1	2	rs4673553	211316624	0	-5.84E-02	5.50E-03	2.57E-26	4.39E-25	8.25E-24	gly/trp							
CPS1	2	rs4673553	211316624	0	-5.68E-02	6.02E-03	4.11E-21	2.14E-20	1.30E-19	gly/h1							
CPS1	2	rs4673553	211316624	0	-6.25E-02	5.51E-03	8.51E-30	3.68E-28	1.61E-26	gly/arg							

CPS1	2	rs4673553	211316624	0	-6.69E-02	6.05E-03	1.99E-28	6.93E-27	2.44E-25	gly/thr							
CPS1	2	rs4673553	211316624	0	-6.27E-02	6.64E-03	3.63E-21	2.22E-20	1.54E-19	gly/tyr							
CPS1	2	rs4673553	211316624	0	-6.03E-02	5.38E-03	3.46E-29	1.42E-27	5.84E-26	gly/met							
CPS1	2	rs4673553	211316624	0	-6.13E-02	5.49E-03	6.29E-29	2.13E-27	7.51E-26	gly/his							
FADS1	11	rs174536	61308503								0	7.70E-02	7.55E-03	1.95E-24	3.46E-23	6.32E-22	1-arachidonoylglycerophosphocholine /1-palmitoleoylglycerophosphocholine
FADS1	11	rs174536	61308503								0	6.57E-02	6.97E-03	4.06E-21	4.27E-20	4.57E-19	1-arachidonoylglycerophosphocholine /2-palmitoylglycerophosphocholine
FADS1	11	rs174546	61326406								0	6.15E-02	8.00E-03	1.58E-14	4.54E-14	1.43E-13	1-arachidonoylglycerophosphocholine /1-stearoylglycerophosphocholine
FADS1	11	rs174546	61326406								0	7.15E-02	7.39E-03	3.87E-22	3.81E-21	4.09E-20	1-arachidonoylglycerophosphocholine /2-oleoylglycerophosphocholine
FADS1	11	rs174547	61327359	1	7.04E-02	4.58E-03	3.06E-53	1.19E-47	1.14E-42	lysopc_a_c20_4 /lysopc_a_c18_1	2	9.24E-02	6.82E-03	8.37E-42	3.31E-38	7.61E-35	1-arachidonoylglycerophosphocholine /1-eicosatrienoylglycerophosphocholine
FADS1	11	rs174547	61327359	1	7.04E-02	4.58E-03	3.06E-53	1.19E-47	1.14E-42	lysopc_a_c20_4 /lysopc_a_c18_1.1							
FADS1	11	rs174547	61327359	1	8.82E-02	6.14E-03	9.41E-47	1.37E-42	9.66E-39	lysopc_a_c20_4 /lysopc_a_c18_2							
FADS1	11	rs174547	61327359	1	8.82E-02	6.14E-03	9.41E-47	1.37E-42	9.66E-39	lysopc_a_c20_4 /lysopc_a_c18_2.1							
FADS1	11	rs174547	61327359	1	5.98E-02	5.65E-03	3.63E-26	6.84E-25	1.37E-23	lysopc_a_c20_4 /lysopc_a_c18_0							
FADS1	11	rs174547	61327359	1	5.98E-02	5.65E-03	3.63E-26	6.84E-25	1.37E-23	lysopc_a_c20_4 /lysopc_a_c18_0.1							
FADS1	11	rs174549	61327958	2	9.53E-02	4.86E-03	1.56E-85	8.01E-72	1.02E-60	lysopc_a_c20_4 \lysopc_a_c20_3							
FADS1	11	rs174549	61327958	2	7.58E-02	5.77E-03	2.19E-39	2.04E-36	1.39E-33	lysopc_a_c20_4 /lysopc_a_c16_1							
FADS1	11	rs174549	61327958	2	6.86E-02	5.04E-03	2.76E-42	1.17E-38	2.60E-35	lysopc_a_c20_4 /lysopc_a_c16_0							
FADS1	11	rs1535	61354548								0	7.23E-02	6.39E-03	1.04E-29	1.45E-27	1.51E-25	1-arachidonoylglycerophosphocholine /1-oleoylglycerophosphocholine
FADS1	11	rs1535	61354548								0	9.34E-02	8.28E-03	1.67E-29	2.83E-27	3.55E-25	1-arachidonoylglycerophosphocholine /2-linoleoylglycerophosphocholine
FADS1	11	rs1535	61354548								0	8.06E-02	6.38E-03	1.52E-36	1.19E-33	6.31E-31	1-arachidonoylglycerophosphocholine /1-linoleoylglycerophosphocholine
FADS1	11	rs1535	61354548								0	6.48E-02	7.31E-03	7.97E-19	5.62E-18	4.07E-17	1-arachidonoylglycerophosphocholine /1-palmitoylglycerophosphocholine
FADS1	11	rs174574	61356918	49	7.23E-02	7.09E-03	2.06E-24	3.74E-23	6.82E-22	lysopc_a_c20_4 /lysopc_a_c14_0							
ACADS	12	rs2066938	119644998	0	1.16E-01	6.05E-03	1.02E-81	2.79E-68	1.25E-57	c3/c4	0	2.22E-01	8.64E-03	3.56E-146	1.83E-109	3.36E-84	carnitine/butyrylcarnitine
ACADS	12	rs2066938	119644998	0	1.10E-01	1.00E-02	4.24E-28	2.01E-26	8.72E-25	c5_dc/c4	0	2.19E-01	8.35E-03	1.85E-151	1.04E-112	3.34E-86	acetylcarnitine/butyrylcarnitine
ACADS	12	rs2066938	119644998	0	1.13E-01	6.93E-03	3.37E-60	8.97E-53	2.02E-46	c0/c4	0	2.21E-01	8.38E-03	6.83E-153	1.13E-113	7.66E-87	propionylcarnitine/butyrylcarnitine
ACADS	12	rs2066938	119644998	0	-1.17E-01	7.95E-03	6.05E-49	5.97E-44	1.62E-39	c4/val	0	2.26E-01	8.77E-03	7.57E-147	2.07E-108	7.92E-83	hexanoylcarnitine/butyrylcarnitine

Table S3-4. p-gain calculations for 101 overlap mGWAS Ratio Results for Biocrates and Metabolon platforms

BIOCRATES RATIO	p-GAIN_BIOC	p-value-RATIO-BIOC	MIN(p-value BIOC)	METABOLON RATIO	p-GAIN_METAB	p-value_RATIO_MET	MIN(p-value MET)
lysopc_a_c20_4/lysopc_a_c20_3	1.28205E+71	1.56E-85	2.00E-14	1-arachidonoylglycerophosphocholine_1-eicosatrienoylglycerophosphocholine	6.37993E+31	8.37E-42	5.34E-10
lysopc_a_c20_4/lysopc_a_c18_2.1	2.1254E+32	9.41E-47	2.00E-14	1-arachidonoylglycerophosphocholine_1-linoleoylglycerophosphocholine	3.51316E+26	1.52E-36	5.34E-10
lysopc_a_c20_4/lysopc_a_c18_1	6.53595E+38	3.06E-53	2.00E-14	1-arachidonoylglycerophosphocholine_1-oleoylglycerophosphocholine	5.13462E+19	1.04E-29	5.34E-10
lysopc_a_c20_4/lysopc_a_c18_2	2.1254E+32	9.41E-47	2.00E-14	1-arachidonoylglycerophosphocholine_2-linoleoylglycerophosphocholine	3.1976E+19	1.67E-29	5.34E-10
c3/c4	2.88235E+37	1.02E-81	2.94E-44	propionylcarnitine_butyrylcarnitine	1.10688E+18	6.83E-153	7.56E-135
c2/c4	1.46269E-15	2.01E-29	2.94E-44	acetylcarnitine_butyrylcarnitine	4.08649E+16	1.85E-151	7.56E-135
lysopc_a_c20_4/lysopc_a_c16_1	9.13242E+24	2.19E-39	2.00E-14	1-arachidonoylglycerophosphocholine_1-palmitoleoylglycerophosphocholine	2.73846E+14	1.95E-24	5.34E-10
lysopc_a_c20_4/lysopc_a_c18_1.1	6.53595E+38	3.06E-53	2.00E-14	1-arachidonoylglycerophosphocholine_2-oleoylglycerophosphocholine	1.37984E+12	3.87E-22	5.34E-10
c6/c4	1.58065E-07	1.86E-37	2.94E-44	hexanoylcarnitine_butyrylcarnitine	9.98679E+11	7.57E-147	7.56E-135
c0/c4	8.72404E+15	3.37E-60	2.94E-44	carnitine_butyrylcarnitine	2.1236E+11	3.56E-146	7.56E-135
lysopc_a_c20_4/lysopc_a_c18_0	5.50964E+11	3.63E-26	2.00E-14	1-arachidonoylglycerophosphocholine_2-palmitoylglycerophosphocholine	1.31527E+11	4.06E-21	5.34E-10
lysopc_a_c20_4/lysopc_a_c18_0.1	5.50964E+11	3.63E-26	2.00E-14	1-arachidonoylglycerophosphocholine_1-palmitoylglycerophosphocholine	670012547.1	7.97E-19	5.34E-10
c2/c6	34789915.97	1.19E-18	4.14E-11	acetylcarnitine_hexanoylcarnitine	461073825.5	1.49E-21	6.87E-13
lysopc_a_c20_4/lysopc_a_c16_0	7.24638E+27	2.76E-42	2.00E-14	1-arachidonoylglycerophosphocholine_1-stearoylglycerophosphocholine	33797.46835	1.58E-14	5.34E-10
c8/c2	452.4312896	4.73E-11	2.14E-08	octanoylcarnitine_acetylcarnitine	3222.632226	8.13E-14	2.62E-10
c4/val	48595.04132	6.05E-49	2.94E-44	butyrylcarnitine_valine	517.8082192	1.46E-137	7.56E-135
c6/phe	1012.224939	4.09E-14	4.14E-11	hexanoylcarnitine_phenylalanine	253.5055351	2.71E-15	6.87E-13
arg/lysopc_a_c20_4	2534.854246	7.89E-18	2.00E-14	arginine_1-arachidonoylglycerophosphocholine	48.10810811	1.11E-11	5.34E-10
c6/val	646.875	6.40E-14	4.14E-11	hexanoylcarnitine_valine	18.22281167	3.77E-14	6.87E-13
lysopc_a_c20_4/lysopc_a_c14_0	9708737864	2.06E-24	2.00E-14	1-arachidonoylglycerophosphocholine_1-myristoylglycerophosphocholine	16.6875	3.20E-11	5.34E-10
c6/c0	812.173913	5.75E-15	4.67E-12	hexanoylcarnitine_carnitine	16.39618138	4.19E-14	6.87E-13
gly/met	1.52312E+12	3.46E-29	5.27E-17	glycine_methionine	6.4	2.00E-27	1.28E-26
c6/tyr	16.23529412	2.55E-12	4.14E-11	hexanoylcarnitine_tyrosine	3.195348837	2.15E-13	6.87E-13
gly/gln	1.12848E+12	4.67E-29	5.27E-17	glycine_glutamine	2.746781116	4.66E-27	1.28E-26
gly/his	8.37838E+11	6.29E-29	5.27E-17	glycine_histidine	1.693121693	7.56E-27	1.28E-26
c6/c5_oh	1.140495868	3.63E-11	4.14E-11	hexanoylcarnitine_hydroxyisovaleroyl carnitine	1.033082707	6.65E-13	6.87E-13

c3/c6	9.241071429	4.48E-12	4.14E-11	propionylcarnitine_hexanoylcarnitine	0.660576923	1.04E-12	6.87E-13
c6/c5	4456.404736	9.29E-15	4.14E-11	hexanoylcarnitine_isovalerylcarnitine	0.116243655	5.91E-12	6.87E-13
c6/met	2.539877301	1.63E-11	4.14E-11	hexanoylcarnitine_methionine	0.115268456	5.96E-12	6.87E-13
c6/c5_dc	1109.919571	3.73E-14	4.14E-11	hexanoylcarnitine_glutaroyl carnitine	0.066699029	1.03E-11	6.87E-13
c4/phe	0.060995851	4.82E-43	2.94E-44	butyrylcarnitine_phenylalanine	0.05641791	1.34E-133	7.56E-135
c6/gln	655.0632911	6.32E-14	4.14E-11	hexanoylcarnitine_glutamine	0.055853659	1.23E-11	6.87E-13
c6/trp	11.0106383	3.76E-12	4.14E-11	hexanoylcarnitine_tryptophan	0.05452381	1.26E-11	6.87E-13
c6/h1	1.761702128	2.35E-11	4.14E-11	hexanoylcarnitine_glucose	0.047708333	1.44E-11	6.87E-13
gly/ser	1596969697	3.30E-26	5.27E-17	glycine_serine	0.047583643	2.69E-25	1.28E-26
c6/his	3.980769231	1.04E-11	4.14E-11	hexanoylcarnitine_histidine	0.016514423	4.16E-11	6.87E-13
c4/met	0.355072464	8.28E-44	2.94E-44	butyrylcarnitine_methionine	0.002088398	3.62E-132	7.56E-135
gly/thr	2.64824E+11	1.99E-28	5.27E-17	glycine_threonine	0.000579186	2.21E-23	1.28E-26
gly/phe	2497630.332	2.11E-23	5.27E-17	glycine_phenylalanine	0.000579186	2.21E-23	1.28E-26
gly/val	2960.674157	1.78E-20	5.27E-17	glycine_valine	2.21071E-05	5.79E-22	1.28E-26
gly/trp	2050583658	2.57E-26	5.27E-17	glycine_tryptophan	7.61905E-06	1.68E-21	1.28E-26
gly/tyr	14517.90634	3.63E-21	5.27E-17	glycine_tyrosine	3.01176E-06	4.25E-21	1.28E-26
gly/h1	12822.38443	4.11E-21	5.27E-17	glycine_glucose	1.84971E-06	6.92E-21	1.28E-26
gly/pro	1.138228942	4.63E-17	5.27E-17	glycine_proline	5.76577E-08	2.22E-19	1.28E-26
gly/lysopc_a_c18_0.1	0.000151003	3.49E-13	5.27E-17	glycine_1-palmitoylglycerophosphocholine	1.6732E-08	7.65E-19	1.28E-26
gly/c0	0.393283582	1.34E-16	5.27E-17	glycine_carnitine	1.21905E-08	1.05E-18	1.28E-26
gly/c5_dc	1.091097308	4.83E-17	5.27E-17	glycine_glutaroyl carnitine	4.83019E-09	2.65E-18	1.28E-26
trp/c4	3307.086614	8.89E-48	2.94E-44	tryptophan_butyrylcarnitine	1.16308E-09	6.50E-126	7.56E-135
gln/c4	1.36744186	2.15E-44	2.94E-44	glutamine_butyrylcarnitine	4.15385E-10	1.82E-125	7.56E-135
c2/gly	0.075609756	6.97E-16	5.27E-17	acetylcarnitine_glycine	3.03318E-10	4.22E-17	1.28E-26
gly/lysopc_a_c18_1.1	5.94138E-05	8.87E-13	5.27E-17	glycine_2-oleoylglycerophosphocholine	1.2549E-10	1.02E-16	1.28E-26
gly/lysopc_a_c18_2.1	0.000270256	1.95E-13	5.27E-17	glycine_1-linoleoylglycerophosphocholine	8.15287E-11	1.57E-16	1.28E-26
gly/c5_oh	0.024285714	2.17E-15	5.27E-17	glycine_hydroxyisovaleroyl carnitine	2.03498E-11	6.29E-16	1.28E-26
gly/lysopc_a_c18_2	0.000270256	1.95E-13	5.27E-17	glycine_2-linoleoylglycerophosphocholine	3.45013E-12	3.71E-15	1.28E-26
gly/lysopc_a_c18_0	0.000151003	3.49E-13	5.27E-17	glycine_2-palmitoylglycerophosphocholine	1.63057E-12	7.85E-15	1.28E-26

gly/arg	6.19271E+12	8.51E-30	5.27E-17	glycine_arginine	1.04065E-12	1.23E-14	1.28E-26
c4/tyr	1.23013E-05	2.39E-39	2.94E-44	butyrylcarnitine_tyrosine	5.90625E-13	1.28E-122	7.56E-135
gly/lysopc_a_c18_1	5.94138E-05	8.87E-13	5.27E-17	glycine_1-oleoylglycerophosphocholine	2.19554E-13	5.83E-14	1.28E-26
c10/gly	1.43207E-05	3.68E-12	5.27E-17	decanoylcarnitine_glycine	1.54031E-13	8.31E-14	1.28E-26
c3/gly	1.9812E-05	2.66E-12	5.27E-17	propionylcarnitine_glycine	5.63877E-14	2.27E-13	1.28E-26
c6/gly	1.09336E-05	4.82E-12	5.27E-17	hexanoylcarnitine_glycine	3.60563E-14	3.55E-13	1.28E-26
gly/lysopc_a_c16_0	5.82965E-05	9.04E-13	5.27E-17	glycine_1-stearoylglycerophosphocholine	2.35727E-14	5.43E-13	1.28E-26
gly/c16	0.698938992	7.54E-17	5.27E-17	glycine_palmitoylcarnitine	2.11221E-14	6.06E-13	1.28E-26
c8/gly	0.000202692	2.60E-13	5.27E-17	octanoylcarnitine_glycine	1.47636E-14	8.67E-13	1.28E-26
gly/lysopc_a_c26_1	0.000590807	8.92E-14	5.27E-17	glycine_1-docosahexaenoylglycerophosphocholine	1.08475E-14	1.18E-12	1.28E-26
h1/c4	0.013611111	2.16E-42	2.94E-44	glucose_butyrylcarnitine	3.78E-15	2.00E-120	7.56E-135
gly/c18_1	11.73719376	4.49E-18	5.27E-17	glycine_oleoylcarnitine	2.5498E-15	5.02E-12	1.28E-26
c5/gly	1.11416E-05	4.73E-12	5.27E-17	isovalerylcarnitine_glycine	6.5641E-16	1.95E-11	1.28E-26
c4/his	6.606741573	4.45E-45	2.94E-44	butyrylcarnitine_histidine	1.82169E-19	4.15E-116	7.56E-135
c8/c4	6.47577E-08	4.54E-37	2.94E-44	octanoylcarnitine_butyrylcarnitine	2.16619E-23	3.49E-112	7.56E-135
pro/c4	2.26154E-14	1.30E-30	2.94E-44	proline_butyrylcarnitine	4.47337E-31	1.69E-104	7.56E-135
c5_dc/c4	9.76415E+16	4.24E-28	4.14E-11	glutaroyl carnitine_butyrylcarnitine	1.75814E-31	4.30E-104	7.56E-135
c4/lysopc_a_c18_0.1	3.69811E-11	7.95E-34	2.94E-44	butyrylcarnitine_1-palmitoylglycerophosphocholine	4.1087E-33	1.84E-102	7.56E-135
c10/c4	6.17647E-15	4.76E-30	2.94E-44	decanoylcarnitine_butyrylcarnitine	1.39227E-38	5.43E-97	7.56E-135
c4/c5_oh	483.5526316	6.08E-47	2.94E-44	butyrylcarnitine_hydroxyisovaleroyl carnitine	3.72414E-41	2.03E-94	7.56E-135
ser/c4	1.34247E-07	2.19E-37	2.94E-44	serine_butyrylcarnitine	1.02578E-41	7.37E-94	7.56E-135
c5/c4	0.945337621	3.11E-44	2.94E-44	isovalerylcarnitine_butyrylcarnitine	4.60976E-45	1.64E-90	7.56E-135
gly/c4	2.19403E-10	1.34E-34	2.94E-44	glycine_butyrylcarnitine	4.97368E-47	1.52E-88	7.56E-135
c16/c4	1.37383E-07	2.14E-37	2.94E-44	palmitoylcarnitine_butyrylcarnitine	2.52E-47	3.00E-88	7.56E-135
c4/c18	8.32861E-14	3.53E-31	2.94E-44	butyrylcarnitine_stearoylcarnitine	9.53342E-49	7.93E-87	7.56E-135
thr/c4	1.77108E-12	1.66E-32	2.94E-44	threonine_butyrylcarnitine	3.12397E-49	2.42E-86	7.56E-135
arg/c4	1.126436782	2.61E-44	2.94E-44	arginine_butyrylcarnitine	4.52695E-55	1.67E-80	7.56E-135
c4/lysopc_a_c18_1.1	2.72222E-11	1.08E-33	2.94E-44	butyrylcarnitine_2-oleoylglycerophosphocholine	2.72924E-57	2.77E-78	7.56E-135
c4/lysopc_a_c18_2.1	1.27826E-12	2.30E-32	2.94E-44	butyrylcarnitine_1-linoleoylglycerophosphocholine	8.82147E-59	8.57E-77	7.56E-135

c18_1/c4	2.39024E-11	1.23E-33	2.94E-44	oleoylcarnitine_butyrylcarnitine	5.72727E-60	1.32E-75	7.56E-135
c4/lysopc_a_c16_1	3.64764E-15	8.06E-30	2.94E-44	butyrylcarnitine_1-palmitoleoylglycerophosphocholine	1.15596E-61	6.54E-74	7.56E-135
c4/lysopc_a_c18_1	2.72222E-11	1.08E-33	2.94E-44	butyrylcarnitine_1-oleoylglycerophosphocholine	1.63283E-62	4.63E-73	7.56E-135
c4/lysopc_a_c18_0	3.69811E-11	7.95E-34	2.94E-44	butyrylcarnitine_2-palmitoylglycerophosphocholine	2.23009E-64	3.39E-71	7.56E-135
c4/lysopc_a_c16_0	2.72222E-09	1.08E-35	2.94E-44	butyrylcarnitine_1-stearoylglycerophosphocholine	5.32394E-65	1.42E-70	7.56E-135
c4/lysopc_a_c18_2	1.27826E-12	2.30E-32	2.94E-44	butyrylcarnitine_2-linoleoylglycerophosphocholine	1.02997E-66	7.34E-69	7.56E-135
c4/lysopc_a_c14_0	1.52332E-09	1.93E-35	2.94E-44	butyrylcarnitine_1-myristoylglycerophosphocholine	5.86047E-67	1.29E-68	7.56E-135
c4/lysopc_a_c20_3	4.10615E-10	7.16E-35	2.94E-44	butyrylcarnitine_1-eicosatrienoylglycerophosphocholine	1.27703E-67	5.92E-68	7.56E-135
c12/c4	5.74219E-13	5.12E-32	2.94E-44	laurylcarnitine_butyrylcarnitine	5.00662E-73	1.51E-62	7.56E-135
c4/orn	1.13953E-09	2.58E-35	2.94E-44	butyrylcarnitine_ornithine	1.79572E-73	4.21E-62	7.56E-135
c4/lysopc_a_c20_4	1.20492E-13	2.44E-31	2.94E-44	butyrylcarnitine_1-arachidonoylglycerophosphocholine	1.53347E-79	4.93E-56	7.56E-135
c4/lysopc_a_c26_1	0.036842105	7.98E-43	2.94E-44	butyrylcarnitine_1-docosahexaenoylglycerophosphocholine	1.39741E-80	5.41E-55	7.56E-135
c14_1/c4	2.37097E-09	1.24E-35	2.94E-44	2-tetradecenoyl carnitine_butyrylcarnitine	5.68421E-84	1.33E-51	7.56E-135
c4/lysopc_a_c16_0.1	2.72222E-09	1.08E-35	2.94E-44	butyrylcarnitine_2-stearoylglycerophosphocholine	5.55882E-87	1.36E-48	7.56E-135
c4/lysopc_a_c17_0	4.91639E-12	5.98E-33	2.94E-44	butyrylcarnitine_1-heptadecanoylglycerophosphocholine	6.35294E-98	1.19E-37	7.56E-135

APPENDIX B Supplementary Tables for Chapter 5

Table S5-1 121 DMPs identified from 42 T2D Metabolite-EWAS

PROBE	BETA	t-stat	p-value	METABOLITE	PATHWAY	CHR	POSITION	GENE
cg00007810	0.12	4.69	3.28E-06	3-methyl-2-oxovalerate	Amino acid	12	49113738	-
cg00237904	0.11	4.80	1.90E-06	15-methylpalmitate	Lipid	11	2020314	H19
cg00601916	0.14	4.56	6.04E-06	15-methylpalmitate	Lipid	1	16069388	TMEM82
cg00875191	-0.29	-4.73	2.65E-06	arabinose	Carbohydrate	14	105487648	CDCA4
cg01132893	-0.13	-4.63	4.18E-06	X - 11550	NA	17	46712086	-
cg01182926	-0.28	-4.97	8.06E-07	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	17	48614046	EPN3
cg01240049	0.12	4.63	4.25E-06	10-heptadecenoate (17:1n7)	Lipid	9	138068091	-
cg01274929	-0.22	-4.64	4.10E-06	X - 10510	NA	16	4318696	TFAP4
cg01408860	-0.36	-4.67	3.53E-06	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	16	29242368	-
cg01419914	0.12	4.54	6.36E-06	valine	Amino acid	17	79374691	BAHCC1
cg01448610	0.20	4.64	4.15E-06	X - 11550	NA	15	74525447	-
cg01502743	0.18	4.78	2.03E-06	X - 11550	NA	19	18883801	CRTC1
cg01541565	0.15	4.76	2.33E-06	X - 13215	NA	6	32606385	HLA-DQA1
cg01817965	-0.13	-4.81	1.82E-06	X - 11550	NA	5	139284250	NRG2
cg01878214	0.18	4.50	7.86E-06	X - 11550	NA	7	45145334	TBRG4;SNORA5C
cg01896085	0.16	4.45	9.90E-06	X - 12442	NA	2	9279786	-
cg02303209	0.11	4.49	8.20E-06	myristoleate (14:1n5)	Lipid	5	32223673	-
cg02462443	-0.19	-4.51	7.59E-06	dimethylarginine	Amino acid	6	32947897	BRD2
cg02662495	0.10	4.58	5.26E-06	fructose	Carbohydrate	12	34759279	-

cg02868468	0.14	4.52	6.98E-06	15-methylpalmitate	Lipid	14	105045347	C14orf180
cg02899346	-0.20	-4.47	9.01E-06	X - 11550	NA	19	45996372	RTN2
cg02982237	-0.18	-4.48	8.69E-06	X - 11550	NA	14	100492002	-
cg03184452	-0.25	-4.81	1.85E-06	dimethylarginine	Amino acid	8	13105155	DLC1
cg03331175	0.11	4.46	9.35E-06	X - 11315	NA	7	99069044	ZNF789
cg03333776	0.17	5.12	3.79E-07	15-methylpalmitate	Lipid	1	52455148	RAB3B
cg03532223	-0.27	-4.93	9.81E-07	X - 06246	NA	6	32935858	BRD2
cg03666973	-0.29	-4.99	7.30E-07	arabinose	Carbohydrate	2	27008764	CENPA
cg04252592	0.15	4.54	6.62E-06	X - 11315	NA	16	4387370	GLIS2
cg04483701	0.15	5.00	7.03E-07	X - 12442	NA	16	86253703	-
cg04591648	0.10	4.46	9.24E-06	4-methyl-2-oxopentanoate	Amino acid	7	155791233	-
cg04609439	-0.23	-4.49	8.06E-06	X - 12442	NA	16	32599417	-
cg05466901	-0.24	-4.48	8.45E-06	heptanoate (7:0)	Lipid	7	108095863	NRCAM
cg05542681	-0.12	-4.67	3.54E-06	N-acetylglycine	Amino acid	16	744328	FBXL16
cg05721374	0.15	4.68	3.42E-06	pelargonate (9:0)	Lipid	1	109619308	TAF13
cg05925577	0.22	5.03	6.07E-07	malate	Energy	1	121375906	-
cg06175243	-0.30	-4.66	3.72E-06	X - 13215	NA	3	32443207	CMTM7
cg06270776	0.21	4.55	6.05E-06	X - 10510	NA	6	160932670	LPAL2
cg06304546	-0.20	-4.52	7.23E-06	mannose	Carbohydrate	20	32448765	-
cg06495728	0.18	4.91	1.09E-06	octanoylcarnitine	Lipid	3	187462459	BCL6
cg06928741	-0.13	-4.63	4.22E-06	lactate	Carbohydrate	6	31691430	C6orf25
cg07570055	0.16	4.55	6.24E-06	dimethylarginine	Amino acid	2	233387088	-
cg07702548	0.19	4.56	5.82E-06	dimethylarginine	Amino acid	16	88636912	ZC3H18
cg07707505	0.15	4.56	5.90E-06	X - 12442	NA	1	12185435	TNFRSF8

cg07882648	0.12	4.52	7.26E-06	5-dodecenoate (12:1n7)	Lipid	20	62153067	PPDPF
cg07959747	0.21	4.57	5.77E-06	palmitoleate (16:1n7)	Lipid	10	82050672	MAT1A
cg08526784	-0.36	-5.00	7.21E-07	lactate	Carbohydrate	16	1811246	MAPK8IP3
cg08620267	0.19	4.94	9.52E-07	X - 11550	NA	16	85126437	KIAA0513
cg08636385	-0.13	-4.52	7.02E-06	heptanoate (7:0)	Lipid	6	71816668	-
cg09539496	0.09	4.72	2.80E-06	4-methyl-2-oxopentanoate	Amino acid	8	20164433	-
cg09610766	-0.17	-4.61	4.58E-06	malate	Energy	3	105166570	ALCAM
cg09816420	-0.10	-4.95	8.87E-07	heptanoate (7:0)	Lipid	5	140605851	PCDHB14
cg09816420	-0.11	-4.59	5.21E-06	dimethylarginine	Amino acid	5	140605851	PCDHB14
cg09845604	0.14	4.65	3.97E-06	dimethylarginine	Amino acid	1	3229921	PRDM16
cg11113753	0.20	5.28	1.70E-07	X - 11550	NA	2	44065383	ABCG5
cg11183535	-0.23	-4.59	5.21E-06	octanoylcarnitine	Lipid	7	157654985	PTPRN2
cg11449344	0.19	4.49	8.22E-06	X - 11315	NA	6	31919727	CFB
cg12079548	-0.12	-4.84	1.58E-06	mannose	Carbohydrate	12	8068600	-
cg12159028	0.16	4.58	5.29E-06	10-heptadecenoate (17:1n7)	Lipid	12	133553715	-
cg12585331	0.09	4.46	9.17E-06	myristate (14:0)	Lipid	5	92904377	FLJ42709
cg12612277	0.17	4.73	2.65E-06	glucose	Carbohydrate	16	74455542	CLEC18B
cg12612277	0.17	5.35	1.15E-07	mannose	Carbohydrate	16	74455542	CLEC18B
cg12612277	0.15	4.59	5.24E-06	arabinose	Carbohydrate	16	74455542	CLEC18B
cg12798257	0.11	4.49	8.22E-06	pelargonate (9:0)	Lipid	18	77243594	NFATC1
cg12798257	0.12	4.65	3.84E-06	heptanoate (7:0)	Lipid	18	77243594	NFATC1
cg14154487	0.26	4.50	7.71E-06	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	9	125133314	PTGS1
cg14215666	0.11	4.55	6.05E-06	X - 08402	NA	11	97224825	-
cg14215666	0.12	4.60	5.01E-06	X - 10510	NA	11	97224825	-

cg14321038	0.16	4.64	4.03E-06	malate	Energy	7	31034129	-
cg14540555	-0.20	-4.87	1.31E-06	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	11	86667375	FZD4
cg15370815	-0.28	-4.72	2.76E-06	X - 12696	NA	11	108093335	NPAT;ATM
cg15378866	-0.20	-4.48	8.48E-06	dimethylarginine	Amino acid	17	39723992	KRT9
cg15523060	0.12	4.50	7.71E-06	pentadecanoate (15:0)	Lipid	19	8952029	MBD3L1
cg16264393	-0.29	-4.69	3.26E-06	octanoylcarnitine	Lipid	19	2076579	MOBKL2A
cg16443812	-0.18	-4.53	6.86E-06	X - 10506	NA	22	50987294	KLHDC7B
cg16824024	-0.15	-4.50	7.96E-06	octanoylcarnitine	Lipid	3	73092158	PPP4R2
cg16874580	-0.12	-4.94	9.32E-07	X - 12696	NA	18	55092509	-
cg17009073	0.14	4.47	8.76E-06	arabinose	Carbohydrate	1	148865592	-
cg17426923	0.12	4.79	2.00E-06	3-methyl-2-oxovalerate	Amino acid	10	35464474	CREM
cg17459635	0.18	4.74	2.52E-06	X - 11550	NA	15	89181615	ISG20
cg18001780	-0.29	-4.45	9.74E-06	arabinose	Carbohydrate	11	128500564	-
cg18027903	-0.15	-4.52	7.04E-06	X - 12696	NA	7	15601624	TMEM195
cg18093693	-0.33	-4.47	8.76E-06	X - 12696	NA	15	72668265	HEXA;C15orf34
cg18406570	0.13	4.46	9.37E-06	X - 08402	NA	11	83984576	DLG2
cg18446744	-0.13	-4.45	9.67E-06	dimethylarginine	Amino acid	6	100680152	-
cg18934106	0.13	4.67	3.60E-06	X - 12442	NA	6	33289610	DAXX
cg19055828	0.09	4.46	9.28E-06	X - 11315	NA	12	51139321	DIP2B
cg20707202	-0.21	-4.61	4.70E-06	octanoylcarnitine	Lipid	11	33211414	-
cg21365444	0.13	4.49	8.25E-06	X - 12442	NA	3	138554516	-
cg21383495	-0.17	-5.00	7.19E-07	15-methylpalmitate	Lipid	17	54673193	-
cg21831937	-0.20	-4.47	9.06E-06	X - 10500	NA	15	57519802	TCF12
cg21831937	-0.24	-4.98	7.73E-07	cholesterol	Lipid	15	57519802	TCF12

cg21831937	-0.23	-4.99	7.53E-07	X - 10510	NA	15	57519802	TCF12
cg21831937	-0.22	-5.00	6.96E-07	lactate	Carbohydrate	15	57519802	TCF12
cg22118359	-0.28	-4.54	6.61E-06	arabinose	Carbohydrate	11	33060800	TCP11L1
cg22428762	-0.27	-4.64	4.12E-06	dimethylarginine	Amino acid	4	41984085	DCAF4L1
cg22666438	-0.18	-4.61	4.68E-06	X - 11550	NA	19	919596	KISS1R
cg23138608	-0.38	-4.69	3.27E-06	X - 11315	NA	14	91770145	CCDC88C
cg23260105	-0.31	-4.86	1.40E-06	octanoylcarnitine	Lipid	5	1218653	SLC6A19
cg23324953	-0.39	-4.78	2.04E-06	arabinose	Carbohydrate	8	145013728	PLEC1
cg23427998	-0.32	-4.53	6.80E-06	arabinose	Carbohydrate	5	54522784	-
cg23465749	0.07	4.83	1.62E-06	isoleucine	Amino acid	12	46123553	ARID2
cg23826993	-0.11	-4.55	6.19E-06	malate	Energy	8	58168222	-
cg23867721	0.16	5.12	3.81E-07	X - 11550	NA	18	77631069	KCNG2
cg23971215	0.24	4.59	5.13E-06	5-dodecenoate (12:1n7)	Lipid	8	1359688	-
cg24007312	0.19	4.57	5.67E-06	dimethylarginine	Amino acid	3	52312811	WDR82
cg24077401	0.13	4.47	9.03E-06	X - 12696	NA	19	40372977	FCGBP
cg24154777	0.18	4.61	4.62E-06	X - 11550	NA	14	103566588	C14orf73
cg24197445	-0.24	-4.50	7.87E-06	dimethylarginine	Amino acid	3	167045873	ZBBX
cg24470692	0.09	4.53	6.68E-06	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	8	6565084	AGPAT5
cg25190151	-0.20	-4.61	4.60E-06	glucose	Carbohydrate	7	21463675	-
cg25443560	-0.12	-4.74	2.59E-06	N-acetylglycine	Amino acid	19	35790951	MAG
cg25553198	-0.21	-4.60	4.93E-06	octanoylcarnitine	Lipid	4	54854922	RPL21P44
cg25561382	0.15	4.47	8.78E-06	myristoleate (14:1n5)	Lipid	19	18795207	CRTC1
cg25561382	0.16	4.66	3.71E-06	X - 12442	NA	19	18795207	CRTC1
cg25927227	0.13	4.49	8.11E-06	X - 13215	NA	8	41127218	SFRP1

cg26400325	-0.25	-4.69	3.19E-06	octanoylcarnitine	Lipid	7	1997592	MAD1L1
cg26646527	-0.19	-4.84	1.58E-06	1,5-anhydroglucitol (1,5-AG)	Carbohydrate	18	77333373	-
cg27171194	0.15	4.57	5.72E-06	dimethylarginine	Amino acid	1	3229579	PRDM16
cg27229366	0.17	4.67	3.59E-06	X - 13215	NA	17	43574018	-
cg27310761	-0.23	-4.50	7.73E-06	arabinose	Carbohydrate	14	90798393	C14orf102
cg27535047	-0.28	-4.55	6.32E-06	X - 13215	NA	22	42228699	SREBF2

Table S5-2 240 Associations for BN

GENE	PROBE	BIOCHEMICAL	SNP (81 GWAS)	CHR	BETA (42 Ewas)	p-value (42 Ewas)	BETA-caco	p-value-caco	Score1 MB<--SNP -->MT	Score2- SNP -->MB -->MT	Score3 SNP -->MT -->MB	Model with Max Score	Relative likelohood
TCF7L2	cg07861463	X - 11497	rs7903146	10	0.04	0.009015967	0.06	8.39E-05	-2452.9	-2451.4	-2450.6	3	1.53
PSMD6	cg00629144	X - 12696	rs832571	3	0.03	0.008400409	0.05	0.000188961	-2369.6	-2366.2	-2369.3	2	4.77
PRC1	cg26181196	pentadecanoate (15:0)	rs12899811	15	-0.03	0.027091851	0.05	0.000987813	-2594.6	-2591.7	-2592.8	2	1.68
PRC1	cg26181196	X - 11497	rs12899811	15	0.03	0.042059092	0.05	0.000987813	-2594.9	-2591.9	-2592.7	2	1.48
PRC1	cg26181196	palmitoyl sphingomyelin	rs12899811	15	-0.02	0.0085748	0.05	0.000987813	-2594.1	-2590.7	-2592.3	2	2.17
PRC1	cg26181196	arabinose	rs12899811	15	-0.02	0.045663162	0.05	0.000987813	-2595	-2592.1	-2592.7	2	1.35
PRC1	cg26181196	cholesterol	rs12899811	15	-0.04	0.015051407	0.05	0.000987813	-2594	-2591	-2592.7	2	2.32
PRC1	cg26181196	X - 10510	rs12899811	15	-0.03	0.013777596	0.05	0.000987813	-2594	-2591.1	-2592.8	2	2.26
PRC1	cg26181196	malate	rs12899811	15	-0.03	0.008671852	0.05	0.000987813	-2592.6	-2589.3	-2592.4	2	4.72
PRC1	cg26181196	lactate	rs12899811	15	-0.03	0.041817777	0.05	0.000987813	-2588.7	-2585.9	-2592.8	2	4.23
KCNQ1	cg17128405	glucose	rs231362	11	-0.01	0.040116032	0.03	0.002115251	-3189.4	-3184.2	-3185.1	2	1.64
KCNQ1	cg17128405	glucose	rs2237892	11	-0.01	0.040116032	0.03	0.002115251	-2389.1	-2385.9	-2393.2	2	4.79
KCNQ1	cg17128405	glucose	rs231361	11	-0.01	0.040116032	0.03	0.002115251	-2843	-2840.4	-2841.8	2	2.06
KCNQ1	cg17128405	glucose	rs163184	11	-0.01	0.040116032	0.03	0.002115251	-3006.7	-3005.7	-3005.6	3	1.01
KCNQ1	cg17128405	X - 12442	rs231362	11	-0.01	0.030021468	0.03	0.002115251	-3190.5	-3185.6	-3185.4	3	1.08
KCNQ1	cg17128405	X - 12442	rs2237892	11	-0.01	0.030021468	0.03	0.002115251	-2391	-2388.1	-2393.4	2	4.21

KCNQ1	cg17128405	X - 12442	rs231361	11	-0.01	0.030021468	0.03	0.002115251	-2843.4	-2841	-2842.1	2	1.72
KCNQ1	cg17128405	X - 12442	rs163184	11	-0.01	0.030021468	0.03	0.002115251	-3007	-3006.2	-3005.9	3	1.19
KCNQ1	cg17128405	mannose	rs231362	11	-0.01	0.012457595	0.03	0.002115251	-3190.4	-3185.4	-3185.3	3	1.01
KCNQ1	cg17128405	mannose	rs2237892	11	-0.01	0.012457595	0.03	0.002115251	-2393.3	-2390.4	-2393.4	2	4.38
KCNQ1	cg17128405	mannose	rs231361	11	-0.01	0.012457595	0.03	0.002115251	-2842.1	-2839.7	-2842	2	3.27
KCNQ1	cg17128405	mannose	rs163184	11	-0.01	0.012457595	0.03	0.002115251	-3006.6	-3005.7	-3005.9	2	1.06
KCNQ1	cg17128405	octanoylcarnitine	rs231362	11	-0.01	0.045979641	0.03	0.002115251	-3189.4	-3184.4	-3185.3	2	1.59
KCNQ1	cg17128405	octanoylcarnitine	rs2237892	11	-0.01	0.045979641	0.03	0.002115251	-2391	-2388	-2393.3	2	4.38
KCNQ1	cg17128405	octanoylcarnitine	rs231361	11	-0.01	0.045979641	0.03	0.002115251	-2840.4	-2838	-2842	2	3.39
KCNQ1	cg17128405	octanoylcarnitine	rs163184	11	-0.01	0.045979641	0.03	0.002115251	-3008.4	-3007.5	-3005.8	3	2.36
KCNQ1	cg17128405	lactate	rs231362	11	-0.02	0.019901703	0.03	0.002115251	-3189	-3184.1	-3185.4	2	1.91
KCNQ1	cg17128405	lactate	rs2237892	11	-0.02	0.019901703	0.03	0.002115251	-2392.9	-2390.1	-2393.4	2	4.22
KCNQ1	cg17128405	lactate	rs231361	11	-0.02	0.019901703	0.03	0.002115251	-2844.1	-2841.7	-2842.1	2	1.2
KCNQ1	cg17128405	lactate	rs163184	11	-0.02	0.019901703	0.03	0.002115251	-3008.3	-3007.5	-3005.9	3	2.28
ANK1	cg07759223	valine	rs516946	8	-0.01	0.036954492	0.02	0.002152063	-3188.8	-3187	-3190.5	2	2.54
UBE2E2	cg07208565	malate	rs1496653	3	-0.02	0.006083778	0.02	0.00452229	-2877.6	-2875	-2875.4	2	1.2
KCNJ11	cg22937444	X - 13215	rs5215	11	-0.02	0.01962385	0.02	0.004877806	-2980.3	-2979.9	-2979	3	1.6
KCNQ1	cg20768429	X - 12450	rs231362	11	0.01	0.037460735	0.03	0.006189381	-2989	-2986.1	-2982.9	3	4.96
KCNQ1	cg20768429	X - 12450	rs2237892	11	0.01	0.037460735	0.03	0.006189381	-2194.3	-2192.8	-2191.6	3	1.82
KCNQ1	cg20768429	X - 12450	rs231361	11	0.01	0.037460735	0.03	0.006189381	-2643.5	-2640	-2642.8	2	4.17
KCNQ1	cg20768429	X - 12450	rs163184	11	0.01	0.037460735	0.03	0.006189381	-2804.7	-2804.5	-2804.9	2	1.13
UBE2E2	cg03455225	X - 08402	rs1496653	3	-0.02	0.027600074	0.02	0.006270918	-2790	-2811.1	-2786.9	3	4.65
UBE2E2	cg03455225	X - 10500	rs1496653	3	-0.02	0.037948917	0.02	0.006270918	-2789.8	-2811.9	-2788	3	2.38
UBE2E2	cg03455225	cholesterol	rs1496653	3	-0.02	0.038541781	0.02	0.006270918	-2788.6	-2809.9	-2787.1	3	2.11
MAEA	cg13680752	mannose	rs6815464	4	-0.01	0.028523897	0.03	0.007535898	-2211.8	-2217.2	-2210.2	3	2.2
MAEA	cg13680752	lactate	rs6815464	4	-0.03	0.008611729	0.03	0.007535898	-2209.6	-2214.9	-2210.1	1	1.31

MAEA	cg13680752	fructose	rs6815464	4	-0.02	0.023608136	0.03	0.007535898	-2211.3	-2216.7	-2210.2	3	1.7
MAEA	cg13680752	octanoylcarnitine	rs6815464	4	-0.01	0.014960437	0.03	0.007535898	-2211.7	-2217.1	-2210.1	3	2.32
MAEA	cg13680752	glucose	rs6815464	4	-0.02	0.020130916	0.03	0.007535898	-2209.7	-2215.1	-2210.2	1	1.28
MAEA	cg13680752	pentadecanoate (15:0)	rs6815464	4	-0.02	0.03429492	0.03	0.007535898	-2212.6	-2218	-2210.2	3	3.23
KCNQ1	cg26750319	X - 13215	rs231362	11	0.02	0.009561817	0.03	0.008813629	-3122.1	-3118.6	-3117.7	3	1.53
KCNQ1	cg26750319	X - 13215	rs2237892	11	0.02	0.009561817	0.03	0.008813629	-2317.1	-2313.9	-2327.5	2	5.1
KCNQ1	cg26750319	X - 13215	rs231361	11	0.02	0.009561817	0.03	0.008813629	-2774.2	-2774.5	-2773.2	3	1.92
KCNQ1	cg26750319	X - 13215	rs163184	11	0.02	0.009561817	0.03	0.008813629	-2941.3	-2939.6	-2940.5	2	1.61
FTO	cg04036070	X - 12696	rs9936385	16	-0.01	0.00374644	-0.01	0.009219243	-3367.1	-3366.7	-3367.9	2	1.8
TCF7L2	cg04064032	X - 08402	rs7903146	10	0.01	0.048668426	0.02	0.009853878	-2828.4	-2825.3	-2824.9	3	1.18
ST6GAL1	cg15473502	proline	rs16861329	3	-0.03	0.007275164	0.03	0.010646449	-2373.2	-2371.1	-2371.5	2	1.24
ST6GAL1	cg15473502	X - 10506	rs16861329	3	0.02	0.01989082	0.03	0.010646449	-2372.3	-2368.9	-2370.3	2	2
ST6GAL1	cg15473502	15-methylpalmitate	rs16861329	3	-0.01	0.035593955	0.03	0.010646449	-2363.3	-2361.6	-2372	2	2.3
HCCA2	cg18286323	15-methylpalmitate	rs2334499	11	0.01	0.023768066	-0.02	0.011410126	-2998.9	-2995.9	-2998.8	2	4.26
SLC30A8	cg23338195	X - 10500	rs13266634	8	0.01	0.024822682	-0.01	0.011605501	-3369.5	-3365.8	-3368.9	2	4.61
SLC30A8	cg23338195	X - 12696	rs13266634	8	-0.01	0.008682756	-0.01	0.011605501	-3371.5	-3368.2	-3369.2	2	1.72
SLC30A8	cg23338195	palmitoyl sphingomyelin	rs13266634	8	0.01	0.042647784	-0.01	0.011605501	-3371	-3367.7	-3369.2	2	2.14
WFS1	cg00829753	X - 11423	rs4458523	4	-0.03	0.027805604	0.04	0.011896173	-2584.9	-2578.7	-2584.1	2	15.47
CCND2	cg16994506	mannose	rs11063069	12	-0.02	0.014460541	0.03	0.013095892	-2892.8	-2888	-2886.8	3	1.8
CCND2	cg16994506	dimethylarginine	rs11063069	12	0.03	0.0214259	0.03	0.013095892	-2892.5	-2886.5	-2885.6	3	1.57
CCND2	cg16994506	lactate	rs11063069	12	-0.04	0.001773676	0.03	0.013095892	-2892.8	-2887.9	-2886.6	3	1.86
CCND2	cg16994506	cholesterol	rs11063069	12	-0.04	0.009862066	0.03	0.013095892	-2892.3	-2887.2	-2886.5	3	1.41
CCND2	cg16994506	malate	rs11063069	12	-0.03	0.00177275	0.03	0.013095892	-2890.5	-2885.6	-2886.8	2	1.75
CCND2	cg16994506	X - 10510	rs11063069	12	-0.02	0.045944452	0.03	0.013095892	-2891.5	-2886.2	-2886.3	2	1.04
CCND2	cg16994506	X - 10500	rs11063069	12	-0.02	0.01888673	0.03	0.013095892	-2888.4	-2883.5	-2886.7	2	4.95
CCND2	cg16994506	X - 10506	rs11063069	12	-0.03	0.00542079	0.03	0.013095892	-2892.2	-2887.2	-2886.6	3	1.34

CCND2	cg16994506	X - 13496	rs11063069	12	-0.02	0.040812529	0.03	0.013095892	-2893	-2885	-2883.6	3	2.04
CCND2	cg16994506	glucose	rs11063069	12	-0.02	0.006775431	0.03	0.013095892	-2893.1	-2888.2	-2886.8	3	2.09
CCND2	cg16994506	X - 06246	rs11063069	12	-0.03	0.003616706	0.03	0.013095892	-2890.7	-2885.3	-2886.2	2	1.59
MAEA	cg01418351	X - 12442	rs6815464	4	0.01	0.006846993	0.02	0.01332383	-2539.2	-2536	-2539.3	2	4.94
GLIS3	cg14340481	myristate (14:0)	rs7041847	9	0.01	0.048392761	-0.02	0.013578443	-3194.4	-3191.9	-3192.9	2	1.69
GLIS3	cg14340481	myristoleate (14:1n5)	rs7041847	9	0.01	0.032459432	-0.02	0.013578443	-3192.9	-3190.4	-3193.1	2	3.37
LPP	cg04423294	citrulline	rs6808574	3	0.01	0.030666435	-0.01	0.015744096	-3364.9	-3363.8	-3363.1	3	1.4
KCNQ1	cg26958174	arabinose	rs231362	11	-0.01	0.035182519	0.02	0.017026174	-3059.2	-3054.6	-3055.1	2	1.31
KCNQ1	cg26958174	arabinose	rs2237892	11	-0.01	0.035182519	0.02	0.017026174	-2266.7	-2262.9	-2264.4	2	2.16
KCNQ1	cg26958174	arabinose	rs231361	11	-0.01	0.035182519	0.02	0.017026174	-2715.4	-2710.9	-2714.2	2	5.36
KCNQ1	cg26958174	arabinose	rs163184	11	-0.01	0.035182519	0.02	0.017026174	-2882.8	-2879	-2878.9	3	1.06
IGF2BP2	cg21531679	palmitoyl sphingomyelin	rs4402960	3	-0.01	0.030324353	-0.02	0.017288687	-2786.1	-2784.6	-2782	3	3.67
IGF2BP2	cg21531679	proline	rs4402960	3	-0.02	0.015254048	-0.02	0.017288687	-2785.9	-2786.1	-2783.6	3	3.22
IGF2BP2	cg21531679	X - 12696	rs4402960	3	-0.02	0.031253664	-0.02	0.017288687	-2780.3	-2780.4	-2783.6	1	1.05
RBMS1	cg02048613	1,5-anhydroglucitol (1,5-AG)	rs7593730	2	0.01	0.047756694	-0.01	0.01755332	-2780.3	-2781.4	-2783.6	1	1.73
PPARG	cg23514324	X - 13496	rs1801282	3	-0.02	0.005926219	0.02	0.017709731	-2780.3	-2780.4	-2783.6	1	1.05
NOTCH2	cg05824755	X - 10506	rs1493694	1	0.01	0.021138158	-0.02	0.017913928	-2757.9	-2753.8	-2753.3	3	1.27
NOTCH2	cg05824755	malate	rs1493694	1	0.01	0.020645686	-0.02	0.017913928	-2757.7	-2755.2	-2755	3	1.12
NOTCH2	cg05824755	dimethylargini	rs1493694	1	-0.02	0.023529591	-0.02	0.017913928	-2756.3	-2752.1	-2753.2	2	1.74
NOTCH2	cg05824755	arabinose	rs1493694	1	0.01	0.04746119	-0.02	0.017913928	-2756	-2753.4	-2754.8	2	2.04
NOTCH2	cg05824755	X - 06246	rs1493694	1	0.01	0.019867879	-0.02	0.017913928	-2757.4	-2752.5	-2752.5	2	1
NOTCH2	cg05824755	cholesterol	rs1493694	1	0.02	0.044871094	-0.02	0.017913928	-2755.6	-2753	-2754.9	2	2.5
NOTCH2	cg05824755	X - 10500	rs1493694	1	0.01	0.048393796	-0.02	0.017913928	-2757.6	-2754.6	-2754.4	3	1.11
NOTCH2	cg05824755	lactate	rs1493694	1	0.02	0.014079808	-0.02	0.017913928	-2754.2	-2752.2	-2755.4	2	2.73
NOTCH2	cg05824755	glucose	rs1493694	1	0.01	0.034804872	-0.02	0.017913928	-2757.7	-2754.6	-2754.4	3	1.12
IGF2BP2	cg25127692	X - 10500	rs4402960	3	0.02	0.017763148	0.02	0.018125741	-3104.5	-3102.1	-3103.4	2	1.91

IGF2BP2	cg25127692	X - 10510	rs4402960	3	0.01	0.012417556	0.02	0.018125741	-3105.7	-3103.5	-3103.6	2	1.03
IGF2BP2	cg25127692	cholesterol	rs4402960	3	0.02	0.005962779	0.02	0.018125741	-3106.7	-3104.5	-3103.5	3	1.6
IGF2BP2	cg25127692	X - 11497	rs4402960	3	-0.02	0.027062077	0.02	0.018125741	-3105.9	-3103.7	-3103.5	3	1.1
IGF2BP2	cg25127692	pentadecanoate (15:0)	rs4402960	3	0.01	0.037175207	0.02	0.018125741	-3106.1	-3103.7	-3103.3	3	1.2
IGF2BP2	cg25127692	X - 12450	rs4402960	3	0.01	0.005071247	0.02	0.018125741	-3105.7	-3103.5	-3103.6	2	1.01
IGF2BP2	cg25127692	heptanoate (7:0)	rs4402960	3	-0.01	0.03955623	0.02	0.018125741	-3105.5	-3103.3	-3103.6	2	1.14
IGF2BP2	cg25127692	dimethylargini	rs4402960	3	-0.02	0.045603961	0.02	0.018125741	-3102.1	-3099.9	-3103.6	2	2.96
IGF2BP2	cg25127692	X - 10506	rs4402960	3	0.01	0.032264169	0.02	0.018125741	-3105.8	-3103.3	-3103.3	2	1
IGF2BP2	cg25127692	X - 12442	rs4402960	3	0.01	0.04440806	0.02	0.018125741	-3106.1	-3103.9	-3103.5	3	1.2
IGF2BP2	cg25127692	lactate	rs4402960	3	0.02	0.03152512	0.02	0.018125741	-3099	-3096.3	-3103.1	2	3.85
IGF2BP2	cg25127692	mannose	rs4402960	3	0.01	0.035120653	0.02	0.018125741	-3103	-3100.3	-3103.1	2	3.89
IGF2BP2	cg25127692	glucose	rs4402960	3	0.01	0.046005436	0.02	0.018125741	-3104.9	-3102.4	-3103.2	2	1.52
CDKAL1	cg06512263	arabinose	rs7756992	6	0.01	0.008586571	-0.01	0.018960068	-3341.7	-3340.1	-3341.8	2	2.29
ARL15	cg04530345	X - 12696	rs702634	5	0.01	0.04945006	0.02	0.019046856	-2889.5	-2885.5	-2888.1	2	3.62
HNF1A	cg01394199	1,5-anhydroglucitol (1,5-AG)	rs7957197	12	0.01	0.048207375	0.02	0.019165061	-2682.4	-2677.3	-2675.1	3	3.08
HNF1A	cg01394199	X - 12696	rs7957197	12	0.02	0.014917941	0.02	0.019165061	-2681	-2677	-2676.2	3	1.48
FAF1	cg07911682	4-methyl-2-oxopentanoate	rs17106184	1	-0.01	0.048374901	0.02	0.019270021	-2539.5	-2533.3	-2536.8	2	5.62
FAF1	cg07911682	X - 10506	rs17106184	1	-0.02	0.047965615	0.02	0.019270021	-2543.7	-2536	-2535.3	3	1.4
FAF1	cg07911682	X - 12442	rs17106184	1	-0.02	0.007194181	0.02	0.019270021	-2540.6	-2536.4	-2538.7	2	3.23
FAF1	cg07911682	3-methyl-2-oxovalerate	rs17106184	1	-0.01	0.049473993	0.02	0.019270021	-2539.6	-2535.4	-2538.7	2	5.45
IGF2BP2	cg13618735	X - 12450	rs4402960	3	0.02	0.011771561	0.02	0.021113864	-2798.4	-2796.2	-2796.7	2	1.3
IGF2BP2	cg13618735	proline	rs4402960	3	-0.02	0.044790212	0.02	0.021113864	-2800	-2798.5	-2797.4	3	1.72
TP53INP1	cg20039814	X - 11315	rs896854	8	-0.01	0.021925071	-0.01	0.022885282	-3212.5	-3244.5	-3212.8	1	1.18
WFS1	cg22247194	citrulline	rs4458523	4	-0.02	0.006919983	0.02	0.022945677	-2971.9	-2966.9	-2975.9	2	12.69
WFS1	cg22247194	X - 11315	rs4458523	4	-0.02	0.012720613	0.02	0.022945677	-2979.6	-2974.9	-2976.4	2	2.05
WFS1	cg22247194	1,5-anhydroglucitol (1,5-AG)	rs4458523	4	-0.01	0.012585365	0.02	0.022945677	-2976.9	-2971.3	-2975.4	2	7.69

HCCA2	cg04902924	isoleucine	rs2334499	11	-0.01	0.049353796	0.02	0.024296623	-3048.8	-3051.8	-3053.1	1	4.55
HCCA2	cg04902924	leucine	rs2334499	11	-0.01	0.048270233	0.02	0.024296623	-3050.1	-3053	-3053	1	4.18
HCCA2	cg04902924	valine	rs2334499	11	-0.02	0.037465148	0.02	0.024296623	-3048.5	-3051.6	-3053.2	1	4.65
CCND2	cg07181862	pelargonate (9:0)	rs11063069	12	-0.02	0.031578696	0.03	0.024400664	-2851.1	-2845.6	-2848.4	2	4.17
CCND2	cg07181862	X - 11497	rs11063069	12	-0.02	0.040680081	0.03	0.024400664	-2853.1	-2847.7	-2848.6	2	1.56
CCND2	cg07181862	palmitoyl sphingomyelin	rs11063069	12	-0.02	0.003926466	0.03	0.024400664	-2851.2	-2845.8	-2848.6	2	4.12
CCND2	cg07181862	X - 11550	rs11063069	12	-0.02	0.003567977	0.03	0.024400664	-2850.7	-2845.6	-2848.9	2	5.04
FAF1	cg12278631	X - 10510	rs17106184	1	-0.02	0.025279779	0.02	0.024643124	-2451.7	-2447.2	-2448.2	2	1.63
FAF1	cg12278631	lactate	rs17106184	1	-0.02	0.041766419	0.02	0.024643124	-2453.5	-2449.2	-2448.4	3	1.46
FAF1	cg12278631	fructose	rs17106184	1	-0.02	0.005580859	0.02	0.024643124	-2451.9	-2446.9	-2447.7	2	1.5
FAF1	cg12278631	X - 13496	rs17106184	1	-0.02	0.027794359	0.02	0.024643124	-2448.1	-2443.5	-2448.1	2	9.87
FAF1	cg12278631	octanoylcarnitine	rs17106184	1	-0.01	0.046233567	0.02	0.024643124	-2444.8	-2440.5	-2448.4	2	8.55
FTO	cg03312170	octanoylcarnitine	rs9936385	16	-0.01	0.038521381	0.02	0.024721643	-3060.7	-3059.1	-3060.3	2	1.91
HCCA2	cg00435469	X - 13215	rs2334499	11	-0.03	0.021199428	0.03	0.026510685	-2649.7	-2647.9	-2647.5	3	1.26
HCCA2	cg00435469	proline	rs2334499	11	-0.02	0.02656214	0.03	0.026510685	-2648.1	-2647.5	-2648.6	2	1.38
HCCA2	cg00435469	X - 06246	rs2334499	11	-0.03	0.014867383	0.03	0.026510685	-2649.8	-2648.8	-2648.3	3	1.28
CDKAL1	cg23560765	X - 12696	rs7756992	6	0.01	0.047844542	0.02	0.02711749	-3033.7	-3031.5	-3032.5	2	1.67
IGF2BP2	cg24960291	X - 13215	rs4402960	3	-0.01	0.020281147	-0.01	0.027439181	-3246.9	-3245	-3242.8	3	3.02
IGF2BP2	cg24960291	lactate	rs4402960	3	0.01	0.018083612	-0.01	0.027439181	-3246.5	-3242.4	-3240.5	3	2.51
IGF2BP2	cg24960291	X - 11550	rs4402960	3	-0.01	0.049134376	-0.01	0.027439181	-3244.7	-3244.3	-3244.3	2	1.01
IGF2BP2	cg24960291	10-heptadecenoate (17:1n7)	rs4402960	3	-0.01	0.044255401	-0.01	0.027439181	-3237.1	-3236.8	-3244.4	2	1.17
IGF2BP2	cg24960291	X - 11497	rs4402960	3	-0.02	0.003447156	-0.01	0.027439181	-3242.6	-3241.4	-3243.5	2	1.79
HNF1A	cg07065256	X - 11423	rs7957197	12	-0.03	0.018877851	0.03	0.027506288	-2492.7	-2489.2	-2490.1	2	1.61
HNF1A	cg07065256	3-methyl-2-oxovalerate	rs7957197	12	-0.01	0.048732922	0.03	0.027506288	-2492.9	-2488.7	-2489.5	2	1.46
ZFAND3	cg03003722	arabinose	rs9470794	6	0.01	0.025896405	-0.01	0.029458718	-2869.3	-2872.5	-2870	1	1.41
CDKAL1	cg24273995	3-methyl-2-oxovalerate	rs7756992	6	-0.01	0.014465973	0.03	0.02993546	-2622.3	-2619.1	-2626	2	4.87

CDKAL1	cg24273995	4-methyl-2-oxopentanoate	rs7756992	6	-0.01	0.009416892	0.03	0.02993546	-2620.9	-2617.5	-2625.8	2	5.47
HCCA2	cg25050723	octanoylcarnitine	rs2334499	11	0.01	0.030869026	0.02	0.031301212	-3165.8	-3163.1	-3163.2	2	1.06
KCNQ1	cg24609402	fructose	rs231362	11	0.02	0.026264586	0.02	0.031545812	-3079	-3073.8	-3077.4	2	5.85
KCNQ1	cg24609402	fructose	rs2237892	11	0.02	0.026264586	0.02	0.031545812	-2278.2	-2276.8	-2283.7	2	1.96
KCNQ1	cg24609402	fructose	rs231361	11	0.02	0.026264586	0.02	0.031545812	-2733.1	-2732.1	-2732.4	2	1.16
KCNQ1	cg24609402	fructose	rs163184	11	0.02	0.026264586	0.02	0.031545812	-2904	-2900.2	-2900.7	2	1.3
MAEA	cg14386311	X - 12696	rs6815464	4	0.02	0.035536628	0.02	0.032578624	-2127.2	-2124.5	-2127.1	2	3.63
IGF2BP2	cg24450631	X - 12696	rs4402960	3	0.01	0.049018909	0.01	0.032656191	-3420.5	-3417.4	-3423.8	2	4.89
KLF14	cg09529138	X - 11550	rs972283	7	-0.01	0.015725329	-0.01	0.033795627	-3620.8	-3617.1	-3618.3	2	1.82
ADCY5	cg14844401	N-acetylglycine	rs11717195	3	0.01	0.004065766	-0.01	0.034699627	-3296.4	-3290.6	-3292.4	2	2.43
ADCY5	cg14844401	X - 10510	rs11717195	3	0.01	0.033902512	-0.01	0.034699627	-3294.5	-3288.5	-3292.2	2	6.35
ADCY5	cg14844401	X - 10506	rs11717195	3	0.01	0.019467994	-0.01	0.034699627	-3292.3	-3286.1	-3291.9	2	18.46
ADCY5	cg14844401	X - 13496	rs11717195	3	0.01	0.040217595	-0.01	0.034699627	-3297.8	-3291.5	-3291.8	2	1.18
ADCY5	cg14844401	arabinose	rs11717195	3	0.01	0.04449706	-0.01	0.034699627	-3296.3	-3290.2	-3292.1	2	2.51
ADCY5	cg14844401	lactate	rs11717195	3	0.01	0.047796316	-0.01	0.034699627	-3296.3	-3289.6	-3291.5	2	2.6
ADCY5	cg14844401	malate	rs11717195	3	0.01	0.016594712	-0.01	0.034699627	-3292	-3286	-3292.2	2	21.36
ADCY5	cg14844401	glucose	rs11717195	3	0.01	0.042320012	-0.01	0.034699627	-3296.9	-3290.8	-3292	2	1.87
ADCY5	cg14844401	pentadecanoate (15:0)	rs11717195	3	0.01	0.043189212	-0.01	0.034699627	-3292.9	-3287.2	-3292.4	2	13.93
ADCY5	cg14844401	cholesterol	rs11717195	3	0.01	0.025670726	-0.01	0.034699627	-3295.6	-3289	-3291.5	2	3.58
MTNR1B	cg13171406	X - 12696	rs10830963	11	-0.01	0.033354661	-0.01	0.035355051	-3313.9	-3308.8	-3311.4	2	3.66
PRC1	cg06613755	lactate	rs12899811	15	-0.02	0.006661835	0.02	0.035363912	-3033.6	-3029.9	-3031.4	2	2.09
PRC1	cg06613755	X - 10500	rs12899811	15	-0.02	0.040334966	0.02	0.035363912	-3034.8	-3031.9	-3032.2	2	1.16
PRC1	cg06613755	dimethylargin	rs12899811	15	0.02	0.021548853	0.02	0.035363912	-3028.3	-3025.1	-3031.9	2	5.02
PRC1	cg06613755	arabinose	rs12899811	15	-0.01	0.030034761	0.02	0.035363912	-3035.3	-3031.5	-3031.4	3	1.09
KCNQ1	cg20170839	mannose	rs231362	11	-0.02	0.020227261	0.02	0.03539522	-2981.4	-2976.1	-2977.3	2	1.82
KCNQ1	cg20170839	mannose	rs2237892	11	-0.02	0.020227261	0.02	0.03539522	-2178.5	-2177.9	-2182.8	2	1.37

KCNQ1	cg20170839	mannose	rs231361	11	-0.02	0.020227261	0.02	0.03539522	-2635.1	-2632.4	-2634	2	2.31
KCNQ1	cg20170839	mannose	rs163184	11	-0.02	0.020227261	0.02	0.03539522	-2800.2	-2797.7	-2799.2	2	2.18
TLE1	cg14254562	X - 13496	rs2796441	9	0.02	0.02987329	0.02	0.037060539	-3089.1	-3087.1	-3086.6	3	1.32
JAZF1	cg27102995	X - 13496	rs849135	7	-0.02	0.022723393	-0.02	0.037324519	-3061.6	-3056.9	-3058.9	2	2.78
FTO	cg12495954	myristate (14:0)	rs9936385	16	-0.01	0.026640058	-0.02	0.038265962	-2983	-2986.8	-2978.8	3	8.25
FTO	cg12495954	palmitoleate (16:1n7)	rs9936385	16	-0.01	0.045960443	-0.02	0.038265962	-2984.6	-2986.9	-2977.4	3	37.12
FTO	cg12495954	X - 13215	rs9936385	16	-0.02	0.036707553	-0.02	0.038265962	-2983.9	-2987.6	-2978.7	3	13.7
IGF2BP2	cg23956648	X - 11315	rs4402960	3	0.01	0.045049917	-0.01	0.040498821	-3180.5	-3185.7	-3179.8	3	1.39
ANK1	cg23668222	X - 12696	rs516946	8	-0.02	0.020433541	0.02	0.041369955	-2843.7	-2850.8	-2841.1	3	3.67
KCNQ1	cg19030519	X - 08402	rs231362	11	0.01	0.03398176	0.01	0.041832201	-3392.3	-3387.6	-3386.4	3	1.85
KCNQ1	cg19030519	X - 08402	rs2237892	11	0.01	0.03398176	0.01	0.041832201	-2591.7	-2592.2	-2591.2	3	1.29
KCNQ1	cg19030519	X - 08402	rs231361	11	0.01	0.03398176	0.01	0.041832201	-3043.6	-3043	-3041.5	3	2.14
KCNQ1	cg19030519	X - 08402	rs163184	11	0.01	0.03398176	0.01	0.041832201	-3212	-3209.6	-3208.7	3	1.56
KCNQ1	cg19030519	malate	rs231362	11	0.01	0.011664896	0.01	0.041832201	-3386.4	-3381	-3385.7	2	10.36
KCNQ1	cg19030519	malate	rs2237892	11	0.01	0.011664896	0.01	0.041832201	-2590.6	-2590.5	-2590.5	2	1.01
KCNQ1	cg19030519	malate	rs231361	11	0.01	0.011664896	0.01	0.041832201	-3039.4	-3038.1	-3040.9	2	1.85
KCNQ1	cg19030519	malate	rs163184	11	0.01	0.011664896	0.01	0.041832201	-3209.2	-3206.1	-3208.1	2	2.63
KCNQ1	cg19030519	X - 10506	rs231362	11	0.01	0.012660328	0.01	0.041832201	-3390	-3384.2	-3385.3	2	1.72
KCNQ1	cg19030519	X - 10506	rs2237892	11	0.01	0.012660328	0.01	0.041832201	-2589.3	-2588.8	-2590.1	2	1.27
KCNQ1	cg19030519	X - 10506	rs231361	11	0.01	0.012660328	0.01	0.041832201	-3042.4	-3040.8	-3040.5	3	1.16
KCNQ1	cg19030519	X - 10506	rs163184	11	0.01	0.012660328	0.01	0.041832201	-3210.4	-3207	-3207.7	2	1.41
KCNQ1	cg19030519	X - 13496	rs231362	11	0.01	0.040368502	0.01	0.041832201	-3389.2	-3384.4	-3386.2	2	2.47
KCNQ1	cg19030519	X - 13496	rs2237892	11	0.01	0.040368502	0.01	0.041832201	-2592.3	-2592.7	-2591	3	2.3
KCNQ1	cg19030519	X - 13496	rs231361	11	0.01	0.040368502	0.01	0.041832201	-3041.4	-3040.7	-3041.4	2	1.4
KCNQ1	cg19030519	X - 13496	rs163184	11	0.01	0.040368502	0.01	0.041832201	-3211.4	-3208.8	-3208.6	3	1.15
KCNQ1	cg19030519	arabinose	rs231362	11	0.01	0.038591021	0.01	0.041832201	-3391.9	-3387	-3386.1	3	1.55

KCNQ1	cg19030519	arabinose	rs2237892	11	0.01	0.038591021	0.01	0.041832201	-2590.8	-2591.1	-2590.9	1	1.06
KCNQ1	cg19030519	arabinose	rs231361	11	0.01	0.038591021	0.01	0.041832201	-3043.7	-3042.8	-3041.3	3	2.15
KCNQ1	cg19030519	arabinose	rs163184	11	0.01	0.038591021	0.01	0.041832201	-3207.8	-3205.2	-3208.5	2	3.74
IGF2BP2	cg19952454	leucine	rs4402960	3	-0.01	0.045154865	-0.01	0.043458776	-3321.1	-3315.1	-3316.6	2	2.11
IGF2BP2	cg19952454	valine	rs4402960	3	-0.01	0.044714999	-0.01	0.043458776	-3320	-3314.5	-3317.2	2	3.78
BCAR1	cg01805890	X - 12696	rs7202877	16	0.02	0.014890923	0.02	0.043581866	-2499.6	-2497.2	-2499.1	2	2.68
BCAR1	cg01805890	N-acetylglycine	rs7202877	16	0.01	0.047497504	0.02	0.043581866	-2500.7	-2498	-2498.9	2	1.59
PROX1	cg13921308	octanoylcarnitine	rs2075423	1	-0.01	0.047495126	-0.02	0.044005829	-2997.4	-2992.2	-2993.9	2	2.31
KCNQ1	cg04204548	15-methylpalmitate	rs231362	11	-0.01	0.047000901	0.02	0.044440295	-3072.1	-3068.6	-3070.7	2	2.98
KCNQ1	cg04204548	15-methylpalmitate	rs2237892	11	-0.01	0.047000901	0.02	0.044440295	-2278.2	-2275.2	-2280.2	2	4.33
KCNQ1	cg04204548	15-methylpalmitate	rs231361	11	-0.01	0.047000901	0.02	0.044440295	-2732.1	-2729.2	-2729.4	2	1.13
KCNQ1	cg04204548	15-methylpalmitate	rs163184	11	-0.01	0.047000901	0.02	0.044440295	-2896.4	-2894.6	-2893.6	3	1.65
MPHOSPH9	cg20350484	X - 12450	rs4275659	12	0.01	0.028497686	-0.02	0.044462367	-2883.7	-2878.2	-2878.8	2	1.34
MPHOSPH9	cg20350484	X - 11315	rs4275659	12	0.01	0.022662912	-0.02	0.044462367	-2884.6	-2879.8	-2879.5	3	1.18
MPHOSPH9	cg20350484	dimethylarginine	rs4275659	12	-0.02	0.040372391	-0.02	0.044462367	-2878.9	-2877.8	-2879.5	2	1.73
RBMS1	cg19506623	leucine	rs7593730	2	-0.02	0.036618474	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
RBMS1	cg19506623	X - 13496	rs7593730	2	-0.01	0.046927625	0.02	0.044539174	-2878.9	-2879.8	-2879.6	1	1.42
RBMS1	cg19506623	proline	rs7593730	2	-0.02	0.023933275	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
RBMS1	cg19506623	palmitoyl sphingomyelin	rs7593730	2	-0.01	0.020646416	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
RBMS1	cg19506623	valine	rs7593730	2	-0.02	0.032304512	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.37
RBMS1	cg19506623	octanoylcarnitine	rs7593730	2	-0.01	0.007047227	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
RBMS1	cg19506623	fructose	rs7593730	2	-0.01	0.0217423	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
RBMS1	cg19506623	X - 06246	rs7593730	2	-0.02	0.02098359	0.02	0.044539174	-2878.9	-2879.8	-2879.7	1	1.49
RBMS1	cg19506623	X - 11315	rs7593730	2	-0.02	0.006864053	0.02	0.044539174	-2878.9	-2879.8	-2879.5	1	1.35
HNF4A	cg08407434	palmitoyl sphingomyelin	rs4812829	20	-0.01	0.046885769	-0.01	0.044838842	-2993.9	-3000.7	-2989.5	3	8.93
HNF4A	cg08407434	X - 13496	rs4812829	20	-0.01	0.009969839	-0.01	0.044838842	-2992.7	-2998.1	-2988	3	10.5

HNF4A	cg08407434	malate	rs4812829	20	-0.01	0.027263263	-0.01	0.044838842	-2992	-2999.2	-2989.9	3	2.79
TP53INP1	cg01824466	pentadecanoate (15:0)	rs896854	8	0.01	0.021975132	-0.01	0.045217064	-3476.7	-3473.4	-3475.6	2	3.04
TP53INP1	cg01824466	X - 10510	rs896854	8	0.01	0.043236591	-0.01	0.045217064	-3469	-3466.2	-3476.1	2	3.99
TP53INP1	cg01824466	X - 13496	rs896854	8	0.01	0.023706849	-0.01	0.045217064	-3477.3	-3474.4	-3476	2	2.19
HCCA2	cg18481342	valine	rs2334499	11	0.02	0.016942719	0.02	0.045800576	-3105.5	-3103.3	-3102	3	1.87
HCCA2	cg18481342	palmitoyl sphingomyelin	rs2334499	11	-0.01	0.013978415	0.02	0.045800576	-3099.5	-3097.8	-3102.6	2	2.36
HCCA2	cg18481342	leucine	rs2334499	11	0.02	0.014910949	0.02	0.045800576	-3098.1	-3096.4	-3102.6	2	2.34
HCCA2	cg18481342	isoleucine	rs2334499	11	0.02	0.022289429	0.02	0.045800576	-3097.9	-3095.9	-3102.3	2	2.71
ANK1	cg26172342	palmitoleate (16:1n7)	rs516946	8	0.01	0.046859439	-0.01	0.047536783	-3214.5	-3209.1	-3209.5	2	1.21
ANK1	cg26172342	pelargonate (9:0)	rs516946	8	0.01	0.030532926	-0.01	0.047536783	-3213.8	-3208.9	-3210	2	1.71
ANK1	cg26172342	heptanoate (7:0)	rs516946	8	0.01	0.030092805	-0.01	0.047536783	-3207.5	-3202.5	-3209.8	2	12.49
ANK1	cg26172342	X - 11550	rs516946	8	0.01	0.045652037	-0.01	0.047536783	-3213.3	-3208.4	-3209.9	2	2.15
ANK1	cg26172342	X - 10500	rs516946	8	-0.01	0.010859229	-0.01	0.047536783	-3214.7	-3209.8	-3209.9	2	1.07
ANK1	cg26172342	glucose	rs516946	8	-0.01	0.030653213	-0.01	0.047536783	-3211.3	-3205.5	-3209.1	2	6.06
ZMIZ1	cg14841514	X - 11550	rs12571751	10	-0.01	0.009290396	-0.02	0.048019101	-3061	-3057.2	-3058.1	2	1.55
ANK1	cg08388995	X - 11550	rs516946	8	0.01	0.021247737	0.01	0.049314193	-3151.4	-3146	-3146.4	2	1.25

REFERENCES

Abecasis, G. R., et al. (2002), 'Merlin--rapid analysis of dense genetic maps using sparse gene flow trees', *Nat Genet,* 30 (1), 97-101.

Abecasis, G. R., et al. (2010), 'A map of human genome variation from population-scale sequencing', *Nature,* 467 (7319), 1061-73.

Adamski, J. and Suhre, K. (2013), 'Metabolomics platforms for genome wide association studies--linking the genome to the metabolome', *Curr Opin Biotechnol,* 24 (1), 39-47.

Adrian, T. E., et al. (2012), 'Rectal taurocholate increases L cell and insulin secretion, and decreases blood glucose and food intake in obese type 2 diabetic volunteers', *Diabetologia,* 55 (9), 2343-7.

Agarwal, A. K. (2012), 'Lysophospholipid acyltransferases: 1-acylglycerol-3-phosphate O-acyltransferases. From discovery to disease', *Curr Opin Lipidol,* 23 (4), 290-302.

Akaike, H. (1976), 'An information criterion (AIC)', *Math Sci,* 14 (153), 5-9.

Altmaier, E., et al. (2009), 'Variation in the human lipidome associated with coffee consumption as revealed by quantitative targeted metabolomics', *Mol Nutr Food Res,* 53 (11), 1357-65.

Altmaier, E., et al. (2011), 'Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics', *Eur J Epidemiol,* 26 (2), 145-56.

Amarasekera, M., et al. (2014), 'Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans', *FASEB J,* 28 (9), 4068-76.

Andrew, T., et al. (2001), 'Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women', *Twin Res,* 4 (6), 464-77.

Antequera, F. and Bird, A. (1993), 'Number of CpG islands and genes in human and mouse', *Proc Natl Acad Sci U S A,* 90 (24), 11995-9.

Aulchenko, Y. S., Struchalin, M. V., and van Duijn, C. M. (2010), 'ProbABEL package for genome-wide association analysis of imputed data', *BMC Bioinformatics,* 11, 134.

Aulchenko, Y. S., et al. (2007), 'GenABEL: an R library for genome-wide association analysis', *Bioinformatics,* 23 (10), 1294-6.

Ball, M. P., et al. (2009), 'Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells', *Nat Biotechnol,* 27 (4), 361-8.

Banovich, N. E., et al. (2014), 'Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels', *PLoS Genet,* 10 (9), e1004663.

Barderas, M. G., et al. (2011), 'Metabolomic profiling for identification of novel potential biomarkers in cardiovascular diseases', *J Biomed Biotechnol,* 2011, 790132.

Bates, D, et al. (2015), 'lme4:Linear mixed-effects models using Eigen and S4', (R package version 1.1-8).

Bell, J. T. and Spector, T. D. (2011), 'A twin approach to unraveling epigenetics', *Trends Genet,* 27 (3), 116-25.

Bell, J. T. and Saffery, R. (2012), 'The value of twins in epigenetic epidemiology', *Int J Epidemiol,* 41 (1), 140-50.

Bell, J. T., et al. (2011), 'DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines', *Genome Biol,* 12 (1), R10.

Bell, J. T., et al. (2012), 'Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population', *PLoS Genet,* 8 (4), e1002629.

Besingi, W. and Johansson, A. (2014), 'Smoke-related DNA methylation changes in the etiology of human disease', *Hum Mol Genet,* 23 (9), 2290-7.

Bibikova, M., et al. (2009), 'Genome-wide DNA methylation profiling using Infinium® assay', *Epigenomics,* 1 (1), 177-200.

Bictash, M., et al. (2010), 'Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology', *J Clin Epidemiol,* 63 (9), 970-9.

Billings, L. K. and Florez, J. C. (2010), 'The genetics of type 2 diabetes: what have we learned from GWAS?', *Ann N Y Acad Sci,* 1212, 59-77.

Bird, A. (2002), 'DNA methylation patterns and epigenetic memory', *Genes Dev,* 16 (1), 6-21.

Bjornsson, H. T., et al. (2008), 'Intra-individual change over time in DNA methylation with familial clustering', *JAMA,* 299 (24), 2877-83.

Bock, C. (2012), 'Analysing and interpreting DNA methylation data', *Nat Rev Genet,* 13 (10), 705-19.

Boker, S., et al. (2011), 'OpenMx: An Open Source Extended Structural Equation Modeling Framework', *Psychometrika,* 76 (2), 306-17.

Boks, M. P., et al. (2009), 'The relationship of DNA methylation with age, gender and genotype in twins and healthy controls', *PLoS One,* 4 (8), e6767.

Bolstad, B. M., et al. (2003), 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics,* 19 (2), 185-93.

Breitling, L. P., et al. (2011), 'Tobacco-smoking-related differential DNA methylation: 27K discovery and replication', *Am J Hum Genet,* 88 (4), 450-7.

Breton, C. V., et al. (2014), 'Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation', *PLoS One,* 9 (6), e99716.

Britschgi, A., et al. (2013), 'Calcium-activated chloride channel ANO1 promotes breast cancer progression by activating EGFR and CAMK signaling', *Proc Natl Acad Sci U S A,* 110 (11), E1026-34.

Brown, A. A., et al. (2014), 'Genetic interactions affecting human gene expression identified by variance association mapping', *Elife,* 3, e01381.

Bryois, J., et al. (2014), 'Cis and trans effects of human genomic variants on gene expression', *PLoS Genet,* 10 (7), e1004461.

Buro-Auriemma, L. J., et al. (2013), 'Cigarette smoking induces small airway epithelial epigenetic changes with corresponding modulation of gene expression', *Hum Mol Genet,* 22 (23), 4726-38.

Büscher, J. M., et al. (2009), 'Cross-platform comparison of methods for quantitative metabolomics of primary metabolism', *Anal Chem,* 81 (6), 2135-43.

Cao, H., et al. (2008), 'Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism', *Cell,* 134 (6), 933-44.

Chasman, D. I., et al. (2009), 'Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis', *PLoS Genet,* 5 (11), e1000730.

Cokus, S. J., et al. (2008), 'Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning', *Nature,* 452 (7184), 215-9.

Cosgrove, M. S. and Wolberger, C. (2005), 'How does the histone code work?', *Biochem Cell Biol,* 83 (4), 468-76.

Currie, E., et al. (2013), 'Cellular fatty acid metabolism and cancer', *Cell Metab,* 18 (2), 153-61.

de Bakker, P. I., et al. (2008), 'Practical aspects of imputation-driven meta-analysis of genome-wide association studies', *Hum Mol Genet,* 17 (R2), R122-8.

Demirkan, A., et al. (2015), 'Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses', *PLoS Genet,* 11 (1), e1004835.

Demirkan, A., et al. (2012), 'Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations', *PLoS Genet,* 8 (2), e1002490.

Dempster, E. L., et al. (2011), 'Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder', *Hum Mol Genet,* 20 (24), 4786-96.

DIAGRAM, et al. (2014), 'Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility', *Nat Genet,* 46 (3), 234-44.

Dogan, M. V., et al. (2014), 'The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women', *BMC Genomics,* 15, 151.

Doi, A., et al. (2009), 'Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts', *Nat Genet,* 41 (12), 1350-3.

Dominguez-Salas, P., et al. (2014), 'Maternal nutrition at conception modulates DNA methylation of human metastable epialleles', *Nat Commun,* 5, 3746.

Draisma, H. H., et al. (2015), 'Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels', *Nat Commun,* 6, 7208.

Drong, A. W., et al. (2013), 'The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue', *PLoS One,* 8 (2), e55923.

Eichler, E. E., et al. (2010), 'Missing heritability and strategies for finding the underlying causes of complex disease', *Nat Rev Genet,* 11 (6), 446-50.

Elliott, H. R., et al. (2014), 'Differences in smoking associated DNA methylation patterns in South Asians and Europeans', *Clin Epigenetics,* 6 (1), 4.

Evans, A. M., et al. (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), 6656-67.

Feinberg, A. P. and Irizarry, R. A. (2010), 'Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease', *Proc Natl Acad Sci U S A,* 107 Suppl 1, 1757-64.

Fiehn, O., et al. (2000), 'Metabolite profiling for plant functional genomics', *Nat Biotechnol,* 18 (11), 1157-61.

Flanagan, J. M., et al. (2006), 'Intra- and interindividual epigenetic variation in human germ cells', *Am J Hum Genet,* 79 (1), 67-84.

Fraser, H. B., et al. (2012), 'Population-specificity of human DNA methylation', *Genome Biol,* 13 (2), R8.

Frayling, T. M., 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), 889-94.

Gamazon, E. R., et al. (2013), 'Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants', *Molecular Psychiatry,* 18 (3), 340-6.

Gervin, K., et al. (2011), 'Extensive variation and low heritability of DNA methylation identified in a twin study', *Genome Res,* 21 (11), 1813-21.

Gibbs, J. R., et al. (2010), 'Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain', *PLoS Genet,* 6 (5), e1000952.

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

Gordon, L., et al. (2012), 'Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence', *Genome Res,* 22 (8), 1395-406.

Grundberg, E., et al. (2013), 'Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements', *Am J Hum Genet,* 93 (5), 876-90.

Guan, W., et al. (2008), 'Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium', *Hum Hered,* 66 (1), 35-49.

Guida, F., et al. (2015), 'Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation', *Hum Mol Genet,* 24 (8), 2349-59.

Gutierrez-Arcelus, M., et al. (2013), 'Passive and active DNA methylation and the interplay with genetic variation in gene regulation', *Elife,* 2, e00523.

Hackett, N. R., et al. (2012), 'RNA-Seq quantification of the human small airway epithelium transcriptome', *BMC Genomics,* 13, 82.

Harlid, S., et al. (2014), 'CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study', *Environ Health Perspect,* 122 (7), 673-8.

Heijmans, B. T., et al. (2007), 'Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus', *Hum Mol Genet,* 16 (5), 547-54.

Heuberger, A. L., et al. (2014), 'Application of nontargeted metabolite profiling to discover novel markers of quality traits in an advanced population of malting barley', *Plant Biotechnol J,* 12 (2), 147-60.

Hicks, A. A., et al. (2009), 'Genetic determinants of circulating sphingolipid concentrations in European populations', *PLoS Genet,* 5 (10), e1000672.

Hill, W. G. and Mulder, H. A. (2010), 'Genetic analysis of environmental variation', *Genet Res (Camb),* 92 (5-6), 381-95.

Hindorff, L. A., et al. (2009), 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc Natl Acad Sci U S A,* 106 (23), 9362-7.

Holliday, R. (1994), 'Epigenetics: an overview', *Dev Genet,* 15 (6), 453-7.

Holliday, R. and Pugh, J. E. (1975), 'DNA modification mechanisms and gene activity during development', *Science,* 187 (4173), 226-32.

Holmes, E., Wilson, I. D., and Nicholson, J. K. (2008a), 'Metabolic phenotyping in health and disease', *Cell,* 134 (5), 714-7.

Holmes, E., et al. (2008b), 'Human metabolic phenotype diversity and its association with diet and blood pressure', *Nature,* 453 (7193), 396-400.

Hulse, A. M. and Cai, J. J. (2013), 'Genetic variants contribute to gene expression variability in humans', *Genetics,* 193 (1), 95-108.

Illig, T., et al. (2010), 'A genome-wide perspective of genetic variation in human metabolism', *Nat Genet,* 42 (2), 137-41.

Irizarry, R. A., et al. (2009), 'The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores', *Nat Genet,* 41 (2), 178-86.

Ivorra, C., et al. (2015), 'DNA methylation patterns in newborns exposed to tobacco in utero', *J Transl Med,* 13 (1), 25.

Javierre, B. M., et al. (2010), 'Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus', *Genome Res,* 20 (2), 170-9.

Jonsson, A., et al. (2009), 'A variant in the KCNQ1 gene predicts future type 2 diabetes and mediates impaired insulin secretion', *Diabetes,* 58 (10), 2409-13.

Joubert, B. R., et al. (2012), '450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy', *Environ Health Perspect,* 120 (10), 1425-31.

Kaminsky, Z. A., et al. (2009), 'DNA methylation profiles in monozygotic and dizygotic twins', *Nat Genet,* 41 (2), 240-5.

Kastenmuller, G., et al. (2015), 'Genetics of human metabolism: an update', *Hum Mol Genet*.

Katada, S., Imhof, A., and Sassone-Corsi, P. (2012), 'Connecting threads: epigenetics and metabolism', *Cell,* 148 (1-2), 24-8.

Kettunen, J., et al. (2012), 'Genome-wide association study identifies multiple loci influencing human serum metabolite levels', *Nat Genet,* 44 (3), 269-76.

Kinoshita, M., et al. (2013), 'DNA methylation signatures of peripheral leukocytes in schizophrenia', *Neuromolecular Med,* 15 (1), 95-101.

Klein, R. J., et al. (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science,* 308 (5720), 385-9.

Koal, T. and Deigner, H. P. (2010), 'Challenges in mass spectrometry based targeted metabolomics', *Curr Mol Med,* 10 (2), 216-26.

Koller, D. and Friedman, N. (2009), *Probabilistic graphical models: principles and techniques* (MIT press).

Krumsiek, J., et al. (2012), 'Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information', *PLoS Genet,* 8 (10), e1003005.

Kuratomi, G., et al. (2008), 'Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins', *Molecular Psychiatry,* 13 (4), 429-41.

Laird, P. W. (2010), 'Principles and challenges of genomewide DNA methylation analysis', *Nat Rev Genet,* 11 (3), 191-203.

Lander, E. S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860-921.

Larsen, F., et al. (1992), 'CpG islands as gene markers in the human genome', *Genomics,* 13 (4), 1095-107.

Lawton, K. A., et al. (2008), 'Analysis of the adult human plasma metabolome', *Pharmacogenomics,* 9 (4), 383-97.

Lee, K. W., et al. (2015), 'Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age', *Environ Health Perspect,* 123 (2), 193-9.

Lee, Y. and Nelder, J. A. (2006), 'Double hierarchical generalized linear models (with discussion)', *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 55: 139–185.

Lemaitre, R. N., et al. (2011), 'Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium', *PLoS Genet,* 7 (7), e1002193.

Li, E., Bestor, T. H., and Jaenisch, R. (1992), 'Targeted mutation of the DNA methyltransferase gene results in embryonic lethality', *Cell,* 69 (6), 915-26.

Li, H., Ruan, J., and Durbin, R. (2008), 'Mapping short DNA sequencing reads and calling variants using mapping quality scores', *Genome Res,* 18 (11), 1851-8.

Lindon, John C., Nicholson, Jeremy K., and Holmes, Elaine (2007), *The handbook of metabonomics and metabolomics* (1st edn.; Amsterdam ; Boston: Elsevier) x, 561 p.

Lister, R., et al. (2008), 'Highly integrated single-base resolution maps of the epigenome in Arabidopsis', *Cell,* 133 (3), 523-36.

Lister, R., et al. (2009), 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature,* 462 (7271), 315-22.

Liu, Y., et al. (2013), 'Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis', *Nat Biotechnol,* 31 (2), 142-7.

Lu, C. and Thompson, C. B. (2012), 'Metabolic regulation of epigenetics', *Cell Metab,* 16 (1), 9-17.

Luijk, R., et al. (2015), 'An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs', *Bioinformatics,* 31 (3), 340-5.

Magi, R. and Morris, A. P. (2010), 'GWAMA: software for genome-wide association meta-analysis', *BMC Bioinformatics,* 11, 288.

Maher, B. (2008), 'Personal genomes: The case of the missing heritability', *Nature,* 456 (7218), 18-21.

Maier, E. M., et al. (2005), 'Population spectrum of ACADM genotypes correlated to biochemical phenotypes in newborn screening for medium-chain acyl-CoA dehydrogenase deficiency', *Hum Mutat,* 25 (5), 443-52.

Maksimovic, J., Gordon, L., and Oshlack, A. (2012), 'SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips', *Genome Biol,* 13 (6), R44.

Malet-Martino, M. and Holzgrabe, U. (2011), 'NMR techniques in biomedical and pharmaceutical analysis', *J Pharm Biomed Anal,* 55 (1), 1-15.

Mandal, R., et al. (2012), 'Multi-platform characterization of the human cerebrospinal fluid metabolome: a comprehensive and quantitative update', *Genome Med,* 4 (4), 38.

Manolio, T. A., et al. (2009), 'Finding the missing heritability of complex diseases', *Nature,* 461 (7265), 747-53.

Markunas, C. A., et al. (2014), 'Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy', *Environ Health Perspect,* 122 (10), 1147-53.

McCarthy, M. I. (2003), 'Growing evidence for diabetes susceptibility genes from genome scan data', *Curr Diab Rep,* 3 (2), 159-67.

McRae, A. F., et al. (2014), 'Contribution of genetic variation to transgenerational inheritance of DNA methylation', *Genome Biol,* 15 (5), R73.

Meissner, A., et al. (2008), 'Genome-scale DNA methylation maps of pluripotent and differentiated cells', *Nature,* 454 (7205), 766-70.

Menni, C., et al. (2013a), 'Targeted metabolomics profiles are strongly correlated with nutritional patterns in women', *Metabolomics,* 9 (2), 506-14.

Menni, C., et al. (2013b), 'Metabolomic markers reveal novel pathways of ageing and early development in human populations', *Int J Epidemiol,* 42 (4), 1111-9.

Menni, C., et al. (2013c), 'Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach', *Diabetes,* 62 (12), 4270-6.

Mittelstrass, K., et al. (2011), 'Discovery of sexual dimorphisms in metabolic and genetic biomarkers', *PLoS Genet,* 7 (8), e1002215.

Moayyeri, A., et al. (2012), 'Cohort Profile: TwinsUK and Healthy Ageing Twin Study', *Int J Epidemiol*.

--- (2013a), 'Cohort Profile: TwinsUK and healthy ageing twin study', *Int J Epidemiol,* 42 (1), 76-85.

Moayyeri, A., et al. (2013b), 'The UK Adult Twin Registry (TwinsUK Resource)', *Twin Res Hum Genet,* 16 (1), 144-9.

Monick, M. M., et al. (2012), 'Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers', *Am J Med Genet B Neuropsychiatr Genet,* 159B (2), 141-51.

Moran, S., Arribas, C., and Esteller, M. (2016), 'Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences', *Epigenomics,* 8 (3), 389-99.

Morris, A. P., et al. (2012), 'Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes', *Nat Genet,* 44 (9), 981-90.

Morris, T. J., et al. (2014), 'ChAMP: 450k Chip Analysis Methylation Pipeline', *Bioinformatics,* 30 (3), 428-30.

Naeem, H., et al. (2014), 'Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array', *BMC Genomics,* 15, 51.

Nagrath, D., et al. (2011), 'Metabolomics for mitochondrial and cancer studies', *Biochim Biophys Acta,* 1807 (6), 650-63.

Neale, Michael C., Cardon, Lon R., and Organization., North Atlantic Treaty (1992), *Methodology for genetic studies of twins and families* (NATO ASI series Series D, Behavioural and social sciences; Dordrecht ; Boston: Kluwer Academic Publishers) xxv, 496 p.

Nicholson, G., et al. (2011), 'A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection', *PLoS Genet,* 7 (9), e1002270.

Nicholson, J. K. and Lindon, J. C. (2008), 'Systems biology: Metabonomics', *Nature,* 455 (7216), 1054-6.

Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999), "Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data', *Xenobiotica,* 29 (11), 1181-9.

Nitert, M. D., et al. (2012), 'Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes', *Diabetes,* 61 (12), 3322-32.

Ordas, B., Malvar, R. A., and Hill, W. G. (2008), 'Genetic variation and quantitative trait loci associated with developmental stability and the environmental correlation between traits in maize', *Genet Res (Camb),* 90 (5), 385-95.

Pare, G., et al. (2010), 'On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study', *PLoS Genet,* 6 (6), e1000981.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann).

Pearl, Judea (2000), *Causality: models, reasoning and inference* (29: Cambridge Univ Press).

Petersen, A. K., et al. (2012), 'On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies', *BMC Bioinformatics,* 13, 120.

Petersen, A. K., et al. (2014), 'Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits', *Hum Mol Genet,* 23 (2), 534-45.

Philibert, R. A., et al. (2013), 'Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking', *Clin Epigenetics,* 5 (1), 19.

Philibert, R. A., et al. (2012), 'The impact of recent alcohol use on genome wide DNA methylation signatures', *Front Genet,* 3, 54.

Pidsley, R., et al. (2013), 'A data-driven approach to preprocessing Illumina 450K methylation array data', *BMC Genomics,* 14, 293.

Psychogios, N., et al. (2011), 'The human serum metabolome', *PLoS One,* 6 (2), e16957.

Purcell, S., et al. (2007), 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *Am J Hum Genet,* 81 (3), 559-75.

Raffler, J., et al. (2013), 'Identification and MS-assisted interpretation of genetically influenced NMR signals in human plasma', *Genome Med,* 5 (2), 13.

Ragvin, A., et al. (2010), 'Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3', *Proc Natl Acad Sci U S A,* 107 (2), 775-80.

Rakyan, V. K., et al. (2011a), 'Epigenome-wide association studies for common human diseases', *Nat Rev Genet,* 12 (8), 529-41.

Rakyan, V. K., et al. (2011b), 'Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis', *PLoS Genet,* 7 (9), e1002300.

Reik, W. (2007), 'Stability and flexibility of epigenetic gene regulation in mammalian development', *Nature,* 447 (7143), 425-32.

Reik, W. and Walter, J. (2001a), 'Genomic imprinting: parental influence on the genome', *Nat Rev Genet,* 2 (1), 21-32.

--- (2001b), 'Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote', *Nat Genet,* 27 (3), 255-6.

Reik, W., Dean, W., and Walter, J. (2001), 'Epigenetic reprogramming in mammalian development', *Science,* 293 (5532), 1089-93.

Rhee, E. P., et al. (2013), 'A genome-wide association study of the human metabolome in a community-based cohort', *Cell Metab,* 18 (1), 130-43.

Richards, E. J. (2006), 'Inherited epigenetic variation--revisiting soft inheritance', *Nat Rev Genet,* 7 (5), 395-401.

Richmond, R. C., et al. (2015), 'Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC)', *Hum Mol Genet,* 24 (8), 2201-17.

Ried, J. S., et al. (2014), 'Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses', *Hum Mol Genet,* 23 (21), 5847-57.

Römisch-Margl, Werner, et al. (2012), 'Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics', *Metabolomics,* 8 (1), 133-42.

Rönn, T., et al. (2013), 'A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue', *PLoS Genet,* 9 (6), e1003572.

Ronnegard, L. and Valdar, W. (2011), 'Detecting major genetic loci controlling phenotypic variability in experimental crosses', *Genetics,* 188 (2), 435-47.

--- (2012), 'Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability', *BMC Genet,* 13, 63.

Rueedi, R., et al. (2014), 'Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links', *PLoS Genet,* 10 (2), e1004132.

Sanghera, D. K. and Blackett, P. R. (2012), 'Type 2 Diabetes Genetics: Beyond GWAS', *J Diabetes Metab,* 3 (198).

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics,* 6 (2), 461-64.

Scutari, Marco (2009), 'Learning Bayesian networks with the bnlearn R package', *arXiv preprint arXiv:0908.3817*.

Shabalin, A. A. (2012), 'Matrix eQTL: ultra fast eQTL analysis via large matrix operations', *Bioinformatics,* 28 (10), 1353-8.

Shen, X., et al. (2012), 'Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana', *PLoS Genet,* 8 (8), e1002839.

Shenker, N. S., et al. (2013), 'DNA methylation as a long-term biomarker of exposure to tobacco smoke', *Epidemiology,* 24 (5), 712-6.

Shi, J., et al. (2014), 'Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue', *Nat Commun,* 5, 3365.

Shin, S. Y., et al. (2014), 'An atlas of genetic influences on human blood metabolites', *Nat Genet,* 46 (6), 543-50.

Shungin, D., et al. (2015), 'New genetic loci link adipose and insulin biology to body fat distribution', *Nature,* 518 (7538), 187-96.

Smith, A. K., et al. (2014), 'Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type', *BMC Genomics,* 15, 145.

Sonntag, D., et al. (2011), 'Targeted metabolomics for bioprocessing', *BMC Proc,* 5 Suppl 8, P27.

Spector, T. D. and Williams, F. M. (2006), 'The UK Adult Twin Registry (TwinsUK)', *Twin Res Hum Genet,* 9 (6), 899-906.

Steemers, F. J. and Gunderson, K. L. (2007), 'Whole genome genotyping technologies on the BeadArray platform', *Biotechnol J,* 2 (1), 41-9.

Storey, J (2015), '*qvalue: Q-value estimation for false discovery rate control*. R package version 2.0.0 '.

Strachan, T., Read, Andrew P., and Strachan, T. (2011), *Human molecular genetics* (New York: Garland Science) xxv, 781 p.

Struchalin, M. V., et al. (2010), 'Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations', *BMC Genet,* 11, 92.

Struchalin, M. V., et al. (2012), 'An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity', *BMC Genet,* 13, 4.

Suhre, K. and Gieger, C. (2012), 'Genetic variation in metabolic phenotypes: study designs and applications', *Nat Rev Genet,* 13 (11), 759-69.

Suhre, K., et al. (2010), 'Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting', *PLoS One,* 5 (11), e13953.

Suhre, K., et al. (2011a), 'A genome-wide association study of metabolic traits in human urine', *Nat Genet,* 43 (6), 565-9.

Suhre, K., et al. (2011b), 'Human metabolic individuality in biomedical and pharmaceutical research', *Nature,* 477 (7362), 54-60.

Sun, Y. V., et al. (2013), 'Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans', *Hum Genet,* 132 (9), 1027-37.

Surakka, I., et al. (2012), 'A Genome-Wide Association Study of Monozygotic Twin-Pairs Suggests a Locus Related to Variability of Serum High-Density Lipoprotein Cholesterol', *Twin Res Hum Genet*, 1-9.

Suter, M., et al. (2011), 'Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression', *Epigenetics,* 6 (11), 1284-94.

Szyf, M. (2013), 'DNA methylation, behavior and early life adversity', *J Genet Genomics,* 40 (7), 331-8.

Tanaka, T., et al. (2009), 'Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study', *PLoS Genet,* 5 (1), e1000338.

Teschendorff, A. E., 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), 189-96.

Thapar, M., et al. (2012), 'DNA methylation patterns in alcoholics and family controls', *World J Gastrointest Oncol,* 4 (6), 138-44.

Thompson, J. R., Attia, J., and Minelli, C. (2011), 'The meta-analysis of genome-wide association studies', *Brief Bioinform,* 12 (3), 259-69.

van Buuren, S and Groothuis-Oudshoorn, K (2011), 'mice: Multivariate Imputation by Chained Equations in R', *Journal of Statistical Software,* 45 (3), 1-67.

van Eijk, K. R., et al. (2012), 'Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects', *BMC Genomics,* 13, 636.

Venter, J. C., et al. (2001), 'The sequence of the human genome', *Science,* 291 (5507), 1304-51.

Visscher, P. M. and Posthuma, D. (2010), 'Statistical power to detect genetic Loci affecting environmental sensitivity', *Behav Genet,* 40 (5), 728-33.

Voisin, S., et al. (2015), 'Dietary fat quality impacts genome-wide DNA methylation patterns in a cross-sectional study of Greek preadolescents', *Eur J Hum Genet,* 23 (5), 654-62.

Wagner, J. R., et al. (2014), 'The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts', *Genome Biol,* 15 (2), R37.

Wan, E. S., et al. (2012), 'Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome', *Hum Mol Genet,* 21 (13), 3073-82.

Wang, G., et al. (2014), 'Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort', *Genetics,* 196 (2), 413-25.

Wang, T. J., et al. (2011), 'Metabolite profiles and the risk of developing diabetes', *Nat Med,* 17 (4), 448-53.

Wang, Y., et al. (2012), 'Metastasis-associated gene, mag-1 improves tumour microenvironmental adaptation and potentiates tumour metastasis', *J Cell Mol Med,* 16 (12), 3037-51.

Ward, M. C., et al. (2013), 'Latent regulatory potential of human-specific repetitive elements', *Mol Cell,* 49 (2), 262-72.

Weber, M., et al. (2005), 'Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells', *Nat Genet,* 37 (8), 853-62.

Weber, M., et al. (2007), 'Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome', *Nat Genet,* 39 (4), 457-66.

Welter, D., et al. (2014), 'The NHGRI GWAS Catalog, a curated resource of SNP-trait associations', *Nucleic Acids Res,* 42 (Database issue), D1001-6.

Willer, Cristen J., Li, Yun, and Abecasis, Goncalo R. (2010), 'METAL: fast and efficient meta-analysis of genomewide association scans', *Bioinformatics,* 26 (17).

Wilson, I. D., et al. (2005), 'HPLC-MS-based methods for the study of metabonomics', *J Chromatogr B Analyt Technol Biomed Life Sci,* 817 (1), 67-76.

Wong, C. C., et al. (2010), 'A longitudinal study of epigenetic variation in twins', *Epigenetics,* 5 (6), 516-26.

Yang, J., et al. (2010), 'Common SNPs explain a large proportion of the heritability for human height', *Nat Genet,* 42 (7), 565-9.

Yang, J., et al. (2012), 'FTO genotype is associated with phenotypic variability of body mass index', *Nature,* 490 (7419), 267-72.

Yu, Z., et al. (2012), 'Human serum metabolic profiles are age dependent', *Aging Cell,* 11 (6), 960-67.

Yuan, W., et al. (2014), 'An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins', *Nat Commun,* 5, 5719.

Zeggini, E. and Ioannidis, J. P. (2009), 'Meta-analysis in genome-wide association studies', *Pharmacogenomics,* 10 (2), 191-201.

Zeggini, E., et al. (2008), 'Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes', *Nat Genet,* 40 (5), 638-45.

Zeilinger, S., et al. (2013), 'Tobacco smoking leads to extensive genome-wide changes in DNA methylation', *PLoS One,* 8 (5), e63812.

Zhang, D., et al. (2010), 'Genetic control of individual differences in gene-specific methylation in human brain', *Am J Hum Genet,* 86 (3), 411-9.

Zhang, H., et al. (2013), 'Array-based profiling of DNA methylation changes associated with alcohol dependence', *Alcohol Clin Exp Res,* 37 Suppl 1, E108-15.

Zhang, Y., et al. (2014), 'F2RL3 methylation as a biomarker of current and lifetime smoking exposures', *Environ Health Perspect,* 122 (2), 131-7.

Zhao, R., et al. (2013), 'Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence', *Asia Pac Psychiatry,* 5 (1), 39-50.

Zhou, X. and Stephens, M. (2012), 'Genome-wide efficient mixed-model analysis for association studies', *Nat Genet,* 44 (7), 821-4.