Imperial College London
Department of Epidemiology and Biostatistics
School of Public Health

Statistical association networks as complex phenotypes: new methods and applications

Beatriz Valcárcel Salamanca

Supervised by Dr Maria de Iorio

Professor Marjo-Riitta Jarvelin and Dr Timothy M Ebbels

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy of Imperial College London and Diploma of Imperial College

To my parents

for their endless love and support

Abstract

Recent advances in 'omics technologies and the development of new computational techniques have greatly contributed to the identification of factors influencing the onset and progression of many common diseases. Yet, despite this great success, it is unlikely that the independent analysis of these data will elucidate the complex web of mechanisms involved in disease development. To enhance our knowledge of disease aetiology, new approaches for linking the large amount of available data need to be developed.

As a step towards this goal, the aim of this thesis is to investigate and develop novel statistical methods for the integrative analysis of 'omic data. In particular, in this project we analyse genomic and metabolomic data in relation to the health outcomes in three study populations. To investigate how genetic and metabolic variables act as risk factors in the development of complex disorders, we have developed three novel analytical methodologies, namely 'Differential Network', '**GEMINi**: **G**Enome **M**etabolome **I**ntegrated **N**etwork analys**i**s' and 'Variance and Covariance regression' and illustrate their use on real data sets.

The results demonstrate the applicability of the new methodologies to identify key molecular changes undetectable with standard approaches. The approaches introduce here have the potential of providing insight into the biological basis of phenotypic variation and aid the generation of new hypotheses about molecular control and regulation in the context of systems biology.

Contents

Abstract **3**

Declaration of originality **12**

Acknowledgements **14**

List of Abbreviations **16**

Summary **18**

1 Genomics and Metabolomics **21**

1.1 Genomics 21

1.1.1 The human genome 22

1.1.2 Human genetic variations 23

1.1.3 The structure of variations in the human genome . . . 25

1.1.4 Genome-wide association studies 26

1.2 Metabolomics 28

1.2.1 The human metabolome 29

1.2.2 Analysis of the metabolome 29

1.2.3 Exploring complex traits through metabolomics 31

1.3 Genomics with metabolomic traits 34

1.4 Summary 35

2 The Data Sets **37**

2.1 Study populations 37

2.1.1 Northern Finland Birth Cohort 1966 38

2.1.2 Northern Finland Birth Cohort 1986 39

2.1.3 Cadmium study population 39

2.2 Metabolomic data sets 40

2.2.1 ^{1}H NMR serum metabolites 40
2.2.2 ^{1}H NMR urinary metabolites 44
2.2.3 Biological context of ^{1}H NMR metabolites 44
2.3 Genetic data 52
2.3.1 Data quality controls 52
2.4 General Notation 53

3 Statistical methods for GWAS 54
3.1 Single-SNP analysis 54
3.1.1 Single-SNP association analysis 56
3.2 Multi-SNP analysis 61
3.2.1 Lasso 62
3.2.2 HyperLasso 64
3.2.3 The HyperLasso regression analysis 65
3.3 Multiple response analysis 70
3.3.1 SUR model 71
3.3.2 The SUR analysis 72
3.4 Summary 74

4 Differential Networks 76
4.1 Introduction 77
4.2 Basic Concepts of Complex Networks 78
4.3 Statistical Methods 80
4.3.1 Individual network analysis 81
4.3.2 Differential network analysis 84
4.4 Simulation study 85
4.4.1 Simulation Study Results 87
4.5 Case study: Prediabetes dyslipidemias 88
4.5.1 Study population and metabolic dataset 88
4.5.2 Results and Discussion 89
4.6 Summary 100

5 GEMINi 101
5.1 Introduction 101

5.2 Materials . . . 102
5.3 Statistical Methods . . . 104
5.3.1 Conventional analysis . . . 104
5.3.2 GEMINi . . . 105
5.4 Results and Discussion . . . 106
5.4.1 Conventional analysis of metabolic and genetic data . . 106
5.4.2 GEMINi . . . 111
5.5 Summary . . . 120

6 Variance and Covariance Regression 121
6.1 Introduction . . . 122
6.2 Material . . . 123
6.3 Statistical Methods . . . 124
6.4 Results and Discussion . . . 128
6.5 Summary . . . 135

7 Discussion 139
7.1 Summary . . . 139
7.2 Main achievements . . . 141
7.3 Challenges and limitations . . . 143

Bibliography 146

Appendices 167

A Data Sets 168

B Statistical methods for GWAS 173

C Differential Networks 192

D GEMINi 199

E Variance and Covariance Regression 203

F Permission for third party copyright 213

List of Figures

1.1 A graphic illustrating the topographical levels of human genome 23
1.2 A graphical representation of genetic variations 25
1.3 Graphical representation of genetic variants identified by GWAS for obesity and type 2 diabetes (T2DM) 28
1.4 Graphical representation of flow from the genetic code (DNA) to metabolites 30

2.1 Illustration of ^{1}H NMR spectra of serum obtained following a three-molecular-window approach 42
2.2 Overall structure of a serum lipoprotein 46
2.3 Overview of the lipoproteins metabolism 48
2.4 Diagram of the glycolysis pathway 49
2.5 Diagram of the citric acid cycle 51

3.1 The DE density distribution 63
3.2 Logarithms of NEG and DE densities 65

4.1 The individual networks obtained using three measures of correlation 92
4.2 The individual networks inferred from male data 93
4.3 Topological properties of the individual networks 94
4.4 The differential network inferred from males with normal fasting glucose (NFG) and impaired fasting glucose (IFG) 97
4.5 Topological properties of the differential network 98

5.1 GEMINi analysis outline 107
5.2 Differential network analysis for serum metabolites inferred from normal and obese individuals 113

6.1 Trace plots 130
6.2 Variance regression for dimethylglycine 131
6.3 Changes in pair-wise associations between urinary metabolites across the range of values of urinary cadmium 132
6.4 Graphic representation of choline metabolism 133
6.5 Correlation Networks inferred from the covariance regression analysis 137
6.6 Betweenness-centrality of the correlation networks 138

A.1 A flowchart for NFBC data analysis 172

B.1 The Manhattan plots of P-values for the leading traits obtained in the single-SNP association analysis. 185
B.1 The Manhattan plots of P-values for the leading traits obtained in the single-SNP association analysis. 186
B.2 Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the single-SNP analysis. 187
B.2 Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the single-SNP analysis. 188
B.3 Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the HyperLasso regression analysis. 189
B.3 Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the HyperLasso regression analysis. 190

C.1 Correlation maps using Pearson's correlation 195
C.2 Correlation maps using standard partial correlation 196
C.3 Correlation maps using shrinkage partial correlation 197
C.4 The differential network inferred from females with normal fasting glucose (NFG) and impaired fasting glucose (IFG) . . . 198

D.1 Differential network analysis for serum metabolites inferred from normal and obese individuals (BMI) 202

List of Tables

2.1 Lipoprotein subclasses and particle diameter. 43
2.2 Molecular properties of the five major lipoprotein classes. . . 46

3.1 Summary of the results from the single-SNP association analysis of serum metabolic traits adjusted for sex. 60

4.1 Results for the simulation study. 88
4.2 Global properties of the individual networks obtained using shrinkage partial correlation as measure of pair-wise association. 91

5.1 Linear regression analysis in the NFBC1966 between body adipose index (BAI) and the serum metabolite measures 109
5.2 Linear regression analysis in the NFBC1986 between body adipose index (BAI) and the serum metabolite measures 110
5.3 Genetic loci associated with differences in metabolite-metabolite correlations between obese and non-obese individuals. 116
5.4 Pathway analysis of the whole set of genetic variants identified by the GWC analysis of variation in metabolic associations. . 119

6.1 Multiple linear regression analysis between exposure to cadmium (U-Cd) and the urinary metabolite measures. 129

A.1 List of all serum metabolic measures included in this thesis . 168
A.2 List of all urinary metabolic measures included in this thesis . 171

B.1 Results from the multi-SNP analysis of serum metabolic traits 173
B.2 Results of the SUR analysis of VLDL system 191
B.3 Results of the SUR analysis of HDL system 191

B.4 Results of the SUR analysis of metabolites involved in the citric acid cycle 191
B.5 Results of the SUR analysis of metabolites involved in the urea cycle 191

C.1 Clinical characteristic of the NFBC1966 study participants. . . 192
C.2 Lipoprotein subclass measures and their mean concentrations for normal (NFG) and impaired fasting glucose (IFG). 193

D.1 Linear regression analysis in the NFBC1966 between body mass index (BMI) and the serum metabolite measures 199
D.2 Linear regression analysis in the NFBC1986 between body mass index (BMI) and the serum metabolite measures 200

E.1 Connectivity of the correlation networks. 203
E.2 Results from the variance regression analysis 204
E.3 Results from the covariance regression analysis 205

Declaration of originality

I hereby declare that this thesis is my own work and that where any material could be construed as the work of others, it is fully cited and referenced, and/or with appropriate acknowledgement given.

Beatriz Valcárcel Salamanca

Copyright Statement

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commericial No Derivitives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Acknowledgements

In the preparation of this thesis I would like to express my sincere gratitude to my supervisor Dr Maria De Iorio, for her endless patience, advice and guidance throughout this project. Thank you for leading me forward. I could have not asked for better supervision.

Thanks are also due to my co-supervisors Professor Marjo Riita Jarvelin and Dr Timothy M Ebbels. Thanks Marjo Riita for giving me the opportunity to work with such an excellent data set and for always having a smile, willing to help and provide advice. Thanks Tim for your constant support and positive encouragement. This thesis would certainly not have been the same without your great collaboration.

I would like to gratefully acknowledge Professor Mika Ala-Korpela and Dr Jake G Bundy for generously supplying the metabonomic data sets used in this thesis and for their invaluable critical-review of my work.

I would also like to thank my colleagues of the Department of Epidemiology and Biostatistics for their constant help and support and Ms Tuula Ylitalo for her help. Special thanks are due to Dr Paul O'Reilly who patiently guiding me during the first stages of my research and Dr Queenie Chan for her generous support and guidance.

Thanks to my fellow students, for making the often lonely process of PhD much more enjoyable. I would like to acknowledge the informal support and encouragement of many friends which has been indispensable for the completion of this thesis.

A special mention goes to Calvin. I can not express in words how much your love and care mean to me, without your support the path leading to this thesis would not have been possible.

A final mention must go to my family. My parents, Asuncion Salamanca and Nemesio Valcárcel, my bother Fran Valcárcel have been the foundation

of love and integrity throughout my life and have inspired me to become a better person, professionally and personally. Thanks for always being there when I need you the most!.

Funding Statement

This project was supported by the Economic Social Research Council (ESRC award no. ES/H016058/1).

NFBC1966 received financial support from the Academy of Finland (project Grant no. 24300796, 104781, 120315, 129269, 1114194, Center of Excellence in Complex Disease Genetics and SALVE), University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R0 1MH63706:02), ENGAGE project and grant agreement HEALTH-F4-2007-201413, and the Medical Research Council, UK (G0500539, G0600705, PrevMetSyn/SALVE).

The DNA extractions, sample quality controls, biobank up-keeping and aliquotting was performed in the National Public Health Institute, Biomedicum Helsinki, Finland and supported financially by the Academy of Finland and Biocentrum Helsinki. We thank Professor (emerita) Paula Rantakallio (launch of NFBC1966 and 1986), and Ms Outi Tornwall and Ms Minttu Jussila (DNA biobanking). The authors would like to acknowledge the contribution of the late Academian of Science Leena Peltonen.

List of Abbreviations

GWAS	Genome-wide association studies
^{1}H NMR	Proton Nuclear Magnetic Resonance
BAI	Body adipose index
BCAA	Branched-chain amino acids
BMI	Body mass index
CE	Cholesterol ester
CHD	Coronary heart diseases
CNV	Copy number variation
CVD	Cardiovascular diseases
DE	Double exponential
DNA	Deoxyribonucleic acid
FC	Free Cholesterol
GEMINi	GEnome Metabolome Integrated Network analysis
HDL	High density lipoprotein
HWE	Hardy-Weinberg Equilibrium
IBD	Identical by descent
IDL	Intermediate density lipoprotein
IFG	Impaired fasting glucose
Lasso	The Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LMWM	low molecular weight metabolites
MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
MS	Mass Spectrometry
$\boldsymbol{N}$	Sample size
NEG	Normal-Exponential Gamma

NFBC1966	Northern Finland Birth Cohort 1966
NFBC1986	Northern Finland Birth Cohort 1986
NFG	normal fasting glucose
NMR	Nuclear Magnetic Resonance
P	P-value
PL	Phospholipids
SNP	Single nucleotide polymorphism
SUR	The Seemingly Unrelated Regression
T2DM	Type 2 diabetes mellitus
TG	Triacylglycerols
U-Cd	Urinary cadmium
VLDL	Very low density lipoprotein

Summary

The major goal for modern biomedicine is to elucidate the complex web of interactions through which genes and their products, in combination with environmental factors, influence the onset and progression of many complex disorders. The knowledge of mechanisms is expected to be translated into clinical settings to improve prediction of individuals' disease risk and develop better and more personalized treatments for complex diseases.

The development of high throughput technologies for assaying molecular characteristics of biological systems has been transforming biomedical research for the past decade. Molecular profiling technologies such as single nucleotide polymorphism (SNP) arrays, gene expression microarrays and high-throughput metabolomics provide a wealth of information about molecular changes underlying diseases development. For example, SNP datasets have aided the identification of genetic variants strongly associated with the risk of disease development. Similarly, the rapidly developing field of metabolomics has made remarkable progress in understanding metabolic changes as response to genetic modifications and environmental stimuli.

While separated analysis of these different types of data has provided valuable information, it is unlikely that considering these data sets independently will uncover the complex web of biological mechanisms involved in disease development. To enhance our knowledge of mechanisms involved in disease aetiology, as well as facilitate the prediction of (patho)physiological status more accurately, systems-based approaches for analysing biological data is becoming crucial.

As a step towards this goal, the aim of this thesis is to investigate and develop novel statistical methods for the integrative analysis of metabolomic and genomic data. In particular, in this project we analyse and combine genetic and metabolic information in relation to the health outcomes in the

three study populations. We begin by examining the genetic basis of variations in serum metabolite levels. For that purpose, three different statistical approaches, namely, single-SNP, multiple-SNP and multiple responses, are applied. Secondly, we investigate the use of changes in molecular associations to identify key differences between (patho)physiological states. We present a novel statistical approach for the differential analysis of molecular associations and illustrate its application in the context of diabetic dyslipidemias. Thirdly, we apply the differential network approach to an integrative genome-metabolome analysis and investigate the potential use of variations in association between different molecules as an alternative phenotype to uncover genetic variants involved in complex biological processes. We refer to this statistical analysis as GEMINi: **G**Enome **M**etabolome **I**ntegrated **N**etwork analys**i**s. Finally, we propose to employ Bayesian hierarchical models to investigate changes in mean levels and variances of metabolic traits as response to environmental stimuli. We further investigate changes in correlation between metabolites in relation to these covariates.

Following these integrative approaches, we expect to uncover hidden aspects of the combined genetic and metabolic variables' contribution to the development of complex disorders. This will aid the generation of new hypotheses about molecular control and regulation in the context of systems biology.

Main Achievements of This Thesis

In this research project, we introduce three novel analytical methodologies for the comprehensive analysis of metabolic data in combination with SNP data. Their application is shown on real data sets. They can be summarized as follows:

The **Differential Network** approach is a formal statistical method for the differential analysis of molecular associations via network representation.

GEMINi: **G**Enome **M**etabolome **I**ntegrated **N**etwork analys**i**s is a novel analytical approach which exploits differences in molecular associations to uncover genetic variants involved in phenotype variation.

The **Variance and Covariance regression** is a new analytical procedure based on Bayesian hierarchical models that allows the investigation of continuous changes in mean level, variance and correlation of metabolic traits in relation to covariates.

Thesis Outline

Chapter 1 - Genomics and Metabolomics: In this chapter we provide an introduction to the two main areas of research covered by this thesis: Genomics and Metabolomics. We review basic definitions and key concepts used in the area of research. Moreover, applications of genomics and metabolomics in biomedical research are reviewed.

Chapter 2 - Data Sets: In this chapter we describe the materials used in this thesis. Firstly, we present the three study populations investigated in this thesis. Brief description of the data collection and design is also provided. The two types of 'omic data investigated in this work are described in this chapter. Details of analytical platforms experimental design and protocols are also covered. Moreover, we provide a biological background of the main metabolites investigated.

Chapter 3 - Statistical Methods in Genome-Wide analysis: In this introductory chapter we focus on investigating the genetic basis of serum metabolite traits. Consequently, three different statistical approaches for genome-wide association analysis, namely, single-SNP, multiple-SNP and multiple responses, are applied.

Chapter 4 - Differential Networks: Here, we investigate the use of changes in molecular association to identify key differences between (patho) physiological states. We present a novel statistical approach for the differential analysis of molecular associations. Firstly, a simulation study is performed with the aim of assessing the performance of the proposed differential networks at identifying differences in molecular associations between

two different groups. Subsequently, we show the application of the differential network approach in the context of diabetic dyslipidemias.

Chapter 5 - GEnome Metabolome Integrated Network analysis: In this chapter, we apply the differential network approach to an integrative genome-metabolome analysis and investigate the potential use of variations in association between different molecules as an alternative phenotype to uncover genetic variants involved in complex biological processes. We refer to this statistical analysis as GEMINi: GEnome Metabolome Integrated Network analysis.

Chapter 6 - Variance and covariance regression: In this chapter, the focus is on investigating continuous changes in metabolic traits as response to pathogenic stimuli. In particular, we propose to employ Bayesian hierarchical models to investigate changes in mean levels, variance and correlation of metabolic traits in relation to covariates. To illustrate our approach we investigate changes in urinary metabolic traits in response to cadmium exposure, a toxic environmental pollutant.

Chapter 7 - Conclusions and future work: In the last chapter we draw conclusions on the results obtained in the thesis and we outline some possible future research directions.

Chapter 1

Genomics and Metabolomics

In this chapter we provide an overview of the two main areas of research covered by this thesis: Genomics and Metabolomics. A full review of the entire field of genomics and metabolomics is beyond the scope of this project, therefore the aim of this chapter is to provide a brief review of key concepts and methodological aspects underlying these two research areas. Current applications of genomics and metabolomics to the study of complex traits and diseases will be covered in more detail, highlighting some of their strengths and limitations. In the last part of this chapter we present a number of examples that illustrate how by integrating metabolomics with genomics, studies can gain in power and provide a deeper understanding of mechanisms underlying complex traits.

1.1 Genomics

Genomics is a relatively new scientific discipline which focuses on the global analysis of genes, examining not only the role but also the interplay between genes and the interaction of genes with the organism's environment. The field of genomics is a very broad area of research ranging from structural genomics to functional genomics. In this thesis we focus on genome-wide association studies (GWAS) and their application to elucidate the genetic structure of human complex traits.

1.1.1 The human genome

The term genome refers to the complete set of the genetic material of an organism, which is essential for its correct functioning and survival. In humans, each copy of the genome is constituted of approximately 3 billion base pairs (bp) of deoxyribonucleic acid (DNA), stored in 23 chromosome pairs in the cells' nucleus. An individual inherits one chromosome per pair from the mother and the other copy from the father. In addition to the nuclear DNA, there is also a small proportion of circular DNA (cDNA) stored in mitochondria, which is inherited only from the female line.

The information carried in the DNA is stored as a code made up of chemical building blocks called nucleotides. Nucleotides are made of three components: a phosphate group, a sugar group and one of four types of nitrogenous bases known as adenine (A), thymine (T), guanine (G) and cytosine (C) (Figure 1.1.a). The most common form of DNA consists of two twisted strands that run in opposite directions to each other, so-called the double helix (1) (Figure 1.1.b). The base-pairing rules in this form of DNA specify that adenine pairs with thymine and guanine pairs with cytosine (Figure 1.1.a). The DNA sequences that contain instructions to code for a particular protein (amino acid sequences) or with a particular function (e.g. regulatory) are called genes, which are unevenly distributed across the chromosomes (Figure 1.1.c).

The specific location of a gene or DNA sequence in a chromosome is known as locus. Genes are divided into two types of DNA sequence: exons and introns. An exon is the portion of DNA in a gene that codes for proteins whereas intron refers to the gene sequence that do not code for proteins. It is estimated that the number of protein-coding genes is about 20,000 to 25,000, which represents around the 1.5% of whole genome (2). Although the majority of this genetic material has no known biological function, much of it is expected to be involved in transcriptional and translational regulatory functions (3).

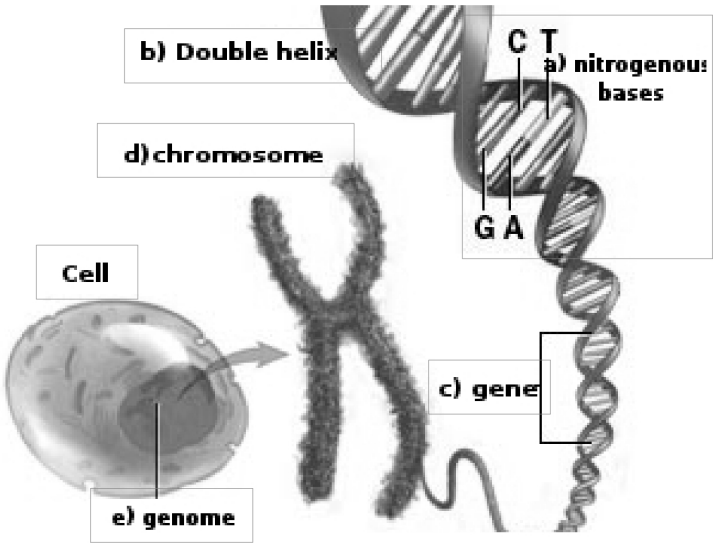


Figure 1.1: A graphic illustrating the topographical levels of human genome. From a) nucleotide sequence progressively folding as b) DNA double helix, c) genes, d) chromosome and finally the e) genome in the cells' nucleus. Figure taken and modified from Riverside Health System's webside. (Original URL: http://www.riversideonline.com/health_reference/Tools/DS00549.cfm)

1.1.2 Human genetic variations

While the human genome is largely similar among different individuals - 99.9% of one individual DNA sequence is identical to that of another individual - there are still a few variations. The genomes of two non-related subjects are estimated to differ about 1 in every 1,200-1,500 bases (4). There are, therefore, more than three million differences between different individuals' genomes. Each version of a DNA sequence in a given locus is called an allele. The complete set of alleles (genetic makeup) of an individual is defined as a genotype.

Genetic variations can be introduced into our genome through two main events: recombination and mutation. Recombination is the major force for generating genetic variation. It is a common phenomenon through which two chromosomes exchange segments of nucleotide sequence, resulting in the production of a new DNA molecule. This shuffling of alleles varies between individuals and across different locations within the human genome (5). The

second most important force that generates genetic variations is genetic mutation. A mutation occurs when the nucleotide sequence of a DNA molecule is altered. It can occur at random regions across the entire genome at any time and in any body cell. The estimated rate of mutation in humans is 2.5^{-8} per nucleotide site per generation (6). Mutations can be classified according to their phenotypic effect as silent or functional mutations. The majority of mutations are functionally silent, that is, no phenotypic effect is observable (7). This can happen either because a mutation has occurred in a non-coding or non-regulatory region with no biological function or because the variation in DNA sequence does not impair the final amino acid sequence. On the other hand, when a functional mutation occurs, it is usually associated with phenotypic changes such as morphological, biochemical, loss/gain of function. A single mutation can have a large effect (e.g. lethal mutation), but in most of the cases, phenotypic changes are based on the accumulation of multiple mutations with only small effects.

Genetic variants can also be discussed in terms of the frequency of their minor allele (less frequent allele) as common or rare variants in a specific population. The term 'common genetic variants' is generally taken to mean genetic variants with a minor allele frequency (MAF) of 5% or more in a population. The most common type of variants are single nucleotide polymorphisms (SNP). This type of genetic variant represents the substitution of a single nucleotide by another (Figure 1.2.a). It is estimated that SNPs make up about 90% of all human genetic variation and occur every 100 to 300 bases. Since the human genome has approximately 3 billion base pairs, it has been estimated that there are approximately 11 million SNPs across the entire genome, of which 7 million have a MAF$>$ 5% (8). Other types of mutations are structural variations (macromutations) which are rarer than SNPs but alter a larger region in the genome. Examples of structural variations are deletions (Figure 1.2.b), duplications (Figure 1.2.c), inversion (Figure 1.2.d), substitution (Figure 1.2.e) and translocation (Figure 1.2.f) of large blocks of DNA sequence. A form of structural genetic variation that has increasingly gained attention in genetic studies is copy number variation (CNV)(9).

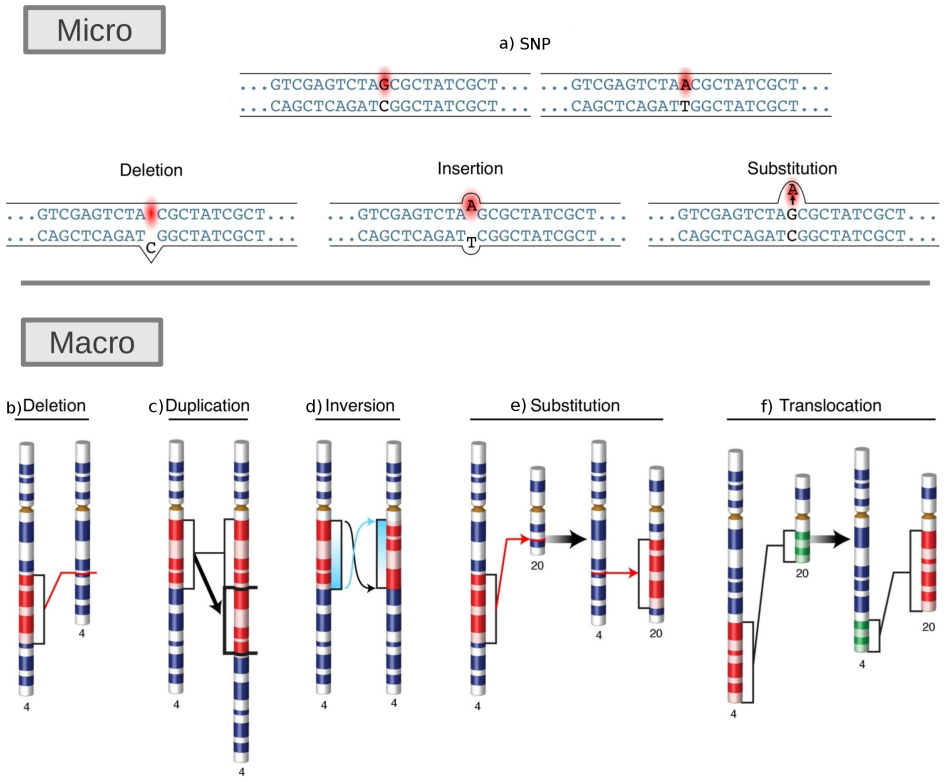


Figure 1.2: A graphical representation of genetic variations. Top panel represents the difference between two DNA molecules at a single base-pair location. Bottom panel illustrates a number of structural variations including deletions, duplications, inversions, substitution and translocations of large blocks of DNA sequence. Image taken and modified from National Human Genome Research Institute glossary.

1.1.3 The structure of variations in the human genome

Due to their abundance, the most commonly used variants are SNPs. Most SNPs are exploited as genetic markers, that is, genetic variants strongly associated with the true causative genetic variant. The statistical measure that defines association between genetic variants is called linkage disequilibrium (LD) (10). Specifically, LD describes a phenomenon through which some combinations of alleles occur together more often than expected by chance, based on their allele frequencies in the population. This phenomenon occurs either because there is a low recombination rate between two loci or be-

cause one (or both) of the alleles arose recently (11). The linear combination of closely linked alleles (generally at adjacent locations) on a chromosome and transmitted together are known as haplotypes. Overall, the genome LD structure can be summarized as a series of regions within which alleles are in strong LD (cold recombination spots) separated by short discrete regions of very low LD (hotspots). In regions of low recombination frequencies, there is a limited number of haplotypes (haplotype blocks) (12).

The information provided by LD and haplotype blocks can be used to generate genetic maps of variants were genotyped SNPs act as direct proxies to other untyped but highly correlated SNPs.

1.1.4 Genome-wide association studies

During the last decade there has been considerable effort devoted to the identification of genetic variants that underlie complex phenotypes. With the development of high throughput genotyping technology and SNPs reference maps (13, 14, 15, 16), the genome-wide association (GWAS) approach has become a powerful tool to investigate the genetic basis of complex traits (17). In GWAS studies typically hundreds of thousands of genetic variants are scanned across the entire genome of thousands of individuals to find genetic regions that are associated with a trait of interest. An important feature of GWAS studies is that they follow a hypothesis-free design, that is, all SNPs are assessed for association without focusing on loci whose biology suggests that a relationship to trait is likely.

The GWAS method relies on the assumption that common variants (with MAF $> 5\%$) are strong candidates for being responsible for most of the genetic contribution to complex traits variability, each of them exerts a small-to-modest effect size ("Common Disease-Common Variant" hypothesis). However, the question of whether disease risk can be explained by these common variants, detectable by GWAS, or is mainly determined by more rare variants with frequency lower than 1% , still remains unanswered (18).

The power to detect genetic variants with weak effect sizes is contingent upon a number of factors which include the sample size of the study pop-

ulation, genome coverage and the strength of LD between the markers and causal variants. The application of appropriate statistical approaches can also have important impact on the detection of significant SNP-phenotype associations. While a number of statistical methods have been proposed for analysing genome-wide data, the per-SNP-per-phenotype approach still dominates the literature. Based on this approach, association between genetic variants and a trait of interest is tested considering only one SNP at a time. SNPs that do not pass the cut-off for significance are discarded. Those SNPs that show a significant association with the traits are generally tested again in an independent dataset. The SNPs that remain significantly associated with the trait are taken as potentially causal variants or if not, at least as a proxy for the true causal variant. Under the univariate approach each tested SNP constitutes a separate hypothesis test, therefore the statistical significance threshold needs to be corrected to reduce the chance a false positive. The most accepted threshold for significance in GWAS is $P < 5 \times 10^{-8}$ which is equivalent to a per-test $\alpha = 0.05$ after a Bonferroni correction for 1 million independent tests (19). With the application of such a stringent cut-off for significance, very strong association between genetic variants and the traits under study should be identified.

Since the first GWAS study published in 2005 (20), more than 200 complex diseases and traits have been studied by GWAS and hundreds of genetic variants have been identified (21). In Figure 1.3 we show a graphical representation of genetic variants identified by GWAS for obesity and type 2 diabetes (T2DM).

Despite the success of GWAS at identifying genetic variants associated with complex traits, the major challenge now is how to translate these results into clinical settings. It is important to bear in mind that currently GWAS studies only provide information about statistical association between a genetic variant and the trait of interest which does not necessarily infer causal relationship to the outcome. Several strategies have been proposed to improve the biological interpretation of GWAS findings. One approach that is gaining popularity is the use of molecular phenotypes such as metabolic profiling in combination with GWAS. It has been suggested that detecting

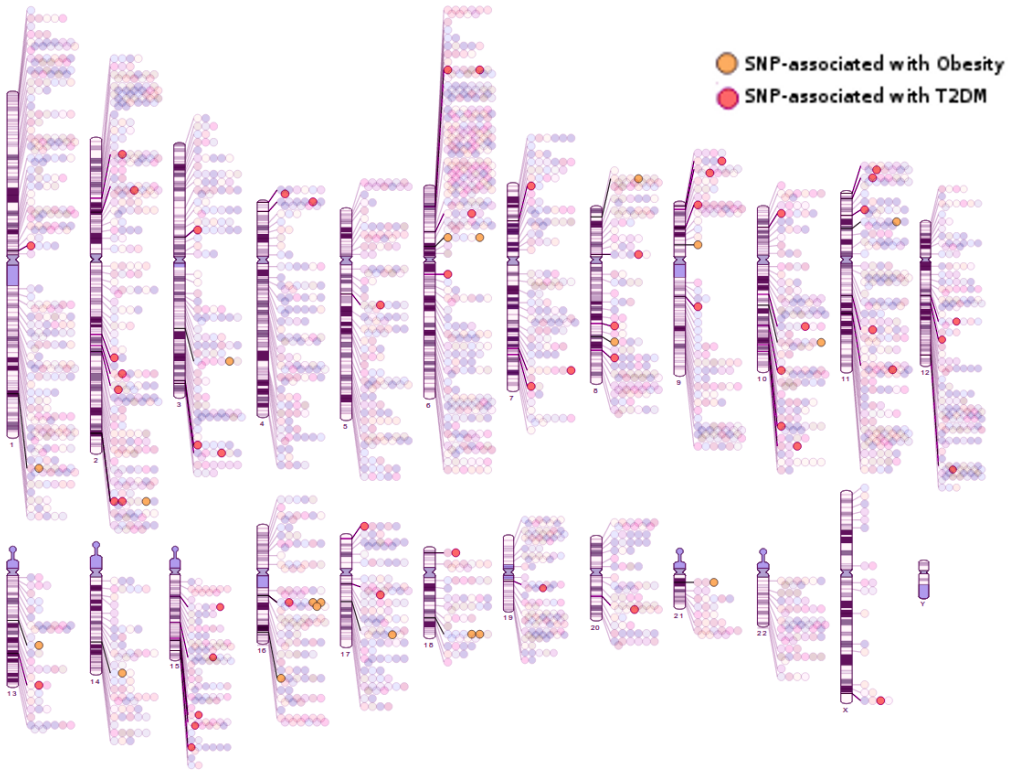


Figure 1.3: Graphical representation of genetic variants identified by GWAS for obesity and type 2 diabetes (T2DM). Genetic variants associated with the traits of interest are shown by their location on each of the 23 human chromosomes. Image was taken and modified from National Human Genome Research Institute: Catalogue of published Genome-Wide association studies (http://www.genome.gov/gwastudies/).

genetic variants associated with molecular phenotypes may provide more detailed information about the specific mechanism involved in disease development, generating in this way a causal link between genes and disease traits.

1.2 Metabolomics

Metabolomics (22), also known as metabonomics (23) or metabolic profiling, is defined as the comprehensive, qualitative and quantitative study of the

repertoire of small molecules. Specifically, metabolomic research combines the application of spectroscopic technology with multivariate statistics to investigate the metabolic composition of cells, tissues or biofluids as well as to gain an overall understanding of how metabolites levels change in response to genetic and environmental stimuli. During the last decade, metabolomics has grown in popularity and has been applied in a wide range of research areas. In biomedical science, metabolomics has been demonstrated to be a valuable tool to elucidate mechanisms underlying disease development.

1.2.1 The human metabolome

Metabolites (biomolecules less than 1.5 kDa in size) are the intermediate and end products of all cellular activities and play critical roles for the correct functioning of the cell. They act as building blocks of many cellular components (e.g. phospolipids), provide energy (e.g. triglyceride) and play active roles in cell signalling and regulation. The full suite of small-molecule metabolites to be found in a biological sample (cell, tissues, or fluids) in a particular physiological or developmental state is known as the metabolome (22). The metabolome of an organism - its metabolic composition, concentration and fluxes - is closely related to and highly affected by its genotype, physiological state and environment (Figure 1.4). Because variations at any level of the biological system are reflected as changes at the metabolic level, the metabolome represents the most integrative profile of biological state. As result, the analysis of the metabolome serves as a link to investigate interactions between the genotype and phenotype as well as to provide a direct read-out of the physiological status of organisms.

1.2.2 Analysis of the metabolome

Due to the large diversity of biochemical species (e.g. ionic inorganic, hydrophilic carbohydrates, hydrophobic lipids) that form the metabolome and differences in composition between biosamples, it is not currently possible to analyse the entire metabolome using a single analytical method. The most commonly used analytical techniques for data acquisition in metabolomics

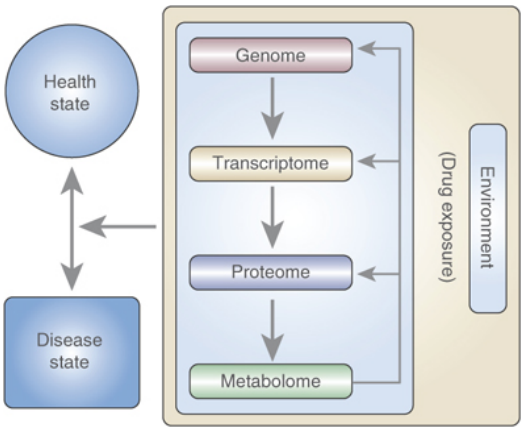


Figure 1.4: Graphical representation of flow from the genetic code (DNA) to metabolites. The environment and DNA affect the end products and influence health disease states. Figure reprinted by permission from Macmillan Publishers Ltd: Neuropsychopharmacology, Kaddurah-Daouk, R. and Krishnan, K. R., Metabolomics: A Global Biochemical Approach to the Study of Central Nervous System Diseases, copyright (2009) (24))

are Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR). The principles and properties of these two techniques are very different, however they are also complementary instruments and are often used in parallel. MS is based on the principle of ions separation according to their mass-to-charge ratio and is a very powerful method for screening large number of metabolites with great sensitivity, providing a great metabolome coverage. MS generally requires cleaning and separation steps prior to analysis, which is usually done using high-performance liquid or gas chromatography. Nuclear Magnetic Resonance (NMR) spectroscopy is an non-destructive analytical technique used for determining the content of a sample as well as its molecular structure, both in vitro or in vivo. NMR-based metabolomics is used to detect metabolites that contain atomic nuclei with a nuclear spin, that is, atomic nuclei that resonate (spin) when an external magnetic field is applied. Example of these compounds are ^{1}H, ^{13}C, ^{15}N. Protons (^{1}H) are one of the most used nuclei for NMR magnetic spectroscopy because they are present in nearly all naturally occurring compounds.

Broadly speaking, analytical approaches for metabolomic analysis can

be placed in two main categories - targeted and untargeted. Untargeted metabolomics is used for the global analysis of all metabolites in a sample including known or unknown compounds. The main aim is to achieve specific fingerprints or metabolic patterns that differentiate samples from different biological origins or treatments. Targeted metabolomics on the other hand focuses on the identification and quantification of a selected group of pre-defined metabolites, that belong to a specific class of compounds such as carbohydrates or amino acids. Because such a set of targeted metabolites are chemically known, it is possible to quantify their absolute concentrations by comparing the biosample spectrum to internal standard compound. Once the compounds have been identified and quantified, classical statistical methods such regression models or correlation analysis can be applied to draw biological conclusions.

1.2.3 Exploring complex traits through metabolomics

Many complex disorders, such as cancer or kidney diseases, manifest themselves as abnormal patterns of metabolic profiles. Such alterations can occur during the early stages of disease development, even decades before they manifest as clinical phenotype. Hence the identification of disease-related metabolic changes can serve as an early diagnostic tool as well as to determine individuals predisposed to a specific disorder. To date, there have been a number of studies that have used metabolomic data to identify diseases-specific molecular mechanisms (25, 26, 27). Since applications of metabolomics are numerous, here we only provide some examples of metabolomics applied to the study of two common metabolic disorders, type 2 diabetes and obesity which are the main clinical phenotypes investigated in this thesis.

Metabolomics applied to diabetes research

Type 2 diabetes mellitus (T2DM) is a heterogeneous metabolic disorder known to be caused by a complex combination of genetic and lifestyle factors (28). Patients clinically diagnosed with T2DM present increased levels of fasting glucose in blood and a relative insulin resistance/insulin deficiency.

In addition, individuals with T2DM often exhibit abnormal patterns of serum metabolites. Association between altered levels of lipids (dyslipidemias) and T2DM has been known for many years (29). More recently, the application of advanced lipoprotein profiling such as NMR spectroscopy of plasma have revealed alterations in particle size and lipids composition of certain lipoprotein subclasses in relation to T2DM-related disorders. Garvey et al. (2003) (30) investigated the effect of insulin resistance on lipoprotein metabolism using NMR-based lipoprotein subclasses. This study demonstrated the association between insulin resistance and accumulation of atherogenic lipoprotein particles, including increased levels of large VLDL, small LDL and small HDL. Such abnormalities in lipoprotein subclasses have been reported even at prediabetic state. Festa et al. (2005) (31) studied lipoproteins subfractions measured by NMR in 830 nondiabetic individuals at baseline (31). A specific pattern of abnormalities in lipoprotein subclasses, which includes size and compositional changes in VLDL and HDL subclasses, was identified for those prediabetic subjects who developed diabetes when compared with subjects who stayed nondiabetic. More recently, the relationship between glucose tolerance and lipoprotein profiles has been investigated in a large population-based study which included a total of 9399 Finnish men (normoglycemic and newly diagnosed T2DM patients) (32). Association between abnormal glucose metabolism and increased levels of all VLDL subclass particles (with the exception of very small VLDL particles) was found. In addition, decrease in levels of large HDL particles and increased concentration of small HDL were found in T2DM patients after adjustment for total triglyceride and HDL-C concentrations.

In addition to diabetes dyslipidemias, there is an increasing interest in understanding the relationship between branched-chain amino acids (BCAAs) and T2DM-related disorders. The first metabolomic study that drew attention to the potential role of BCAAs in diabetes was performed by Newgard et al. (2009) (33). Based on a targeted metabolomic analysis on obese and lean individuals, a distinctive metabolic signature, which related to variations in BCCAS, was identified in obese subjects. These metabolic changes, which correspond with increased levels of valine, isoleucine and leucine, showed high

correlation with insulin resistance. More recently increased concentrations of three branched amino acids and two aromatic amino acids (phenylalanine, tyrosine, valine, isoleucine, leucine) were found in individuals who developed T2DM when compared to individuals who did not develop T2DM (34).

Metabolomics applied to obesity research

Obesity, defined as an excessive body fat accumulation, is a multifactorial metabolic disorder which results from an imbalance between a positive energy intake and a negative energy expenditure. Obesity - especially central obesity - is known to be strong risk factor for the development of cardiometabolic disorders such as insulin resistance and type 2 diabetes. Since obesity is highly associated with T2DM development, many of the previously mention diabetes-related metabolic changes such as dyslipidemias and abnormal levels of BAACs in plasma have been also found in obesity. In addition, a large number of obesity-related metabolic disturbances have been identified by metabolomics, both in animal models and humans (35).
Serkova et al. (36) performed a NMR-based metabolomic analysis of both blood and liver tissue from obese and lean rats. Based on this study, metabolic abnormalities in mitochondrial function and methionine metabolism were related to obesity. In particular, liver extracts from obese rats showed increased levels of fatty acids and triglyceride in liver, increase of methionine concentration as well as decrease of betaine. Based on these findings, authors suggested that the specific metabolic abnormalities in obesity might result from a decreased hepatic energy state (36).

NMR has also been employed to investigate the metabolic effect of high-calorie diets. Example of this is the work performed by Duggan et al. (2011) (37) where NMR metabolic profiling of serum was used to differentiate between obesity-related effects to those effects specific from diet. Based on this study, differences in metabolite levels such as elevated pyruvate, alanine and lactate were identified. Most of these changes were attributed to diet-specific effects.

A recent metabolomic-proteomic approach was used to identify circulat-

ing molecules that differentiate healthy lean individuals from healthy obese (38). Targeted serum metabolomics identified 12 metabolites significantly related to obesity. Of the reported metabolic changes, glutamine and glycine were significantly higher in obese compared to lean individuals.

1.3 Genomics with metabolomic traits

The analysis of genome and metabolome data individually can provide valuable information about molecular changes related to clinical traits. However, to fully understand the biological processes and mechanisms that underlie disease development, it is essential that we jointly consider the different levels of molecular information. Here, we present a number of studies that illustrate the advantages of combining metabolomics and genomics.

The first human study using metabolomic data as a phenotype for GWAS was performed by Gieger et al.(2008) (39). Gieger and colleagues investigated the use of ratios of individual metabolites to explore genetic factors contributing to the modulation of the human metabolism. Despite the modest sample size (N=284), authors were able to identify genetic variants that exhibit large effect sizes on the metabolic phenotypes. Interestingly, nearly all identified SNPs were in genes encoding for proteins functionally involved in the metabolism of the associated metabolites. Through this analysis, it was shown that the use of metabolic measures increases the power of GWAS, allowing the identification of gene-phenotype associations with a very high degree of significance.

In a recent GWAS study on the genetic impact on the human metabolome (40), 37 loci were associated with metabolic concentrations in blood at genome-wide significance level. Of these, 25 loci showed very high effect size, accounting for up to 60% of the differences in metabolic levels per allele copy. Moreover, 30 loci were mapped to a protein biochemically related to the associated metabolite and 15 loci were related to complex diseases or drug response.

Lipidomics - the global study of lipid molecules produced by cells, tissue or organism - in combination with SNP data is an increasingly common

approach. It is used to investigate genetic variants contributing to changes in lipid concentrations and therefore affecting the risk of cardiovascular disease development. Over 90 genetic loci have been so far identified in GWAS of classic lipid measures (41). While some of the identified genetic variants have not been implicated in lipids metabolism, the majority of the genetic determinants are in or near genes previously known to influence plasma lipid levels. More recently Kettunen et al. (2012) (42) performed a GWAS on 8330 Finnish individuals for whom a range of 216 serum metabolic traits were measured by NMR metabolomic platform. They identified 31 loci significantly associated with one or more metabolic measures at a genome-wide significance level. Eleven of the identified loci were newly reported for metabolic measures or cardiometabolic disorders.

Tukiainen et al. (2012) (43) used a detailed dataset on lipids and lipoprotein subclasses measures (216 metabolic concentrations) measured by NMR metabolomic platform to investigate association between NMR-based lipoproteins and the set of loci previously identified for enzymatic total lipid measures (41). Based on this approach, 30 new SNP-metabolic trait associations were uncovered. For the majority of the loci, the strongest association was to a more specific metabolite measure than the originally identified through enzymatic total lipid measures (41). In addition, particle-specific effects were identified in four loci and the smallest high-density lipoprotein (HDL) measures showed effects opposite to the larger ones. The findings indicated that much greater resolution in the analysis can be obtained by looking at more specific phenotypic measures than those from routine clinical lipid measures (43).

1.4 Summary

With technological advances as well as development of novel analytical approaches, genomic and metabolomic studies have become powerful tools to gain a better understanding of biological processes involved in disease development. The information generated by these studies provides us with increased knowledge of how molecular mechanisms are affected by pathogenic

conditions. While genomic and metabolomic studies are very useful individually, these two approaches collectively hold exceptional promise. The literature illustrates that the combination of GWAS with metabolic phenotyping is an effective approach for improving the identification of molecular factors contributing to disease development. Furthermore, these studies revealed that the use of metabolomics in GWAS can increase the power to identify variations, yielding strong statistical associations and more importantly, can provide direct evidence of molecular basis of disease aetiology.

Chapter 2

The Data Sets

This chapter presents the data sets used in this thesis. We begin by presenting the study populations investigated (section 2.1). Next, we describe the two metabolomic data sets (serum and urinary data) employed in this work (section 2.2). The analytical platforms and experimental protocols used for metabolomic data acquisition are briefly described. To better understand the biological significance of the metabolic measures investigated in this work, we provide an overview of the main biological roles played by these sets of metabolites. The genomic data and quality control methods applied are briefly described in section 2.3. We finish this chapter describing the general notation used throughout the thesis.

2.1 Study populations

The primary study population investigated in this thesis is the Northern Finland Birth Cohort 1966 (NFBC1966) (subsection 2.1.1). In addition to the NFBC1966, two other data sets are investigated: The Northern Finland Birth Cohort 1986 (NFBC1986) which is described in subsection 2.1.2 and the Cadmium study population, presented in subsection 2.1.3. All studies were approved by local ethics committees and participants provided informed consent.

2.1.1 Northern Finland Birth Cohort 1966

The Northern Finland Birth Cohort is an epidemiological and longitudinal research program started in 1965. A detailed description of the data can be found on the study web page (http://kelo.oulu.fi/NFBC). In brief, this birth cohort is based on a general, unselected and genetically homogeneous population. It consists of data on individuals who were due to be born between January 1^{st} and December 31^{st} of 1966 in the two northernmost provinces of Finland, Oulu and Lapland. The cohort initially included data on 12,068 mothers and their 12,058 live infants. The majority of the cohort members (99%) are Finns (white Caucasians). According to Finland's central Office of Statistics, the cohort members comprised 96.3% of all births in 1966 in the target area. Data on the cohort members have been collected prospectively since pregnancy and supplemented at the ages of 1, 14 and 31 years, when a wide range of biological, socioeconomic and lifestyle factors as well as health conditions and family background information have been included (44, 45, 46, 47). The data collection and the follow-ups are described below. In Figure A.1 we represent how the NFBC data was used in this thesis.

The follow-ups

All mothers with expected delivery date in 1966 who were living in the study area (Oulu and Lapland), were invited to participate in the study: 12,068 attended. Prenatal data were collected by local midwives in the antenatal clinic visit ($24^{th} - 28^{th}$ gestational weeks)(44). Altogether 12,058 children were born into the cohort. The first follow-up data was collected in 1967, when cohort members were one year old (1-year follow-up). Developmental data and health status of the children were collected during the children's visits to welfare centres (44). The second follow-up was performed in 1980 and in early 1981, when cohort members were 14 years old (14-year follow-up). All cohort members with known addresses received postal questionnaires containing questions on their growth, health habits and family situation. A shortened version of the questionnaire was sent to the families in cases where the adolescent did not respond or to the school health nurse if neither the

child nor the families responded. The postal questionnaire was returned by 94% (N = 11,010) of the adolescents, 52% (N = 389) of the parents and 97% (N = 354) of the nurses (45). The last follow-up was conducted when cohort members were 31 years old (31-year follow-up) during 1997-98. All living cohort members with known addresses (N=11,541) received a postal questionnaire including questions on their health, lifestyle and occupation. The questionnaire was returned by 75% (N = 8,690). At the same time, those subjects living in the original study area - northern Finland or in the capital area (Helsinki) - were invited to participate to a clinical examination and 71% (N = 6,007) of those invited participated. At this point, blood samples were drawn after overnight fasting and samples were stored at -80^{o}C until analysed (46).

2.1.2 Northern Finland Birth Cohort 1986

The Northern Finland Birth Cohort 1986 (NFBC1986) consists of data on individuals with expected date of birth between 1^{st} July 1985 and 30^{th} June 1986, also in the two northernmost provinces of Finland, Oulu and Lapland (http://kelo.oulu.fi/NFBC) (48, 49). The number of mothers and deliveries was 9,362 (99% of all eligible) and there were 9,432 children born alive. Each cohort member has been followed up from early pregnancy until adolescence. Growth measurements were obtained from communal child health clinics. All those alive with known address were invited to a clinical examination at the age of 15 to 16 years. Blood samples were collected from those participants who provided written informed consent (N= 6,266).

2.1.3 Cadmium study population

The study population consists of individuals living near the lead/zinc smelter Avonmouth (Bristol, England), one of the largest source of atmospheric cadmium in the world. Details on the samples collection can be found in (50). Briefly, study subjects were recruited using the National Health Service database of patients registered with general practitioners (the Exeter System) in the area (50). Subjects were stratified by age, sex and esti-

mated exposure prior to random sample selection. A total of 865 people were initially invited to take part to the study. Data collection consisted of a questionnaire on health history, residential area, occupation and living habits such as smoking as well as the consumption of home-grown vegetables and the use of a private well. At the same time participants were asked to provide one spot morning urine sample. Samples were frozen on the day of collection and stored frozen until analysis.

2.2 Metabolomic data sets

In this work we investigate metabolomic data from urinary and serum samples. Both sets of metabolic data have been analysed by ^{1}H NMR spectroscopy but the experiment procedures and data acquisition for each type of data are different. In subsection 2.2.1, we describe the serum metabolomic data and the ^{1}H NMR metabolomic protocol used for its acquisition. The urinary metabolomic data and ^{1}H NMR metabolomic experiment are presented in subsection 2.2.2.

2.2.1 ^{1}H NMR serum metabolites

Serum metabolic profiles were available for two of the study populations: NFBC1966 and NFBC1986. Both cohorts were analysed using the same high-throughput serum NMR metabolomic platform in a single analysis laboratory. All analytical processes - sample collection, sample preparation, data acquisition, data pre-processing and quality assurance - were performed by Professor Mika Ala-Korpela and colleagues at the University of Oulu. Details on the NMR platform and experiment protocol can be found in (51), (52), (53). In brief, serum molecular data were acquired using a high-throughput NMR platform (Bruker AVANCE III spectrometer) operating at 500 MHz. A recently introduced NMR spectroscopy protocol, which optimizes measurement and characterization of serum metabolites, was employed to determine the serum metabolic profiles. To facilitate the identification and quantification of the large variety of compounds in blood serum, measurements were

performed following a three-molecular-window approach (51). The first set of data (LIPO window) provides detailed information about macromolecules such as lipids in lipoprotein subclasses. The second window (LMWM window) provides data on a variety of low-molecular-weight metabolites, such as amino acids, creatinine and ketone bodies. The last window (LIPID window) contains information on a range of lipid molecules along with their degree of (poly)(un)saturation. The methodology currently provides information on 117 directed measured serum metabolites and > 100 derived variables, all of them with high biological significance. Below we describe in more detail each of the molecular windows.

LIPO window: The data for the LIPO window contain information about lipoproteins subclasses distribution, particle concentrations and individual subclass lipids content (Figure 2.1.a). The data were acquired from native serum samples at 37.0°C. A standard NOESY-presat pulse sequence was used with mixing time of 10 ms and irradiation field of 25 Hz to suppress the water peak. The spectra were recorded with 80k data points using 8 transients acquired with an automatically calibrated 90° pulse. The acquisition time was 2.7s and the relaxation delay 3.0s. Prior to lipoprotein quantification the albumin signal was mathematically removed from the spectrum to improve accuracy (51). To isolate the different lipoprotein categories, the NMR-visible lipid amount was converted into information on the lipoprotein subclasses by a regression model. In addition, lipoprotein categories were converted into lipoprotein fractions - cholesterol and triglyceride - by assuming an average lipid composition of the lipoprotein particles at a certain NMR chemical shift range and calibrated to units of mmol/l (e.g. total VLDL triglyceride). Thus, the lipoprotein, lipid and particle concentrations are all different approximations based on the NMR measure. Quantification of particle concentrations and various lipid components in 14 lipoprotein subclasses were obtained using a computationally more efficient modification of the regression approach presented in Vehtari et al. (2007)(54). The 14 lipoprotein subclasses along with the average particle diameter are presented in Table 2.1. In addition, concentrations of apolipoprotein B and apolipoprotein A-1 were calculated using an extended Friedewald algorithm based on

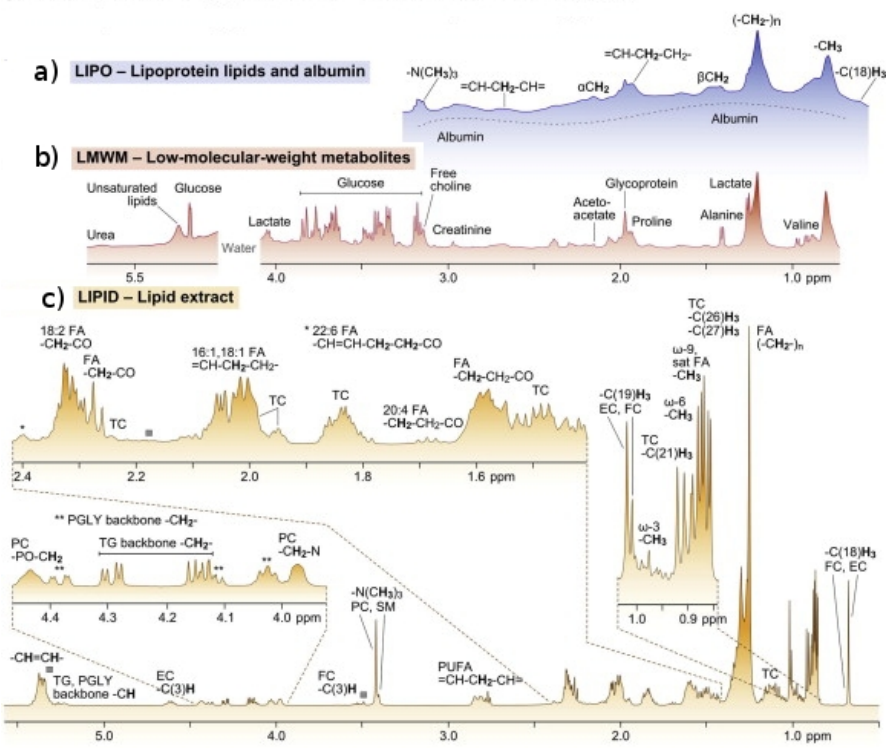


Figure 2.1: Illustration of ^{1}H NMR spectra of serum obtained following a three-molecular-window approach. a) The lipoprotein and albumin spectrum (LIPO window); b) The low-molecular weight metabolites spectrum (LMWM window); c) The serum lipid extracts (LIPID window). **FA**, fatty acid; **EC**, esterified cholesterol; **PGLY**, phosphoglycerides; **FC**, free cholesterol; **PC**, phosphatidylcholine; **PUFA**, polyunsaturated fatty acid; **sat**, saturated; **SM**, sphingomyelin; **TC**, total serum cholesterol; **TG**, total serum triglycerides. Figure reprinted from *Biochemical and Biophysical Research Communications*, 375/3, Tukiainen T, Tynkkynen T et al. A multi-metabolite analysis of serum by ^{1}H NMR spectroscopy: Early systemic signs of Alzheimers disease, 356-361, Copyright (2008), with permission from Elsevier.

triglycerides, total cholesterol and HDL cholesterol measures quantified from the spectral data (55).

LMWM window: The low molecular weight metabolites (Figure 2.1.b) data set was acquired from native serum samples and contains information on various amino acids, a number of metabolites involved in the glycolysis pathway, citric acid cycle and excretion metabolites (51). The chemical shifts from the low-molecular-weight metabolites were uncovered by the applica-

Table 2.1: Lipoprotein subclasses and particle diameter.

Lipoprotein	Subclasses	Diameter(nm/m)
Chylomicrons	extremely large VLDL	75
VLDL	very large VLDL	64
	large VLDL	53.6
	medium VLDL	44.5
	small VLDL	36.8
	very small VLDL	31.3
IDL	IDL	28.6
LDL	large LDL	25.5
	medium LDL	23
	small LDL	18.7
HDL	very large HDL	14.3
	large HDL	12.1
	medium HDL	10.9
	small HDL	8.7

tion of a pulse sequence that suppresses the majority of the signals from macromolecules such as lipoproteins (52). These molecular windows - LIPO window and LMWM window - are likely to contain most of the molecules information currently observable by ^{1}H NMR in native serum (51).

LIPID window: Following a standard lipid extraction protocol (52), the lipoprotein particle structure is broken down. In the LIPID window, the signals in the NMR spectrum arise from various lipid molecules within the lipoprotein particles (Figure 2.1.c). In addition, information on the degree of (poly)(un)saturation is available. The following fifteen lipids-related metabolites are currently quantified from the LIPID spectra: total amount of triglycerides; total amount of phosphoglycerides, total amount of cholesterol esterified; total amount of omega-3; omega-6 and omega-9 fatty acids; total amount of linoleic acid and total amount of docosahexaenoic acid (DHA); total amount of fatty acids that contain more than one double bond in their backbone (Polyunsaturated fatty acids); total amount of cholines; total amount of phosphatidylcholines; total amount of sphingomyelin.

For NFBC1966, data on the three molecular windows were available and used in this thesis. For NFBC1986, data on two molecular windows, LIPO window and LMWM, were available. In appendix (Table A.1) we summarise the data used.

2.2.2 ^{1}H NMR urinary metabolites

Urinary metabolic profiles were available for only one of the study populations: the Cadmium study population. Urinary metabolic data were acquired using ^{1}H NMR spectroscopy. Details on the NMR platform and experiment protocol can be found in Ellis et at. (2012)(56). In brief, the data were acquired from native serum samples at 37.0°C using a high-throughput NMR platform (Bruker AVANCE 600 spectrometer) operating at 600 MHz fitted with a flow-injection mode probehead at a field strength of 14.1 T (600.13 MHz ^{1}H frequency). The probe temperature was set to 300 K. Samples were delivered to the probehead in 96-well plates using an automated flow injection system (Bruker). Following introduction to the probe, samples were left to equilibrate (three minutes) prior to gradient shimming using the ^{1}H channel to ensure good magnetic field homogeneity. All liquid sample handling, automation and acquisition were controlled using XWIN-NMR 3.1 software (Bruker) running on a UNIX workstation (Silicon Graphics, USA). One-dimensional NMR spectra were acquired using a standard pulse sequence. Excitation sculpting with gradients was used to suppress the water peak. Following eight dummy scans, each spectrum was acquired into 64 k datapoints over a spectral width of 20 ppm as the sum of 64 transients. The acquisition time was 2.73 seconds and the relaxation delay two seconds. Peak assignment was performed based on the addition of known standards to the biological samples, together with published literature. Quantification of urinary metabolites were obtained integrating manually placed bins.

This methodology provided information on a total of 22 urinary metabolic measures, including amino acids, 3-hydroxybutyrate, creatinine and urea and information on metabolites involved in gluconeogenesis pathway such as pyruvate, lactate, glycerol. The complete list of metabolites included in this analysis can be found in the appendix (Table A.2).

2.2.3 Biological context of ^{1}H NMR metabolites

The metabolites quantified by the NMR platforms previously described play important roles in a variety of biological functions. Here we provide an

overview of the main roles and metabolic pathways in which they are involved. Since lipoprotein are the most explored molecular data in this thesis, a major emphasis will be given to these metabolites and their metabolism.

Lipids and lipoproteins

The term lipid defines a diverse group of naturally-occurring molecules which includes fats, waxes, sterols, sphingolipids, glycolipids, fat-soluble vitamins and phospholipids. The main biological functions of lipids include energy storage, building-block of cell membranes and molecule signalling. There are two main types of lipids in blood, triglycerides and cholesterol. Due to their low solubility in water, blood lipids are carried in lipoproteins. It is well established that elevated levels of blood lipids are risk factors for the development of cardiovascular diseases and related disorders. Because lipoproteins are the main lipids transporter in blood, abnormalities in lipoprotein profiles are often used as strong indicator of lipid disorders and consequently of the risk of cardiovascular disease development.

The overall structure of a serum lipoprotein is presented in Figure 2.2. Lipoproteins are spherical macromolecular complexes consisting of a water-soluble envelope and lipid-rich core and transport lipids to various body tissues. Phospholipids (PL) are the main compound forming the envelope. PLs have amphipathic character with a polar part oriented to the aqueous surface and an apolar part that interacts with the hydrophobic core of the lipoprotein. Free Cholesterol (FC) also appears forming part of the lipoprotein envelope. The core of the lipoprotein is formed mainly by cholesterol ester (CE) and triacylglycerols (TG). The protein components of the lipoprotein are known as apoproteins (apo) and serve as cofactors for enzymes and ligands for receptors.

Lipoproteins represent a heterogeneous group of particles that differ in density, size, composition, functions and binding affinity. These physical and chemical properties are influenced by the apolipoproteins type and the amount of lipids that forms the lipoprotein structure. Lipoproteins can be identified and classified based on their physical and chemical properties. Den-

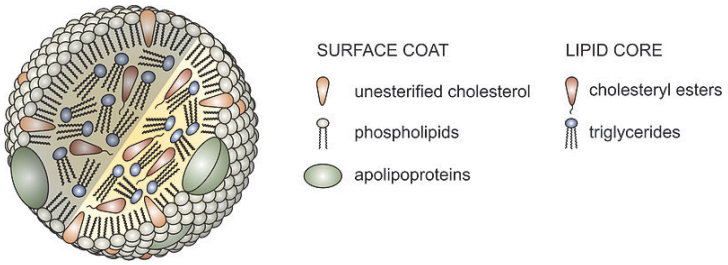


Figure 2.2: Overall structure of a serum lipoprotein. Shown is a schematic representation of a lipoprotein particle comprised of a hydrophilic layer of phospholipids, apoproteins and free cholesterol and a hydrophobic core that contains triglycerides and cholesterol esters. Figure reprinted from Wikimedia Commons, the free media repository with permission from Free Software Foundation.

sity is one of the criteria used for lipoproteins classification. Based on this property, there are five major lipoprotein classes which are chylomicron, very low density lipoproteins (VLDL), intermediate density lipoproteins (IDL), low density lipoproteins (LDL) and high density lipoproteins (HDL), each of which has a different function in lipid metabolism. Properties of the five major lipoprotein classes are listed in Table 2.2. Density and particle size are related by lipid to protein ratio. For instance, when lipoproteins are formed by more lipids than proteins, lipoprotein particles are larger and less dense. In addition to the main classification, specific subfractions of three lipoprotein classes (VLDL, LDL and IDL) can be determined. These different subcategories are defined also by differences in particle size, density and apolipoprotein content.

Table 2.2: Molecular properties of the five major lipoprotein classes.

Class	Density	Diameter	Apo	%Prot	%TL	%Cho	%PL	%TG
Chylomicrons	< 1.006	100-1000	ApoB	2	98	8	7	83
VLDL	0.95-1.006	30-80	ApoA	10	90	22	18	50
IDL	1.006-1.019	25-50	ApoB	18	82	29	22	31
LDL	1.019-1.063	18-28	ApoA	23	77	50	20	7
HDL	1.063-1.210	5-15	ApoA	35	65	30	30	5

Class, main lipoprotein classes; **Density**, density of the particle measured in (g/mL); **Diameter**, diameter of the particle measured in (nmm); **Apo**, main apolipoprotein; **%Prot**, percentage of protein; **%TL**, percentage of total lipids; **%Cho** percentage of cholesterol; **%PL** percentage of phospholipid; **%TG** percentages of triglycerides. Table partially taken from Biochemistry $2^n d$ Ed. 1995 Garrett & Grisham.

Lipoprotein metabolism

The lipoprotein metabolism is divided into two pathways, exogenous and endogenous. In the exogenous pathway (Figure 2.3.a), dietary fats (mainly triglycerides) enter the intestinal mucosa (enterocytes) via the fatty acid transporter (FA). Once inside the intestinal lacteals, triglycerides (TG) are packaged with cholesterol ester and apolipoprotein B-48 (APOB48) into nascent chylomicrons (nCMs). The nCMs are segregated via the intestinal lymphatic system and enter the blood stream. While circulating in blood, the nascent chylomicrons are converted into mature chylomicron (mCM) by exchanging apolipoprotein C-II (APOC2) and apolipoprotein E (APOE) with high-density lipoproteins (HDL). Then TGs in the core of mCM are hydrolyzed into free fatty acids and glycerol by the endothelial-bound lipoprotein lipase with APOC2 as cofactor. This process produces smaller CM, so-called remnants CM which enter the liver by hepatic LDL receptor (LDLR) and breakdown (57).

In the endogenous pathway (Figure 2.3.b), very low density lipoproteins (VLDL) are synthesized by packaging TG with cholesterol and the apolipoprotein B100, in liver cells. VLDLs are segregated into the blood stream (nascent VLDL). As in chylomicron metabolism, nascent VLDLs are converted into mature VLDL by exchanging apolipoprotein C-II (APOC2) and apolipoprotein E (APOE) with HDL. In plasma, mature VLDL particles are hydrolysed into free fatty acids and glyceride by lipoprotein lipase (LPL) with APOC2 as cofactor. Most of the VLDLs are converted to an intermediary lipoprotein, IDL particles. Some of the IDL are removed through interaction of apolipoprotein E with the LDL on the liver surface and the others are hydrolized by hepatic lipase to produce low density lipoproteins (LDL). LDLs are cholesterol-enriched lipoproteins. Circulating LDL are taken up by the liver and other tissues by endocytosis, process that involves the LDL receptor (LDLR). If LDL are oxidized, LDLs enters the peripheral cell such as macrophages via the scavenger receptor CD36 and SR-A(3). LDLs serve to deliver cholesterol to the tissues for energy utilization. Nascent HDLs are synthesized in the liver and intestine by the transfer of cholesterol and phos-

pholipids onto apolipoproteinA-1 (ApoA-1). Nascent HDLs interact with the peripheral cell and remove unesterified cholesterol through the ATP-binding cassette (ABC1). Cholesterol in nascent HDL then is esterified to fatty acid producing spherical mature HDL. Mature HDLs can transfer cholesterol ester (CE) to the liver by interaction with the scavenger receptor B-1(SRB1) and also transfer their CE-rich core to VLDL remnants creating LDL. HDL of small size can be degraded by the kidney.

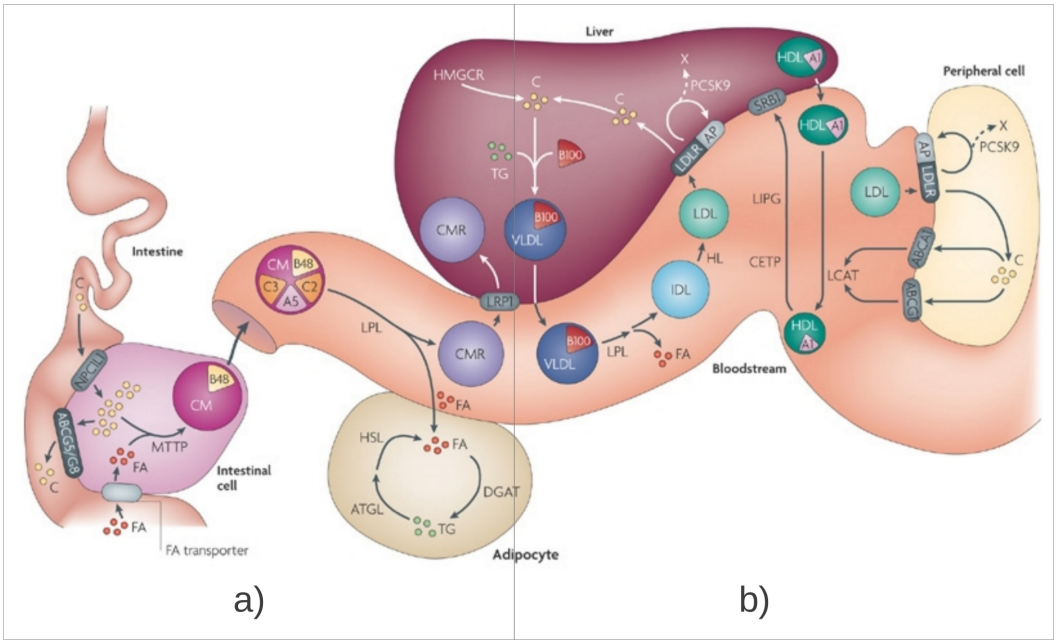


Figure 2.3: a) Exogenous Pathway; b) Endogenous Pathway. **FA**, fatty acids; **CM**, Chylomicrons; **MTTP**, microsomal TG-transfer protein; **B48**, apolipoprotein B isoform B48 **C3**, apoliprotein C3; **A5**, apoliprotein A5; **C2**, apoliprotein C2; **B28**, apolipoprotein B isoform B28; **LPL**, ipoprotein lipase; **CMR**, CM remnants; **TG**, triglyceride; **ATGL**, adipose TG lipase; **LRP1**; LDLR-related protein-1; **VLDL**, very low-density lipoprotein; **B100**, apolipoprotein B isoform B100; **C**, free and esterified cholesterol; **LDLR**, hepatic LDL receptor; **HL**, hepatic lipase; **IDL**, intermediate density lipoprotein; **LDL**, low-density lipoprotein; **LIPG**, endothelial lipase; **CETP**, cholesterol ester transfer protein; **HDL**, high-density lipoprotein; **LCAT**, Lecithin-cholesterol acyltransferase; **A1**, apolipoprotein A1. Figure adapted by permission from Nature Publishing Group (2009): *Nature Reviews Genetics*, Hegele, R. A. Plasma lipoproteins: genetic influences and clinical implications, copyright (2009)(58).

Glycolysis pathway

The metabolomic platform described allow the quantification of a number of metabolites related to the glycolysis pathway (glycerol, glucose, pyruvate and lactate). Glucose is the most important sugar in metabolism and the primary source of energy for the cell. It also acts as intermediate metabolite in many biochemical reactions. In glycolisis one molecule of glucose is converted into two molecules of pyruvate, releasing energy in the form of ATP (Figure 2.4). Pyruvate can be oxidized into acetyl-coenzyme A(acetyl-CoA), which is the main input for a series of reactions such as Krebs cycle. Apart from being the end product of glycolisis, pyruvate plays important roles in a range of metabolic processes. For instance, pyruvate is one of the precursors for the synthesis of glucose via gluconeogenesis pathway. It can be converted into fatty acids through acetyl-CoA or being used as building block of amino acids such as alanine. In absence or short supply of oxygen, pyruvate is broken down anaerobically, process that produces lactate. Glycerol, which is produced from triglycerides oxidation, can enter the glycolysis pathway in the form of D-glyceraldehyde-3-phosphate.

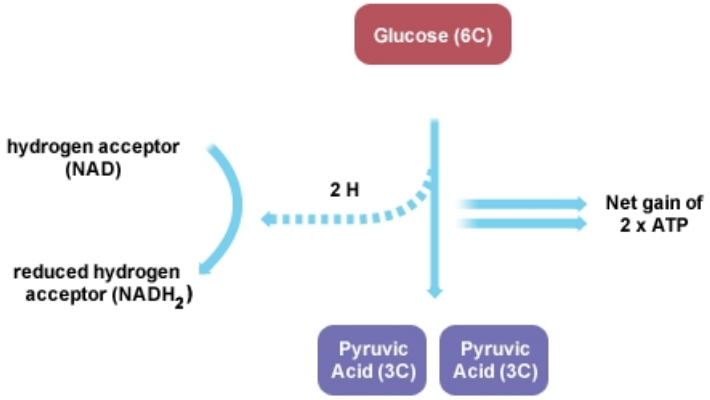


Figure 2.4: Diagram of the glycolysis pathway. Figure representing the conversion of one molecule of glucose into two molecules of pyruvate together with NADH and ATP. Picture taken from http://www.bbc.co.uk/bitesize/higher/biology/cell_biology/respiration/revision/2/

Citric acid cycle

Data on a number of metabolites involved in or related to the citric acid pathway such as 3-hydroxybutyrate, acetate, acetoacetate and citrate are identified and quantified by the NMR platform described in section 2.2.1. The citric acid cycle (also known as Krebs cycle) is the final step of the oxidation of fats, proteins and carbohydrates into $C0_2$, water and NADH. An overview of the citric acid cycle is shown in Figure 2.5. Citric acid cycle starts with the conversion of acetyl-CoA into citrate which goes through a series of chemical transformations. Citrate is converted into a number of citric acid cycle intermediates (e.g. succinate, malate) that can either be used as precursors for the biosynthesis of other molecules or used for the regeneration of a molecule of citrate. Citrate can also be re-transformed to acetyl-CoA in cytosol and be used for fatty acid synthesis. In situations where there is an excess of acetyl-CoA, this component can be converted into ketone bodies such as acetoacetate, 3-hydroxybutyrate and acetone in the liver. Two of these ketone bodies - acetoacetate and 3-hydroxybutyrate - are transported by the blood to extrahepatic tissues such as skeletal and heart muscle. In absence of other sources of energy such as glucose, tissues can obtain energy from the oxidation of acetoacetate and 3-hydroxybutyrate via the citric acid cycle. When the levels of glucose are low, the brain can use 3-hydroxybutyrate to obtain energy.

Amino acids

Amino acids play central roles both as structural units of proteins and as intermediate molecules in metabolism. Amino acids are classified into two categories, essential and non-essential amino acid, depending on whether they can be produced by the body or need to be supplied in the food. Data on four essential (histidine, leucine phenylalanine and valine) and three non-essential amino acids (alanine, glutamine and tyrosine) are available. Apart from their key function as protein building blocks, all essential and non-essential amino acid can be catabolized and used as energy source. In the catabolism of amino acids, the carbon skeletons of amino acids are broken down into metabolites

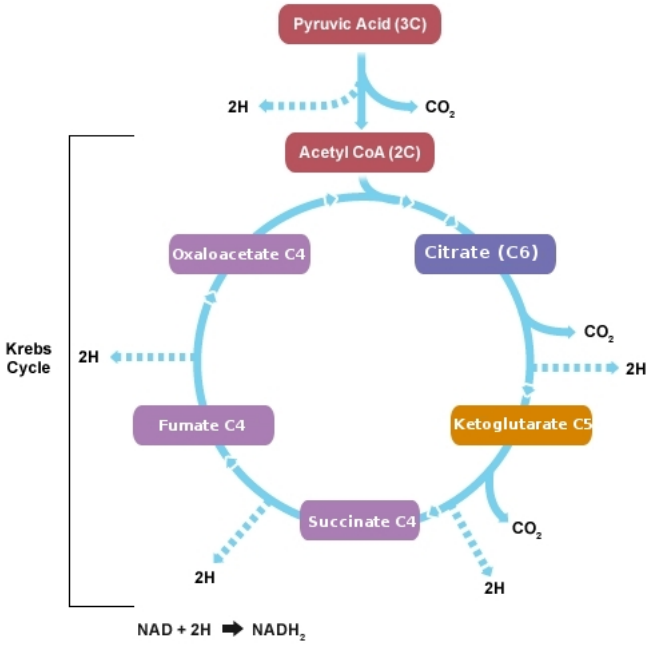


Figure 2.5: Diagram of the citric acid cycle. Picture taken from http://www.bbc.co.uk/bitesize/higher/biology/cell_biology/respiration/ revision/2/

that can either be oxidized into CO_2 and H_2O to generate ATP, or can be used in gluconeogenesis. Most of this process takes place in the liver, but three branched-chain amino acids (BCAA) - isoleucine, leucine and valine - are degraded in muscle or adipose tissues.

Urea cycle

The end products of amino acids' metabolism are ammonia, CO_2, water and energy. Ammonia is a very toxic compound to cells. To be excreted from the body, ammonia is converted into urea in the liver. Urea is transported in the blood to the kidneys, where it is excreted from the body in urine. Another product of metabolic degradation is creatinine, which is formed when creatine breaks down in muscle tissue. As urea, creatinine is excreted by the kidneys.

2.3 Genetic data

Genetic data from only NFBC1966 were used for the genetic analysis since genome-wide data for the other two populations were not available. The details for the genotyping procedure and quality control are presented in (47). In brief, all DNA samples were prepared and genotyped by the Broad Institute Biological Sample Repository (BSP). Genotyping was performed using the HumanCNV370-Duo DNA Analysis BeadChip. This array contains an informative set of tag' SNPs which was derived from the HapMap European-derived (CEU) sample (over 318,000 SNPs). In addition, 52,000 variants were included to target approximately 14,000 copy number variant (CNV) regions. This gives a total of over 370,000 genotyped markers. Genotyping was initially completed for a total of 339,629 SNPs and 4,936 samples.

2.3.1 Data quality controls

Extensive quality controls of the data were performed by Dr. Chiara Sabatti, Dr. Susan Service and colleagues at the University of California, Los Angeles (UCLA) and by Ms. Marika Kaakinen (University of Oulu) and Dr. Inga Prokopenko at the University of Oxford. The per-sample quality control process included analysis of duplicate samples, gender discrepancy, sample contamination and relatedness. Of the related pairs that shared over 20% of the genome identical by descent (IBD), the individuals with less complete genotype data were excluded. In total 173 subjects were excluded after the quality control analysis. Call rate in the final data set was $\sim 99.7\%$ on average (range 95.3-99.9%) and only 1.7% of individuals had call rate $\leq 97.5\%$.

Of the 339,629 genotyped genetic markers, 3,345 SNPs were excluded because the assumption of Hardy-Weinberg Equilibrium (HWE) did not hold ($P < 10^{-4}$). In addition, 55 SNP with call rate $< 95\%$ and 7,681 with minor allele frequency (MAF) $< 1\%$ were removed. The average SNP call rate was 99.7% and only for 1.6% of SNPs it was $95.0 - 97.5\%$. After quality control the data comprised 329,091 genotyped SNPs in 4,763 individuals.

The whole quality control procedure was not repeated in this study. How-

ever, we performed a minor allele frequency test at a 1% threshold, call rate test with threshold 95% and Hardy-Weinberg Equilibrium (HWE) test at threshold $p < 1 \times 10^{-4}$ which left N=318,443 SNPs for analysis. The software SNPtest v.2 was used to perform this analysis.

2.4 General Notation

Throughout this thesis, we will use the following general notation. Random variables are indicated with capital letters and values of the random variable (observations) are denoted in lower letters. In a univariate context, random variables are denoted in italic, non-bold. For instance, Y is a random variable and $y_1, ..., y_n$ are observations of Y. In a multivariate context, vectors and matrices are indicated by bold font and each components of vectors or matrices will be distinguished with subscripts. In this way, $\boldsymbol{Y}$ represents a vector of random variables, $\mathbf{y}$ a vector of observations and $\mathbf{Y}$ a matrix whose (i, j) entry is denoted $y_{i,j}$. Greek characters are used to denote unknown parameters (e.g β). The parameter estimates are represented by placing the symbol $\hat{}$ over the corresponding Greek character (e.g $\hat{\beta}$). We use prime to indicate a transpose. Thus, the transpose of a matrix $\mathbf{Y}$ is $\mathbf{Y}'$.

Chapter 3

Statistical methods for Genome-Wide analysis

In this chapter, we perform an initial exploratory analysis of genomic data in combination with metabolomic data. For that purpose, three statistical methods namely, single-SNP, multi-SNP and multiple responses are applied to investigate genetic variants associated with variations in metabolic phenotypes. The chapter is organised in the following way: In section 3.1 we present a GWAS analysis of NMR serum metabolites which is performed following the standard single-SNP tests of association. In section 3.2 we explore the use of multi-SNP methods in genome-wide association studies. In particular, a variable selection approach, the HyperLasso regression, is used to perform a multi-SNP analysis of metabolic traits. The last section of this chapter (section 3.3) is dedicated to multiple responses methods. Rather than treating each phenotype as a separate variable, in this section we perform the combined analysis of different sets of related metabolic phenotypes. We conclude this chapter by discussing the results (section 3.4) and performance of each of the investigated methods.

3.1 Single-SNP analysis

Due to their fast implementation and easy interpretation, univariate methods play a very important role in the analysis of genetic data. Based on this approach, association between genetic variants and a trait of interest is tested considering only one SNP at a time. The methods applied to test for

association are statistically straightforward - the testing is performed using well established techniques such as univariate regression and the χ^2 test.

Depending on whether the trait is continuous or binary (case-control), different tests can be used for assessing association between a SNP and a phenotype. The most commonly used statistical techniques for testing for association in GWAS are regression models. Linear regression is the method of choice for testing for association with continuous phenotypes. The linear model assumes a linear relationship between genetic factors and the mean value of the trait of interest.

$$y_i = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij} \qquad \epsilon_{ij} \in N(0, \sigma_j^2), \quad i = 1, ...N, \tag{3.1}$$

where y_i is the phenotype value for the ith observation, β_{0j} is the intercept, x_{ij} is the genotype for the jth SNP for the ith observation, β_{1j} is the regression coefficient and ϵ_{ij} is the prediction error. The underlying assumption in linear regression is that the trait distribution is approximately normal, with a common variance and the error terms are normally distributed. If the assumption of data normality does not hold, data distribution can be approximate to normality by transforming the original phenotype values (e.g. log transformation).

The linear model outlined above for continuous phenotypes can not be used for case-control phenotypes because the assumption of data normality does not hold. This problem is overcome with the application of logistic regression in which the transformation logit $(\pi) = log(\frac{\pi}{1-\pi})$ is applied to π, the probability of being a case. The logistic regression assumes that the log-odds of the phenotype is a linear function of the jth genotypes.

$$\log \frac{Pr(y_i = 1|x_{ij})}{Pr(y_i = 0|x_{ij})} = \beta_{0j} + \beta_{1j}x_{ij}, \qquad i = 1, ...N. \tag{3.2}$$

Regression models both in continuous and case-control traits have several advantages compared to other statistical tools for GWAS. One of these advantages is the possibility of easy incorporation of confounders such age, sex into the model. Regression models are also flexible tools that can be used to accommodate multiple predictors simultaneously.

3.1.1 Single-SNP association analysis

Here we present a GWAS analysis of NMR-based metabolic measures where the conventional approach of single-SNP test of association is applied. The aim is to investigate the performance of the per-SNP approach at identifying genetic variants significantly associated with metabolic traits.

- **Materials and Methods**

The analysis was performed using fifty one measures of serum metabolites, including 16 lipoprotein subclasses, 18 lipid-related traits, 13 low-molecular-weight metabolites and 4 variables derived directly from the previous measures (Table A.1). Genotype and serum metabolic data were available on 3,683 individuals (5,747 individuals with serum metabolic data and 4,772 with genotype data) from the NFBC1966.

Test of association was performed by univariate linear regression models assuming an additive contribution to the trait value. To correct for sex-differences in metabolic profiles (59), each of the fifty one metabolic traits was analysed including gender as covariate. In order to reduce skewness of the data, each phenotype was log-transformed prior to the analysis. The threshold for genome-wide statistical significance was set to $P < 9.8 \times 10^{-9}$ which corresponds to a standard genome-wide statistical significance level of $P < 5 \times 10^{-8}$ corrected for the total number of tested metabolic phenotypes ($m = 51$).

- **Results**

On the basis of genome-wide association significance ($P < 9.8 \times 10^{-9}$), we identify thirty eight genetic variants that show significant association with a total of twenty metabolic measures. The results for this analysis are presented in Table 3.1 where we provide a brief description of the identified genetic variants, including their accession numbers, chromosomal positions, the p-values obtained by the most associated metabolic trait (leading trait), other traits associated with the identified genetic makers and the candidate

genes associated with the genetic markers. The Manhattan plots for the leading traits are presented in appendix (Figure B.1).

The identified SNPs are localized in ten different genetic regions across eight chromosomes. Of these loci, three show association with low-molecular weight metabolites and seven loci are associated with measures of lipoprotein subclasses and lipids measures. The most significant association is found between total amount of polyunsaturated fatty acids (TotPUPA) and SNP rs174537 localized at chromosome 11q21.3. Three of the identified loci have not been previously observed to associate with metabolic measures at genome-wide significance. The loci that reached genome-wide significance are presented and discussed in more detail below.

Loci associated with lipids or lipoprotein subclasses

For lipoprotein or lipid measures, we identify a total of thirty one genetic variants localized in seven genetic regions. All loci found through our analysis have been previously reported to be associated with metabolic traits (41, 42). We observe association between SNPs localized at chromosome 2p24.1 and total lipids in very small, small, medium and large VLDL, total lipids in IDL and measure of apolipoprotein B (ApoB). These SNPs are localized within or near a gene encoding for apolipoprotein B *(APOB)*, which is the main apolipoprotein constituent of chylomicrons, VLDL (isoform B-48) and LDL (isoform B-100) (58). Variations in this gene are known to relate to metabolic disorders such as hyperbetalipidemia (60, 61) or hypercholesterolemia (62, 63). SNP rs1260326 localized at chromosome 2p23.3 shows significant association with total lipids in medium, large VLDL as well as measures of monounsaturated fatty acids *(MUFA)*. This variant is localized within a gene encoding for glucokinase regulatory protein *(GCKR)*, a hepatic regulatory protein essential in the metabolism of glucose (64). This protein plays a critical role in controlling the activity of glucokinase (GK) by binding and inhibiting the GK complex (64). This polymorphism (rs1260326) has been related to increased levels of fasting triglycerides (65, 66, 67), which is known to be associated with increased risk of cardiovascular disease devel-

opment. Interestingly, this genetic variant has also been associated with decreased levels of fasting glucose (68), which is related to the reduction in the risk of diabetes and cardiovascular diseases. A variant (rs10096633) localized at 8p21.3 shows significant association with total lipids in medium VLDL. This variant is localized in a gene encoding for lipoprotein lipase *(LPL)*. This enzyme catalyses the hydrolysis of triglycerides in TG-rich lipoproteins such as chylomicrons and very low density lipoproteins (69, 70). It is also known that LPL aids the cellular uptake of chylomicron remnant, cholesterol-rich lipoproteins, and free fatty acids (71, 72). Variants at chromosome 11q21.3 show significant association with several lipid-related traits, namely, total amount of polyunsaturated fatty acids (ToTPUFA), linoleic acid (LA) and ratio of triglycerides to phosphoglycerides (TGtoPG). The identified SNPs are localized in or near genes encoding fatty acid desaturase cluster *(FADS)*, a well-known lipoprotein-related genetic region. This cluster of proteins is component of lipid metabolic pathway and catalyzes the biosynthesis of highly unsaturated fatty acids (73). Association between variants at chromosome 15q21.3 and total lipids in very large and large HDL as well as with measures of phosphatidylcholine and other cholines (PC) are observed. These genetic variants are localized in a gene encoding for hepatic triglyceride lipase *(LIPC)*, which is an important enzyme in HDL metabolism determining the size and density of particles (74). Two genetic variants localized in chromosome 16q13 show significant association with total lipids in medium, large, high density liprotein (HDL) and apoprotein A-I (ApoA-1). These SNPs are localized in a gene encoding for cholesteryl ester transfer protein *(CETL)*, also known as plasma lipid transfer protein. The encoded protein is essential for the transport of lipids, mainly cholesterol ester and triglycerides, between lipoproteins (75). Decreased production and/or activity of CETP has associated with risk of atherosclerosis (76) and CVD (77). A genetic variant localized at chromosome 20q13.12, in or near a gene encoding for phospholipid transfer protein *(PLTP)* is found to be associated with both large and small HDL. The product of this gene is known to be involved in the transfer of phospholipids from triglycerides-rich lipoproteins (VLDL) to high density lipoproteins (HDL) and plays a role in regulating the size of HDL particles

(78).

Loci associated with amino acids and other small molecules

We have identified three loci associated with at least one of the measures of metabolites involved in amino acid metabolism and other low molecular weight metabolites. None of these loci have been previously reported for serum metabolic traits. A variant localized at chromosome 8q12.1 is associated with serum levels of creatinine. This genetic variant is localized near a gene encoding for a member of the inositol monophosphatase family *(IMPAD1)*. This enzyme catalyses the hydrolysis of phosphoadenosine phosphate (PAP) to adenosine monophosphate *(AMP)* (79). Another variant found to be significantly associated with variations in serum levels of creatinine is SNP rs7217872 at chromosome 17q13.3. This genetic variant is localized in gene encoding for rabphilin 3A-like *(RPH3AL)*, which plays an important role in regulating calcium-ion-dependent exocytosis in both endocrine and exocrine cells (80). A number of loci at chromosome 22 show association with serum levels citrate. These genetic variants are localized in a gene encoding for clathrin, heavy chain-like 1 *(CLTCL1)*, a important enzyme in the formation of coated vesicles.

Table 3.1: Summary of the results from the single-SNP association analysis of serum metabolic traits adjusted for sex.

Gene	Marker	Ch	Pos	A1/A2	Leading	P	Other
APOB	rs3923037	2	21069902	A/C	S-VLDL-L	5.8E-10	XS-VLDL-L
APOB	rs6728178	2	21105581	A/G	XS-VLDL-L	1.4E-12	ApoB, M-VLDL-L, S-VLDL-L
APOB	rs6754295	2	21117835	G/T	XS-VLDL-L	1.3E-12	ApoB, M-VLDL-L, S-VLDL-L
APOB	rs676210	2	21143176	A/G	S-VLDL-L	9.3E-12	ApoB, M-VLDL-L, XS-VLDL-L
APOB	rs693	2	21143847	A/G	ApoB	2.4E-9	IDL-L, XS-VLDL-L
APOB	rs673548	2	21149196	A/G	S-VLDL-L	7.4E-12	ApoB, M-VLDL-L, XS-VLDL-L
GCKR	rs1260326	2	27642591	C/T	M-VLDL-L	4.9E-10	MUFA, L-VLDL-L
LPL	rs10096633	8	19875201	C/T	M-VLDL-L	7.7E-11	
IMPAD1	rs16921974	8	58362727	C/T	Crea	1.6E-19	
FADS1-3	rs509360	11	61305135	A/G	TotPUFA	7.5E-19	FAw3-FA, TG-PG
FADS1-3	rs174537	11	61309256	G/T	TotPUFA	9.9E-53	FAw3-FA, LA, TG-PG
FADS1-3	rs102275	11	61314379	C/T	TotPUFA	1.3E-52	FAw3-FA, LA, TG-PG
FADS1-3	rs174546	11	61326406	C/T	TotPUFA	1.5E-52	FAw3-FA, LA, TG-PG
FADS1-3	rs174556	11	61337211	C/T	TotPUFA	1.0E-51	FAw3-FA, LA, TG-PG
FADS1-3	rs174570	11	61353788	C/T	TotPUFA	6.4E-34	FAw3-FA, TG-PG
FADS1-3	rs1535	11	61354548	A/G	TotPUFA	2.6E-52	FAw3-FA, LA, TG-PG
FADS1	rs2072114	11	61361791	A/G	TotPUFA	5.9E-45	FAw3-FA, LA, TG-PG
FADS2	rs2851682	11	61372588	A/G	TotPUFA	2.7E-33	FAw3-FA, TG-PG
FADS2	rs174611	11	61384457	C/T	TotPUFA	2.3E-9	
FADS3	rs174449	11	61396955	A/G	TotPUFA	1.5E-27	FAw3-FA, TG-PG
FADS3	rs174450	11	61398118	G/T	TotPUFA	8.3E-29	FAw3-FA, TG-PG
FADS3	rs1000778	11	61411881	A/G	TotPUFA	3.3E-17	TG-PG
RAB3IL1	rs174478	11	61435152	G/T	TotPUFA	1.1E-22	
LIPC	rs166358	15	56468097	A/G	L-HDL-L	1.6E-13	XL-HDL-L
LIPC	rs1532085	15	56470658	A/G	L-HDL-L	7.0E-25	PC, XL-HDL-L, XS-VLDL-L
LIPC	rs415799	15	56478046	A/G	L-HDL-L	2.3E-14	L-HDL-L, XL-HDL-L
LIPC	rs16940213	15	56482629	C/T	L-HDL-L	1.6E-11	
LIPC	rs473224	15	56524633	G/T	L-HDL-L	1.7E-11	
LIPC	rs261336	15	56529710	A/G	XL-HDL-L	2.0E-12	
CETP	rs3764261	16	55550825	A/C	L-HDL-L	9.3E-18	ApoA1, M-HDL-L
CETP	rs1532624	16	55562980	A/C	ApoA1	3.8E-12	L-HDL-L
RPH3AL	rs7217872	17	88988	C/T	Crea	3.5E-10	
PLTP	rs6065906	20	43987422	C/T	XL-HDL-L	3.2E-11	S-HDL-L
CLTCL1	rs712964	22	17530671	C/T	Cit	1.8E-11	
CLTCL1	rs1061325	22	17558649	C/T	Cit	9.7E-10	
CLTCL1	rs2073740	22	17570666	A/C	Cit	2.6E-9	
CLTCL1	rs885978	22	17572780	A/G	Cit	1.8E-9	
CLTCL1	rs1340958	22	17627886	A/G	Cit	1.2E-9	

Gene, the plausible candidate gene in the locus or the nearest annotated gene to the lead SNP; **Marker**, SNPs with $P < 9.8 \times 10^{-9}$ (standard genome-wide statistical significance level of $P < 5 \times 10^{-8}$ corrected for the total number of tested metabolic phenotypes; $m = 51$); **Ch**, Chromosome; **Pos**, SNPs position at the corresponding chromosome; **A**, listed are 1^{st} major allele, 2^{nd} minor allele; **Leading**, The serum metabolites with best p-value among the traits; **P**, p-value for the leading traits; **Other** Additional traits with $P < 9.8 \times 10^{-9}$.

3.2 Multi-SNP analysis

Analysing one SNP at a time can be a valid approach if a single genetic variant or a number of unrelated variants are responsible for variations in a phenotype. However, for most complex traits - which are known to be controlled by a combination of multiple genes potentially interacting between them - the per-SNP approach is seldom able to provide accurate estimates of genetic effects to complex traits. An alternative class of methods for analysing SNP data are multivariate models. This class of methods takes into account all factors potentially related to the outcome simultaneously. However, when the number of predictors is greater than the number of observations $p >> n$ - which is the typical situation in GWAS studies - the joint analysis of all genetic data becomes infeasible using Ordinary Least Squares methods. Penalized regression methods are an attractive class of techniques for variable selection and has been effectively applied to deal with the high-dimensional data.

In this section HyperLasso (Hlasso) regression (81) - a generalization of the Lasso approach - is applied as a variable selection method in a multi-SNP association analysis. With this method the parameters' estimation is performed while applying a penalty function, the Normal-Exponential Gamma probability density (NEG), that causes shrinkage of the regression coefficients towards zero. This is equivalent to find the posterior mode in a Bayesian regression framework with the NEG prior. The mode at zero for some of the coefficients in the prior reflects the belief that the variable's effect has a high probability of being near zero. Posterior mode with non-zero values are indicative of markers-trait association. Because only the posterior mode is found, the algorithm is Bayesian-inspired rather than fully Bayesian. With this method, we attempt to find the subset of SNPs that best explain the phenotype variation for a specified type I error.

3.2.1 Lasso

The Least absolute shrinkage and selection operator (Lasso), originally proposed by Tibshirani (82) is a widely used shrinkage method in regression analysis for large models. The Lasso procedure performs regularized parameter estimation and variable selection at the same time. It minimizes the residual sum of squares by the addition of a L_1 norm penalty, namely $\sum_{j=1}^{p} \mid \beta_j \mid \leq \lambda$, corresponding to the sum of absolute values of the regression coefficients. This selective shrinkage favours sparse solutions and allows the handling of extremely large datasets.

Suppose we have data $(\mathbf{x}_i, y_i)$, where $i = 1, 2, ..., N$ is the total number of observations, y_i is the dependent variable and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ is a $1 \times p$ vector of explanatory variables. Assume that the observations $y_i s$ are conditionally independent given the x_{ij} and that the data x_{ij} are standardized so that the mean is 0 and the standard deviation is 1. The Lasso estimate of the coefficients $\hat{\boldsymbol{\beta}}^{Lasso} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ can be achieved by:

$$\boldsymbol{\beta}^{Lasso} = \text{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to } \sum_j \mid \beta_j \mid \leq \lambda \, ,$$

where $\boldsymbol{\beta}^{Lasso}$ is the set of shrunk coefficient estimates that minimizes the residual sum of squares subject to constraint. The parameter $\lambda \geq 0$ is the penalty factor and plays an important role controlling the degree of shrinkage. Tibshirani (82) observed that estimating the parameters using the Lasso approach is equivalent to a Bayesian maximun a posteriori (MAP) estimate using the double exponential prior (DE) for all the β_j with mode at zero.

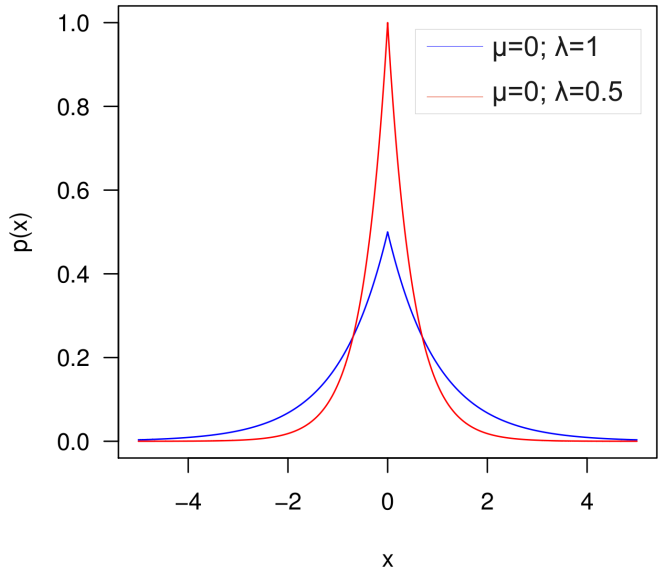
$$p(\beta_j \mid \lambda_j) = (\frac{\lambda}{2}) \exp(-\lambda \mid \beta_j \mid)$$

Assuming that $\lambda_j = \lambda \forall j$, the prior for β can be expressed as the product of p independent DE densities

$$p(\beta \mid \lambda_j) = \prod_{j=1}^{p} (\frac{\lambda}{2}) \exp(-\lambda \mid \beta_j \mid)$$

Figure 3.1 shows the density of the DE distribution for penalty factor λ equal to 0.5 and 1. As the value of λ increases more mass is placed around zero and less in the tails. With large values of λ the probability of the coefficients to be retained by the model decreases. In a Bayesian framework, large values of the penalty factor λ reflect the prior belief that the effects of variables will be near zero. Small values of λ encode no prior assumption about the effect size.

Figure 3.1: The DE density distribution for penalty factors λ equal to 0.5 and 1.



Although the Lasso method has been successfully applied to a range of studies, the Laplace implemented in the Lasso has the limitation of over-shrinking informative variables. The light tails away from zero correspond to a prior belief that non-zero effects are small(81). In this way non-zero parameters tend to be severely shrunk towards zero, producing estimates

that are either small or zero. Ideally the model should retain parameters that are informative and thus, important for the model.

3.2.2 HyperLasso

To overcome the limitation of over-shrinking, Hoggart et al. (2008) (81) proposed the use of a generalization of the double exponential distribution, the normal exponential gamma (NEG) which is sharper at zero and heavier at the tails. Parameters with non-zero effect size will be less shrunk once included in the model.

The normal exponential gamma can be represented as:

$$NEG(\beta \mid \lambda, \gamma) = \int_0^\infty \int_0^\infty N(\beta \mid 0, \sigma^2)\, Ga(\sigma^2|1, \psi)\, Ga(\psi^2|\lambda, \gamma^2)\, d\sigma^2\, d\psi$$

$$= k\, exp\, \frac{\beta^2}{4\gamma^2}\, D_{-2\lambda-1}\, \frac{\mid \beta \mid}{\gamma}.$$

where λ and γ are the shape and the scale parameters, respectively, D is the parabolic cylinder function and k is the normalizing constant. Algorithms for computing the parabolic function and its derivatives are described in Zhang and Jin (83). Figure 3.2 shows three NEG density functions with λ values equal to 0.05, 1 and 10. The shape and scale are the parameters that control the intensity of the peak and the gradient of the tails. When the shape is small the NEG density function presents a sharp spike at zero and very high gradients near zero, but flat tails away from zero. The peak around zero - as with the DE - corresponds to the prior belief that there are few true causal variables and severe penalty on any non-zero coefficient. The NEG has flatter tails than the DE which results in less severe shrinkage of non-zero parameter towards zero. For large values of λ the NEG is peaked at around zero and converges to the DE distribution.

Hoggart et al. (81) implemented a CLG algorithm (84) (85) to obtain MAP estimates. The algorithm is run several times with the order of the SNPs permuted at each run. If a set of SNPs carries the same information, the

Figure 3.2: Logarithms of NEG and DE densities.

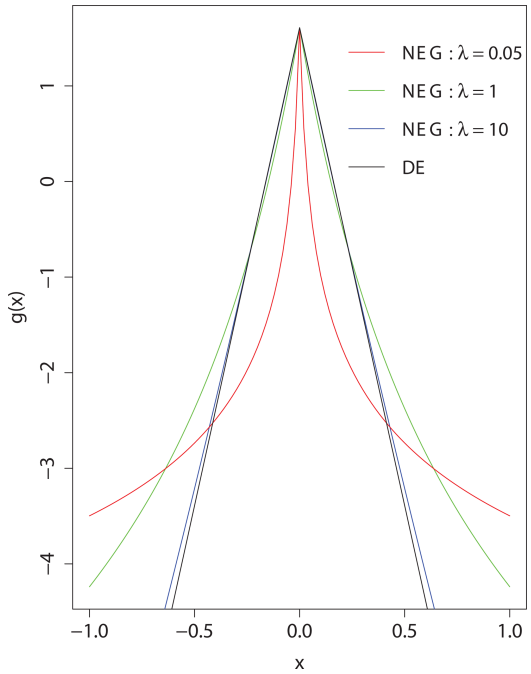


Figure taken with permission from PLoS genetics, 4(7), Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies (81).

algorithm chooses only one of the markers, reducing in this way redundancy.

3.2.3 The HyperLasso regression analysis

Here we present a GWAS analysis of NMR-based metabolites where the HyperLasso regression is applied. With this method, we attempt to find the subset of SNPs that best explains the phenotype variation for a specified type I error.

- **Materials and Methods**

For this analysis we use the same set of metabolic traits and genetic markers that were investigated with the single-SNP approach. Prior to analysis metabolic data were log-transformed to reduce the skewness of the data and

the model was adjusted for sex differences in serum metabolites. As for the per-SNP analysis, we assumed an additive genetic contribution to the phenotype value.

The software runHLasso (86) was used to perform the HyperLasso regression analysis. Although, the case-control phenotype is the default design, runHLasso can also accommodate continuous phenotypes using the linear regression assumption of Gaussian residuals. We considered the NEG distributions as prior distributions for the effect-size parameter. The genotype data were standardized prior to analysis to have mean zero and standard deviation equal to one. As previously described, the NEG probability density function has two parameters, a scale parameter and a shape parameter. In a simulation study, Hoggart et al. (81) showed that shape parameter λ of 0.05 was the most suitable value to select the causal variants given an appropriate Type I error. Large values of λ result in several tightly-linked SNPs being brought to into the model whereas smaller values of λ gave rise to computational problems. Stephens and Balding (2009) (87) suggested that the NEG distributions with lambda values between 1 and 2 were most suitable to identify non-zero effect variants, since small values of λ tend to give more weight to SNPs with large effect sizes. Therefore, the input parameters used to run Hlasso were defined as follows: 1) The NEG shape parameter (λ) was set to 0.05, 1 and 2; 2) the per-SNP type-I error rate (α) was set at 10^{-5}; 3) the number of mode searches was set to 10.

- **Results**

The results for the HyperLasso regression are based on the higher posterior mode found from 10 permutations of the search order using the optimization algorithm described in Hoggart et al. (81). In Table B.1 we present the best-fitting model found for each of the NMR serum metabolites for type I error $\alpha = 10^{-5}$ and shape parameter λ of 1. These results are very similar to those using λ of 0.05 and 2. From a total of 318,443 SNPs, 381 were identified by the HyperLasso regression as associated with at least one of 51 serum metabolites. The number of SNPs included in the best-fitting models

largely vary between metabolites. For instance, 38 genetic variants were returned by the HyperLasso regression as the subset of SNPs that best explain variations in serum levels of glycerol for the specified type I error whereas only two SNPs were returned for serum levels of leucine.

Of the set of SNPs retained by HyperLasso regression, 19 were also identified by the single-SNP test and all of them have been previously associated with metabolic traits (41, 42). These loci correspond to genetic variants localized in or near genes encoding for proteins involved in the metabolism of serum metabolites such as apolipoprotein B-100 *(APOB)*, glucokinase regulatory protein *(GCKR)*, lipoprotein lipase *(LPL)*, fatty acid desaturase cluster *(FADS1,FADS2,FADS3)*, hepatic lipase *(LIPC)*, cholesteryl ester transfer protein *(CETP)*, phospholipid transfer protein *(PLTP)*, clathrin, heavy chain-like 1 *(CLTCL1)*. Interestingly, we observe that SNP rs1260326, which was previously identified only for total lipids in VLDL subclasses, has been retained by HyperLasso regression for a larger number of metabolites including amino acids such as alanine and isoleucine, pyruvate and total triglycerides (TGtoPG). As already mention, GCKR plays an important role in the metabolism of triglycerides and glucose(65, 66, 67, 68). Recently, significant association has been found between SNP rs1260326 and alanine as well as VLDL and total triglycerides (42). Levels of isoleucine have been shown to be associated with the risk for developing type 2 diabetes (34). However, to the best of our knowledge association between this genetic variant and isoleucine, pyruvate has not been previously reported.

Among the set of genetic variants returned by HyperLasso regression but not identified by single-SNP tests, the following SNPs are in or near genes encoding proteins known to be associated with cardiometabolic traits. For total lipids in IDL as well as in small, medium and large LDL, we identify a genetic marker (rs646776) at 1p31. This genetic variant is localized in a gene encoding for cadherin EGF LAG seven-pass G-type receptor 2 *(CELSR2)*. Previous studies have reported a strong association between variants at CELSR2 and levels of total cholesterol, suggesting a potential implication of this gene in cholesterol metabolism and coronary risk (88, 89). Another SNP (rs2056387) also localized at 1q12 was identified for concen-

tration levels of FAw3toFA. This SNP is localized in a gene encoding for ryanodine receptor 2 *(RYR2)*, a protein found in cardiac muscle sarcoplasmic reticulum. The encoded protein is one of the components of a calcium channel and has been found that variations in RYR2 gene are among the causes of familial arrhythmogenic right ventricular dysplasia type 2 (90). A SNP (rs2943641) at 2q36.3 was identified for isoleucine. This marker is localized near a gene encoding for insulin receptor substrate 1 *(IRS1)*. Variations in the gene encoding insulin receptor substrate (IRS)-1 have been reported to be associated with the risk of type 2 diabetes and susceptibility to insulin resistance (91, 92). Additionally, it has been reported that increased concentration of isoleucine along with another four amino acids (valine, phenylalanine, tyrosine and leucine) are associated with the risk type 2 diabetes development (34). However, to our knowledge, association between genetic variants at this gene and isoleucine has not been previously reported.

At 4q32.1, a genetic variant (rs1328925) localized in a gene encoding for transmembrane protein 144 *(TMEM144)* was identified for total lipids in small VLDL. Association between this genetic marker and blood pressure and arterial stiffness has been reported in a previous genome-wide association study (93). A SNP (rs2000571) at 11q23.3 was identified for total lipids in small VLDL which encodes for apolipoprotein A-I *(APOA1)*; apolipoprotein C-III *(APOC3)*; apolipoprotein C-IV *(APOC4)* and apolipoprotein *A-V (APOA5)*. This set of apolipoproteins is the main class of apolipoprotein in HDL and to a lesser extent with VLDL. They are known to be important determinant of plasma triglyceride levels. Furthermore, genetic variants of apolipoprotein *A-V* have been associated with abnormal VLDL-TG metabolism (94). At chromosome 19, the marker rs157580 was associated with total lipids in very small VLDL, IDL and large LDL as well as ApoB. Candidate genes for this variant are: mitochondrial import receptor subunit TOM40 homolog *(TOMM40)* and apolipoprotein E cluster *(APOE)*. Apolipoprotein E is a key protein constituents for cholymicrons, VLDL, IDL and HDL and plays a critical role in lipoprotein metabolism and cardiovascular disorders (95). A SNP (rs1800961) at chromosome 20q13.12 was associated with total lipids in large HDL. This SNP is localized in a gene

encoding for hepatocyte nuclear factor 4, alpha *(HNF4A)*. Genetic variant in HNF4A gene have been associated with total levels of cholesterol in HDL (96).

HyperLasso regression analysis and LD

A typical situation is GWAS studies is that multiple tightly-linked SNPs are identified for the same trait. Such a situation can happen because there are several highly-correlated true causal variants or because there is only one causal variants and all the others identified SNPs show association only through their tight-linkage to the causal SNP. While the selection of highly correlated SNPs can be biologically informative, most of tightly-linked SNPs will provide redundant information about the genetic regions associated with the phenotype of interest. A key feature of the HyperLasso regression is that in presence of high LD between genetic variants, typically only one variant will be chosen to be included in the model, reducing information redundancy.

In order to explore this property of HyperLasso regression, we calculated the correlation structure (LD) among the selected genetic variants included in the best-fitting model and compared them to the results from the single-SNP approach. We simplify this analysis by only looking the ten leading traits (see Table 3.1). The correlation structure for the set of SNPs identified in the single-SNP analysis are presented in Figure B.2 and for HyperLasso regression in Figure B.3. We observe that SNPs selected by HyperLasso regression are in general in low LD ($|r| < 0.5$) (see Figure B.3). Among the SNPs identified for serum levels of apolipoprotein B (APOB), only two SNPs (rs639 and rs6728178 at chromosome 2) were in hig LD ($|r| > 0.8$). In a similar way, we observe that of the thirty four genetic variants included in the best model for total levels of PUPA, only one pair of SNPs (rs10808546 and rs6982636) is in high LD ($|r| > 0.9$). On the contrary, the majority of genetic markers identified by the single-SNP are in high LD among them (Figure B.2).

Based on these results, we can conclude that LD is more effectively handled by HyperLasso. This analysis shows that when a genetic region which is tagged by multiple SNPs in high LD - is found to be associated with a

given phenotype, the Hlasso generally selects just one SNP from this region. Unless the set of SNPs are in perfect LD, the selected genetic variant will typically be the one showing the strongest signal. This feature of the Hlasso has two main benefits: 1) while HLasso might identify the same genetic regions related to the phenotype as the single-SNP analysis, it will return a less redundant set of variants; 2) by selecting a subset of low-LD and strong-signals SNPs, the HyperLasso approach might offer better results in localizing the true causal SNPs.

3.3 Multiple response analysis

Tests of association in GWAS studies are generally conducted one-phenotype at a time, without taking into consideration the correlation structure between phenotypes. However, many complex disorders like diabetes, hypertension or obesity are typically studied based on a cluster of risk factors which are themselves intercorrelated. Therefore, in situations where multiple phenotypes are related by their risks to a given disorder, the joint analysis of all phenotypes can provide a more powerful approach to detect genetic associations (97, 98). Several approaches have been proposed for performing multivariate association tests on GWAS studies (99, 100, 101, 102). Yet, despite methods for multivariate association analyses are available, this approach is still rarely used in GWAS of correlated complex phenotypes.

In this section, we investigate the application of multiple responses methods to jointly analysis different sets of correlated phenotypes. To this aim, we consider the most general case of multiple response models, the Seemingly Unrelated Regression (SUR). The SUR-based analysis is performed for the genetic variants identified for at least one of the fifty one NMR serum metabolites in the single-SNP analysis. The results of the SUR analysis are compared and discussed with those obtained using single-SNP association analysis.

3.3.1 SUR model

Multivariate association analysis for correlated qualitative phenotypes is performed using the Seemingly Unrelated Regression (SUR)(103). SUR is an extension of Ordinary Least Squares estimation which takes into account the residual correlation structure across the univariate linear regressions of each phenotype. The generic expression for a system of SUR applied to genotype data can be illustrated as follows.

Lets N be the total number of unrelated subjects. For each individual observation i, there are P dependent variables $y_{i1}, .., y_{ij}, ..., y_{iP}$. From the same individuals, L SNPs are genotyped. We indicate with $\mathbf{X}_j$ the $N \times (K_j + 1)$ design matrix, with K_j represents the number of terms in the model for phenotype Y_j excluding the intercept, so that $Y_j = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$ with $\boldsymbol{\beta}_j = (\beta_{0j}, ..., \beta_{kj})'$ a $(K_j + 1) \times 1$ vector of coefficients and $\boldsymbol{\epsilon}_j = (\epsilon_{1j}, \epsilon_{2j}, ..., \epsilon_{jN})'$ a $N \times 1$ vector of zero-mean random error terms. Under the SUR model, the system of P equations can be written as (103):

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_P \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_P \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_P \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_P \end{bmatrix} \tag{3.3}$$

The SUR model can be also presented in a more compact way as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.4}$$

with $\mathbf{Y} = (Y_1', Y_2', ..., Y_P')'$, $\boldsymbol{\beta} = (\boldsymbol{\beta'}_1, \boldsymbol{\beta'}_2, ..., \boldsymbol{\beta'}_P)'$,$\boldsymbol{\epsilon} = (\boldsymbol{\epsilon'}_1, \boldsymbol{\epsilon'}_2, ..., \boldsymbol{\epsilon'}_P)'$, and $\mathbf{X}$ a $J \times K$ block-diagonal matrix, $J = N \times P$, $K = \sum_j K_j$. The residual errors $\boldsymbol{\epsilon}$ are assumed to be normally distributed with mean 0 and with covariance matrix given by:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix with size $N \times N$ and $\boldsymbol{\Sigma}_\epsilon$ a $P \times P$ matrix with

generic entry $\sigma_{jj'} = E(\boldsymbol{\epsilon}_{ij}\boldsymbol{\epsilon}_{ij'})$ for $j = 1, ..., P, i = 1, ..., N$. The SUR model allows different genetic effects for different correlated traits as well as the possibility that some loci might be associated with some of the phenotypes but not all of them.

3.3.2 The SUR analysis

The list of SNPs identified by the single-SNP analysis are re-analysed following a multiple response approach. The SUR analysis results are compared and discussed with those obtained using the single-SNP association approach.

- **Materials and Methods**

From our previous analyses, we observed that some of the metabolic measures belonging to the same metabolic class or involved in the same molecular pathway showed association with a similar set of genetic variants. For instance, we identified a set of genetic variants localized at chromosome 2 which were mainly associated with total lipids in VLDL subclasses. In a similar way, genetic markers at chromosomes 15, 16, 20 showed stronger association with total lipids in HDL subclasses than with any other class of lipoproteins. Therefore, in the SUR-based analysis, the previously identified genetic markers are re-analysed according to the metabolic group to which they were significantly associated.

We initially constructed four systems of linear equations namely 'VLDL subclasses', 'HDL subclasses', 'metabolites involved in the citric acid cycle' and 'metabolites involved in the urea cycle'. They represent the four classes of serum metabolites that showed significant association with genetic variants in the single-SNP analysis. In the SUR-based analysis, all metabolic traits included in the same systems of equation are tested for association with the same set of SNPs. For instance, from the single-SNP association tests we identified eight SNPs significantly associated with at least one of the six VLDL-related measures. Therefore, in the SUR analysis, we investigate the association between these eight genetic markers and all VLDL-related variables.

The package **systemfit** in R environment was used to test for association between the systems of linear equations and genetic variants. For the SUR analysis, association between genetic variants and the metabolic traits are determined using the significance threshold of $P < 0.05/\mathrm{K}$ where K is equal to the number genetic variants tested in each system of equations.

- **Results**

The results for the SUR analysis are showed in Table B.2-B.5. For the VLDL system (Table B.2), eight loci were tested for association. Five of these SNPs, (rs3923037, rs6728178, rs6754295, rs676210, rs673548) do not show significant association with any metabolic traits under the combined SUR-analysis. Two SNPs (rs1260326 and rs10096633) are found to be significantly associated with all the VLDL subclasses included in the analysis, with the exception of total lipids in very small VLDL. As indicated before, the SNP rs1260326 is located in a gene encoding glucokinase regulatory protein *(GCKR)* and rs10096633 is located in a gene encoding lipoprotein lipase *(LPL)*. Interestingly, a strong association is observed between SNP rs1532085 and total lipids in very small VLDL but not with all the other VLDL subclasses. These findings may indicate a differential genetic effect between very small VLDL subclass and all the other VLDL subclasses, which can relate to a differential role of very small VLDL subclass within the VLDL metabolism.

For the total lipids in HDL system (Table B.3), eight loci were tested for association. SNP rs261336, which is localized in a gene encoding for hepatic lipase *(LIPC)*, shows strong association with very large HDL subclasses but is not associated with large, medium and small HDL subclasses. On the other hand SNP rs1532085 also localized within hepatic lipase gene shows association with large HDL subclasses but does not show any association with all the other HDL-related traits. In a similar way, the SNP rs3764261, which is in a gene encoding cholesteryl ester transfer protein *(CETP)*, shows significant association with large and medium HDL but not with very large and small HDL. Only one genetic marker (rs6065906) is significantly associated with all the HDL subclasses included in the analysis. The combined analysis

of all HDL subclasses indicates a heterogeneous effect of the genetic variants to the different HDL subclasses. These results may reflect the differential function that each HLD subclass has in the lipoproteins metabolism.

For the remaining two systems of equations ('metabolites involved in the citric acid cycle' and 'metabolites involved in the urea cycle'), we observe that under the multiple response approach the tested SNPs failed to reach significant levels (Tables B.4,B.5).

3.4 Summary

In this chapter, we have performed an initial exploratory analysis of genomic data in combination with metabolomic data. Three different statistical approaches have been employed to investigate the genetic basis of metabolic traits. Firstly, we have performed a GWAS analysis based on the standard single-SNP approach. Following this approach, we have identified a number of genetic variants significantly associated with metabolic traits. The majority of these loci have been previously associated with metabolic measures at genome-wide significance and most of them matched physiological roles of the associated metabolic traits. Secondly, we have explored the use of multi-SNP methods such as penalized regression in genome-wide association studies. Specifically, we have applied the HyperLasso regression as a variable selection method which allows the identification of the subset of SNPs that best explain variation in metabolic traits at the specified type I error. Based on the HyperLasso regression, we have identified genetic variants known to be related to metabolic traits, some of these finding were not reported by the single-SNP approach. The HyperLasso regression selected SNPs that were less correlated with the other SNPs included in the model. In this manner, HyperLasso identified the same genetic regions associated with the phenotype as the single-SNP analysis but returning less redundant genetic information. The last part of this chapter has been dedicated to the use of multiple response methods to investigate SNP-phenotype associations. The top SNPs identified in the univariate analysis have been further investigated following a multiple response approach, which was based on the Seemingly

Unrelated Regression (SUR) model. Based on this analysis, we have been able to discriminate between genetic variants that have a significant effect on all the traits as well as the genetic variants that have significant effect on several phenotypes but not on all of them.

Chapter 4

Differential Networks

The aim of this chapter is to investigate the use of variations in patterns of molecular associations as a tool to uncover key differences between different physiological conditions. To this end, we have developed a formal statistical approach for the differential analysis of molecular associations via network representation. The chapter starts with an introduction to the topic under investigation (section 4.1). A brief description of the basic concepts of network analysis is provided in section 4.2. In section 4.3, first we review a number of statistical methods commonly used in correlation analysis and then we present the differential network methodology. We assess the performance of the proposed differential network at identifying differences in molecular associations between two different groups by performing a simulation study (section 4.4). Subsequently, we illustrate the potential of the method to reveal molecular differences related to physiological states on a real data application (section 4.5). We conclude this chapter (section 4.6) discussing the main results.

The research described in this chapter has been published in: Valcárcel, B., Würtz, P., Seich, a. B., Tukiainen, T., Kangas, A. J., Soininen, P., Järvelin, M., Ala-Korpela, M., Ebbels, T. M. and de Iorio, M. (2011) **A Differential Network Approach to Exploring Differences between Biological States: An Application to Prediabetes**. Plos One. 6 (9), e24702.

4.1 Introduction

Abnormal levels of certain molecules in blood or other biofluids are traditionally used as indicators of changes in physiological conditions. While the study of variations in mean levels of single molecules is of great importance, looking at those changes is not enough to fully understand the biological basis of disease development. For instance, in the transition process from health towards disease status, in addition to individual changes in molecules' abundances, variations in the pattern of molecular associations are also observed (104). That is, under different physiological conditions, different patterns of molecular associations can occur (105). The comprehensive analysis of molecular association patterns and their changes in relation to different physiological conditions can yield insight into the biological basis of phenotype variation in the process of disease development, yet the analysis of these molecular changes is often still underused.

Molecular profiles are complex and information-rich, requiring effective tools to analyse, visualize and compare the underlying molecular features in detail. Association measures of molecules concentrations such as correlation and partial correlation are useful approaches to explore underlying features of phenotypic variations. However, the biological interpretation of the correlation matrix is very complex, especially in the analysis of highly correlated molecular processes and data such as in the case of serum metabolites.

With advances in the theory of complex networks (106) and its application to describe architectural features of molecular systems (107), network-based approaches have been increasingly used to capture underlying properties of biological systems. An appealing feature of this approach is the ability to create a visual simplification of the observed pattern of molecular associations. Network theory is also of interest to identify variations between different physiological stages as well as biological systems. Several studies based on gene expression data have already explored the idea of using differential network-based methods for a variety of purposes. Examples include the use of association networks to identify species-specific network connections (108), gene expression associations to detect body weight-related genes (109)

or the assessment of changes in association patterns in patients with chronic fatigue syndrome (110). A more formal statistical methodology for differential analysis of gene associations was presented in Gill et al. (2010)(111). In this study all the statistical tests were based on a set of scores that measure the strength of the genetic associations in two different networks. An early analysis of metabolic association networks employed metabolic correlation networks to distinguish silent phenotypes in plants (112). More recently, Szymanski et al. (2009)(113) demonstrated the potential of metabolic network analysis to identify important molecules for stress adaptation. Yet, despite being a promising approach, few studies have employed metabolic association networks to investigate underlying features of phenotypic variations.

In this study, we describe a formal statistical approach for the differential analysis of metabolic associations and demonstrate the potential of the methodology to reveal metabolic differences related to (patho) physiological states in the context of lipoprotein metabolism.

4.2 Basic Concepts of Complex Networks

Many natural phenomena such as gene expression, protein-protein interaction networks and metabolic networks are all examples of complex systems and can be naturally modelled using network methods. A network is represented by a set of nodes, which in biology stand for each of the biological entities under study such as genes, proteins or metabolites and a set of edges between nodes which denote interaction between components (e.g. protein-protein interaction, metabolic reactions). Connections between nodes in biological systems can exhibit very complicated structural patterns. With the detailed analysis of the networks' topological properties, key features of the organizing principles of the complex systems can been uncovered. These networks measurements allow the quantifiable description as well as comparison of different complex networks. The most basic and commonly used topological properties for network analysis are listed and described below.

Degree, k_i, (or connectivity) indicates the number of connections that a node i has in the network. The degree of a node is a very elementary network measure, but it is a very important property because it helps to characterize nodes within the network. The number of connections reflects the relative importance of a node in a network. Highly connected nodes also known as hubs generally serve fundamental functions within the system, indicating therefore key points of the represented system (114) .

Networks density, D, measures how well connected a network is by comparing the number of connections in a given network with the maximum possible number of connections between all pairs of nodes. The density is calculated as follows:

$$D = \frac{2 \times E}{N \times (N-1)}$$

where N is the total number of nodes and E denodes the numbers of links in a given graph.

The average node degree, A, is a connectivity property which measures how many connections and nodes are in a given graph. The average degree is calculated as follows:

$$A = \frac{2 \times E}{N}$$

Degree distribution, $P(k)$, is the measure that gives the probability that a chosen node i has k links. $P(k)$ can be calculated by computing the number of nodes with degree $k = 1, 2, ...$ and dividing by the total number of nodes N.

Slope of degree distribution is the value of the scaling exponent $(-\gamma)$ for a power-law degree distribution. Estimation of this value can be made through the log-log plot of the connectivity distribution, in which a straight line with slope $-\gamma$ is expected if the networks is scale-free.

Path length is the measure that gives the number of nodes that need to be crossed while going from one node i to another node j. The path with the less number of steps between two nodes is called the **shortest path** and has a special role in the description of network structure (115). Another measure related to path length is the **average path length** which gives the average length over the shortest paths between all pairs of nodes.

Clustering coefficient, c_i**,** is the network measure that indicates to what extent the neighbours of a selected node i are connected to each other. In other words, in many networks it is likely that if node i is linked to node j, and node j is connected to node g, then i and g are also linked. Clustering coefficient can be obtained for each node i by:

$$c(i) = \frac{2K_i}{k_i(k_i - 1)}$$

where K_i denotes the links observed among the neighbours of node i and k_i represents its degree. For $k_i < 2$, c_i , is defined to be zero.

Betweenness centrality, b_i**,** is the measure that indicates how central a node i is in the graph and it can be calculated by:

$$b(i) = \frac{\sigma(i)}{\sigma}$$

where σ denotes the total number of pairs of nodes and $\sigma(i)$ denotes the number of shortest paths between nodes that pass through i.

4.3 Statistical Methods

In this section we describe the statistical methods applied to investigate differences in patterns of molecular association between different groups. Firstly, a number of methods commonly used for correlation analysis will be reviewed and then we will describe the differential network methodology.

4.3.1 Individual network analysis

The pattern of metabolic associations for each group is investigated by constructing individual networks. To measure the underlying interdependencies between variables, we consider three different correlation methods namely, Pearson's correlation, partial correlation and shrinkage partial correlation. These correlation procedures compute the correlation coefficients which describe the degree and direction of linear relationship between two variables. The correlation coefficient takes a value ranging between -1 to 1. The sign of the correlation coefficient indicates the direction of the relationship, either positive or negative. A positive association (> 0) indicates that two variables tend to increase or decrease simultaneously whereas a negative value of (< 0) indicates that one variable tends to increase when the other decreases. Perfect correlation between two variables is indicated by an absolute value of 1.

Pearson product moment correlation coefficient (Pearson's correlation): Pearson's correlation (r) is the most commonly used measure of correlation. The Pearson's correlation between two given variables, Y_i and Y_j, can be obtaining as follows:

$$r_{Y_i,Y_j} = \frac{\text{cov}(Y_i, Y_j)}{\sigma_{Y_i}\sigma_{Y_j}} \tag{4.1}$$

where $\text{cov}(Y_i, Y_j)$ is the covariance between variables Y_i and Y_j and $\sigma_{Y_i}, \sigma_{Y_j}$ the standard deviation for such variables. Significant correlation between two variables is defined by computing the p-value for the two-sided test with null hypothesis:"Correlation between variables Y_i and Y_j is equal to zero"versus the alternative hypothesis:"Correlation between variables Y_i and Y_j is not equal to zero". The p-value is computed using the Fisher's z' transformation (116):

$$z_{Y_i,Y_j} = \frac{1}{2}\ln(\frac{1 + r_{Y_i,Y_j}}{1 - r_{Y_i,Y_j}}) \tag{4.2}$$

where ln is the natural logarithm. The Z-statistics is:

$$Z = \frac{z_{Y_i,Y_j}}{\sqrt{N-3}} \tag{4.3}$$

where N is the sample size. We reject the null hypothesis at significance level $\alpha = \frac{0.01}{I}$, where $I = \frac{M(M-1)}{2}$ represents the total number of possible metabolite pairs.

Partial correlations: Partial correlation quantifies the correlation between two variables conditional on the remaining variables. In this way, we can determine to what extend correlations between two variables, Y_i and Y_j, are direct and therefore do not originate via intermediate variables. When the number of observation is greater than the number of variables ($n > p$), the covariance matrix ($\boldsymbol{\Sigma}$) is invertible and the partial correlation matrix can be calculated from the inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1}$. If we define $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$, the partial correlation $r^{partial}_{Y_i,Y_j}$ between two variables Y_i and Y_j can be defined as follows:

$$r^{partial}_{Y_i,Y_j} = -\frac{c_{Y_i,Y_j}}{\sqrt{c_{Y_i,Y_i}c_{Y_j,Y_j}}} \tag{4.4}$$

where c_{Y_i,Y_j} is the $ij-th$ element of the inverse of the covariance matrix ($\mathbf{C}$). To interpret zero partial correlation as an indicator of conditional independence between two variables, the assumption that all variables included in the calculation have a joint multivariate normal distribution must hold. To determine the significance of the partial correlations, the Fisher's transformation in equation 4.2 and the Z-statistics in equation 4.3 become:

$$Z = \frac{z^{partial}_{Y_i,Y_j}}{\sqrt{N-q-3}} \tag{4.5}$$

where N is the sample size and q is the number of variables on which the calculation is conditioned.

Shrinkage partial correlations: The calculation of the partial correlation matrix can be a straightforward process in situations where the number of variables is lower than the number of observations $n > p$. However, the calculation of the partial correlation is problematic when the number of variables is very large. If the number of variables is greater than the number of observations $n << p$, the covariance matrix looses its full rank and its inverse does not exist. Moreover, due to the sheer number of pairwise correlations in biological systems, the interpretation of the correlation matrix is very complex. To overcome these difficulties and avoid the potential problem of over-fitting, we propose the use of a linear shrinkage estimator to calculate the covariance matrix and hence the partial correlations between variables. Details about the method employed for the calculation of the partial correlation matrix can be found in (117). The method proposed by the authors is based on the combination of the sample covariance matrix U and the shrinkage target matrix T to obtain a regularized estimate of the covariance matrix U^*, as follows:

$$U^* = \lambda T + (1 - \lambda)U, \tag{4.6}$$

where λ is the shrinkage intensity. Here, we use the identity matrix as the shrinkage target T to shrink to zero the covariance elements. In this way, correlations remaining significantly different from zero are indicators of strong dependence between the variables.

The optimal shrinkage intensity was calculated from the data using the R package **GeneNet**. Details about the shrinkage intensity selection and shrinkage procedure can be found in (117).

Individual networks: Once we have calculated association between variables, we build the individual networks. The networks are constructed by drawing the edges between those pairs of metabolites whose association (Pearson's correlation, partial correlation and shrinkage partial correlation) exceeds the chosen criteria for edge inclusion. For our analysis we use a conservative Bonferroni corrected threshold, $p < \frac{0.01}{I}$, where $I = \frac{M(M-1)}{2}$ represents the total number of possible metabolite pairs.

4.3.2 Differential network analysis

The main aim of this study is to test the applicability of association networks to reveal key metabolic differences between two groups, A and B. To this end, we propose a statistical framework to perform a differential analysis of molecular associations. In this analysis, we test the null hypothesis:"The partial correlations of two metabolites in the two groups are the same"versus the alternative hypothesis:"The partial correlations of two metabolites in the two groups are different"by performing a two-sample permutation test. Given N_A samples from the first group (A) and N_B samples from the second group (B) all the samples are pooled together in one set and their labels are permuted. Permutation was performed for all metabolic variables simultaneously. At each iteration, i, of the permutation test two new datasets are obtained, A_i^* and B_i^* of size N_A and N_B, respectively. For each of the new datasets we compute the partial correlation matrix r_A^* and r_B^*. We follow the same approach described in Equation 4.6 for the estimation of the covariance matrix and hence the computation of the partial correlation matrix. Finally, we calculate the absolute difference between the partial correlations among the new groups, $\mid r_A^* - r_B^* \mid$. We repeat this procedure 100,000 times.

In order to find those metabolite associations that significantly change across the two groups we compute the absolute difference between the partial correlations among the two original groups, $\mid r_A^{obs} - r_B^{obs} \mid$. The p-values that indicate which interactions differ significantly between states are obtained by:

$$p = \Pr(\mid r_A^* - r_B^* \mid > \mid r_A^{obs} - r_B^{obs} \mid)$$

To define which edges are included in the differential network we set a cut-off on the two-tailed p-value. For the same sample size, the power to estimate correlations is lower than that to estimate a change in a single variable mean level. Therefore, we choose to set the cut-off for the inference of the differential network based on the uncorrected threshold $P < 0.01$. An edge in the differential network is included if the partial correlations between two given metabolites are significantly different across the two groups. The association network inferred under this set-up is referred to as the differential

network.

Computation of Pearson's correlation, partial correlation and shrinkage partial correlation were performed using the R package **GeneNet**. The individual and differential networks were built and visualized using the R package **igraph**. All analysis was carried out using R version 2.10.0.

4.4 Simulation study

A simulation study was designed to assess the performance of the differential network at identifying differences in pair-wise association between variables across different groups. Firstly, the differential network method was tested when the null hypothesis:"The partial correlation of two variables in the two groups is the same" is known to be false. For this study, we simulated two data sets (A, B) each of them consisting of five variables and N=1,000 observations per variables. For group A, each of the variables' values were simulated by sampling from a multivariate normal distribution with means (μ) equal to 0 and Σ equal to the covariance matrix. The identity matrix was initially used to define partial correlations between variables. Then, for two randomly selected variables (Y_i and Y_j) the partial correlation coefficient $r^A_{Y_i,Y_j}$ was simulated by sampling from a uniform distribution. The variables' variances (σ^2_A) were generated from an inverse-gamma distribution with shape $\alpha = 1$ and scale $\beta = 1$. The diagonal entries of the covariance matrix were increased sufficiently to ensure positive definiteness.

To simulate the variable's values for group B, we followed similar procedure as previously described, that is, the values of each observation were simulated by sampling from a multivariate normal distribution with means (μ) equal to 0 and Σ equal to the covariance matrix of the variables. The partial correlation between all variables in group B were initially set to be equal to those in group A. Then for the two previously selected variables Y_i and Y_j a change in their partial correlation coefficient was introduced so the partial correlations of variables Y_i and Y_j between the two groups A and B are different. The change in partial correlations was simulated by sampling from a uniform distribution. As with group A, an inverse-gamma distribu-

tion with shape $\alpha = 1$ and scale $\beta = 1$ was used to generate the variable's variances (σ_B^2). We constructed the covariance matrix using each variables standard deviation (σ_B) and ensured positive definiteness of the covariance matrix by increasing sufficiently the diagonal entries.

In addition, the differential network performance was tested when the null hypothesis:"The partial correlation of two variables in the two groups is the same" is known to be true. The data sets for A and B for this study were simulated following the same procedure as previously described. In this simulation, the change in partial correlation between Y_i and Y_j between groups A and B was equal to zero.

Once data sets were simulated, we analysed them using the differential network approach presented in subsection 4.3.2. For the inference of the differential network, we used a significance threshold of $\alpha = 0.01$. The simulation process was repeated 100 times. Each of the times the total number of connections captured by the differential network were recorded.

Scenarios for data simulation

With this simulation study, we are interested in investigating: (1) whether the strength of association between variables influences the performance of the differential network, (2) whether a small and/or large differences in association between two variables and across the two groups are captured by the differential network method. To evaluate these aspects of our methodology, we simulate the data sets following six scenarios. For variables, Y_i and Y_j :

1. **Scenario 1:** The partial correlation between Y_i and Y_j in group A is low ($\mid r^A_{Y_i,Y_j} \mid \leq 0.1$). The difference in association between these two variables across the two groups, A and B, is small ($\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid \leq 0.1$).

2. **Scenario 2:** For group A, the partial correlation between variables Y_i and Y_j is low ($\mid r^A_{Y_i,Y_j} \mid \leq 0.1$). The difference in correlation between these two variables across the two groups is large ($\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid \geq 0.5$).

3. **Scenario 3:** The pairwise correlations between Y_i and Y_j is high for group A ($\mid r^A_{Y_i,Y_j} \mid \geq 0.5$). The difference in correlation between these two variables across the two groups is small ($\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid \leq 0.1$).

4. **Scenario 4:** For group A, the pairwise correlations between two variables is high ($\mid r^A_{Y_i,Y_j} \mid \geq 0.5$). The difference in correlation between these two variables across the two groups is large ($\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid \geq 0.5$).

5. **Scenario Null 1:** The partial correlations between variables Y_i and Y_j is low $\mid r_{Y_i,Y_j} \mid \leq 0.1$ and equal between the two for groups $\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid = 0$.

6. **Scenario Null 2:** The partial correlations between variables Y_i and Y_j is high $\mid r_{Y_i,Y_j} \mid \geq 0.5$ and equal between the two for groups $\mid r^A_{Y_i,Y_j} - r^B_{Y_i,Y_j} \mid = 0$.

4.4.1 Simulation Study Results

Table 4.1 shows the results from the simulation study. We show the total number of connections captured by the differential network at each simulation and under each scenarios. A connection is considered to be true if the difference in association between variables Y_i and Y_j is correctly identified by the differential network. A false connection represents any connection wrongly identified by the differential network.

From this study, we observe that all differences in association are correctly identified by the differential network when the difference in partial correlation between two variables is large (scenario 2 and 4). However, if the difference in association between the variables across the two groups is lower (scenario 1 and 3), the differential network is not able to identify changes in association. The results between simulation null 1 and null 2, scenario 1 and 3, scenario 2 and 4 are very similar which indicates that the performance of the differential network is not affected by the strength of association between variables for each group, A and B, and it is only affected by the difference in association. In addition, for all simulated scenarios the number of false connections identified by our method is very low. Based on this results of the

simulations, we can conclude that the differential network is able to identify differential associations between variables.

Table 4.1: Results for the simulation study.

	Total Connections	TC	FC
Null 1	5	0	5
Null 2	7	0	7
Scenario 1	12	1	11
Scenario 2	105	100	5
Scenario 3	11	5	6
Scenario 4	101	100	1

TC, true positive ; **FC**,false positive. Results are based on 100 simulation processes per scenario.

4.5 Case study: Prediabetes dyslipidemias

We illustrate the differential network methodology using a real data set. Specifically, we explore differences in metabolic associations between individuals with normal fasting glucose (NFG) and individuals with impaired fasting glucose (IFG). Impaired fasting glucose also known as prediabetes refers to a physiological state at which individuals are enhanced risk for developing T2DM (31). To evaluate metabolic changes related to prediabetes we use extensive data on lipoprotein subclasses, enabling the assessment of changes in lipoprotein metabolism, which are known to be strongly connected to the risk and development of T2DM (58).

4.5.1 Study population and metabolic dataset

For this study we use the NFBC1966 metabolic data. In particular we use sixty measures of lipoprotein subclasses measured by NMR spectroscopy (see Table A.1). The data is initially divided into two groups according to levels of fasting glucose and following the categorical criteria of the American Diabetes Society (ADA) for prediabetes classification (118). The normal fasting

glucose (NFG) group represents individuals with fasting glucose levels lower than 5.6 mmol/L. The impaired fasting glucose (IFG) group includes individuals with fasting glucose levels from 5.6 mmol/L to 6.9 mmol/L. Subjects with fasting glucose levels greater than 6.9 mmol/L, clinically considered as having type 2 diabetes Mellitus, or missing data on lipoprotein subclass profile and/or glucose measures were excluded which left 4,937 subjects for the analysis. The sample size for each group is N_{NFG} = 4,406 (2,021 males and 2,385 females) and N_{IFG} = 531 (390 males and 141 females). Due to the well-known sex-specific difference in lipoprotein profiles (59) we perform the analysis separately for men and women. Due to a predominance of males in the IFG group, we base our biological interpretation on the male data; the differential network for the female data is shown in appendix Figure C.4. The clinical characteristics of the study groups are presented in appendix Table C.1.

4.5.2 Results and Discussion

To explore differences in metabolite concentrations between the two defined groups (NFG and IFG) we perform a non-parametric two-sample Mann-Whitney test. We determine significant differences in the metabolic concentrations by setting a conservative Bonferroni corrected threshold, $P < 0.01/M$, where $M = 60$ denotes the total number of measures. The results for males are shown in Table C.2. This initial inspection of the data showed no significant differences in concentration levels of the 60 lipoprotein subclass components between individual with normal fasting glucose and impaired fasting glucose at the Bonferroni corrected threshold $P < 0.01/M$. When the significance threshold is not corrected for multiple testing, $P < 0.01$, eleven measures, extremely large VLDL (phospholipids), very large VLDL (phospholipids and particle concentration), large LDL (phospholipids and particle concentration), medium LDL (cholesterol, phospholipids and particle concentration), small LDL (cholesterol and particle concentration) and Apolipoprotein B show significant differences in concentration levels between the two groups (NFG vs. IFG).

The few significant differences in lipoprotein concentrations between the two groups may reflect a generally healthy status of the young population studied. Nevertheless, the prediabetic state represents increased risk for the development of T2DM and it has been suggested that lipid abnormalities are associated with the progression of T2DM (31). Despite there being only a few significant variations in the concentration levels of lipoprotein subclass components between the two groups, we hypothesize that the lipoprotein subclass dependencies could be affected and assess this by building association networks.

- **Individual Networks**

The individual networks are built to examine the specific pattern of associations exhibited under each condition (NFG and IFG). The aim of this analysis is to explore variations in the observed pattern of metabolic associations between the groups as well as to define some topological features of the metabolic networks. The construction of the individual networks is based on a binary representation of underlying partial correlations (significant and non-significant partial correlation) and therefore the correlation strength between the measures is not represented here.

We present the individual networks obtained using three measures of correlation, Pearson's correlation (Figure 4.1a), standard partial correlation (Figure 4.1b) and shrinkage partial correlation(Figure 4.1c). To better illustrate associations between metabolites, in the appendices we also show correlation maps where the pair-wise associations have been quantified using the three measures of correlation (Figure C.1, Figure C.2, Figure C.3). Using partial correlation and shrinkage methods for the network construction, the number of connections is largely decreased. This implies the reduction in the network complexity and thus, the improvement on the biological interpretation of the association networks. These results show the advantage of using partial correlations in combination to shrinkage methods instead of other correlation measures, especially when the variables under study are highly interrelated, such as lipoprotein subclasses.

Table 4.2: Global properties of the individual networks obtained using shrinkage partial correlation as measure of pair-wise association.

Topological Properties	NFG	IFG
Connected nodes	54	54
Isolated nodes	6	6
Total number of edges	134	126
Networks density	0.075	0.071
Average degree	4.41	4.2
Slope of degree distribution	-0.12	-0.14
Clustering coefficient	0.26	0.29

The connectivity, k, for each of the nodes can be used to quantitatively define differences between the two individual networks. The global properties of NFG and IFG networks are presented in Table 4.2. In terms of total number of connections, both individual networks represent similar information (NFG = 134; IFG =126). Six VLDL-related measures are also isolated in both individual networks. A large number of these links are in common between the two individual networks. For the IFG network, 83% of the connections overlap with connections in the NFG network. Despite the large similarities between the two individual networks, some differences in the pattern of association between measures can be identified in Figure 4.2). Overall, LDL-related measures show a tendency to reduce their associations (light-blue nodes in Figure 4.2) from NFG to IFG state whereas measures in large to extra large HDL show a relative increase in their associations in the IFG condition (cluster of extra large HDL, left in Figure 4.2). In relation to VLDL-related measures, a more heterogeneous pattern of changes in their pair-wise association is observed. Several measures in medium VLDL (top and central-right in Figure 4.2) increase their number of associations in the IFG condition whereas very small VLDL (central in Figure 4.2) and free cholesterol in medium VLDL present a large reduction in their pair-wise associations from NFG to IFG state. Overall, the differences between the two networks represent a reduction in pair-wise association for the IFG network. The reduced metabolic connectivity can indicate that the inverse relation

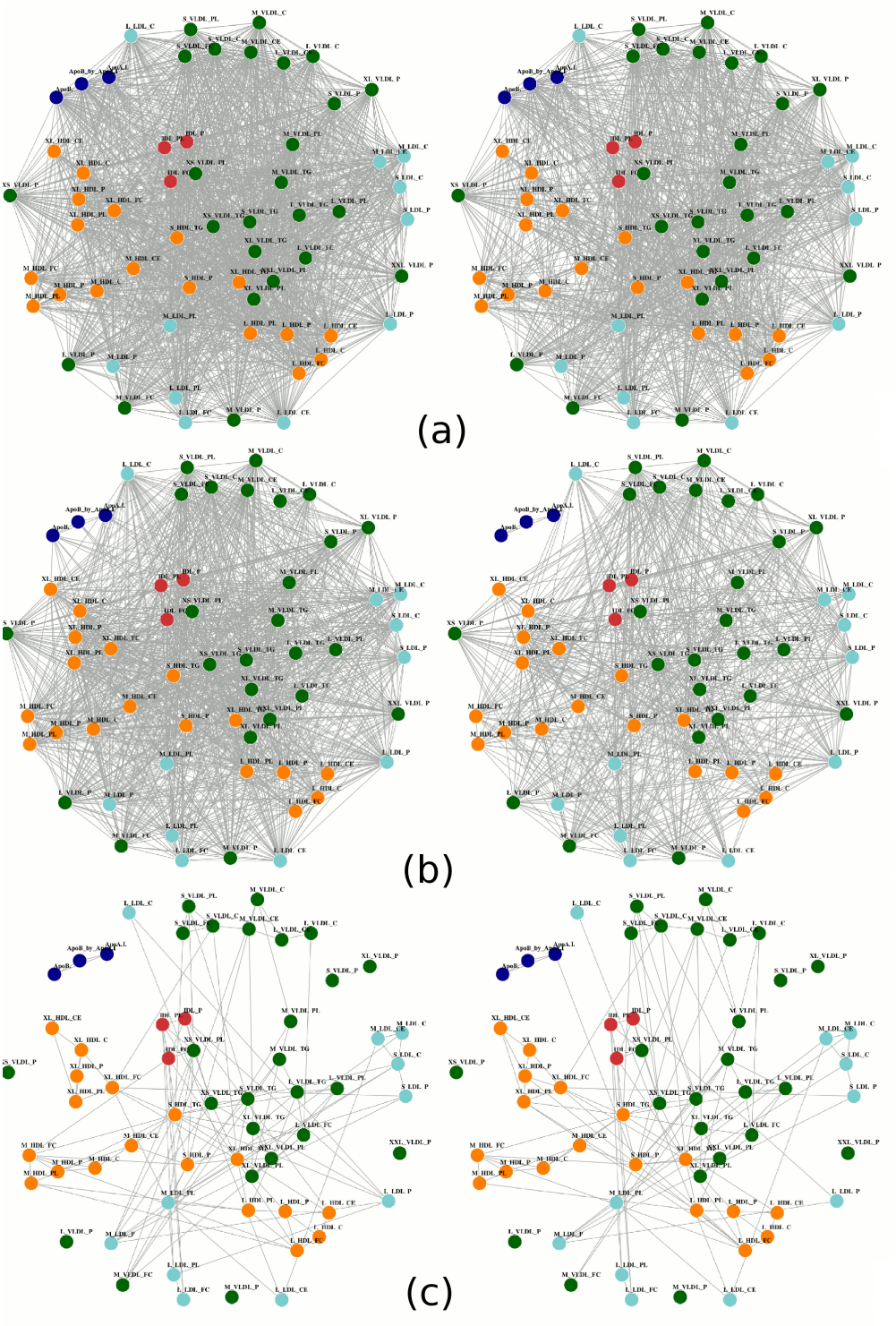


Figure 4.1: The individual networks obtained using: (a) Pearsons correlation; (b) partial correlation; (c) shrinkage partial correlation as measure of pair-wise association between the lipoprotein components.

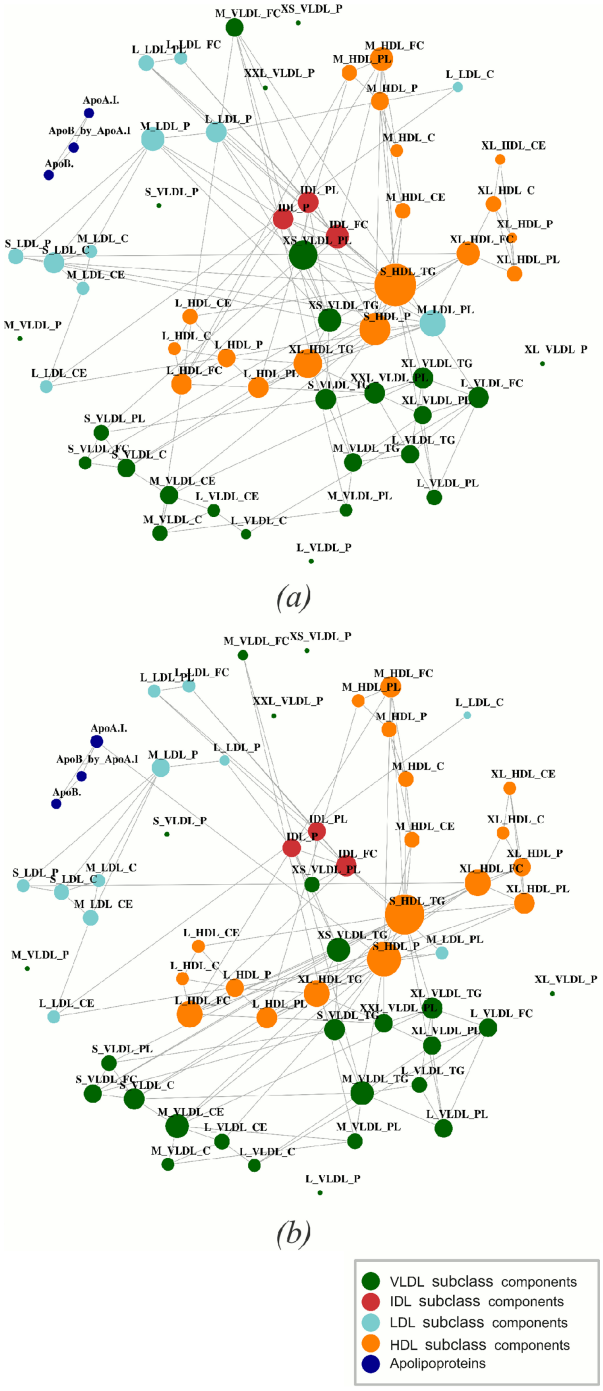


Figure 4.2: The individual networks inferred from male data. Each of the connections indicates significant pair-wise association between two lipoprotein measures using the Bonferroni corrected threshold $p < 0.01/z$, where $z = M(M1)/2$ is the total number of possible interactions and M=60 is the total number of metabolites. The NFG network indicates the pattern of pair-wise association for the normal fasting glucose group. The IFG network reflects the pattern of association for the impaired fasting glucose group. The node size is proportional to the degree of connectivity (number of connections).

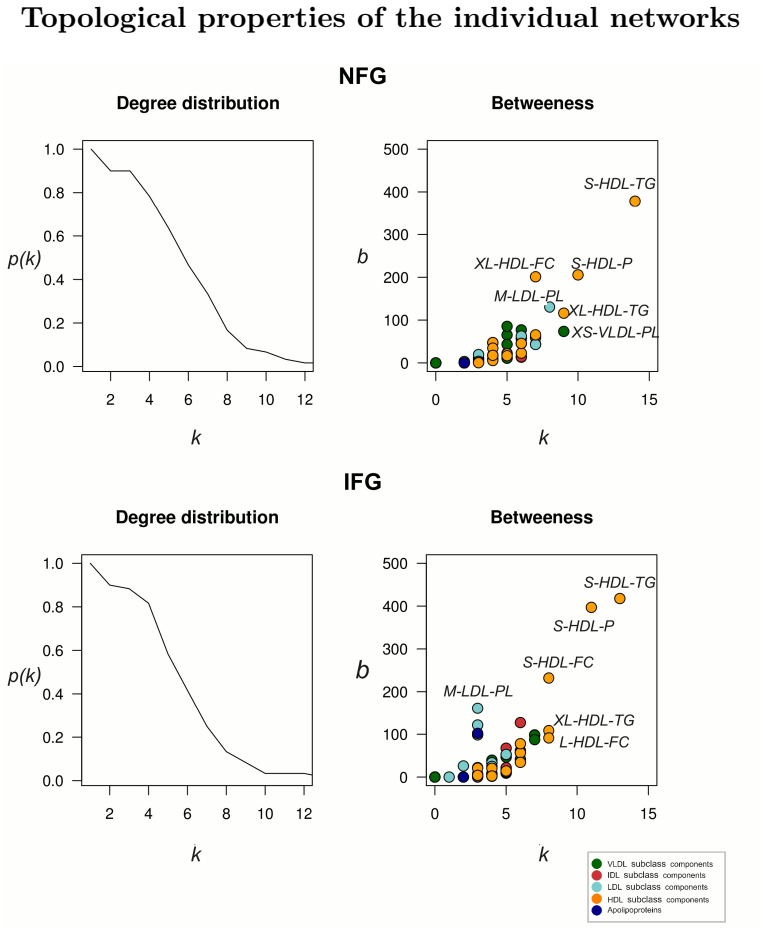


Figure 4.3: The degree distribution shows the probability that a node i has k connections in the network. For both individual networks the degree distribution is well approximated by a power-law. The betweenness indicates how central a node is in the network. For both individual networks, the nodes with higher degree connectivity are also the most central nodes.

between small VLDL and large HDL subclasses is less tightly regulated in the prediabetic state.

Other measures that can be used to quantitatively define the network's features are the degree distributions, and the network centrality. The degree distributions of the individual networks are shown in Figure 4.3. It appears that both individual networks are scale-free networks, class of networks that is characterised by a power-law distribution where the probability of a node

having k links is $P(k) \sim k^{\gamma}$. An important property of scale-free networks is the existence of a relatively small number of highly correlated nodes, so-called hubs (106). Triglycerides in small HDL and particle concentration in small HDL clearly represent the main hubs for both individual networks (central in Figure 4.2). These highly connected nodes are expected to be the central components of the networks, reflecting important points of the network organization (107). This central property can be quantified by computing betweenness centrality in relation to the degree connectivity (Figure 4.3). Although this analysis shows that some low-connected lipoprotein measures have high betweenness scores, the expected property 'high connectivity-high betweeness' holds for both individual networks. The high connectivity and the central position that small HDL (triglycerides and particle concentration) have in the individual networks indicate an important role of small HDL in the lipoprotein metabolism. It is widely accepted that large size HDL have a protective role in the premature development of coronary heart diseases (CHD) and related traits (119). However, the role that small HDL have in the lipoprotein metabolism is still unclear. It has been suggested that small HDL may contribute to the production of medium and large HDL particles during maturation of HDL in plasma. In this manner, small HDL may connect HDL metabolism to VLDL metabolism and also contribute to atheroprotective the role of HDL in CHD (120).

The global clustering is another measure used to describe the organization in complex networks. High values of this measure indicate modular organization of the network. Both individual networks exhibit this modular property where values of clustering coefficients are $c \geq 0.25$. Furthermore, metabolites involved in the same clusters are mainly components belonging to the same lipoprotein subclass e.g. large HDL subclasses with clustering coefficients $c \geq 0.5$. The tendency to form clusters among metabolites participating in the same metabolism has been reported previously (121).

Based on the theory of complex networks, the organization of metabolic systems has been described as large sets of densely interconnected modules that combine into larger and less interconnected units, with both degree of clustering and degree distribution following a power law (122). Thus, explor-

ing the network's topological properties certain features of metabolic systems can be determined. However, individual networks do not enable clear identification of key metabolic differences between the two groups. To approach this aspect we construct the differential network.

- **Differential Network: NFG-IFG**

The differential network inferred from male data is presented in Figure 4.4. Each of the connections in the differential network indicates a significant change in partial correlation between two lipoprotein measures across the (patho)physiological conditions. A set of 22 connected measures, organized in 16 pair-wise interactions are observed. The remaining 38 metabolites show no differential connections. The differential network indicates that the differential connections occur throughout a significant increase or decrease in the pair-wise partial correlation between the lipoprotein measures across the two groups, as well as significant alteration in the partial correlations sign (with NFG as reference group). For instance, the largest cluster in the differential network (right most in Figure 4.4) indicates that large and extra large HDL (phospholipids and triglycerides, respectively) are differentially connected to very small VLDL throughout an increase in their pair-wise associations (Figure 4.4; red edges). In the same way, this connected component (main cluster) reflects the decrease in partial correlation between measures of small VLDL (phospholipids and free cholesterol) and phospholipids in very small VLDL (blue edges). Of biological relevance is the double variation in pair-wise association that is observed between medium VLDL (ester cholesterol) and two measures of very small VLDL (in the largest cluster, central connections). This cluster reflects the shift from positive (NFG) to negative (IFG) partial correlation between medium VLDL and phospholipids in very small VLDL and the change from negative (NFG) to positive (IFG) pair-wise association between medium VLDL and particle concentration in VLDL. The interpretation of these variations points to a change in the underlying regulation of the system.

As with the individual networks, the network's topological properties

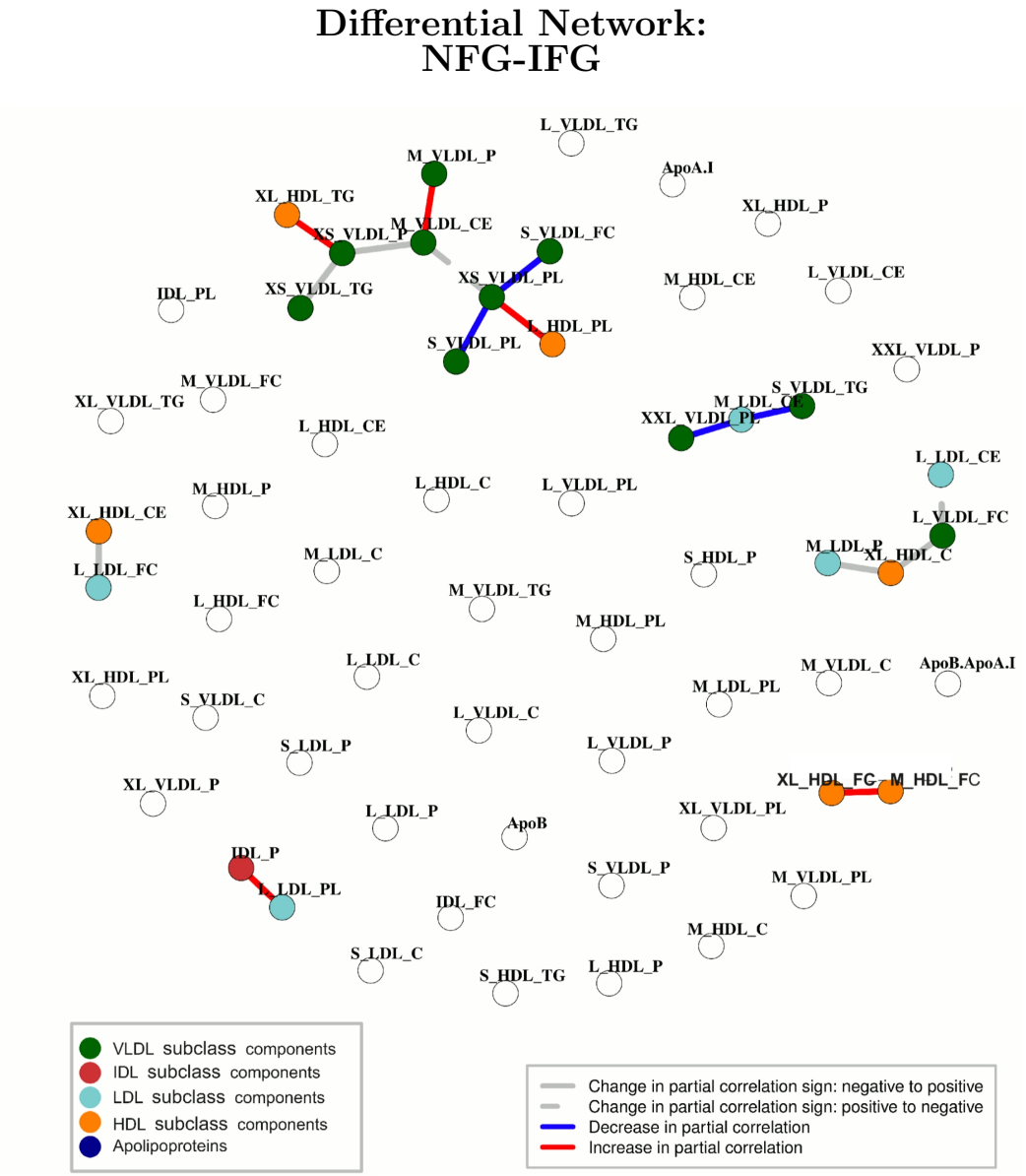


Figure 4.4: The differential network inferred from males with normal fasting glucose (NFG) and impaired fasting glucose (IFG). Each of the connections indicates a significant change in partial correlation between two lipoprotein measures across the physiological conditions. Edge colours represent how the partial correlation between two measures has changed across the two groups.

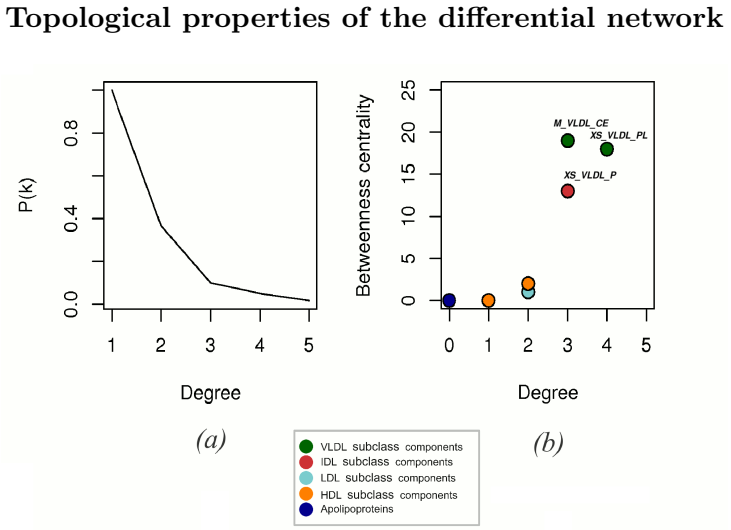


Figure 4.5: The degree distribution shows the probability that a node i has k connections in the network. For the differential network the degree distribution is well approximated by a power-law. The betweenness indicates how central a node is in the network. For the differential network, the nodes with higher degree connectivity are also the most central nodes.

can be used to quantitatively define the differential network. Three VLDL-related measures present the highest connectivity scores. These highly connected lipoprotein measures are the central components of the main cluster in the differential network (Figure 4.5). The first of these measures, very small VLDL (phospholipids) is differentially connected to large HDL (phospholipids), small VLDL (free cholesterol), small VLDL (phospholipids) and very small VLDL (particle concentration). The second measure, medium VLDL (ester cholesterol) is connected to medium VLDL (particle concentration) and very small VLDL (particle concentration). The last measure, very small VLDL (particle concentration) is connected to very large HDL (triglycerides) and very small VLDL (triglycerides).

The differential network indicates that change in metabolic association occur throughout the size-range of lipoprotein subclasses, except the smallest HDL. For instance, both extremely large and small VLDL particles decrease their partial correlation with medium LDL cholesterol (cholesterol

ester), possibly reflecting decreased regulation of the lipoprotein cascade, as observed for the individual networks. The most prominent cluster in the differential network reflects differential changes (in the correlation pattern) between very small to medium VLDL subclasses with large and very large HDL subclasses. For instance, there is an increase in partial correlation of phospholipids in very small VLDL with phospholipids in large HDL. Phospholipids are key metabolites in shaping HDL (123). In addition, we find that large VLDL and very large HDL change in partial correlation (upper most cluster), and that cluster also involves LDL subclasses. While high levels of large VLDL particles are known to cause increased risk for diabetes (31, 124), the pathogenic role of small VLDL is less recognized. The finding of decreased metabolic regulation between small VLDL and large HDL emphasises the continuous nature of VLDL to LDL cascade. In addition, the lack of differential associations for composite measures such as apolipoprotein B and A-1 illustrates the benefit of analysing the metabolic data on the lipoprotein subclass levels. Overall, our results show that lipoprotein association patterns are significantly different from normal fasting glucose already at prediabetic state. Although even more pronounced differences would be expected for diagnosed diabetics, the changes observed at the early stage of disease progression, with less metabolic regulation observed, reflect the gradual disease progression towards diabetes, which is associated with decreased insulin resistance for prediabetics and changes in lipoprotein metabolism.

The differential network results are consistent with existing knowledge on the development of T2DM. Studies based on lipoprotein profiles have resulted in the recognition of a complex pattern of atherogenic-associated changes in size and particle concentration within the major lipoprotein classes in patients with diabetes (124, 125). These disorders are reported as increased levels of medium and large VLDL and small LDL and reduction in size of HDL. It has been suggested that elevated glucose can lead to an increase in the levels of triglycerides in liver, resulting in the shift from small VLDL particle to enlarged VLDL (124, 126). These variations in size range in lipoprotein subclasses, which relate to diabetic dyslipidemias, have been captured by the differential network. The large presence of VLDL-related measures

in the differential connections reflects the early variations in the lipoprotein metabolism during the continuous process towards diabetes condition.

4.6 Summary

In this chapter, our main aim was to investigate the potential use of molecular associations to identify key molecular differences across different groups. We have developed a formal statistical approach for the differential analysis of metabolic associations via network representation. Firstly, we have assessed the performance of the differential network at identifying differences in pair-wise association between groups in a simulation study. Secondly, we have demonstrated the application of our method in the context of lipoprotein metabolism by identifying differences between individuals with normal fasting glucose and individuals with impaired fasting glucose (prediabetes). The results illustrate the advantage of using the network-based approaches to explore topological features of biological systems and to identify differences between physiological groups. Initial inspection based on standard statistical methods showed relatively few significant differences in concentration levels of lipoprotein subclasses between subjects with normal fasting glucose and those with impaired fasting glucose (prediabetes). With the construction of the individual networks, we were able to examine the specific pattern of associations exhibited under each condition (NFG and IFG). Although no major differences were found between the two individual networks, several notable topological properties of lipoproteins system were identified throughout this analysis such as free-scale organization, hubs metabolites and cluster of nodes. By performing the differential analysis of networks, we were able to identify various changes in lipoprotein metabolism between the two groups. Many of the lipoprotein variations highlighted by the differential network analysis are known to be related to diabetic dyslipidemias. The results indicate the suitability of network-method to investigate subtle molecular variation indicative of (patho)physiological status.

Chapter 5

GEMINi: GEnome Metabolome Integrated Network analysis

In this chapter, we investigate the application of differences in molecular dependencies to uncover genetic variants associated with phenotype variation. To this end, we propose a novel analytical approach, which we refer to as GEMINi, (GEnome Metabolome Integrated Network analysis) and illustrate its use in the context of obesity. The chapter starts with section 5.1, where we give a short introduction to the topic. In section 5.2, we describe the datasets used and define the statistical framework proposed in this work (section 5.3). The results of the analysis are presented in section 5.4. We conclude the chapter summarizing the main results, strengths and limitations of the GEMINi approach (section 5.5).

5.1 Introduction

Despite the success of GWAS studies at identifying loci associated with complex traits, a large proportion of the genetic variation contributing to the observed variability in phenotypes still remains unmapped (127, 128). In trying to discover the "missing variants", a number of strategies have been proposed (18). It has been suggested that considering alternative models of genetic effect and their influence on complex traits can help discovering some of the missing variants (18). For instance, some studies have already shown that testing for equality and heterogeneity of variances of the phenotypes among different genotypes can be used as an alternative approach to

uncover unknown information about the system state beyond that reported by changes in mean levels alone (129, 130, 131).

In this study, we hypothesise that some of the undetected variants might be associated not with changes in mean levels of a phenotype, as studied by conventional GWAS, but with alterations that occur at a higher system-level. In previous chapter, we have shown that in addition to mean level alterations of single phenotypes, differences in association between different traits are observed across different physiological states. Such differences in molecular correlations between states can reveal information about the system state beyond that reported by changes in mean levels alone.

With this in mind, our aim is to investigate the use of differences in molecular dependencies to uncover genetic variants associated with phenotype variation. We present a novel analytical approach to integrate metabolic and genetic information, which we refer to as GEMINi, (GEnome Metabolome Integrated Network analysis). It consists of a combination of correlation network analysis, genome-wide correlation study and pathway analysis. To demonstrate the use of this approach, we investigate metabolic changes in relation to obesity.

5.2 Materials

• Study Populations

We used data from two large population-based cohorts, the Northern Finland Birth Cohort 1966 (NFBC1966) and the Northern Finland Birth Cohort 1986 (NFBC1986). The methods and aims of these studies have been described previously on section 2.1). In brief, the NFBC1966 includes 12,068 children who were due to be born in the provinces of Oulu and Lapland in 1966. Data were collected since pregnancy and supplemented at the age of 1,14 and 31 years. Blood samples were taken when individuals were 31 years old. The NFBC1986 includes 9,432 live born children with expected dates of birth between 1st July 1985 and 30th June 1986 also in the above area. The cohort has been monitored since early pregnancy until adolescence. All those alive

with known address were invited to a clinical examination at the age of 15 to 16 years when blood samples were taken.

- **Phenotype measures**

For obese and non-obese group classification, we use a newly introduced method for estimating body fat, the body adipose index (BAI) (132). In comparison to BMI as a measure of obesity, BAI allows us to take into consideration sex-specific differences in total body fat and is also considered to offer an improved classification in subjects with elevated lean body mass. Additionally, the multiple cardiometabolic risk factors related to obesity are known to be mainly associated with adipose accumulation - particularly around the waist (133, 134) - thus making BAI a particularly relevant physiological end point to consider. The BAI is calculated as follows:

$$BAI = \frac{hip}{height^{1.5}} - 18$$

where hip refers to circumference in cm and height is measured in meters. For the differential network analysis, individuals are classified into two groups, obese and non-obese. To classify individuals into obese and non-obese groups, we use the cut-off values for body fat as proposed in by Gallagher et al. (2000) (135). In this way, the non-obese group includes women with measures of BAI equal or lower than 33 and equal or greater than 21 and men with BAI equal or lower than 19 and equal greater than 8. The obese group includes women with BAI greater than 39 and men with BAI greater than 25. Individuals were excluded from analysis based on the following criteria: non-fasting, diagnosed type I or II diabetes, pregnancy, lipid lowering medication, missing data on glucose and/or hip measures and/or height. For the differential network analysis, the sample size for NFBC1966 is 2,794 subjects (N_{normal} = 2,208 and N_{obese} = 586) and for NFBC1986 it is 2,645 subjects (N_{normal} = 2,356 and N_{obese} = 289).

- **Metabolomic data**

For GEMINi analysis, we included 38 metabolic measures representing a wide variety of metabolic processes, such as amino acid, energy and lipoprotein metabolism. Metabolites were selected to allow a consistent analysis across the two cohorts.

- **Genetic data**

Only genetic data from NFBC1966 were used for the genetic analysis since genome-wide data for the NFBC1986 was not available. The final genetic data set included 4,028 individuals.

5.3 Statistical Methods

In this study we compare two statistical procedures. First, we perform a standard analysis of association between metabolic levels and BAI using univariate linear regression and between genetic variants and BAI following a conventional genome-wide association study (GWAS) approach. The second approach, which we refer to as GEMINi, is a combination of (i) differential network analysis of metabolic associations (ii) genome-wide correlation study and (iii) pathway-level analysis. All analyses were adjusted for the effect sex differences on serum metabolites.

5.3.1 Conventional analysis

We use linear regression to test the effect of serum metabolites on body adipose index (BAI). The analyses are performed following a univariate approach where metabolites are tested one at a time. We included gender as covariate to correct for sex-differences in serum metabolic profiles. To identify significant association between serum metabolites and BAI, we use a conservative Bonferroni corrected significance level, $P < 0.01/M$, where $M = 38$ denotes the total number of serum metabolite measures.

A standard GWAS approach is used to test the association between genetic variants and BAI. The analysis is performed following the standard single-SNP approach where SNPs are tested one at a time. Associations were investigated using linear regression assuming an additive effect on the trait and including sex as a covariate. Significant associations between genetic variants and BAI measures were assessed by setting genome-wide significance to $P < 5 \times 10^{-8}$(19).

The above analyses were conducted including all the individuals for which genetic and metabolic data were available (N=4,028).

5.3.2 GEMINi

An outline of the GEMINi analysis is presented in Figure 5.1. Initially we perform a differential network analysis that allows us to test whether the pattern of pair-wise associations between metabolites is the same in two physiological groups (non-obese and obese) or whether it significantly differs across groups. To eliminate the confounding effect of sex on the serum metabolites, the data used for the analysis are the residuals from a linear regression model of each metabolite on sex. To build the differential networks, we use the same methodology presented in subsection 4.3.2. To validate the differential network analysis results, we compare the network structure between the two cohorts. The replicated results between cohorts are further investigated in the next step of the analysis.

In the second step, we perform a genome-wide correlation analysis to identify genetic variants associated with variations in metabolic associations. As for the standard GWAS study, all individuals for which genetic data are available were included in the analysis. To find the desired associations, we first classify individuals according to the number of copies of the less frequent allele carried, giving genotype groups A (0 copies) and B (1 or 2 copies). SNPs are tested one at a time. Subsequently, the correlation between metabolites m_1 and m_2 is calculated for each group, r_A and r_B, and variation in correlation between groups A and B is tested. As in the differential network analysis, we eliminate the confounding effect of sex on the serum metabolites

by using the residuals of a linear regression of the metabolite level on sex. To test if the two correlation coefficients, r_A and r_B are the same, we use the z transform method as described in (136). To correct for multiple testing, significant associations between genetic variants and variations in metabolic association are determined using the genome-wide significance threshold ($p < 5^{-8}/D$), where D refers to the number of differential connections identified in step one.

To assess the biological significance of our findings, identified SNPs are assigned to the nearest gene (maximum distance 1 Mb). We use standard over representation analysis, implemented within the web tool IMPaLA (137), to highlight pathways which are significantly enriched in these genes. IMPaLA tests 3,000 pre-annotated pathways from 11 databases and employs the classical hypergeometric distribution to test for enrichment. Over representation analysis was performed independently for each pathway using the list of genes associated with each metabolite pair (e.g. Tyr-Glu) as well as using the combined list of genes identified by the genome-wide correlation study.

All analyses were conducted in R version 2.14.1, apart from the conventional GWAS that was carried out in SNPtest v2.0 and the pathway analysis which used **IMPaLA** version 4. Computation of the partial correlation matrix was performed using the R package **GeneNet**. The differential networks were built and visualized using the R package **igraph**. The package psych in R was used to test for differences between correlation coefficients.

5.4 Results and Discussion

5.4.1 Conventional analysis of metabolic and genetic data

We performed an initial inspection of the data using linear regression to explore the association between serum metabolites and BAI (Table 5.1, Table 5.2). The analysis showed that all very low density lipoprotein subclasses (VLDL), intermediate density lipoprotein (IDL), all low density lipopro-

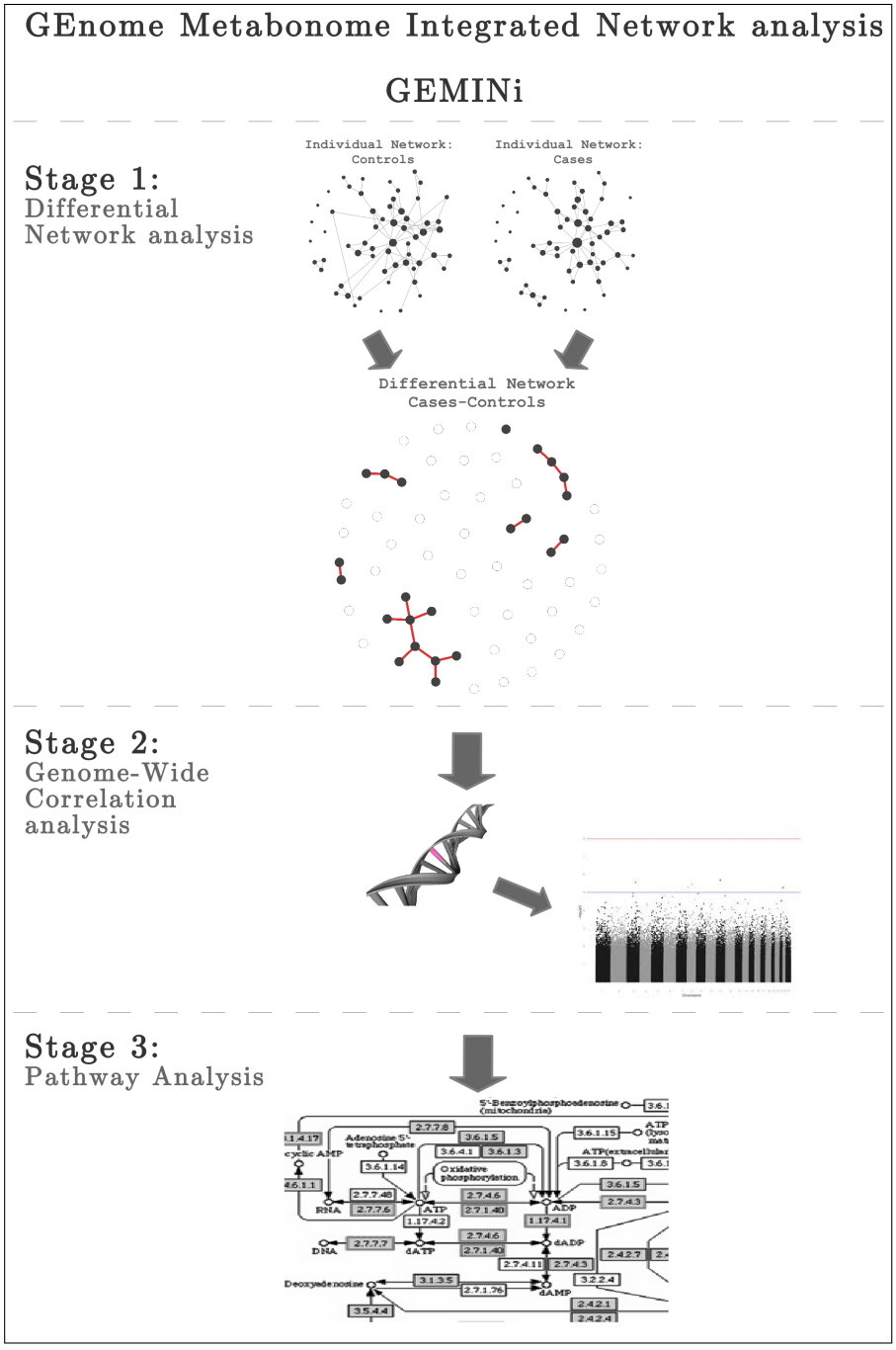


Figure 5.1: GEMINi analysis outline. Stage 1: We perform the differential analyses of metabolic associations with the aim of identifying a set of metabolic associations most affected by BAI/obesity condition. Replicated results between populations are investigated in stage 2 to highlight genetic mechanisms. Stage 2: Genome-wide correlation study aims to determine the existence of genetic variants associated with variations in metabolic associations. Stage 3: Pathway-level analysis is performed to identify the subsets of genes acting in the same functional pathways.

tein subclasses (LDL), small high density lipoprotein subclass (HDL), and apolipoprotein B were positively associated with the BAI and very large, large HDL were inversely associated with BAI in both cohorts. In the NFBC1966, medium HDL was inversely associated with BAI but shows no association in the NFBC1986, which may due to the age difference between cohorts (138). No association was found between BAI and apolipoprotein A in either of the cohorts. Tests of association between BAI and low-molecular-weight metabolites showed that alanine, glycerol, isoleucine, leucine, phenyalanine, pyruvate, tyrosine and valine were positively associated with BAI in both cohorts whereas citrate and glutamine were inversely associated with BAI. No association was found between BAI and 3-hydroxybutyrate or urea. We observed that glucose and lactate were positively associated with BAI and creatinine was negatively associated with BAI in the NFBC1966 but these three metabolites showed no association in the NFBC1986. On the other hand, acetate was negatively associated with BAI and histidine positively associated in NFBC1986 but showed no association in NFBC1966. The same analyses were performed using the body mass index (BMI) (Table D.1,Table D.2) instead of BAI. The results from BAI and BMI analyses were similar, apart from three metabolites (apolipoprotein A, 3-hydroxybutyrate and creatinine) which showed different associations with BAI and BMI in the NFBC1986.

Furthermore, we performed a standard GWAS analysis where associations between genetic variants and BAI (as a continuous variable) were tested. Using this approach, however, we did not find any association that was significant at the genome-wide significance level. The fact that we found no association at a genome-wide significance level might be due to the lack of power in this cohort to identify loci with key roles in determining variations in BAI.

The large number of significant associations between BAI and the metabolite concentrations reflects close relationship between obesity and systemic metabolism. Body fat especially around the waist - is known to represent an increased risk for the development of many chronic disorders associated to obesity (133, 139, 140). It is not only the concentration levels of serum

Table 5.1: Linear regression analysis in the NFBC1966 between body adipose index (BAI) and the serum metabolite measures

Metabolic measure	Abbreviation	Estimate (β)	P
Total lipids in chylomicrons	XXL-VLDL-L	26.8	**9.05E-50**
Total lipids in very largeVLDL	XL-VLDL-L	10.86	**5.47E-50**
Total lipids in large VLDL	L-VLDL-L	3.29	**2.69E-54**
Total lipids in medium VLDL	M-VLDL-L	2.54	**3.18E-62**
Total lipids in small VLDL	S-VLDL-L	3.81	**1.69E-77**
Total lipids in very small VLDL	XS-VLDL-L	5.22	**1.64E-59**
Total lipids in IDL	IDL-L	2.05	**1.59E-42**
Total lipids in large LDL	L-LDL-L	1.59	**4.72E-44**
Total lipids in medium LDL	M-LDL-L	2.77	**5.42E-52**
Total lipids in small LDL	S-LDL-L	4.88	**1.69E-67**
Total lipids in very large HDL	XL-HDL-L	-1.98	**5.22E-16**
Total lipids in large HDL	L-HDL-L	-2.17	**1.20E-59**
Total lipids in medium HDL	M-HDL-L	-1.01	**3.23E-8**
Total lipids in small HDL	S-HDL-L	0.99	**1.82E-4**
Apolipoprotein A-I	ApoA1	-0.66	2.51E-3
Apolipoprotein B	ApoB	3.99	**4.43E-70**
Mean diameter for VLDL particles	VLDL-D	0.55	**9.06E-46**
Mean diameter for LDL particles	LDL-D	-2.2	**3.81E-13**
Mean diameter for HDL particles	HDL-D	-3.75	**2.92E-62**
3-hydroxybutyrate	bOHBut	-0.22	0.6
Acetate	Ace	-7.05	2.46E-3
Acetoacetate	AcAce	-0.49	0.66
Alanine	Ala	5.99	**1.42E-20**
Citrate	Cit	-5.92	0.02
Creatinine	Crea	-14.67	**1.18E-5**
Glucose	Glc	0.22	**1.25E-4**
Glutamine	Gln	-3.59	**7.19E-12**
Glycerol	Glol	18.85	**1.65E-39**
Glycoprotein acetyls	Gp	3.21	**2.40E-62**
Histidine	His	3.4	0.38
Isoleucine	Ile	54.41	**9.72E-62**
Lactate	Lac	0.83	**5.70E-15**
Leucine	Leu	41	**6.62E-54**
Phenylalanine	Phe	66.87	**1.34E-89**
Pyruvate	Pyr	12.91	**5.44E-12**
Tyrosine	Tyr	57.1	**2.32E-43**
Urea	Urea	0.3	0.87
Valine	Val	15.89	**4.73E-45**

A P-value given in bold indicates a statistically significant association between BAI and a metabolite measure with a Bonferroni corrected threshold, $P < 0.01/M$, where $M = 38$ is the total number of metabolites included in the analyses. Estimate(β), estimate BAI effect size.

Table 5.2: Linear regression analysis in the NFBC1986 between body adipose index (BAI) and the serum metabolite measures

Metabolic measure	Abbreviation	Estimate (β)	P
Total lipids in chylomicrons	XXL-VLDL-L	37.09	**1.51E-45**
Total lipids in very large VLDL	XL-VLDL-L	14.88	**6.25E-52**
Total lipids in large VLDL	L-VLDL-L	4.43	**6.05E-55**
Total lipids in medium VLDL	M-VLDL-L	3.3	**2.59E-60**
Total lipids in small VLDL	S-VLDL-L	4.9	**8.89E-72**
Total lipids in very small VLDL	XS-VLDL-L	6.59	**1.31E-51**
Total lipids in IDL	IDL-L	2.9	**1.02E-44**
Total lipids in large LDL	L-LDL-L	2.41	**6.28E-50**
Total lipids in medium LDL	M-LDL-L	4.11	**4.62E-56**
Total lipids in small LDL	S-LDL-L	6.69	**4.93E-64**
Total lipids in very large HDL	XL-HDL-L	-3	**8.86E-29**
Total lipids in large HDL	L-HDL-L	-2.15	**5.87E-36**
Total lipids in medium HDL	M-HDL-L	0.89	7.45E-4
Total lipids in small HDL	S-HDL-L	4.16	**4.46E-29**
Apolipoprotein A-I	ApoA1	0.03	0.91
Apolipoprotein B	ApoB	5.19	**2.02E-74**
Mean diameter for VLDL particles	VLDL-D	0.36	**5.16E-22**
Mean diameter for LDL particles	LDL-D	-1.4	**2.91E-7**
Mean diameter for HDL particles	HDL-D	-3.5	**1.07E-53**
3-hydroxybutyrate	bOHBut	-4.24E-6	3.06E-3
Acetate	Ace	-6.32E-5	**2.62E-5**
Acetoacetate	AcAce	-7.55E-6	0.03
Alanine	Ala	1.19E-5	**1.28E-6**
Citrate	Cit	-5.92E-5	**9.21E-24**
Creatinine	Crea	-9.76E-6	0.55
Glucose	Glc	1.30E-6	0.07
Glutamine	Gln	-1.77E-5	**4.82E-15**
Glycerol	Glol	4.03E-5	**1.69E-13**
Glycoprotein acetyls	Gp	1.24E-5	**1.17E-64**
Histidine	His	9.24E-5	**1.39E-8**
Isoleucine	Ile	1.46E-4	**1.20E-32**
Leucine	Leu	1.17E-4	**6.91E-33**
Phenylalanine	Phe	2.17E-4	**1.66E-52**
Pyruvate	Pyr	3.38E-5	**3.25E-5**
Tyrosine	Tyr	1.37E-4	**2.99E-21**
Urea	Urea	-9.62E-6	0.15
Valine	Val	3.36E-5	**1.71E-15**

A P-value given in bold indicates a statistically significant association between BAI and a metabolite measure with a Bonferroni corrected threshold, $P < 0.01/M$, where $M = 38$ is the total number of metabolites included in the analyses. Estimate (β), estimate BAI effect size.

metabolites that can change due to variations in total body fat percentage but also the dependencies between metabolites can be affected. To investigate how metabolic associations differ between non-obese and obese individuals and to what extent these differences can serve as an indicator of important molecular alteration between these physiological states, we employ the GEMINi approach.

5.4.2 GEMINi

Differential Networks: Normal and Obese individuals

Differential network analysis is performed with the aim of determining the subset of metabolic associations that significantly differ between non-obese and obese individuals. We examine the same metabolic variables from the two independent cohorts (NFBC1966 and NFBC1986), and use estimates of partial correlation as measures of metabolic dependency. Metabolic associations that are consistently different across physiological conditions in the two study cohorts are interpreted as obesity-related molecular characteristics and are further investigated in the genome-wide correlation study and pathway analysis.

The differential analysis for BAI is presented in Figure 5.2. Each connection in the differential network indicates that a significant difference is found between the two physiological conditions in the pair-wise correlation between two variables. To assess the robustness of differences in the serum metabolite associations, we compare the two differential networks (Figure 5.2c). The differential network for NFBC1966 (Figure 5.2a) consists of a set of 30 connected variables, which are organized in 46 pair-wise differential interactions. The differential network for NFBC1986 (Figure 5.2b) comprises a group of 28 differentially connected variables, which are organized in 27 pair-wise differential interactions. We observe similar patterns of differences in the metabolic associations in these two differential networks. For both networks (Figure 5.2), the majority of the connected variables represent measures of very-low-density lipoproteins (dark green nodes), low-density lipoproteins (light blue nodes), high-density lipoproteins (light-blue nodes) and amino

acids (pink nodes). In addition, three metabolites - 3-hydroxybutyrate , acetoacetate and valine - show no connections in either of the two networks. Ten overlapping connections are observed between the two networks (Figure 5.2c). This overlap is significantly higher than expected at random. The overlapping connections highlight differences in the pair-wise associations between metabolites which occur throughout all size ranges of VLDL and HDL (apart from very-large HDL), two measures of low-density lipoproteins (S-LDL-L and LDL-D) and two amino acids (tyrosine and glutamine). Moreover, not only are a significant number of connected components overlapping between the two networks but also the signs of the difference in partial correlations are identical (represented as connection colours). We observe that the correlation is significantly lowered (blue edges) in the obese group for the pairs: very-large and large VLDL, extremely-large VLDL and small HDL, mean diameter for VLDL and large HDL and tyrosine and glutamine. On the other hand, the correlation is higher in the obese group for small LDL and mean diameter for LDL, ApoB and mean diameter for VLDL, large VLDL and small VLDL and very large and medium VLDL (red edges). VLDL-related measures present the highest number of connections in overlapping connected component. The equivalent analysis for BMI is presented in Figure D.1 and shows broadly similar features.

Differences in metabolic associations and obesity

A number of obesity-related differences in metabolic associations identified here are in line with prior studies on obesity-related traits. For instance, the central position and abundance of VLDL-related measures in the differential networks may reflect disregulation in VLDL metabolism. These alterations may result from the shift from small VLDL particles to enlarged VLDL (126). There is also evidence to suggest that in obesity-related dyslipidemia an overproduction of large VLDL particles is the initiator of all sequences of lipoprotein changes, resulting in a decrease in levels and particle size of HDL and smaller-denser LDL particles (34). Previous studies on obesity-related dyslipidaemias have recognised a complex pattern of change

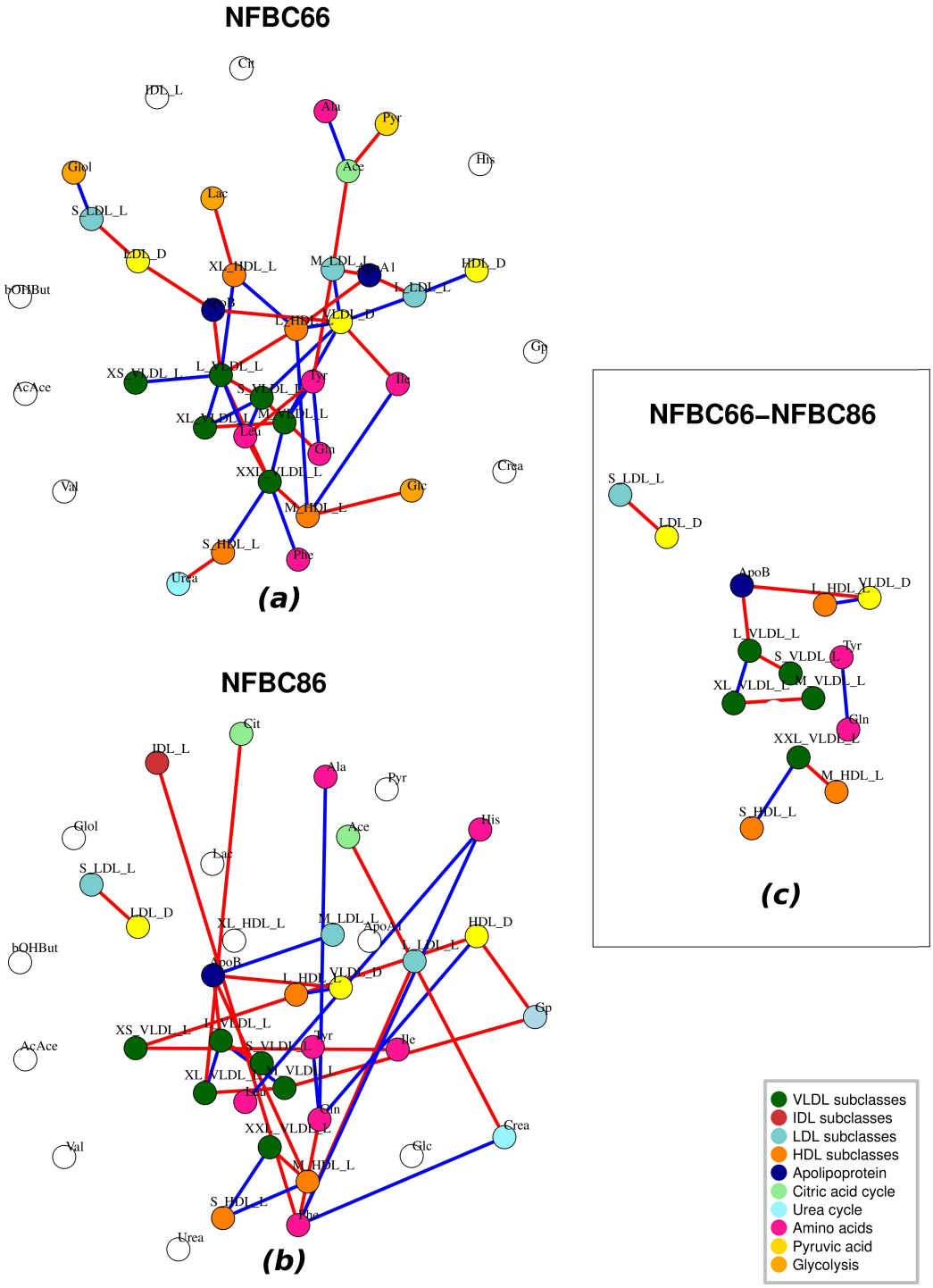


Figure 5.2: Differential network analysis for serum metabolites inferred from normal and obese individuals. (a) Differential Network inferred from NFBC1966; (b) differential network inferred from NFBC1986; (c) Replicated differential connection between NFBC1966 and NFBC1986 differential networks. Individuals are classified using body adipose index (BAI) definitions.

in size and particle concentration within the major lipoprotein classes in patients with obesity, especially abdominal obesity (141, 142). These disorders are related to elevated VLDL triglycerides that trigger the shift from small VLDL to medium and large VLDL, increased levels of small LDL and reduction in concentration and size of HDL. These variations in lipoprotein particle size subclasses, which relate to dyslipidaemias associated with obesity, are captured by the differential network of lipoprotein associations.

Of the low-molecular-weight metabolites, six amino acids were repeatedly involved in the differential connections. Of these, two are branched chain amino acids (BCAAs) leucine and isoleucine, two are aromatic amino acids - phenylalanine and tyrosine, and two represent non- essential amino acids - alanine and glutamine. Although only one of these connections is overlapping between the two cohort (tyrosine and glutamine), the repeated presence of this set of amino acids in the networks may indicate the robust association between obesity and alterations in amino acid metabolism. In fact, the implication of the BCAAs in obesity-related metabolic disorders has been previously reported. For instance, several studies in animal models have suggested that BCAAs play a role in a variety of important biochemical functions such as regulation of satiety, leptin, glucose and cell signing (143, 144). Moreover, studies on humans have highlighted the increased levels of BCAAs in obese individuals (diabetic and non-diabetic) (33, 145). It has been suggested that the specific BCAAs perturbed in obese individuals may be due to a loss of catabolism capacity, which may lead to the increased levels of BCAAs (146). Another interesting connection, although not replicated between cohorts, is the differential associations between alanine and glutamine observed in NFBC1986 network (Figure 5.2b). Among other functions, glutamine has been reported to play a role in glucose and glycogen metabolism (147). It has been suggested that transamination of pyruvate to alanine may be increased by a high rate of flux through BCAA catabolic pathways and accumulation of glutamate (33). Given that alanine is a gluconeogenic amino acid, its disregulation could be a contributor to the development of glucose intolerance in obesity.

The large number of independently replicated results and the known rel-

evance of the identified metabolites to obesity provide strong support for differential association analysis in improving our understanding of the biological mechanisms of disease. Assuming that these metabolic differences are related to obesity, we hypothesized that these measures could be used as an additional phenotype to identify genetic variants associated to obesity-related traits. To test this hypothesis, we performed a genome-wide correlation analysis.

Genome-wide correlation analysis

The associations of 318,443 SNPs with the metabolite-metabolite partial correlations were individually tested. Our analysis revealed 37 loci significantly associated ($P < \frac{5\times10^{-8}}{D}$), where $D = 10$, with differences in partial correlation between two pairs of metabolites, namely total lipids in large VLDL and very-large VLDL, as well as total lipids in medium VLDL and very-large VLDL. A brief description of these genetic variants, including their accession numbers, chromosomal positions, p-values obtained, along with candidate genes associated with the genetic marker is presented in Table 5.3.

Of the 37 loci identified, 16 are related to differences in the association between large and very large VLDL total lipids and 30 are associated with the association between total lipids in medium and very-large VLDL. Nine loci are shared between these two metabolic associations. Several of the genes associated with these loci have been reported to be involved in obesity-related processes. For instance, among the genetic variants associated with differences in the association between large and very-large VLDL, we identify a SNP (rs1078248) at chromosome position 3p21.33, which is localized within 1 Mb from a gene encoding the abhydrolase domain containing 5 protein (ABHD5/CG1-58). The protein coded by this gene is known to be linked to triglyceride (TAG) metabolism, the primarily energy source in vertebrates (148). TAGs are stored in adipose tissue and hydrolyzed into fatty acids and glycerol and can be used by body tissues during times of fast or energy deprivation (149). Mutations in the gene encoding for ABHD5 have been associated with a triglyceride storage disease with impaired long-chain

Table 5.3: Genetic loci associated with differences in metabolite-metabolite correlations between obese and non-obese individuals.

^{1}P	^{2}P	SNP	Gene	Chr	Pos
1.19E-9	2.56E-7	rs1334806	USP1	1	62875313
9.37E-14	$< 2.00E-16$	rs3817222	PPP1R12B	1	202464760
1.01E-13	$< 2.00E-16$	rs12743401	PPP1R12B	1	202476648
7.33E-14	$< 2.00E-16$	rs3881953	PPP1R12B	1	202528021
1.93E-5	**8.50E-10**	rs4669444	YWHAQ	2	9869267
1.23E-9	**7.44E-9**	rs7592040	OTOF	2	26741551
2.11E-7	**3.02E-9**	rs966859	DPP10	2	115828493
2.02E-9	3.05E-6	rs1078248	ABHD5	3	43794357
9.96E-9	**1.19E-10**	rs2175730	TRH	3	129677772
1.38E-8	**1.64E-10**	rs9856606	TRH	3	129680366
1.72E-6	**5.15E-10**	rs10517060	CCDC149	4	24929280
1.85E-6	**1.71E-10**	rs7719640	IRX1	5	3752812
1.30E-12	**8.88E-16**	rs34868670	PTGER4	5	40237843
1.64E-9	**1.33E-10**	rs573496	GRIA1	5	152893433
2.11E-6	**9.07E-11**	rs11957360	SCGB3A1	5	180010639
2.72E-8	**8.04E-11**	rs681283	CRISP2	6	49673195
6.85E-8	**1.64E-11**	rs1319087	EYA4	6	133897107
7.77E-10	**2.97E-10**	rs2500542	UST	6	149331204
1.69E-5	**3.98E-10**	rs1403185	DLX6	7	96540655
5.36E-10	4.17E-7	rs1454953	DLC1	8	13315983
2.00E-10	2.14E-7	rs7814428	DLC1	8	13324439
4.29E-9	**4.44E-10**	rs7357357	FAM84B	8	127189138
6.21E-9	**4.54E-10**	rs10094775	FAM84B	8	127197554
1.92E-9	**4.89E-9**	rs4557669	FAM84B	8	127306602
4.95E-6	**6.07E-12**	rs10811316	MLLT3	9	20095103
1.20E-6	**6.41E-10**	rs1014807	PIP5K1B	9	71302196
5.08E-6	**9.78E-10**	rs2386636	ASB13	10	5712087
2.58E-7	**4.79E-11**	rs7923418	ZNF365	10	64230159
7.77E-7	**1.47E-9**	rs1727	LIPA	10	91066769
3.81E-10	1.38E-5	rs17315588	RP11-494M8.4	11	7913911
1.57E-9	**1.43E-9**	rs9971532	EED	11	85929490
9.48E-8	**2.68E-9**	rs7110414	GRIA4	11	105861507
3.38E-8	**1.20E-11**	rs484580	CLMP	11	123109469
3.36E-8	**1.23E-11**	rs558362	CLMP	11	123114359
2.86E-9	3.89E-8	rs1650307	LDH8	12	21803770
2.82E-7	**5.85E-10**	rs11175596	TBC1D30	12	65340828
4.49E-13	**2.22E-16**	rs2451078	TPTE2	13	20098289
8.60E-6	**1.57E-9**	rs2224442	VRK1	14	97272382
2.67E-9	**5.33E-10**	rs2302553	NAGPA	16	5056295
8.15E-10	**1.70E-10**	rs3815490	NAGPA	16	5060568
1.15E-7	**1.99E-9**	rs2072620	ZP2	16	21216033
4.26E-5	**2.89E-10**	rs9303619	RP11.173M1.1	17	25446389
9.26E-10	1.47E-8	rs8106239	ATP4A	19	36085358
1.90E-6	**2.87E-9**	rs844879	CD93	20	23102665
7.22E-10	2.57E-7	rs2093066	BPIFB3	20	31652596
6.84E-10	1.47E-7	rs378098	BPIFB3	20	31660543

^1P, p-value for association between locus and variation in correlation between XL-VLDL/L-VLDL; ^2P, p-value for association between locus and variation in correlation between XL-VLDL/M-VLDL; **Chr**, chromosome; **Pos**, SNP position in NCBI human genome build 36; **Gene**, potential candidate gene in the region. Bold indicates significant association between genetic variant and variation in association between pairs of metabolites using genome-wide significance level threshold.

fatty acid oxidation called Chanarin-Dorfman syndrome (150, 151). Despite the fact that ABHD5 is involved in TGAs metabolism and highly expressed in adipose tissue, patients with Chanarin-Dorfman syndrome are not obese (152). No previous association between ABHD5 and obesity-related traits has been reported.

Among the genetic variants associated with differences in association between medium and very-large VLDL, we found a marker (rs966859) at 2q14.1, which is located in a gene encoding for inactive dipeptidyl peptidase 10 (DPP10). Genetic variants in DPP10 are associated with susceptibility to asthma (153). Asthma is a complex disorder with a heterogeneous phenotype, largely attributed to the combined effect of multiple genes (154) and environment. A number of studies have shown a consistent association between asthma and obesity (155, 156). This strong relationship has led to the hypothesis that both disorders may be related due to shared genetic factors (157, 158). This hypothesis was tested in a candidate gene study where known genes for BMI and asthma were jointly analysed (158). Several of these genes, including DPP10, showed shared associations between these two conditions, which supports the hypothesis of shared genetics between asthma and obesity. However, after correction for multiple testing none of the identified associations remained significant (158). Two other markers (rs2175730, rs9856606) at 3q22.1 are also associated with the difference in the association between medium and very-large VLDL. These SNPs are localized near a gene encoding thyrotropin-releasing hormone (TRH). A number of studies have indicated the important role that TRH plays in central regulation of the cardiovascular system and hypertension induced by body weight gain (159, 160). This function appears to be mediated by leptin (161), an adipocyte-released hormone, and it has been suggested that the interaction between TRH and leptin is an important contributor to the strong association existing between hypertension and obesity (161). Another variant associated with the difference in the association between medium and very-large VLDL is SNP rs1727 at chromosome 10q23.3. This marker is localized within a gene encoding for lipase A, lysosomal acid, cholesterol esterase (LIPA). This enzyme is found in the lysosomes and plays a role in the hydrolysis of lipids such as cholesteryl

esters and triglycerides. Cholesteryl ester storage disease (CESD) is a rare recessive disorder associated with lysosomal acid lipase deficient activity. Patients with CESD are characterized by a massive accumulation of cholesteryl esters and triglycerides in most tissues of the body (162). This disorder is known to be due to mutations in LIPA, which is gene encoding the lysosomal acidic lipase (163). Moreover, a recent study on subjects with severe obesity has reported significant associations between a number of LIPA genetic variants and obesity-related metabolic complications such as cholesterol in HDL (164). Two more genetic variants were associated with the difference in the association between medium and very-large VLDL (rs1192198 and rs558362) at chromosome 11q24.1. These genetic variants are localized near a gene encoding for a CXADR-like membrane protein (CLMP). It has been suggested that CLMP may play a role in adipose tissue differentiation (165) and be involved in cell-cell adhesion (166). Of the genes associated with the difference in the association between both pairs of lipoprotein fractions, we identified SNP rs573496 at chromosome 5q33.2. This SNP is localized within a gene encoding for glutamate receptor 1 (GRIA1). In addition to this SNP, we also identified SNP rs7110414 at chromosome 11 which is within 1Mb distance from a gene encoding for glutamate receptor 4 GRIA4. Glutamate receptors are protein complexes formed by the combination of four different subunits (GRIA1-GRIA4) and are known to be the main excitatory neurotransmitter receptors in the mammalian brain. Regulation of glutamate receptors is influenced by leptin levels (167), as discussed below.

Pathway-level analysis

To further the interpretation of the genetic findings, the set of identified genetic markers was tested in aggregate using pathway-level approaches. The aim is to assess the combined contribution of the identified genetic variants and to determine to what extent the products of the associated genes act within obesity-related pathways. For this purpose we apply an over- representation analysis using the web tool IMPALA. The top three pathways for the general analysis are presented in Table 5.4. Of these, the top-ranked

pathway is the activation of AMPA receptors pathway, with 4 genes in the pathway ($P = 2.34 \times 10^{-5}$).

Examination of the activation of AMPA receptors pathway shows that a subset of two genes from our selected list of genes GRIA1, GRIA4 - belong to this pathway. AMPA receptors are members of the ionotropic class of glutamate receptors. It is well established that AMPA receptors are the predominant regulators of the fast synaptic transmission in the central nervous system (CNS) (168). Evidence is growing that glutamate receptor trafficking processes can be regulated by the hormone leptin (167, 168, 169). This hormone is predominantly synthesized in adipose tissue and it is known that leptin plays an important role in the regulation of appetite and control of body weight (170). It is also well established that there is an inverse relationship between plasma levels of leptin and food intake (171). In addition, leptin secretion is proportional to body fat content (172). Alterations in the leptin system are also linked to a number of CNS-driven diseases and neurodegenerative disorders (173). It has been hypothesized that regulation of AMPA receptor trafficking by leptin is likely to have important implications not only in health but also in diseases associated with leptin dysfunction (172). In addition, it has been hypothesized that AMPA receptors may play a pivotal role in the comorbidity between depression and cardiovascular disorders (174).

Table 5.4: Pathway analysis of the whole set of genetic variants identified by the GWC analysis of variation in metabolic associations.

Pathway	**Pathway source**	**N.o**	**Overlapping genes**	**N.p**	**P**	**Q**
Activation of AMPA receptors	Reactome	2	GRIA1; GRIA4	4	2.34e-5	0.0752
Trafficking of GluR2-containing AMPA receptors	Reactome	2	GRIA1; GRIA4	16	0.000461	0.559
Unblocking of NMDA receptor	Reactome	2	GRIA1; GRIA4	17	0.000522	0.559

N.o, Number of overlapping genes; **N.p,** Number of all pathway genes;**P,** p-value; **Q,** q-value

5.5 Summary

Our aim in this chapter was to investigate the use of additional phenotypic variations such as differences in associations between metabolic measures to aid the discovery of genetic factors involved in complex traits. We have developed a novel analytical approach for the combined analysis of metabolic and genetic information.

Initially, we have examined the relationship between obesity and serum metabolites by performing a differential analysis of metabolic associations. Using this approach a number of obesity-related characteristics were identified comprising several metabolites linked by previous studies to the development of obesity-related disorders. Metabolites involved in such differential connections included all sizes of very low density lipoprotein (VLDL) and high density lipoprotein (HDL) as well as several amino acids such as tyrosine and glutamine. A significant proportion of these findings overlapped between two independent cohorts. In the light of these results, we postulated that differences in metabolite associations could potentially be used as an additional phenotype to investigate genetic factors in obesity. We further investigated this metabolic differences by performing a genome-wide correlation analysis where genetic variants were tested for association with the metabolic correlations. These analysis revealed a large set of associations between genetic variants and these differential metabolic dependencies. Several of the associated genes have been previously implicated in obesity-related traits.

A clear limitation of our study is lack of a replication cohort for our genetic findings. However, the fact that we found no associations following standard GWAS, but did find a large number of associations using our genome-wide correlation approach, many of which are implicated in obesity, is very promising. This study therefore highlights the importance of investigating differences in correlation between metabolites or other phenotypic indicators as an additional phenotype to identify novel loci with key roles in the pathogenesis of complex traits.

Chapter 6

Variance and Covariance Regression Models

In chapter 4, we have highlighted the importance of exploring differences in molecular association patterns in relation to different physiological conditions to identify key features of phenotype variations. In chapter 5, we have showed the applicability of using changes in correlation between metabolites as an additional phenotype to identify novel loci with key roles in the pathogenesis of complex traits. Yet, despite their potential, these two approaches have the limitation of being based on a two-groups comparison - healthy and non-healthy - where individuals are classified based on cut points for disease state such as impaired fasting glucose (IFG) for prediabetes or body adipose index (BAI) for obesity. Considering the data as binary phenotype might ignore some important details on the underlying biological mechanisms of phenotype variation, providing a limited view into the continuous process of disease development.

With this in mind, the focus in this chapter is on investigating continuous changes in molecular traits as response to a pathogenic stimulus. In particular, we propose to employ Bayesian hierarchical models to investigate changes in mean level, variance and correlation of metabolic traits in relation to covariates. To illustrate our approach we investigate changes in urinary metabolic traits in response to cadmium exposure, a toxic environmental pollutant. The chapter starts with section 6.1, where we give a short introduction to the problem under consideration. Next, we describe the dataset used for statistical inference (section 6.2) and then define the

statistical framework proposed in this work (section 6.3). The results of the analysis are summarized in section 6.4.

6.1 Introduction

Over the last decade, our understanding of the pathogenesis of many complex disorders has expanded considerably. It has become clear that conditions such as obesity, diabetes or chronic kidney disorders stem from a web of complex interactions between genetic and environmental factors, resulting in increased predisposition to a clinical phenotype. In the presence of pathogenic factors, a progressive cascade of subtle abnormalities at molecular level can be observed. Many of these abnormalities begin early during the course of the disease development and progressively continue until they manifest as clinical disease. Since the pathogenesis of complex conditions are continuous in nature (175), considering data as continuous phenotype, rather than the conventional two-groups comparison, might provide a much detailed view into biological processes and potentially reveal key features of disease development.

From the statistical point of view, it can be expected that if a particular factor (physiological or environmental) has an effect on a specific trait, the marginal mean of such a trait would be modified at increasing exposure to such a stimulus. In addition to effects on mean values of a quantitative trait, effects on other statistical properties such as the variance or covariance could also be expected. Hence investigating changes of additional molecular properties can be used as an approach to detect potentially novel effects on a trait of interest. Several genetic studies have already explored the idea of testing for equality and heterogeneity of variances of quantitative traits to further characterize the consequence and risk of genetic variations (129, 130). The problem of characterising and modelling the conditional mean is a well-studied topic, both in the univariate and multivariate cases. While there is also a relatively rich literature in methods allowing the variance to vary across different values of an explanatory variable (176), (177), little has been done in the case of multivariate covariance regression. Recently Hoff and

Niu (178) introduced a covariance regression method which parametrises the covariance of a multivariate response vector as a function of explanatory variables. Unlike methods which characterize the heteroscedasticity in the form of the mean-variance relationship, the covariance regression model proposed by Hoff and Niu (178) allows for the mean function μ_x and the covariance function Σ_x to be separately parameterized.

In this study, we propose a novel statitical framework for detecting progressive changes in molecular traits as response to a pathogenic stimulus. To illustrate our approach, we use urinary metabolic data on individuals living near an industrial source of cadmium. Cadmium (Cd) is a highly toxic metal commonly found in industrial workplaces and it is known to have important health effects. We employ Bayesian hierarchical models to investigate variations in mean levels and variances of a set of urinary metabolic traits across a range of values of urinary cadmium. In addition, we investigate variations in correlations between pairs of metabolic traits in relation to covariates. To assess the potential of this approach to reveal key molecular changes related to pathogenic stimuli, we compare our results with existing knowledge of cadmium-induced physiological and biochemical effects.

6.2 Material

• Study Populations

We used data from the Cadmium study population. Urinary metabolic data were available on 127 individuals from the Cadmium study population.

• ^{1}H NMR urinary metabolites

The analysis was performed using a total of 22 urinary metabolic measures including amino acids, 3-hydroxybutyrate, creatinine and urea and information on metabolites involved in gluconeogenesis pathway such as pyruvate, lactate, glycerol. The complete list of metabolites included in this analysis can be found in Table A.2.

- **Cadmium exposure**

Data on urinary cadmium (U-Cd) was used to assess lifetime kidney accumulation from chronic cadmium exposure. Details on data acquisition can be found in Thomas et al. 2009 (50). Since urinary cadmium excretion levels increase with age, as consequence of increased body burden and also urinary cadmium excretion levels are higher in smokers than non-smokers and in women than men (50), data on biological and lifestyle factors including age, gender, smoking status (current, past, never) were used in the analysis.

6.3 Statistical Methods

Firstly, we perform a multiple linear regression to explore associations between levels of urinary metabolites and urinary levels of cadmium. Secondly, we perform a variance regression analysis for estimating changes in the variance of urinary metabolites across different values of urinary cadmium. Thirdly, we employ the covariance regression model proposed by Hoff and Niu (178) to estimate changes in pair-wise correlation between metabolites at different levels of U-Cd. To ensure normality of the metabolic data, each metabolic phenotype was log-transformed prior to analysis.

- **Association between levels of urinary metabolites and urinary cadmium**

Multiple linear regression was applied to investigate the association between levels of urinary metabolites and urinary cadmium. A set of lifestyle and biological factors - sex, age, and smoking status - expected to have a confounding effect on the concentration of levels of urinary cadmium was included into the model. We determine significant association between levels of urinary metabolites and urinary cadmium by setting the significance threshold to $P < 0.05$.

• Variance Regression Model

To investigate the potential effect of cadmium exposure on the mean and the variance of traits of interest across different values of an explanatory variable, we propose to use Bayesian hierarchical models. The variance regression model employed to estimate changes in the mean and variance of a set of urinary metabolic traits over a range of cadmium exposure can be defined as follows.

Let $\mathbf{Y}$ denote the $N \times K$ set of normally distributed random variables (metabolites) where N is the number of observations and K is the number of variables. For the purpose of this analysis, $Y_k = (y_{1k}, ..., y_{Nk})$ represents each of the urinary metabolic traits and y_{ik} is the ith observation for the variable Y_k. We assume

$$y_{ik} \sim N(\mu_{ik}, \sigma^2_{ik}) \quad \text{for} \quad i = 1, ..., N; k = 1, ..., K. \tag{6.1}$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with parameters μ and σ^2. The conditional mean $\mu_{ik} = E[y_{ik}|\mathbf{x}_i]$ can be expressed as follows:

$$E[y_{ik}|\mathbf{x}_i] = \boldsymbol{\beta}_k \mathbf{x}_i \quad \text{for} \quad i = 1, ..., N; k = 1, ..., K. \tag{6.2}$$

where $\mathbf{x}_i$ represents a q-dimensional vector of covariates for individual i whose first entry is the intercept and $\boldsymbol{\beta}_k$ is a q-dimensional vector of parameters. Similarly to the conditional mean, the variance $\sigma^2_{ik} = Var[y_{ik}|\mathbf{z}_i]$, can be expressed in terms of a linear regression model as:

$$\log[Var[y_{ik}|\mathbf{z}_i]] = \boldsymbol{\lambda}_k \mathbf{z}_i \quad \text{for} \quad i = 1, ..., N; k = 1, ..., K. \tag{6.3}$$

where $\mathbf{z}_i$ is a p-dimensional vector of covariates for individual ith whose first entry is the intercept and $\boldsymbol{\lambda}_k$ is a p-dimensional vector of parameters. The $\mathbf{z}_i$ of equation 6.3 and $\mathbf{x}_i$ of equation 6.2 can be the same. Each regression is fitted individually per metabolite. To estimate the model parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_k)$, we perform a Bayesian analysis. In Bayesian statistical inference, a prior distribution for all unknowns parameters $p(\boldsymbol{\beta}, \boldsymbol{\lambda})$

has to be specified. In this analysis we follow a non-informative prior for all unknowns parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. The prior density for the parameters $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta}_k \sim N_q(\mu_\beta, \Sigma_\beta)$$

with $\mu_\beta = 0$ and $\Sigma_\beta = \tau_\beta I$. The prior density of $\boldsymbol{\lambda}$ is:

$$\boldsymbol{\lambda}_k \sim N_p(\mu_\lambda, \Sigma_\lambda)$$

with $\mu_\lambda = 0$ and $\Sigma_\lambda = \tau_\lambda I$. The prior for τ_β and τ_λ for the Inverse-Gamma variance were set to 0.1 and 1. To check posterior sensitivity to these pre-specified values, we analysed the data with several other values (0.01, 0.01 and 0.001, 0.001). From the prior distribution and the likelihood function together, the posterior distribution of the parameters can then be obtained as:

$$p(\boldsymbol{\beta}_k, \boldsymbol{\lambda}_k | Y_k, \mathbf{X_k}, \mathbf{Z_k}) \propto p(Y_k | \mathbf{X_k}, \mathbf{Z_k}, \boldsymbol{\beta}_k, \boldsymbol{\lambda}_k) p(\boldsymbol{\beta}_k) p(\boldsymbol{\lambda}_k)$$

where $\mathbf{X_k}$ is the $N \times q$ matrix of predictors for the mean and $\mathbf{Z_k}$ is the $N \times p$ matrix of predictors for the variance. Because the posterior distribution is not available in closed-form, a MCMC approximation of the joint posterior distribution can obtained via Gibbs sampling. We used JAGS linked to R.2.14.1 via the *R2jags* package to perform this analysis. We run 50,000 iterations of the Gibss sampler with the first 1,000 discarded as burn-in. To determine if the variance of a urinary metabolite Y_k changes with cadmium exposure we calculate the posterior 95% credible intervals of the parameters $\boldsymbol{\lambda}_k$.

• Covariance Regression Model

The effect of urinary levels of cadmium on the pair-wise association between metabolic variables is investigated by using the covariance regression model developed by Hoff and Niu (178). This methodology provides a parsimonious model and estimation method for the conditional covariance matrix of a multiple response vector $\boldsymbol{Y}$ over a range of values of an explanatory variable

$\mathbf{x}$, here levels of urinary cadmium.
Let now consider two continuous random variables $Y_{k_1} = (y_{1,k_1}, ..., y_{N,k_1})$ and $Y_{k_2} = (y_{1,k_2}, ..., y_{N,k_2})$ where $1, ..., N$ is the number of observations. The covariance regression model proposed by Hoff and Niu (178) is based on an analogy with linear regression and starts by expressing the conditional mean in terms of linear regression:

$$E[Y_{k_J}|\mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta}_J \quad \text{for} \quad J = 1, 2 \tag{6.4}$$

where $\mathbf{x}$ is a q-dimensional vector of explanatory variables whose first entry is the intercept and $\boldsymbol{\beta}$ is q-dimensional vector of parameters. The simplest version of the proposed covariance model can be expressed as $\text{Cov}[\mathbf{y}|\mathbf{x}] = \boldsymbol{\Sigma}_\mathbf{x}$ with:

$$\text{Var}[Y_{k_1}|\mathbf{x}] = \psi_{k_1,k_1} + \mathbf{b}_{k_1}^T\mathbf{x}\mathbf{x}^T\mathbf{b}_{k_1} \tag{6.5}$$

$$\text{Cov}[Y_{k_1}, Y_{k_2}|\mathbf{x}] = \psi_{k_1,k_2} + \mathbf{b}_{k_1}^T\mathbf{x}\mathbf{x}^T\mathbf{b}_{k_2} \tag{6.6}$$

where ψ_{k_1,k_1} and ψ_{k_1,k_2} are elements of $\boldsymbol{\Psi}$, a 2×2 positive-definite baseline covariance matrix and $\mathbf{b}_{k_1}, \mathbf{b}_{k_2}$ are the rows of $\mathbf{B}$, a $2 \times q$ matrix of parameters. As described in Hoff and Niu (178), the parametrization of the variance suggests that the models requires the variance of each element of $\boldsymbol{Y}$ to increase in the elements of $\mathbf{x}$, as the minimum variance is obtained when $\mathbf{x} = 0$. As proposed in Hoff and Niu (178), this constraint can be relaxed by including an intercept term.

To estimate the model parameters represented in Equation 6.6, we use a MCMC approximation via a Gibbs sampling. We follow Hoff and Niu (178) and specified the prior distribution for the unknown parameters $p(\boldsymbol{\beta}, \mathbf{B}, \boldsymbol{\Psi})$ in terms of the following semi-conjugate priors:

$$\begin{aligned}
\boldsymbol{\Psi} &\sim \text{Inverse} - \text{Wishart}(\boldsymbol{\Psi}_0^{-1}, v_0), \\
\boldsymbol{\beta} &\sim N(0, 100 \cdot \mathbf{I}_2), \\
\mathbf{B} &\sim M(\mathbf{B}_0, \boldsymbol{\Psi} \cdot \mathbf{V}_0),
\end{aligned}$$

where $v_0 = p + 2$ degrees of freedom, Ψ_0 is the sample covariance matrix,

$\mathbf{B}_0 = 0$ and $\mathbf{V}_0 = N \times (\mathbf{X}^T\mathbf{X})^{-1}$. We run 50,000 iterations of the Gibbs sampler with the first 1,000 discarded. From the the estimated covariance, we calculate the correlation between metabolic traits. To determine variations in correlation between metabolic traits, we calculate the 95% credible intervals of the estimated correlation.

- **Covariance regression model networks**

To improve the interpretation of the covariance regression results, we construct seven correlation networks. Each network indicates the pair-wise correlation between metabolites under seven different levels of exposure to cadmium (U-Cd(nmol/mmol creatinine)=0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8). We define the edges in the networks by computing the 95% credible interval of the posterior probability distributions for the correlation conditioned on the urinary cadmium concentration. The networks are built by drawing the edges between those pairs of measures if the credible interval does not contain zero.

The networks were built and visualized using the R package **igraph**. All analysis were carried out using R version R.2.14.1.

6.4 Results and Discussion

Our first aim is to determine whether measures of cadmium exposure are associated with variations in urinary metabolic levels (U-Cd). Initial inspection of the data using a multiple linear regression shows that three of the urinary metabolites are associated with cadmium urinary levels after adjusting for sex, age and smoking status ($P < 0.05$) (Table 6.1). Two metabolites - succinate and dimethylglycine - are negatively associated with levels of U-Cd whereas citrate shows positive association with U-Cd. Two of these metabolites are involved in the citric acid cycle (succinate and citrate) whereas dimethylglycine is related to choline metabolism.

Our next objective is to evaluate if the variance of urinary metabolites changes across a range of values of urinary cadmium. The trace plot from the MCMC for the parameter $\boldsymbol{\lambda}$ is presented in Figure 6.1. It show that the

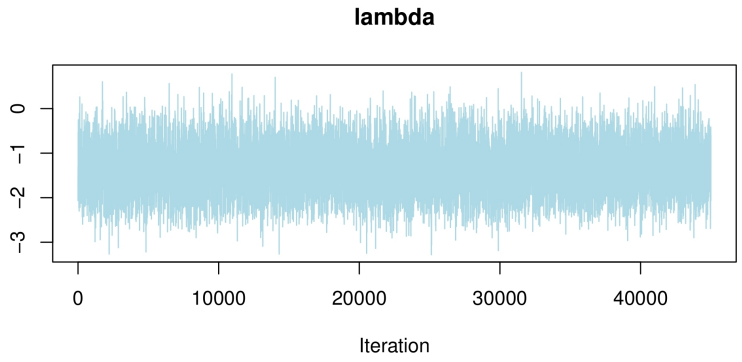
Table 6.1: Multiple linear regression analysis between exposure to cadmium (U-Cd) and the urinary metabolite measures.

Metabolites	Abbreviation	B	P
4-deoxyerythronic acid	4-DEA	144.8	0.49
3-hydroxyisovalerate	3-HV	-411.8	0.12
Lactate	Lac	-32.46	0.41
Alanine	Ala	152.5	0.22
Acetate	AcO	32.85	0.09
Pyruvate	Pyr	-96.1	0.34
Succinate	Suc	-316.6	**0.04**
Citrate	Cit	18.76	**0.04**
Dimethylamine	DMA	141.2	0.28
Trimethylamine	TMA	20.16	0.94
Dimethylglycine	DMG	-558.4	**0.03**
Creatine	Crea	43.12	0.26
Creatinine	Creat	-13.85	0.22
Proline betaine	PB	-6.34	0.81
P-cresol sulphate	PCS	73.56	0.69
Phenylacetylglutamine	PAG	37.02	0.53
Hippurate	Hip	9.35	0.49
N-methyl-nicotinic acid	NMNA	154.3	0.42
Trimethylamine oxide	TMAO	0.38	0.93
Glycine	Gly	10.99	0.78
Urea	Urea	-1.1	0.21
Formate	For	-386.4	0.35

B, Effect size; **P**, P-values for the multiple linear regression; Bold font indicates association between concentration levels of urinary metabolites and urinary metabolic levels ($P < 0.05$) adjusted for sex, age and smoking status.

parameter converges rapidly to the stationary distribution.

Figure 6.1: The trace plot from the MCMC for the parameter λ from 50,000 iterations after 1,000 burn-in period

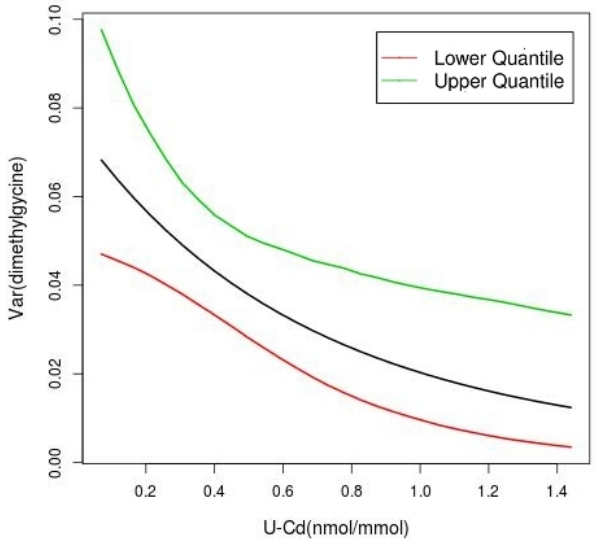


The variances regression fits are plotted as a function of urinary cadmium (Figure 6.2). Results for the 22 urinary metabolites are presented in appendix Table E.2. From this analysis, we observe that the variance of one of the urinary metabolites, dimethylglycine (DMG), changes across the range of values of urinary cadmium. Specifically, we observe a decrease in the variance of dimethylglycine as the levels of urinary cadmium increase (Figure 6.2).

In addition to changes in mean level and variance, we are interested in investigating whether correlation between metabolic traits changes in relation to urinary cadmium. To this end, we have employed the covariance regression model(178). Results for the 22 urinary metabolites are presented in the appendix Table E.3. The empirical correlation between metabolites as a function of urinary cadmium are showed in Figure 6.3 (blue lines). The pair-wise correlations calculated from the covariance regression fits are also presented in Figure 6.3 (black lines) with the 95% credible interval (green and red lines). Using the covariance regression model, we find that the associations between four pairs of urinary metabolites change across the range of values of urinary cadmium (Figure 6.3).

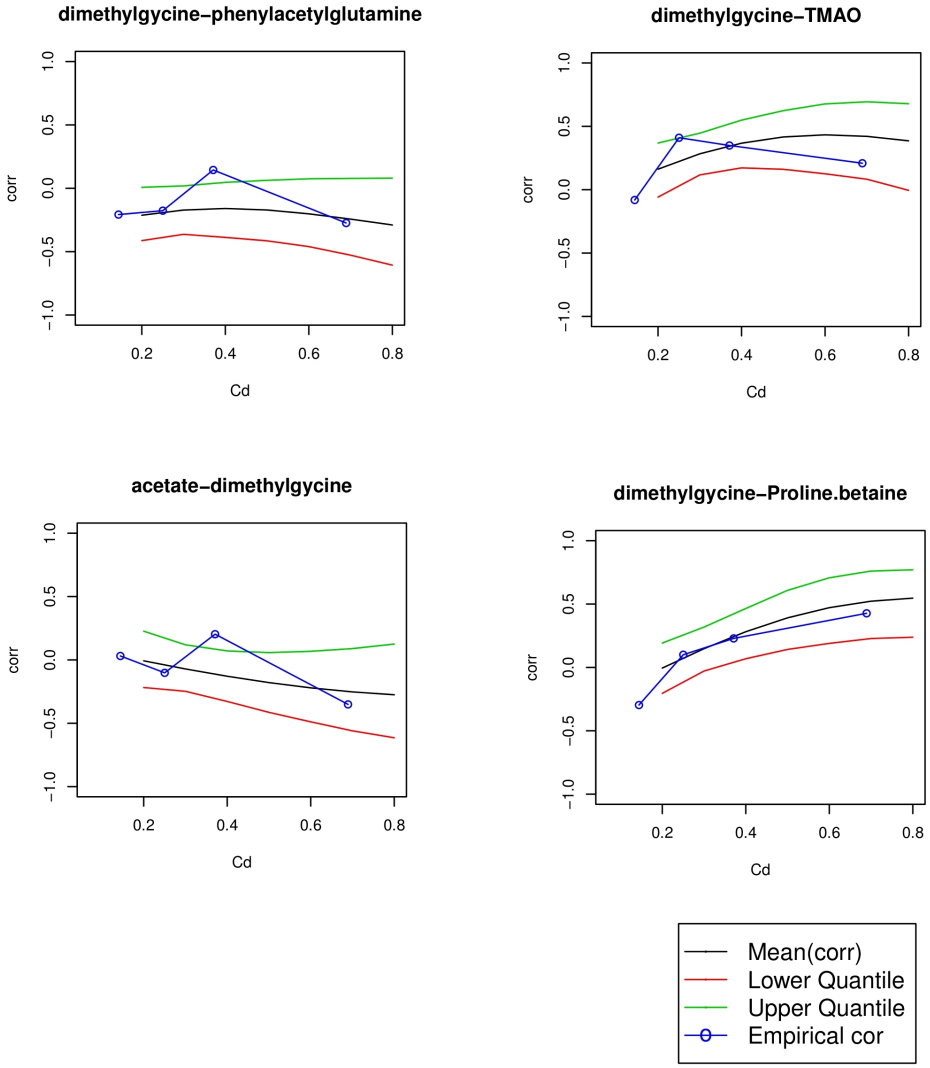
Interestingly, we observe that dimethylglycine is consistently involved in the four pairs of metabolites identified. This analysis indicates that correlation between dimethylglycine and phenylacetylglutamine and acetate de-

Figure 6.2: Variance regression for dimethylglycine



crease as levels of urinary cadmium increase. On the other hand, the pairwise association between dimethylglycine and trimethylamine oxide (TMAO) and dimethylglycine and Proline betaine (PB) increase as a function of urinary cadmium. As previously indicated, dimethylglycine is related to choline metabolism. In addition, two of the metabolites identified by the covariance regression analysis - acetate and trimethylamine oxide - are also related to choline metabolism (Figure 6.4). Specifically, in this metabolic process choline is metabolised to betaine (Figure 6.4.a). Betaine donates a methyl group to homocysteine to form methionine (Figure 6.4). This methylation reaction also generates dimethylglycine (Figure 6.4.b), which is further metabolized to glycine (Figure 6.4.c). Moreover, dietary choline forms trimethylamine (TMA) in a reaction catalyzed by enzymes within gut bacteria (179)(Figure 6.4.d). TMA together with acetate biosynthesizes trimethylamine-N-oxide (TMAO)(Figure 6.4.e). Dimethylglycine that is not metabolised in the liver is transported by the circulatory system to body tissue or it is excreted by urine. While association between mean levels of urinary dimethylglycine and urinary cadmium have been reported previously

Figure 6.3: Changes in pair-wise associations between urinary metabolites across the range of values of urinary cadmium



in a human population (56), these additional variations in the metabolism of dimethylglycine and related metabolites are novel.

To further explore the covariance regression results, we have constructed seven correlation networks (Figure 6.5) which indicate the pair-wise correlation between pairs of metabolites under different levels of urinary cadmium

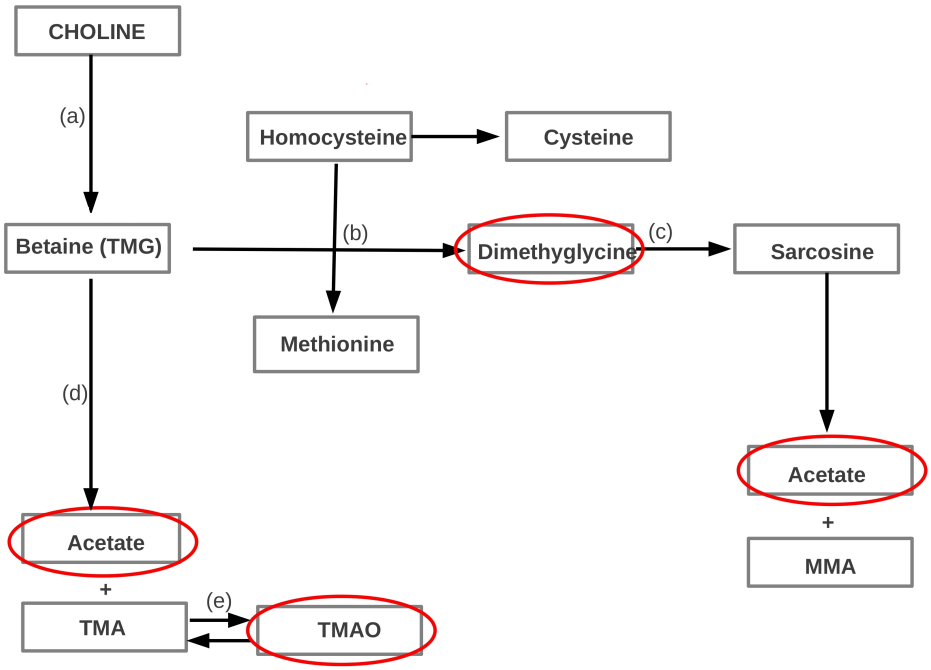


Figure 6.4: Graphic representation of choline metabolism as described in (180) and (181). Metabolites identified by the covariance regression analysis are highlighted in the graph with red continuous line. TMAO, trimethylamine oxide; TMA, trimethylamine; MMA, Methylamine.

(U-Cd (nmol/mmol creatinine)= 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8). To quantitatively describe the topological features of the correlation networks, we calculate the connectivity, centrality and betweenness for each node of the seven correlation networks. The connectivity for each of the nodes for different levels of urinary cadmium are presented in Table E.1. We can see that while a large number of these links are in common between the seven networks (Figure 6.5), there is an overall tendency to decrease the number of links as the level of urinary cadmium increases (Total Connectivity, Table E.1). Of the 22 urinary metabolites, dimethylglycine is the metabolic measure that shows the largest reduction in associations, from seven connections at urinary cadmium levels 0.2 nmol/mmol creatinine to only two at U-Cd of 0.8 nmol/mmol creatinine. Interestingly, the largest decrease in connectivity occurs at U-Cd of 0.5 nmol/mmol creatinine, which is the lower limit of potential risk of renal damage suggested by the Scientific Committee on Toxicity, Ecotoxicity and the Environment (50). Despite the overall tendency to decrease their connectivity, several metabolites exhibit a tendency to increase their connectivity as levels of U-Cd increase. Of these metabolites alanine shows the large increase in connectivity, from four connections at cadmium levels of 0.2 nmol/mmol creatinine to nine at U-Cd 0.8 nmol/mmol creatinine.

In addition to connectivity, there are other networks' topological feature that can reflect important points of the networks' organization. The betweenness-centrality of a node is very important property because it indicates how central a node is in the graph. It is known that nodes with central position in the network have great influence in the overall structure of network (114). We observe that for the lower level of urinary cadmium (U-Cd (0.2 nmol/mmol creatinine), dimethylglycine shows the highest betweenness-centrality however as levels of urinary cadmium increase, dimethylglycine loses its central position in the network and other urinary metabolites become more central in the structure of the correlation networks Figure 6.6. We observe that under high levels of urinary cadmium, alanine becomes the most central metabolite in the structure of the correlation networks

It is well-established that prolonged exposure to cadmium (Cd) is asso-

ciated with number of adverse health effects(182, 183). Due to its long biological half-life, cadmium accumulates with age in human tissues, especially the liver and kidney, causing their damage. Increased excretion of alanine aminopeptidase (AAP) and other enzymes of renal tubular origin, such as N-acetyl-β-D-glucosaminidase (NAG) are often used as biomarkers of early symptom of cadimum-induced renal lesions (184). In particular, positive association between urinary concentration of AAP and urinary cadmium has been previously reported (184). The main role of alanine aminopeptidase is to cleave the amino end of an amino acid from various peptides, amide or arylamide. Although most amino acids can be used by the enzyme, alanine is the preferable amino acid released in reaction catalysed by alanine aminopeptidase. Our analysis indicates that as levels of urinary cadmium, alanine becomes the most central component in the correlation networks. While the interpretation of these large changes in topological properties of the network - large changes in nodes' connectivity and betweenness-centrality - is not obvious, these variations may point to a marked change in the underlying regulation of the system as levels of urinary cadmium increase.

6.5 Summary

Our aim in this study was to investigate continuous changes in metabolic traits in response to pathogenic stimuli. This study describes a methodological approach to investigate changes in mean levels, variance and correlation of metabolic traits in relation to urinary cadmium, a common environmental and occupational pollutant. The results illustrate the advantage of using the proposed methodology to investigate the continuous effect of pathogenic stimulus in molecular traits. Following our approach we have been able to identify previously unreported variations in the metabolism of urinary metabolites in relation to urinary cadmium. For instance, exploring the potential effect of urinary cadmium on the variance and correlation of urinary metabolites we have identified changes in a number of metabolites involved in the metabolism of choline. Moreover, by conducting a detailed examination of the changes in the topological properties of association networks, we

were able to identify pattern of molecular changes across different levels of cadmium exposure.

Throughout these comprehensive analyses of network changes, we have been able to evaluate not only the structural principles of the system but also to identify what are the key structural points of the system and how their role change under increasing pathogenic influence. From this analysis, inferences can be drawn about the importance of individual molecules in disease processes. This study demonstrates the benefit of applying network-based methods and hierarchical modelling for a more in depth analysis of the changes induced on molecular traits by potential pathogenic stimulus. In this context, we can conclude that analysis of molecular changes across the whole spectrum can provide a much more detailed view into biological processes and potentially reveal key features of disease development.

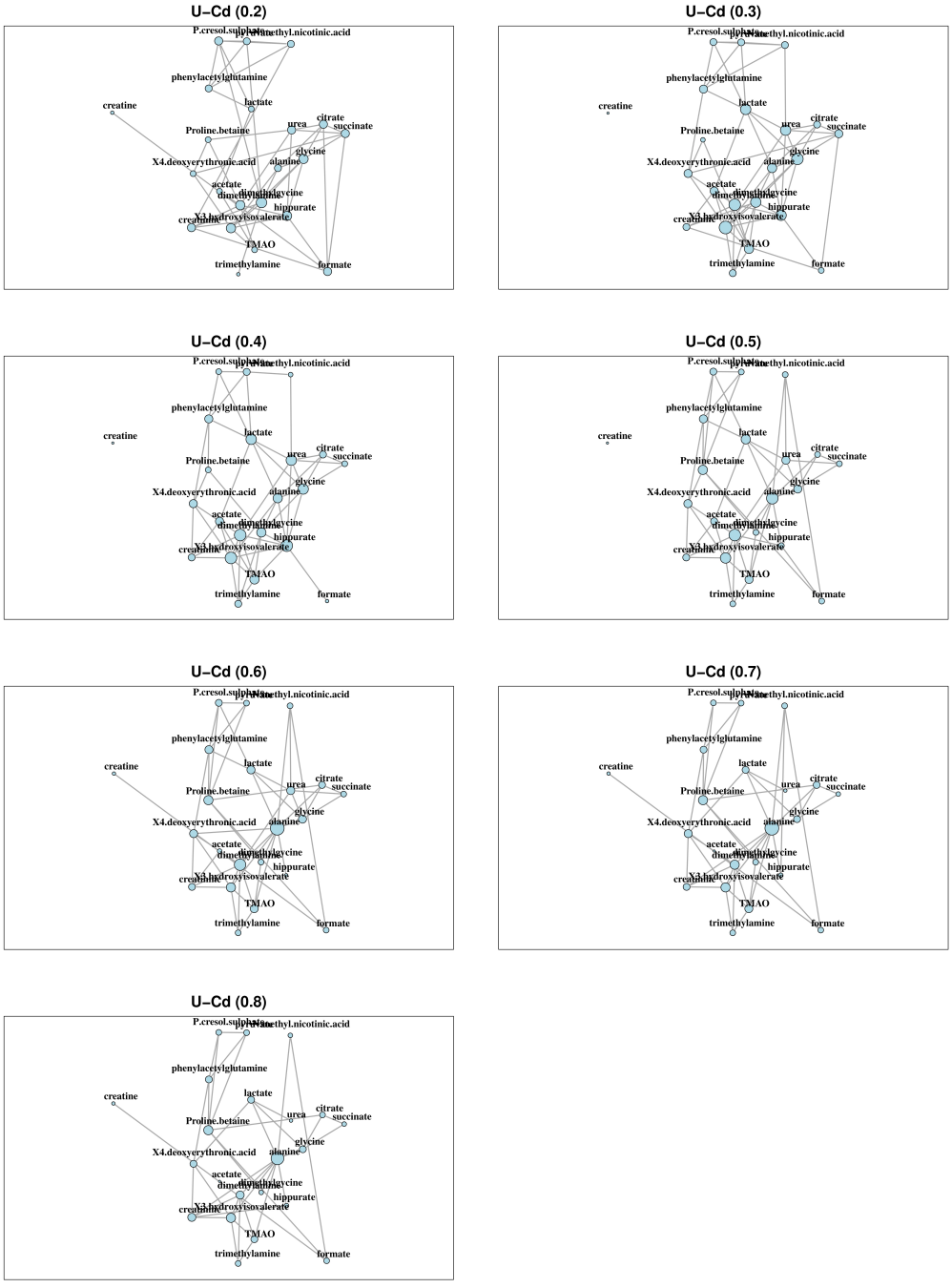


Figure 6.5: Correlation Networks inferred from the covariance regression analysis. Each of the connections indicated pair-wise correlations between pairs of metabolites which was calculated from the covariance regression fits under seven different levels of urinary cadmium different levels of urinary cadmium (U-Cd (nmol/mmol creatinine)= 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8). The node size is proportional to the degree of connectivity (number of connections).

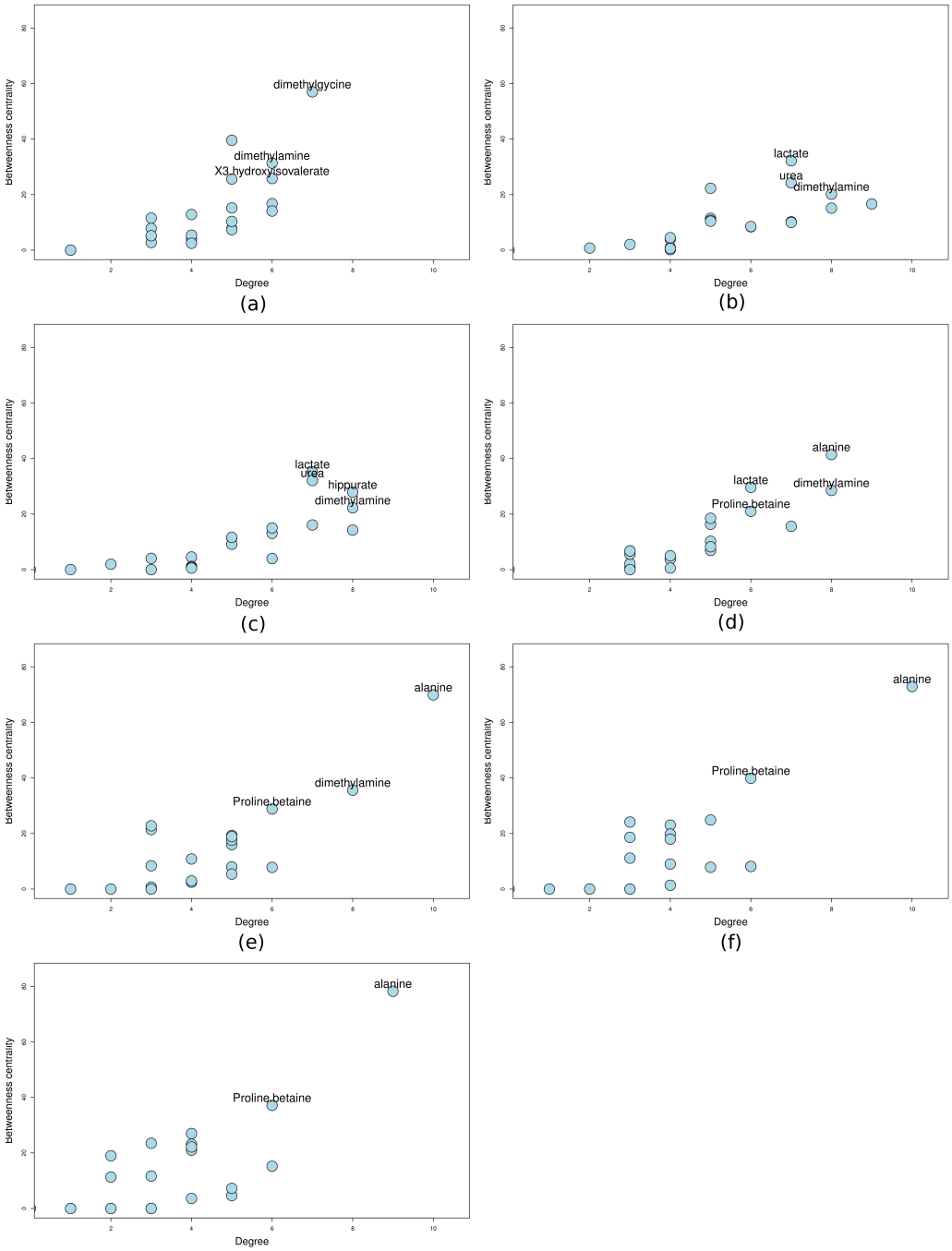


Figure 6.6: Betweenness-centrality of the correlation networks. The betweenness indicates how central a node is in the network under seven different levels of urinary cadmium (U-Cd (nmol/mmol creatinine)= a) 0.2, b) 0.3, c) 0.4, d) 0.5, e)0.6, f) 0.7, g) 0.8).

Chapter 7

Discussion

The two main aims of this thesis were 1) to investigate and develop new statistical methodologies for the integrative analysis of omics data; 2) to explore the potential of use of association networks as an additional phenotype to uncover underlying features of phenotype variation. Throughout this work we presented three novel network-based methodologies - differential network, GEMINi and Variance and Covariance regression - and illustrated their applicability in the context of genomic (SNP data) and metabolomic data analysis. The methods showed promising results, highlighting novel variations in molecular systems in relation to (patho)physiological conditions, which were undetectable by more conventional analytical approaches.

In this final chapter, we first summarise the main objectives of this research project thesis. Then we briefly discuss what we believe are the main achievements of this research project. Some of the challenges and limitations encountered throughout this work will also be presented. We conclude the chapter with suggestions regarding future works and how the investigation presented in this thesis can be further developed.

7.1 Summary

We initiate our work by performing an exploratory analysis of SNP data in combination with metabolomic phenotypes. Our objective in this chapter was to show the value of accounting for correlation in 'omics data analysis, where multiple levels of molecular data, such as genetic variants or metabolic phenotypes, are generally highly correlated. Consequently, two multivariate

statistical methods, in addition to the traditional one-SNP-one-phenotype at-a-time approach, were employed. Each of the three applied approaches providing access to different levels of biological information. For instance, the well-established univariate approach provided us with information about genetic variants that individually have a large effect on each of the serum metabolic traits. With the HyperLasso regression, we were able to explore the potential joint effect of genetic variants on each of the metabolic phenotypes and selected sets of SNPs that in combination best explain the trait's value. Finally, with the SUR-based analysis we were able to identify genetic variants that have a significant effect on a set of related traits as well as the genetic variants that have a significant effect on several phenotypes but not on all of them.

For the following three chapters our research was focused on the use of association networks. In particular, we were interested in investigating the use of variations in molecular associations as a tool to yield new insight into the biological basis of disease-specific phenotype changes. First, we presented a novel statistical methodology, the differential network. The performance of the differential network method at identifying differences in pair-wise association in relation to a binary outcome was tested in a simulation study. Additionally the use of the differential network method was investigated in a real data set in the context of diabetic dyslipidemias. Based on a standard statistical analysis, a few significant differences in concentration levels of lipoprotein subclasses between subjects with normal fasting glucose and those with impaired fasting glucose (prediabetes) were found. Interestingly, with the application of the differential network approach we were able to identify significant differences in lipoprotein metabolism between the two groups, many of which were known to be related to the development of diabetic dyslipidemias. The results of this analysis are potentially of clinical relevance, showing that alterations in the lipoproteins metabolism can occur early in the progression towards disease onset, preceding the clinical diagnosis.

In chapter 5, our objetive was to extend the differential network methodology to jointly investigate metabolomic and genetic variations in relation to

phato-physiological conditions. We started this work postulating that differences in metabolite associations could potentially be used as an additional phenotype to investigate genetic factors that influence complex traits. To investigate this hypothesis, we developed a new analytical approach (GEMINi) which combines correlation network analysis with genome-wide correlation study and pathway analysis. We illustrated the GEMINi approach by exploring molecular changes related to obesity. Application of our approach revealed a set of serum metabolite associations which differ between normal and obese individuals. Most importantly further investigation of these metabolic differences identified a number of loci, several of which have been previously implicated in obesity-related traits such as LIPA, a recently reported gene for obesity-related metabolic complications.

In our last chapter, we proposed to apply network analysis in combination with Bayesian hierarchical models to explore changes in mean level, variance and correlation of variables in relation to covariates. This methodology was illustrated in the context of urinary metabolic traits in response to cadmium exposure. The results show the potential of the proposed approach to investigating the continuous effect of pathogenic stimuli in molecular traits. With the application of this statistical procedure we were able to identify previously unreported variations in the metabolism of urinary metabolites in relation to urinary cadmium.

7.2 Main achievements

From the results obtained, we can conclude that the major contributions of this thesis to the field of omics data analysis are:

- **It highlights the need for a more in depth analysis of molecular features, beyond those detectable using mean level changes.** Throughout this thesis we have shown that to fully understand how molecular systems change in relation to phatho-physiological stimuli, it is essential to consider multiple and more complex features of the available data. Unlocking the full potential of 'omics data requires new

approaches that move beyond the convectional analysis of mean level changes. Systematic identification of meaningful relationships between different entities of the system can yield new insights into underlying biological basis of disease-specific phenotype variation.

- **It demonstrates the use of differential association networks as a tool to uncover new features of disease development.** This work showed that the comprehensive assessment of changes in molecular associations via network analysis represents a promising tool to better understand the biological basis of phenotype variation in the process of disease development. Moreover, the combination of methods proposed in this work - shrinkage estimates of partial correlations and network methods - provides a simple approach to investigate the complex web of molecular interactions and consequently helps to improve the biological interpretation of the results. Our findings demonstrated the potential of the differential network approach to identify subtle molecular variations that occur at early stages of disease development.

- **It shows the potential use of variations in molecular associations as an alternative phenotype.** In this work, we have presented and investigated the idea of using molecular associations as a phenotype when studying the genetic basis of complex traits and diseases. This novel approach shows promising power gains over the conventional genetic and metabolomic analyses.

- **It reinforces the need to combine multiple levels of 'omics data to characterize the mechanisms implicated in the development of complex disorders.** This work illustrates how the various types of 'omic data capture different features of the biological system, being complementary to each other. To enhance our knowledge of mechanisms involved in disease aetiology, an integrative approach to omics data analysis is becoming essential.

7.3 Challenges and limitations

Several challenges and limitations of the proposed methods need to be acknowledged. A first clear limitation of this study is the lack of an independent cohort that can be used to replicate our genetic findings. Although the application of the proposed methodologies showed promising results, we are unable to draw more final conclusions without the adequate validation study.

A second limitation is related to the relatively small sample size of the NFBC1966 and the young age of the cohort members which makes difficult to investigate cardiometabolic disorders. This issue has a clear impact on the performance of the convectional GWAS in chapter 5, where we found no association at a genome-wide significance level. Although this limitation was overcome by the application of the GEMINI approach, we could expect that with larger sample sizes, undetected associations could have been uncovered.

A main limitation for the inference of the differential networks was the unbalanced design of the study groups when clinical criteria were used for the classification of phenotypic groups (e.g.fasting glucose levels). With unbalanced data, the power of the test is lower than when sample sizes are equal and therefore higher significance level was required. Thus, to define which edges were to be included in the differential network we had to set the cut-off for significance to an uncorrected threshold.

Another limitation encountered throughout this project was the often limited biological knowledge available to interpret results from omics data sets. For instance, to assess the potential of the differential networks to reveal key differences related to different physiological conditions, we compare our results with existing knowledge of differences in lipoprotein composition related to the development of T2DM. However, we found that in some cases there was very limited biological information regarding the metabolic measures investigated in this thesis (e.g. components of lipoprotein subclasses). This problem arises from the fact that while the development of high throughput technologies has provided a large amount of valuable molecular data, advances in theoretical biology have not kept pace.

A similar limiting issue is the current lack of functional knowledge of

genetic variants localized away from protein-coding gene or well-known regulatory region. The common practice in GWAS to interpret genetic finding is to associate SNPs to the closest protein-coding gene. While it is certainly true that some disease causing variants alter protein function, there are other modes of action such as modification of gene expression through alteration of regulatory regions. In our study, there was a number of identified SNPs that fell into genetic regions where the closest protein-coding gene was several Mb away. Although this is the standard approach in GWAS study, evaluation of the results based on proximity to protein-coding gene do certainly lead to less sound conclusions.

Future Work

There is a number of ways in which the work in this thesis can be extended. One potential area of future research is the use of pathway information to improve the construction of molecular networks. In particular, it would be interesting to extend our methodologies to incorporate prior knowledge of molecular pathways. The prior information about network structure could be extracted from a publicly available database or literature. An example of use of pathway information is the boosting approach proposed by Shazia et al. (2008) for learning the structure of networks. This study showed the value of using pathways information to increase the statistical power and improve the inference.

Another extension could be the use of different measures of associations for the inference of the networks, such as rank correlations - which does not depend on linearity and normality of the data - and nonlinear measures of association.

The methods proposed in this thesis have been presented based on two types of omic data: SNP and metabolomic data. However, these analytical approaches can easily be applied to the analysis of other molecular profiling techniques such as transcriptomics and proteomics. Thus, a natural extension of this research is to use the proposed methods to analyse additional types of molecular information to obtain a complete picture of how biological

systems change in relation to disease development. One interesting analysis would be the systematic application of the differential network methodology to investigate molecular changes at different levels of biological systems. The subsequent superimposition of the inferred differential networks can help to identify the subset of molecular mechanisms involved in particular biological process. Moreover, this integrative approach can allow the identification of specific points along the chain of biological processes that become dysregulated under disease condition. This superimposing analysis might also present the benefit of allowing the validation of the results without the need of incorporating data on new study populations. For instance, the identification of an altered pathway at one level of the system (e.g. gene expression) can be validated by the observation of altered molecules (e.g enzymes and metabolites) related to such pathway. In addition, it may be useful to construct combined gene-protein-metabolic-phenotype interaction networks. Complex outcomes are known to be defined by multiple molecular factors, which are interconnected between them. Therefore, with the construction of networks that simultaneously use data on phenotypic characteristic (BMI, fasting glucose levels), gene expression, enzymatic activity and metabolic profiles, we can get an understanding of how the patterns of molecular association change across different levels of biological systems in disease, which can potentially allow the identification of dysfunctional regions in the systems. These multiple- system analyses can provide an in-depth understanding of the cascade of biochemical changes associated with disease development aiding the generation of biologically meaningful hypotheses which can be the basis for further detailed analysis (e.g. to design lab experiments that can help to establish the causal influences of the system dysregulation).

Another avenue for future work is to further develop our methodologies to facilitate the analysis of longitudinal data. Association networks applied to longitudinal data can be used to characterize dysregulated part of molecular systems over time. An example of this approach could be the application of the differential network approach to analyse data from a study whose aim is the evaluation of the effects of medication/treatment used on patients with a given disorder. The differential networks can be constructed at different

time points of the trial to evaluate molecular differences between e.g. responders and non-responders to treatment. The examination of network changes with time can provide information which may be useful in determining the molecular mechanism underlying a given phenotype. This information can help the design of further investigations of properties of biochemical reactions and give us information about pathways that can be modified through more targeted therapies.

Bibliography

[1] Watson, J. D. and Crick, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738.

[2] Human Genome, S. C. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–945.

[3] Consortium, T. E. P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.

[4] Venter, J. C., et al. (2001) The sequence of the human genome. *Science (New York, NY)*, 291(5507), 1304–1351.

[5] Stumpf, M. P. H. and McVean, G. A. T. (2003) Estimating recombination rates from population-genetic data. *Nature reviews Genetics*, 4(12), 959–968.

[6] Nachman, M. W. and Crowell, S. L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1), 297–304.

[7] Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, 217(5129), 624–626.

[8] Kruglyak, L. and Nickerson, D. A. (2001) Variation is the spice of life. *Nature genetics*, 27(3), 234–236.

[9] McCarroll, S. A. and Altshuler, D. M. (2007) Copy-number variation and association studies of human disease. *Nature genetics*, 39(7 Suppl), S37–42.

[10] Lewontin, R. C. and Kojima, K.-i. (1960) The Evolutionary Dynamics of Complex Polymorphisms. *Society for the Study of Evolution*, 14(4), pp. 458–472. ISSN 00143820.

[11] Slatkin, M. (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature reviewsGenetics*, 9(6), 477–485.

[12] Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature genetics*, 29, 217–222.

[13] The International, H. C. (2005) A haplotype map of the human genome. *Nature*, 437(7063), 1299–1320.

[14] Consortium, I. H., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 851–861.

[15] Consortium, I. H. ., et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52–58.

[16] Consortium, . G. P. (2010) A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–1073.

[17] Ku, C. S., Loy, E. Y., Pawitan, Y., and Chia, K. S. (2010) The pursuit of genome-wide association studies: where are we now? *Journal of human genetics*, 55(4), 195–206.

[18] Manolio, T. A., et al. (2009) Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753.

[19] Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science (New York, NY)*, 273(5281), 1516–1517.

[20] Klein, R. J., et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720), 385–389.

[21] Johnson, A. and O'Donnell, C. (2009) An open access database of genome-wide association results. *BMC Med Genet*, 10, 6.

[22] Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. (1998) Systematic functional analysis of the yeast genome. *Trends in biotechnology*, 16(9), 373–378.

[23] 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; the fate of foreign compounds in biological systems*, 29(11), 1181–1189.

[24] Kaddurah-Daouk, R. and Krishnan, K. R. (2008) Metabolomics: A Global Biochemical Approach to the Study of Central Nervous System Diseases. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 34(1), 173–186.

[25] Ramanathan, A., Wang, C., and Schreiber, S. L. (2005) Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17), 5992–5997.

[26] Shaham, O., et al. (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Molecular systems biology*, 4, 214.

[27] Wang, Y., et al. (2008) Global metabolic responses of mice to Trypanosoma brucei brucei infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5), 6127–32. ISSN 1091-6490.

[28] Chen, L., Magliano, D. J., and Zimmet, P. Z. (2012) The worldwide epidemiology of type 2 diabetes mellitus - present and future perspectives. *Nat Rev Endocrinol*, 8(4), 228–236.

[29] Mooradian, A. D. (2009) Dyslipidemia in type 2 diabetes mellitus. *Nat Clin Pract End Met*, 5(3), 150–159.

[30] Garvey, W. T., et al. (2003) Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance. *Diabetes*, 52(2), 453–462.

[31] Festa, A., et al. (2005) Nuclear magnetic resonance lipoprotein abnormalities in prediabetic subjects in the Insulin Resistance Atherosclerosis Study. *Circulation*, 111, 3465–3472.

[32] Wang, J., et al. (2012) Lipoprotein subclass profiles in individuals with varying degrees of glucose tolerance: a population-based study of 9399 Finnish men. *Journal of Internal Medicine*, 272(6), 562–572. ISSN 1365-2796.

[33] Newgard, C. B., et al. (2009) A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism*, 9(4), 311–326.

[34] Wang, T. J., et al. (2011) Metabolite profiles and the risk of developing diabetes. *Nature medicine*, 17(4), 448–453.

[35] Xie, B., Waters, M. J., and Schirra, H. J. (2012) Investigating potential mechanisms of obesity by metabolomics. *Journal of biomedicine and biotechnology*, 2012, 805683.

[36] Serkova, N. J., et al. (2006) Metabolic profiling of livers and blood from obese Zucker rats. *Journal of hepatology*, 44(5), 956–962.

[37] Duggan, G. E., Hittel, D. S., Hughey, C. C., Weljie, A., Vogel, H. J., and Shearer, J. (2011) Differentiating short- and long-term effects of diet in the obese mouse using 1H-nuclear magnetic resonance metabolomics. *Diabetes, Obesity and Metabolism*, 13(9), 859–862. ISSN 1463-1326.

[38] Oberbach, A., et al. (2011) Combined Proteomic and Metabolomic Profiling of Serum Reveals Association of the Complement System with Obesity and Identifies Novel Markers of Body Fat Mass Changes. *Journal of Proteome Research*, 10(10), 4769–4788. doi:10.1021/pr2005555.

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

[40] Suhre, K., et al. (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362), 54–60.

[41] Teslovich, T. M., et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307), 707–713.

[42] Kettunen, J., et al. (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, 44, 269–276.

[43] Tukiainen, T., et al. (2012) Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Human molecular genetics*, 21(6), 1444–1455.

[44] Rantakallio, P. (1969) Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, 193, Suppl 193:1+.

[45] Rantakallio, P. (1979) Social background of mothers who smoke during pregnancy and influence of these factors on the offspring. *Social science & medicine*, 13A(4), 423–429.

[46] Jarvelin, M. R., et al. (2004) Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension*, 44(6), 838–846.

[47] Sabatti, C., et al. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1), 35–46.

[48] Jarvelin, M. R., Hartikainen-Sorri, A. L., and Rantakallio, P. (1993) Labour induction policy in hospitals of different levels of specialisation. *British journal of obstetrics and gynaecology*, 100(4), 310–315.

[49] Jarvelin, M. R., et al. (1997) Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986. *Paediatric and perinatal epidemiology*, 11(3), 298–312.

[50] Thomas, L. D., Hodgson, S., Nieuwenhuijsen, M., and Jarup, L. (2009) Early kidney damage in a population exposed to cadmium and other heavy metals. *Environmental health perspectives*, 117(2), 181–184.

[51] Soininen, P., et al. (2009) High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *The Analyst*, 134(9), 1781–1785.

[52] Tukiainen, T., et al. (2008) A multi-metabolite analysis of serum by 1H NMR spectroscopy: early systemic signs of Alzheimer's disease. *Biochemical and biophysical research communications*, 375(3), 356–361.

[53] Inouye, M., et al. (2010) Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular systems biology*, 6, 441.

[54] Vehtari, A., et al. (2007) A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data. *BMC bioinformatics*, 8 Suppl 2, S8.

[55] Niemi, J., et al. (2009) Estimation of VLDL, IDL, LDL, HDL2, apoA-I, and apoB from the Friedewald inputs–apoB and IDL, but not LDL, are associated with mortality in type 1 diabetes. *Annals of Medicine*, 41(6), 451–461.

[56] Ellis, J. K., et al. (2012) Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. *BMC medicine*, 10, 61–7015–10–61.

[57] Redgrave, T. G. (2004) Chylomicron metabolism. *Biochemical Society transactions*, 32(Pt 1), 79–82.

[58] Hegele, R. A. (2009) Plasma lipoproteins: genetic influences and clinical implications. *Nature reviews Genetics*, 10, 109–121.

[59] Bertram, H. C., et al. (2009) Nuclear magnetic resonance-based metabonomics reveals strong sex effect on plasma metabolism in 17-year-old Scandinavians and correlation to retrospective infant plasma parameters. *Metabolism: clinical and experimental*, 58, 1039–1045.

[60] Schonfeld, G. (2003) Familial hypobetalipoproteinemia: a review. *Journal of lipid research*, 44(5), 878–883.

[61] Schonfeld, G., Lin, X., and Yue, P. (2005) Familial hypobetalipoproteinemia: genetics and metabolism. *Cellular and molecular life sciences : CMLS*, 62(12), 1372–1378.

[62] Innerarity, T. L., et al. (1990) Familial defective apolipoprotein B-100: a mutation of apolipoprotein B that causes hypercholesterolemia. *Journal of lipid research*, 31, 1337–1349.

[63] Huijgen, R., et al. (2012) Genetic variation in APOB, PCSK9, and ANGPTL3 in carriers of pathogenic autosomal dominant hypercholesterolemic mutations with unexpected low LDL-Cl Levels. *Human mutation*, 33(2), 448–455.

[64] de la Iglesia, N., Mukhtar, M., Seoane, J., Guinovart, J. J., and Agius, L. (2000) The role of the regulatory protein of glucokinase in the glucose sensory mechanism of the hepatocyte. *The Journal of biological chemistry*, 275(14), 10597–10603.

[65] Shen, H., Pollin, T. I., Damcott, C. M., McLenithan, J. C., Mitchell, B. D., and Shuldiner, A. R. (2009) Glucokinase regulatory protein gene polymorphism affects postprandial lipemic response in a dietary intervention study. *Human genetics*, 126(4), 567–574.

[66] Willer, C. J., et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*, 40(2), 161–169.

[67] Orho-Melander, M., et al. (2008) Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. *Diabetes*, 57(11), 3112–3121.

[68] Vaxillaire, M., et al. (2008) The common P446L polymorphism in GCKR inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. *Diabetes*, 57(8), 2253–2257.

[69] Beisiegel, U. (1996) New aspects on the role of plasma lipases in lipoprotein catabolism and atherosclerosis. *Atherosclerosis*, 124(1), 1 – 8. ISSN 0021-9150.

[70] Santamarina-Fojo, S. and Haudenschild, C. (2000) Role of hepatic and lipoprotein lipase in lipoprotein metabolism and atherosclerosis: studies in transgenic and knockout animal models and somatic gene transfer. *International journal of tissue reactions*, 22(2-3), 39–47.

[71] Merkel, M., et al. (2002) Inactive lipoprotein lipase (LPL) alone increases selective cholesterol ester uptake in vivo, whereas in the presence of active LPL it also increases triglyceride hydrolysis and whole particle lipoprotein uptake. *The Journal of biological chemistry*, 277(9), 7405–7411.

[72] Mead, J. R., Irvine, S. A., and Ramji, D. P. (2002) Lipoprotein lipase: structure, function, regulation, and role in disease. *Journal of Molecular Medicine (Berlin, Germany)*, 80(12), 753–769.

[73] Ge, L., Gordon, J. S., Hsuan, C., Stenn, K., and Prouty, S. M. (2003) Identification of the delta-6 desaturase of human sebaceous glands: expression and enzyme activity. *The Journal of investigative dermatology*, 120(5), 707–714.

[74] Frazier-Wood, A. C., et al. (2013) Genetic variants associated with VLDL, LDL and HDL particle size differ with race/ethnicity. *Human genetics*, 132(4), 405–413.

[75] Tall, A. R. (1993) Plasma cholesteryl ester transfer protein. *Journal of lipid research*, 34(8), 1255–1274.

[76] Zhong, S., et al. (1996) Increased coronary heart disease in Japanese-American men with mutation in the cholesteryl ester transfer protein gene despite increased HDL levels. *The Journal of clinical investigation*, 97(12), 2917–2923.

[77] Vasan, R. S., et al. (2009) Association of circulating cholesteryl ester transfer protein activity with incidence of cardiovascular disease in the community. *Circulation*, 120(24), 2414–2420.

[78] Tu, A. Y., Nishida, H. I., and Nishida, T. (1993) High density lipoprotein conversion mediated by human plasma phospholipid transfer protein. *The Journal of biological chemistry*, 268(31), 23098–23105.

[79] Frederick, J. P., et al. (2008) A role for a lithium-inhibited Golgi nucleotidase in skeletal development and sulfation. *Proceedings of the National Academy of Sciences*, 105(33), 11605–11612.

[80] Kato, M., et al. (1996) Physical and functional interaction of rabphilin-3A with alpha-actinin. *The Journal of biological chemistry*, 271(50), 31775–31778.

[81] Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7), e1000130.

[82] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58, 267–288.

[83] Zhang, J., S. Jin (1996) *Computation of Special Functions*. New York, Wiley.

[84] Zhang, T. and Oles, F. J. (2000) Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval*, 4, 5–31.

[85] Genkin, Alexander, Lewis, David, D., Madigan, and David (2007) Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, 49(3), 291–304.

[86] Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). runHLasso. URL `http://www.ebi.ac.uk/projects/BARGEN/`.

[87] Stephens, M. and Balding, D. J. (2009) Bayesian statistical methods for genetic association studies. *Nature reviews Genetics*, 10(10), 681–690.

[88] Samani, N., et al. (2008) The novel genetic variant predisposing to coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. *Journal of Molecular Medicine*, 86, 1233–1241.

[89] Kathiresan, S., et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2), 189–197. M3: 10.1038/ng.75; 10.1038/ng.75.

[90] Laitinen, P. J., et al. (2001) Mutations of the cardiac ryanodine receptor (RyR2) gene in familial polymorphic ventricular tachycardia. *Circulation*, 103(4), 485–490.

[91] Kovacs, P., et al. (2003) The role of insulin receptor substrate-1 gene (IRS1) in type 2 diabetes in Pima Indians. *Diabetes*, 52(12), 3005–3009.

[92] Sesti, G., Federici, M., Hribal, M. L., Lauro, D., Sbraccia, P., and Lauro, R. (2001) Defects of the insulin receptor substrate (IRS) system in human metabolic disorders. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 15(12), 2099–2111.

[93] Levy, D., et al. (2007) Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC medical genetics*, 8 Suppl 1, S3.

[94] Aouizerat, B. E., et al. (2003) Genetic analysis of a polymorphism in the human apoA-V gene: effect on plasma lipids. *Journal of lipid research*, 44(6), 1167–1173.

[95] Mahley, R. W. (1988) Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science (New York, NY)*, 240(4852), 622–630.

[96] Chen, Z., et al. (2010) Variants in hepatocyte nuclear factor 4alpha gene promoter region and type 2 diabetes risk in Chinese. *Experimental biology and medicine (Maywood, NJ)*, 235(7), 857–861.

[97] Amos, C. I. and Laing, A. E. (1993) A comparison of univariate and multivariate tests for genetic linkage. *Genetic epidemiology*, 10, 671–676.

[98] Jiang, C. and Zeng, Z. B. (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140, 1111–1127.

[99] Liu, J., Pei, Y., Papasian, C. J., and Deng, H. W. (2009) Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic epidemiology*, 33(3), 217–227.

[100] Zhang, L., Pei, Y. F., Li, J., Papasian, C. J., and Deng, H. W. (2009) Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PloS one*, 4(8), e6502.

[101] Petretto, E., et al. (2010) New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS computational biology*, 6(4), e1000737.

[102] O'Reilly, P. F., et al. (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one*, 7(5), e34861.

[103] Zellner, A. (1962) An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298), 348–368.

[104] Steuer, R. (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Briefings in bioinformatics*, 7(2), 151–158.

[105] Camacho, D., de, l. F., and Mendes, P. (2005) The origin of correlations in metabolomics data. *Metabolomics*, 1, 53–63.

[106] Barabasi, A. L. and Albert, R. (1999) Emergence of scaling in random networks. *Science (New York, NY)*, 286, 509–512.

[107] Barabasi, A. L. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nature reviewsGenetics*, 5, 101–113.

[108] Oldham, M. C., Horvath, S., and Geschwind, D. H. (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47), 17973–17978.

[109] Fuller, T. F., A, G., Aten, J. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2007) Weighted gene expression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18, 463–472.

[110] Fuite, J., Vernon, S. D., and Broderick, G. (2008) Neuroendocrine and immune network re-modeling in chronic fatigue syndrome: an exploratory analysis. *Genomics*, 92, 393–399.

[111] Gill, R., Datta, S., and Datta, S. (2010) A statistical framework for differential network analysis from microarray data. *BMC bioinformatics*, 11, 95.

[112] Weckwerth, W., Loureiro, M. E., Wenzel, K., and Fiehn, O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20), 7809–7814.

[113] Szymanski, J., et al. (2009) Stability of metabolic correlations under changing environmental conditions in Escherichia coli–a systems approach. *PloS one*, 4(10), e7441.

[114] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000) The large-scale organization of metabolic networks. *Nature*, 407, 651–654.

[115] Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.

[116] Fisher, R. A. (1915) Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4), 507–521.

[117] Schafer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)*, 21(6), 754–764.

[118] on the Diagnosis, E. C. and of Diabetes Mellitus, C. (2003) Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes care*, 26 Suppl 1, S5–20.

[119] Kontush, A. and Chapman, M. J. (2006) Antiatherogenic small, dense HDL-guardian angel of the arterial wall? *Nat Clin Pract Cardiovasc Med*, 3(3), 144–153.

[120] Colvin, P. L., Moriguchi, E., Barrett, P. H., Parks, J. S., and Rudel, L. L. (1999) Small HDL particles containing two apoA-I molecules are precursors in vivo to medium and large HDL particles containing three and four apoA-I molecules in nonhuman primates. *Journal of lipid research*, 40, 1782–1792.

[121] Ursem, R., Tikunov, Y., Bovy, A., van Berloo, R., and van Eeuwijk, F. (2008) A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica*, 161(1), 181–193.

[122] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002) Hierarchical organization of modularity in metabolic networks. *Science (New York, NY)*, 297(5586), 1551–1555.

[123] Kumpula, L. S., Kumpula, J. M., Taskinen, M. R., Jauhiainen, M., Kaski, K., and Ala-Korpela, M. (2008) Reconsideration of hydrophobic lipid distributions in lipoprotein particles. *Chemistry and physics of lipids*, 155(1), 57–62.

[124] Mora, S., Otvos, J. D., Rosenson, R. S., Pradhan, A., Buring, J. E., and Ridker, P. M. (2010) Lipoprotein particle size and concentration by nuclear magnetic resonance and incident type 2 diabetes in women. *Diabetes*, 59(5), 1153–1160.

[125] Hodge, A. M., Jenkins, A. J., English, D. R., ODea, K., and Giles, G. G. (2009) NMR-determined lipoprotein subclass profile predicts type 2 diabetes. *Diabetes Research and Clinical Practice*, 83(1), 132 – 139.

[126] Adiels, M., et al. (2006) Overproduction of large VLDL particles is driven by increased liver fat content in man. *Diabetologia*, 49, 755–765.

[127] Ala-Korpela, M., Kangas, A. J., and Inouye, M. (2011) Genome-wide association studies and systems biology: together at last. *Trends in genetics : TIG*, 27, 493–498.

[128] Inouye, M., et al. (2012) Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS genetics*, 8(8), e1002907.

[129] Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C., and Aulchenko, Y. S. (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC genetics*, 11, 92–2156–11–92.

[130] Pare, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (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 genetics*, 6(6), e1000981.

[131] Hothorn, L. A., Libiger, O., and Gerhard, D. (2012) Model-specific tests on variance heterogeneity for detection of potentially interacting genetic loci. *BMC genetics*, 13, 59–2156–13–59.

[132] Bergman, R. N., et al. (2011) A Better Index of Body Adiposity. *Obesity*, 19, 1083–1089.

[133] Tanaka, H., Clevenger, C. M., Jones, P. P., Seals, D. R., and DeSouza, C. A. (1998) Influence of body fatness on the coronary risk profile of physically active postmenopausal women. *Metabolism: clinical and experimental*, 47(9), 1112–1120.

[134] Terry, R. B., Stefanick, M. L., Haskell, W. L., and Wood, P. D. (1991) Contributions of regional adipose tissue depots to plasma lipoprotein concentrations in overweight men and women: possible protective effects of thigh fat. *Metabolism: clinical and experimental*, 40(7), 733–740.

[135] Gallagher, D., Heymsfield, S. B., Heo, M., Jebb, S. A., Murgatroyd, P. R., and Sakamoto, Y. (2000) Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *The American Journal of Clinical Nutrition*, 72, 694–701.

[136] Revelle, W. (2010) psych:Procedures for Personality and Psychological Research. *R package version 10-91.*

[137] Kamburov, A. F., Cavill R FAU Ebbels, M., Timothy, FAU, E. T., Herwig R FAU Keun, C., Hector, and Keun, H. C. (2006). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.

[138] Abbott, R. D., et al. (1998) Changes in total and high-density lipoprotein cholesterol over 10- and 20-year periods (the Honolulu Heart Program). *The American Journal of Cardiology*, 82, 172–178.

[139] Katzmarzyk, P. T., et al. (2001) Fitness, fatness, and estimated coronary heart disease risk: the HERITAGE Family Study. *Medicine and science in sports and exercise*, 33, 585–590.

[140] Canoy, D., et al. (2007) Body fat distribution and risk of coronary heart disease in men and women in the European Prospective Investigation Into Cancer and Nutrition in Norfolk cohort: a population-based prospective study. *Circulation*, 116(25), 2933–2943.

[141] Magkos, F., Mohammed, B. S., and Mittendorfer, B. (2008) Effect of obesity on the plasma lipoprotein subclass profile in normoglycemic and normolipidemic men and women. *International journal of obesity (2005)*, 32(11), 1655–1664.

[142] Okazaki, M., et al. (2005) Identification of unique lipoprotein subclasses for visceral obesity by component analysis of cholesterol profile in high-performance liquid chromatography. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 25, 578–584.

[143] Lynch, C. J., Gern, B., Lloyd, C., Hutson, S. M., Eicher, R., and Vary, T. C. (2006) Leucine in food mediates some of the postprandial rise in plasma leptin concentrations. *American journal of physiologyEndocrinology and metabolism*, 291(3), E621–30.

[144] Cota, D., et al. (2006) Hypothalamic mTOR signaling regulates food intake. *Science (New York, NY)*, 312, 927–930.

[145] Felig, P., Marliss, E., and Cahill, J., G. F. (1969) Plasma amino acid levels and insulin secretion in obesity. *The New England journal of medicine*, 281, 811–816.

[146] She, P., Van Horn, C., Reid, T., Hutson, S. M., Cooney, R. N., and Lynch, C. J. (2007) Obesity-related elevations in plasma leucine are associated with alterations in enzymes involved in branched-chain amino acid metabolism. *American journal of physiologyEndocrinology and metabolism*, 293(6), E1552–63.

[147] Antonio, J. and Street, C. (1999) Glutamine: a potentially useful supplement for athletes. *Canadian journal of applied physiology = Revue canadienne de physiologie appliquee*, 24, 1–14.

[148] Lass, A., et al. (2006) Adipose triglyceride lipase-mediated lipolysis of cellular fat stores is activated by CGI-58 and defective in Chanarin-Dorfman Syndrome. *Cell metabolism*, 3(5), 309–319.

[149] Jaworski, K., Sarkadi-Nagy, E., Duncan, R. E., Ahmadian, M., and Sul, H. S. (2007) Regulation of triglyceride metabolism. IV. Hormonal regulation of lipolysis in adipose tissue. *American journal of physiologyGastrointestinal and liver physiology*, 293, G1–4.

[150] Akiyama, M., Sawamura, D., Nomura, Y., Sugawara, M., and Shimizu, H. (2003) Truncation of CGI-58 protein causes malformation of lamellar granules resulting in ichthyosis in Dorfman-Chanarin syndrome. *The Journal of investigative dermatology*, 121, 1029–1034.

[151] Schleinitz, N., Fischer, J., Sanchez, A., Veit, V., Harle, J. R., and Pelissier, J. F. (2005) Two new mutations of the ABHD5 gene in a new adult case of Chanarin Dorfman syndrome: an uncommon lipid storage disease. *Archives of Dermatology*, 141(6), 798–800.

[152] Steinberg, G. R., Kemp, B. E., and Watt, M. J. (2007) Adipocyte triglyceride lipase expression in human obesity. *American journal of physiologyEndocrinology and metabolism*, 293(4), E958–64.

[153] Gao, J., et al. (2010) Polymorphisms of PHF11 and DPP10 are associated with asthma and related traits in a Chinese population. *Respiration; international review of thoracic diseases*, 79, 17–24.

[154] Los, H., Koppelman, G. H., and Postma, D. S. (1999) The importance of genetic influences in asthma. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*, 14(5), 1210–1227.

[155] Weiss, S. T. (2005) Obesity: insight into the origins of asthma. *Nature immunology*, 6(6), 537–539.

[156] Thomsen, S. F., et al. (2007) Association between obesity and asthma in a twin cohort. *Allergy*, 62(10), 1199–1204.

[157] Hallstrand, T. S., Fischer, M. E., Wurfel, M. M., Afari, N., Buchwald, D., and Goldberg, J. (2005) Genetic pleiotropy between asthma and obesity in a community-based sample of twins. *The Journal of allergy and clinical immunology*, 116, 1235–1241.

[158] Melen, E., et al. (2010) Analyses of shared genetic factors between asthma and obesity in children. *The Journal of allergy and clinical immunology*, 126(3), 631–7.e1–8.

[159] Landa, M. S., et al. (2007) Knocking down the diencephalic thyrotropin-releasing hormone precursor gene normalizes obesity-induced hypertension in the rat. *American journal of physiology Endocrinology and metabolism*, 292(5), E1388–94.

[160] Garcia, S. I., et al. (1997) Central overexpression of the TRH precursor gene induces hypertension in rats: antisense reversal. *Hypertension*, 30, 759–766.

[161] Garcia, S. I., et al. (2001) Thyrotropin-releasing hormone receptor (TRHR) gene is associated with essential hypertension. *Hypertension*, 38, 683–687.

[162] Boldrini, R., Devito, R., Biselli, R., Filocamo, M., and Bosman, C. (2004) Wolman disease and cholesteryl ester storage disease diagnosed by histological and ultrastructural examination of intestinal and liver biopsy. *Pathology, research and practice*, 200, 231–240.

[163] Redonnet-Vernhet, I., Chatelut, M., Basile, J. P., Salvayre, R., and Levade, T. (1997) Cholesteryl ester storage disease: relationship between molecular defects and in situ activity of lysosomal acid lipase. *Biochemical and molecular medicine*, 62(1), 42–49.

[164] Guenard, F., et al. (2012) Association of LIPA Gene Polymorphisms With Obesity-Related Metabolic Complications Among Severely Obese Patients. *Obesity (Silver Spring, Md)*, 20, 2075–2082.

[165] Eguchi, J., et al. (2005) Identification of adipocyte adhesion molecule (ACAM), a novel CTX gene family, implicated in adipocyte maturation and development of obesity. *The Biochemical journal*, 387, 343–353.

[166] Raschperger, E., Engstrom, U., Pettersson, R. F., and Fuxe, J. (2004) CLMP, a novel member of the CTX family and a new component of epithelial tight junctions. *The Journal of biological chemistry*, 279(1), 796–804.

[167] Moult, P. R. and Harvey, J. (2009) Regulation of glutamate receptor trafficking by leptin. *Biochemical Society transactions*, 37(Pt 6), 1364–1368.

[168] Choquet, D. (2010) Fast AMPAR trafficking for a high-frequency synaptic transmission. *The European journal of neuroscience*, 32, 250–260.

[169] Moult, P. R., et al. (2010) Leptin regulates AMPA receptor trafficking via PTEN inhibition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(11), 4088–4101.

[170] Masuzaki, H., et al. (1995) Human obese gene expression. Adipocyte-specific expression and regional differences in the adipose tissue. *Diabetes*, 44(7), 855–858.

[171] Licinio, J., et al. (2004) Phenotypic effects of leptin replacement on morbid obesity, diabetes mellitus, hypogonadism, and behavior in leptin-deficient adults. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13), 4531–4536. doi: 10.1073/pnas.0308767101.

[172] Harvey, J. (2007) Leptin regulation of neuronal excitability and cognitive function. *Current opinion in pharmacology*, 7, 643–647.

[173] Schwartz, M. W., Peskind, E., Raskind, M., Boyko, E. J., and Porte, J., D. (1996) Cerebrospinal fluid leptin levels: relationship to plasma levels and to adiposity in humans. *Nature medicine*, 2(5), 589–593.

[174] Chen, H. (2009) Possible Role of Platelet GluR1 Receptors in Comorbid Depression and Cardiovascular Disease. *Cardiovascular psychiatry and neurology*, 2009, 424728.

[175] Plomin, R., Haworth, C. M. A., and Davis, O. S. P. (2009) Common disorders are quantitative traits. *Nature reviews Genetics*, 10(12), 872–878.

[176] Rao, C. R. (1970) Estimation of Heteroscedastic Variances in Linear Models. *Journal of the American Statistical Association*, 65(329), 161–172.

[177] Rutemiller, H. C. and Bowers, D. A. (1968) Estimation in a Heteroscedastic Regression Model. *Journal of the American Statistical Association*, 63(322), 552–557.

[178] Hoff, P. D. and Niu, X. (2011) A covariance regression model. *ArXiv e-prints*.

[179] Zeisel, S. H., Wishnok, J. S., and Blusztajn, J. K. (1983) Formation of methylamines from ingested choline and lecithin. *The Journal of pharmacology and experimental therapeutics*, 225(2), 320–324.

[180] Wood, A. P., Warren, F. J., and Kelly, D. P. (2010) Methylotrophic Bacteria in Trimethylaminuria and Bacterial Vaginosis. In K. Timmis, editor, *Handbook of Hydrocarbon and Lipid Microbiology*, volume Part 29 of *Biomedical and Life Sciences*, pages 3227–3240. Springer, Heidelberg.

[181] Lever, M., George, P. M., Dellow, W. J., Scott, R. S., and Chambers, S. T. (2005) Homocysteine, glycine betaine, and N,N-dimethylglycine in patients attending a lipid clinic. *Metabolism*, 54(1), 1 – 14. ISSN 0026-0495.

[182] Jarup, L. and Akesson, A. (2009) Current status of cadmium as an environmental health problem. *Toxicology and applied pharmacology*, 238(3), 201–208.

[183] Nawrot, T. S., et al. (2010) Cadmium exposure in the population: from health risks to strategies of prevention. *Biometals : an international journal on the role of metal ions in biology, biochemistry, and medicine*, 23(5), 769–782.

[184] Noonan, C. W., Sarasua, S. M., Campagna, D., Kathman, S. J., Lybarger, J. A., and Mueller, P. W. (2002) Effects of exposure to low levels of environmental cadmium on renal biomarkers. *Environmental health perspectives*, 110(2), 151–155.

Appendix A

Data Sets

Table A.1: List of all serum metabolic measures included in this thesis

Full name	Abbreviation	Ch3	Ch4	Ch5	Ch6
Total lipids in extremely large VLDL	XXL-VLDL-L	✓		✓	
Phospholipids extremely large VLDL	XXL-VLDL-PL		✓		
Concentration extremely large VLDL particles	XXL-VLDL-P		✓		
Total lipids in very large VLDL	XL-VLDL-L	✓		✓	
Phospholipids in very large VLDL	XL-VLDL-PL		✓		
Triglycerides in very large VLDL	XL-VLDL-TG		✓		
Concentration of very large VLDL particles	XL-VLDL-P		✓		
Total lipids in large VLDL	L-VLDL-L	✓		✓	
Cholesterol in large VLDL	L-VLDL-C		✓		
Free cholesterol in large	L-VLDL-FC		✓		
Phospholipids in large VLDL	L-VLDL-PL		✓		
Triglycerides in large VLDL	L-VLDL-TG		✓		
Cholesterol esters in large VLDL	L-VLDL-CE		✓		
Concentration of large VLDL particles	L-VLDL-P		✓		
Total lipids in medium VLDL	M-VLDL-L	✓		✓	
Cholesterol in medium VLDL	M-VLDL-C		✓		
Free cholesterol in medium VLDL	M-VLDL-FC		✓		
Phospholipids in medium VLDL	M-VLDL-PL		✓		
Triglycerides in medium VLDL	M-VLDL-TG		✓		
Cholesterol esters in medium VLDL	M-VLDL-CE		✓		
Concentration of medium VLDL particles	M-VLDL-P		✓		
Total lipids in small VLDL	S-VLDL-L	✓		✓	
Cholesterol in small VLDL	S-VLDL-C		✓		
Free cholesterol in small VLDL	S-VLDL-FC		✓		
Phospholipids in small VLDL	S-VLDL-PL		✓		
Triglycerides in small VLDL	S-VLDL-TG		✓		

Continued on next page

Table A.1 – continued from previous page

Full name	Abbreviation	Ch3	Ch4	Ch5	Ch6
Concentration of small VLDL particles	S-VLDL-P		✓		
Total lipids in very small VLDL	XS-VLDL-L	✓		✓	
Phospholipids in very small VLDL	XS-VLDL-PL		✓		
Triglycerides in very small VLDL	XS-VLDL-TG		✓		
Concentration of very small VLDL particles	XS-VLDL-P		✓		
Total lipids in IDL	IDL-L	✓		✓	
Free cholesterol in IDL	IDL-FC		✓		
Phospholipids in IDL	IDL-PL			✓	
Concentration of IDL particles	IDL-P		✓		
Total lipids in large LDL	L-LDL-L	✓		✓	
Cholesterol in large LDL	L-LDL-C		✓		
Free cholesterol in large LDL	L-LDL-FC		✓		
Phospholipids in large LDL	L-LDL-PL		✓		
Cholesterol esters in large LDL	L-LDL-CE		✓		
Concentration of large LDL particles	L-LDL-P		✓		
Total lipids in medium LDL	M-LDL-L	✓		✓	
Cholesterol in medium LDL	M-LDL-C		✓		
Phospholipids in medium LDL	M-LDL-PL		✓		
Cholesterol esters in medium LDL	M-LDL-CE		✓		
Concentration of medium LDL particles	M-LDL-P		✓		
Total lipids in small LDL	S-LDL-L	✓		✓	
Cholesterol in small LDL	S-LDL-C		✓		
Concentration of small LDL particles	S-LDL-P		✓		
Total lipids in in very large HDL	XL-HDL-L	✓		✓	
Cholesterol in very large HDL	XL-HDL-C		✓		
Free cholesterol in very large HDL	XL-HDL-FC		✓		
Phospholipids in very large HDL	XL-HDL-PL		✓		
Triglycerides in very large HDL	XL-HDL-TG		✓		
Cholesterol esters in very large HDL	XL-HDL-CE		✓		
Concentration of very large HDL particles	XL-HDL-P		✓		
Total lipids in large HDL	L-HDL-L	✓		✓	
Cholesterol in large HDL	L-HDL-C		✓		
Free cholesterol in large HDL	L-HDL-FC		✓		
Phospholipids in large HDL	L-HDL-PL		✓		
Cholesterol esters in large HDL	L-HDL-CE		✓		
Concentration of large HDL particles	L-HDL-P		✓		
Total lipids in medium HDL	M-HDL-L	✓		✓	
Cholesterol in medium HDL	M-HDL-C		✓		
Free cholesterol in medium HDL	M-HDL-FC		✓		

Continued on next page

Table A.1 – continued from previous page

Full name	Abbreviation	Ch3	Ch4	Ch5	Ch6
Phospholipids in medium HDL	M-HDL-PL		✓		
Cholesterol esters in medium HDL	M-HDL-CE		✓		
Concentration of medium HDL particles	M-HDL-P		✓		
Total lipids in small HDL	S-HDL-L	✓		✓	
Triglycerides in small HDL	S-HDL-TG		✓		
Concentration of small HDL particle	S-HDL-P		✓	✓	
Apolipoprotein A-I	ApoA.I.	✓	✓	✓	
Apolipoprotein B	ApoB.	✓	✓	✓	
Apolipoprotein B by Apolipoprotein A-I	ApoB/ApoA.I		✓		
3-hydroxybutyrate	bOHBut	✓		✓	
Acetate	Ace	✓		✓	
Acetoacetate	AcAce	✓		✓	
Alanine	Ala	✓		✓	
Citrate	Cit	✓		✓	
Creatinine	Crea	✓		✓	
Glucose	Glc	✓		✓	
Glutamine	Gln	✓		✓	
Glycerol	Glol	✓		✓	
Histine	His	✓		✓	
Isoleucine	Ile	✓		✓	
Lactate	Lac	✓		✓	
Leucine	Leu	✓		✓	
Phenylalanine	Phe	✓		✓	
Pyruvate	Pyr	✓		✓	
Tyrosine	Tyr	✓		✓	
Urea	Urea	✓		✓	
Valine	Val	✓		✓	
esterified cholesterol	EstC	✓			
Free cholesterol	FreeC	✓			
Omega-3 fatty acids	FAw3	✓			
Omega-6 fatty acids	FAw6	✓			
omega-9 and saturated fatty acids	FAw79S	✓			
Total fatty acids	TotFA	✓			
18:2, linoleic acid (LA)	LA	✓			
Polyunsaturated fatty acids 18:2	TotPUFA	✓			
22:6, docosahexaenoic acid (DHA)	DHA	✓			
Monounsaturated fatty acids	MUFA	✓			
Total phosphoglycerides	TotPG	✓			
Phosphatidylcholine and other cholines	PC	✓			

Continued on next page

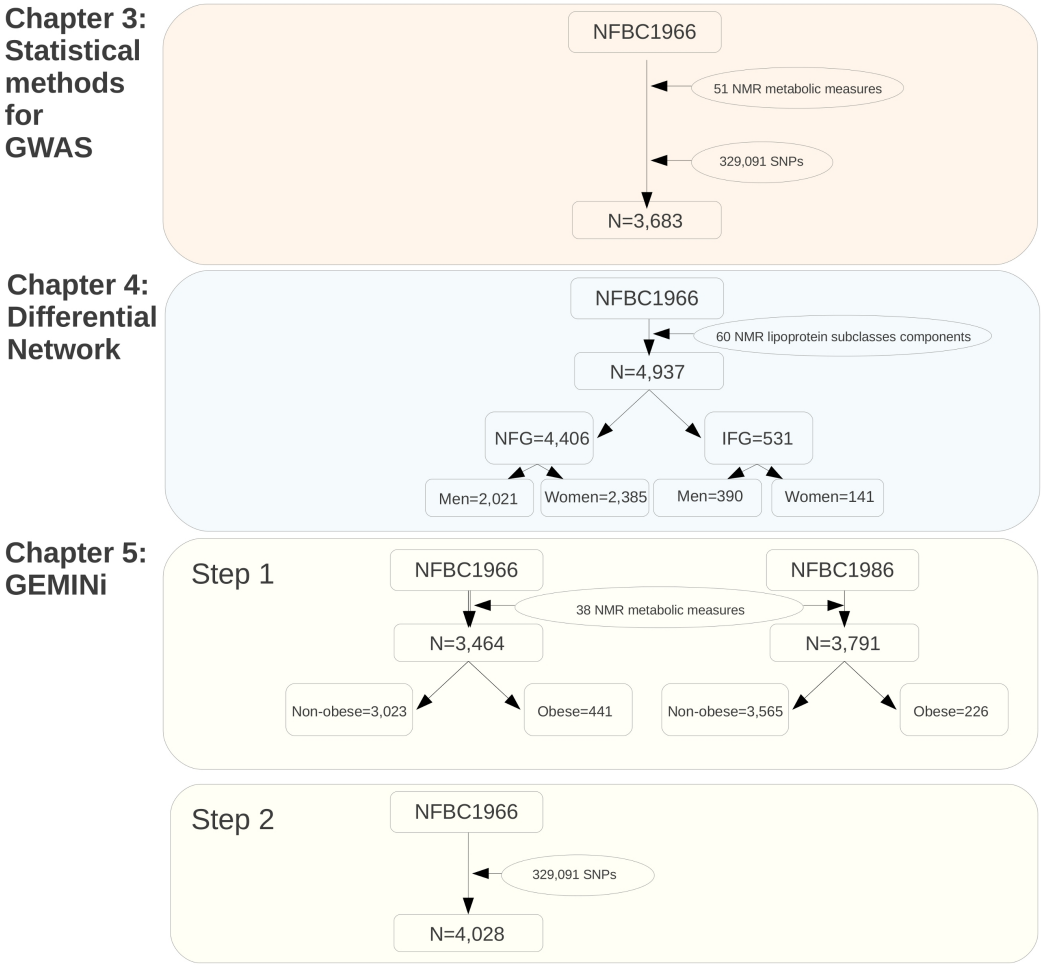
Table A.1 – continued from previous page

Full name	Abbreviation	Ch3	Ch4	Ch5	Ch6
Sphingomyelins	SM	✓			
Ratio of omega-3 fatty acids to total fatty acids	FAw3toFA	✓			
Ratio of omega-6 fatty acids to total fatty acids	FAw6toFA	✓			
Ratio of omega-9 and saturated fatty acids to total fatty acids	FAw79StoFA	✓			
Total phosphoglycerides to Total triglycerides	TGtoPG	✓			

Table A.2: List of all urinary metabolic measures included in this thesis

Full name	Abbreviation	Ch3	Ch4	Ch5	Ch6
4-deoxyerythronic acid	4-DEA				✓
3-hydroxyisovalerate	3-HV				✓
Lactate	Lac				✓
Alanine	Ala				✓
Acetate	AcO				✓
Pyruvate	Pyr				✓
Succinate	Suc				✓
Citrate	Cit				✓
Dimethylamine	DMA				✓
Trimethylamine	TMA				✓
Dimethylglycine	DMG				✓
Creatine	Crea				✓
Creatinine	Creat				✓
Proline betaine	PB				✓
P-cresol sulphate	PCS				✓
Phenylacetylglutamine	PAG				✓
Hippurate	Hip				✓
N-methyl-nicotinic acid	NMNA				✓
Trimethylamine oxide	TMAO				✓
Glycine	Gly				✓
Urea	Urea				✓
Formate	For				✓

Figure A.1: A flowchart for NFBC data analysis



Appendix B

Statistical methods for GWAS

Table B.1: Results from the multi-SNP analysis of serum metabolic traits

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Genes
XXL-VLDL-L	rs780094	2	27652888	C/T	4.98E-5	GCKR
	rs6472088	8	64381899	G/T	1.38E-5	YTHDF3
	rs12542080	8	88775120	A/G	1.71E-3	DCAFAL2
XL-VLDL-L	rs1260326	2	27642591	C/T	1.98E-7	GCKR
	rs6472088	8	64381899	G/T	7.38E-6	YTHDF3
	rs12542080	8	88775120	A/G	5.24E-4	DCAFAL2
	rs2248141	15	51959844	A/G	3.16E-4	UNC13C
	rs4985124	16	15032942	G/T	2.50E-5	PDXDC1
	rs7289448	22	31913986	A/G	2.00E-3	LARGE
L-VLDL-L	rs1260326	2	27642591	C/T	1.87E-9	GCKR
	rs10096633	8	19875201	C/T	1.11E-7	LPL
	rs4985124	16	15032942	G/T	4.74E-6	PDXDC1
M-VLDL-L	rs673548	2	21149196	A/G	**7.75E-10**	APOB
	rs780094	2	27652888	C/T	3.89E-9	GCKR
	rs2054046	2	70028725	C/T	3.09E-2	GMCL1
	rs10096633	8	19875201	C/T	1.78E-9	LPL
	rs2304821	12	69249886	A/G	3.42E-3	PTPR3
S-VLDL-L	rs6728178	2	21105581	A/G	**1.83E-12**	APOB
	rs1260326	2	27642591	C/T	5.45E-8	GCKR
	rs2054046	2	70028725	C/T	1.82E-2	GMCL1
	rs1328925	4	159547895	C/T	1.05E-3	TMEM144
	rs10096633	8	19875201	C/T	3.18E-8	LPL
	rs10984096	9	118418644	A/G	1.40E-4	
	rs2000571	11	116090743	A/G	1.12E-5	APOcluster
	rs2248141	15	51959844	A/G	2.40E-4	UNC13C
	rs4243298	18	62688709	G/T	1.11E-3	CDH19
	rs2303180	19	8401765	A/G	3.02E-6	MARCH
XS-VLDL-L	rs6728178	2	21105581	A/G	**1.45E-12**	APOB

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs2054046	2	70028725	C/T	2.18E-2	GMCL1
	rs1264542	6	30494975	A/G	1.00E-3	TRIM39
	rs4986947	6	52957189	A/G	6.29E-3	GSTA4
	rs6937974	6	133682141	C/T	5.01E-4	EYA4
	rs11793564	9	108613870	A/G	1.70E-4	ACTL7B
	rs1904683	10	52619168	G/T	3.75E-4	PRKG1
	rs10498404	14	48015708	A/G	5.64E-4	
	rs1532085	15	56470658	A/G	**5.53E-10**	LIPC
	rs261336	15	56529710	A/G	8.34E-7	LIPC
	rs4436850	18	25841455	A/G	1.58E-3	
	rs157580	19	50087106	A/G	1.22E-8	TOMM40, APOE
IDL-L	rs646776	1	109530572	C/T	4.00E-6	CELSR2
	rs6754295	2	21117835	G/T	3.07E-8	APOB
	rs693	2	21143847	A/G	3.68E-9	APOB
	rs2464698	2	70394102	C/T	4.45E-2	PCYOX1
	rs4384906	3	33215318	G/T	5.86E-4	SUSD5
	rs7655585	4	181174294	A/G	4.18E-4	
	rs1264542	6	30494975	A/G	1.12E-3	TRIM39
	rs3011286	9	132923544	C/T	6.61E-4	GSTA4
	rs1904694	10	52575500	A/G	1.52E-4	PRKG1
	rs6575501	14	94710082	A/C	9.48E-6	DICER1-AS1
	rs1532085	15	56470658	A/G	2.10E-7	LIPC
	rs12904245	15	77303043	C/T	1.16E-2	ANKRD34C
	rs11877591	18	26069966	A/G	1.16E-3	
	rs157580	19	50087106	A/G	1.29E-8	TOMM40, APOE
	rs405509	19	50100676	G/T	2.95E-1	APOE
	rs4821240	22	32947519	C/T	2.09E-2	LARGE
L-LDL-L	rs646776	1	109530572	C/T	2.55E-7	CELSR2
	rs693	2	21143847	A/G	1.56E-8	APOB
	rs4384906	3	33215318	G/T	1.63E-3	SUSD5
	rs7655585	4	181174294	A/G	2.87E-4	
	rs1264542	6	30494975	A/G	3.33E-3	TRIM39
	rs1904683	10	52619168	G/T	1.92E-4	GSTA4
	rs7159503	14	88458634	A/G	2.53E-3	TTC8
	rs10409243	19	10193988	C/T	2.79E-5	DNMT1
	rs157580	19	50087106	A/G	2.55E-7	TOMM40, APOE
	rs405509	19	50100676	G/T	1.92E-2	APOE
	rs5767585	22	45923688	A/C	2.13E-3	TBC1D22A
M-LDL-L	rs646776	1	109530572	C/T	1.42E-7	CELSR2
	rs693	2	21143847	A/G	5.06E-8	APOB
	rs1904683	10	52619168	G/T	2.39E-4	PRKG1
	rs2965101	19	49929652	C/T	1.14E-5	BCL3
S-LDL-L	rs646776	1	109530572	C/T	2.78E-7	CELSR2
	rs693	2	21143847	A/G	2.53E-7	APOB
	rs1904683	10	52619168	G/T	2.88E-4	PRKG1

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs2965101	19	49929652	C/T	5.30E-6	BCL3
XL-HDL-L	rs7650071	3	137460003	A/G	9.36E-3	PCCB
	rs13156109	5	40941833	C/T	7.72E-5	C7
	rs1935420	6	82745959	A/G	5.60E-3	IBTK
	rs9285492	6	140837328	C/T	2.85E-4	
	rs11005804	10	58887377	C/T	8.21E-3	
	rs1532085	15	56470658	A/G	**4.00E-16**	LIPC
	rs261336	15	56529710	A/G	**1.94E-10**	LIPC
	rs7178362	15	56634555	C/T	3.09E-8	LIPC
	rs1532624	16	55562980	A/C	1.16E-6	CETP
	rs7499892	16	55564091	C/T	4.87E-7	CETP
	rs6065906	20	43987422	C/T	**3.18E-11**	PLTP
L-HDL-L	rs6697783	1	221934995	A/G	4.25E-2	LBR
	rs4654103	1	242851660	G/T	2.55E-3	SMYD3
	rs17148957	5	121551075	C/T	7.74E-4	ZNF474
	rs11154522	6	130483157	A/G	5.15E-4	L3MBTL3
	rs9397900	6	150757866	A/G	2.26E-2	IYD
	rs2977469	8	141618473	A/G	1.19E-2	EFI2C2
	rs4836625	9	128596392	C/T	6.68E-3	ZER1
	rs766606	13	94452168	A/C	5.95E-2	ABCC4
	rs1532085	15	56470658	A/G	**2.27E-18**	LIPC
	rs473224	15	56524633	G/T	5.10E-6	LIPC
	rs3764261	16	55550825	A/C	**9.32E-18**	CETP
	rs8054916	16	56089134	A/G	9.48E-4	CCDC102A
	rs1121985	16	77920580	A/C	1.88E-4	WWOX
	rs1800961	20	42475778	C/T	4.38E-9	HNF4A
	rs4810479	20	43978455	C/T	2.03E-7	PLTP
M-HDL-L	rs12547884	8	10032925	A/G	1.05E-5	MSRA
	rs2047554	12	24945476	C/T	1.45E-2	BCAT1
	rs3764261	16	55550825	A/C	**4.89E-11**	CETP
	rs1799281	19	53640582	A/G	7.88E-4	GRIN2D
S-HDL-L	rs4453935	4	13585054	C/T	9.16E-7	BOD1L1
	rs1390498	6	140586160	A/G	9.07E-4	
	rs3739942	9	137052237	C/T	2.78E-5	TRAF2
	rs4917911	10	102549411	A/G	1.11E-3	PAX2
	rs17152324	10	126547196	A/G	4.65E-4	FAM175B
	rs6065906	20	43987422	C/T	7.34E-9	PLTP
ApoA1	rs6697783	1	221934995	A/G	4.90E-4	LBR
	rs10495736	2	23131509	A/G	1.53E-4	KLHL29
	rs1940309	4	116444565	C/T	1.15E-3	NDST4
	rs425974	4	189752867	A/C	3.44E-6	RP11-219G10
	rs652888	6	31959213	A/G	8.96E-4	EHMT2
	rs10499028	6	100459334	A/G	9.12E-5	MCHR2
	rs2171981	6	168344796	C/T	1.85E-6	FRMD1
	rs2350842	7	9207348	A/C	1.52E-2	

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs174570	11	61353788	C/T	7.07E-5	FADS2-3
	rs1532085	15	56470658	A/G	5.30E-7	LIPC
	rs3764261	16	55550825	A/C	**2.10E-15**	CETP
	rs1465354	17	50789174	C/T	1.64E-3	HLF
	rs1948557	19	36410329	A/G	1.92E-4	TSHZ3
	rs8110189	19	61522069	A/C	7.99E-3	ZSCAN5A
ApoB	rs6728178	2	21105581	A/G	**1.79E-10**	APOB
	rs693	2	21143847	A/G	**2.43E-9**	APOB
	rs1260326	2	27642591	C/T	6.53E-5	GCKR
	rs2054046	2	70028725	C/T	3.03E-3	GMCL1
	rs1904683	10	52619168	G/T	3.29E-5	PRKG1
	rs2717440	12	69245373	C/T	5.31E-4	PTPRB
	rs7186194	16	82183721	G/T	1.68E-4	CDH13
	rs10409243	19	10193988	C/T	6.73E-6	DNMT1
	rs157580	19	50087106	A/G	3.21E-7	TOMM40, APOE
3-hydroxybutyrate	rs10510515	3	21695378	A/G	2.11E-3	ZNF385D
	rs2328943	13	74918850	C/T	1.62E-3	TBC1D4
	rs2242629	14	76403716	C/T	1.77E-3	C14orf166B
Acetate	rs848792	3	126057032	C/T	6.84E-2	ITGB5
	rs2968505	7	71339280	C/T	6.13E-3	CALN1
	rs10765031	10	132754466	C/T	6.51E-2	TCERG1L
	rs1564580	17	30587907	C/T	2.20E-2	SLFN5
Acetoacetate	rs4778908	15	79632208	A/G	9.97E-3	TMC3
	rs2328559	20	20726180	A/G	1.98E-4	RALGAPA2
Alanine	rs1260326	2	27642591	C/T	5.72E-7	GCKR
	rs492272	3	162947528	G/T	2.65E-3	OTOL1
	rs6829634	4	45453629	A/G	3.79E-4	GABRG1
	rs4415952	14	72258906	C/T	7.03E-4	DPF3
Citrate	rs6677954	1	172445619	A/G	1.75E-2	TNR
	rs701157	1	226981561	C/T	2.91E-2	LOC729257
	rs1437313	2	139659908	A/G	2.58E-6	NXPH2
	rs4667819	2	166113483	C/T	5.50E-2	CSRNP3
	rs12485763	3	27558321	A/G	1.62E-1	SLC4A7
	rs10236225	7	9999670	C/T	1.08E-4	NA
	rs8074277	17	78382917	C/T	1.39E-2	TBCD
	rs712964	22	17530671	C/T	**1.84E-11**	CLTCL1
Creatinine	rs1113958	2	148314711	A/G	6.97E-4	ACVR2A
	rs12506374	4	8272605	C/T	1.00E-3	ABLIM2
	rs11751605	6	160933641	C/T	1.67E-3	LPA
	rs2078980	7	28350789	A/G	3.26E-3	CREB5
	rs1939470	11	95485565	A/G	6.75E-3	MAML2
	rs10777233	12	89109577	A/G	9.26E-4	
	rs9573940	13	34916882	A/C	2.02E-3	NBEA
	rs2239990	15	89624441	A/G	2.80E-5	SV2B

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs7505939	18	7987833	C/T	1.87E-3	PTPRM
	rs565798	18	8125770	A/G	1.39E-3	UQCRFS1
	rs6022671	20	51743446	A/G	9.25E-4	ZNF217
Glucose	rs4666204	2	29457029	C/T	8.95E-4	ALK
	rs7623833	3	64314296	C/T	2.53E-4	PRICKLE2
	rs4831130	3	117522532	G/T	3.30E-4	LSAMP
	rs11714053	3	133332108	C/T	1.54E-3	CPNE4
	rs6788064	3	144959109	A/G	1.16E-3	SLC9A9
	rs1864372	4	147501177	A/G	8.91E-5	LSM6
	rs2069297	5	93609173	C/T	1.17E-2	KIAA825
	rs9275495	6	32781552	A/T	4.21E-3	HLA-DQB1
	rs1373130	9	99233349	C/T	5.12E-4	SEC61B
	rs17267324	10	14805582	C/G	1.48E-4	FAM107B
	rs543970	11	125421262	C/T	9.66E-7	CDON
	rs6574591	14	80152153	C/T	7.14E-2	CEP128
	rs6022671	20	51743446	A/G	1.21E-5	ZNF217
	rs768832	21	27843561	A/C	3.18E-3	
Glutamine	rs1000226	4	63199692	C/T	2.88E-3	LPHN3
	rs3802037	7	90161256	A/G	9.70E-4	CDK14
	rs7845972	8	69683092	A/G	5.25E-3	C8orf34
	rs7015712	8	75643058	C/T	1.18E-2	GDAP1
	rs2620636	8	145748936	C/T	2.27E-3	ARHGAP39
	rs11817730	10	99348501	A/G	1.18E-4	PI4K2A
	rs4919305	10	101123989	G/T	9.53E-4	CNNM1
	rs1366906	11	44454928	G/T	7.56E-2	CD82
	rs774211	12	55207206	C/T	4.02E-8	RBMS2
	rs4778464	15	20537129	G/T	1.81E-5	CYFIP1
	rs1194463	15	53060288	C/T	2.27E-2	RSL24D1
Glycerol	rs3737593	1	3719693	A/G	1.43E-2	LRRC47
	rs213473	1	54642991	A/G	5.13E-1	SSBP3
	rs17016934	1	106573537	C/T	7.00E-3	
	rs7538977	1	108465875	A/G	1.98E-1	NBPF4
	rs12060029	1	208959970	C/T	1.28E-1	TMEM206
	rs9288032	2	180143698	G/T	2.45E-3	ZNF385B
	rs7639655	3	99853530	C/T	5.67E-1	CPOX
	rs4422306	3	108282550	G/T	6.22E-1	RP11-280F2
	rs1242064	3	144854964	C/T	1.02E-3	SLC9A9
	rs3843376	3	198302734	C/T	8.26E-3	DLG1
	rs13125250	4	8128183	C/T	9.15E-2	ABMIM2
	rs11734226	4	178536007	C/T	3.66E-2	NEIL3
	rs2102575	4	187506653	A/G	5.89E-4	CYP4V2
	rs29560	5	123116099	A/G	2.39E-2	CSNK1G3
	rs17067004	5	165847033	C/T	2.28E-1	
	rs9354652	6	68738367	C/T	3.09E-2	
	rs3793227	7	158385404	A/G	3.86E-4	VIPR2

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs4260902	8	23426356	A/G	5.58E-2	SLC25A37
	rs11774796	8	25094151	A/C	3.48E-2	DOCK5
	rs1535658	9	10361281	A/G	2.73E-1	PTPRD
	rs4509708	10	41934567	A/G	4.25E-2	CCNYL2
	rs3844492	10	49492807	A/G	4.34E-4	ARHGAP22
	rs7075347	10	76203313	C/T	3.52E-1	KAT6B
	rs7911910	10	126021464	A/G	2.42E-2	OAT
	rs1792549	11	15182234	A/G	2.74E-2	INSC
	rs2306365	11	65183922	A/G	2.16E-2	RELA
	rs10743621	12	28636597	A/G	3.78E-1	CCDC91
	rs4293188	12	38522359	A/C	6.01E-4	SLC2A13
	rs2044747	14	42393956	A/C	6.58E-3	
	rs6573242	14	58623902	A/G	1.27E-2	DAAM1
	rs922178	15	62056463	C/T	3.52E-3	RSL24D1
	rs7237794	18	9941304	C/T	2.03E-1	VAPA
	rs2287671	19	11469931	A/C	1.14E-2	ZNF653
	rs6068407	20	51013811	A/G	9.16E-1	TSHZ2
	rs394484	21	19299986	A/G	2.87E-1	PP1AP22
	rs457779	21	27790008	A/G	1.62E-2	
	rs2092331	22	33956098	C/T	1.22E-1	HMGXB4
	rs2267341	22	35301475	C/T	1.07E-3	CACNG2
Histine	rs2731672	5	176775080	A/G	3.86E-2	GRK6
	rs7800178	7	1400873	C/T	5.06E-5	PSMG3
	rs10486100	7	10599434	C/T	2.25E-4	NDUFA4
	rs190925	10	70990949	A/G	1.00E-4	NEUROG3
Isoleucine	rs1334336	1	88104461	C/T	1.78E-3	
	rs9287724	2	10939186	A/C	7.96E-3	
	rs1260326	2	27642591	C/T	1.47E-7	GCKR
	rs1399627	2	226908916	A/G	1.58E-2	
	rs2943641	2	226919250	C/T	1.59E-2	IRS1
	rs2943645	2	226924685	C/T	1.48E-2	
	rs908265	2	239998541	C/T	4.47E-6	HDAC4
	rs688136	6	153172175	C/T	3.55E-3	YIP
	rs4719934	7	28261492	A/G	1.34E-6	CREB5
	rs1421320	7	154801369	A/C	5.04E-3	CNPY1
	rs7839522	8	94909310	A/G	4.73E-5	TMEM67
	rs4905214	14	94067248	A/G	8.28E-4	SERPINA12
	rs12460985	19	50382729	C/T	1.83E-3	MARK4
Lactate	rs12069195	1	202947755	A/G	1.75E-1	SRGAP2
	rs3008206	1	210334156	C/T	4.76E-2	PROX1
	rs17047306	1	214632092	C/T	3.66E-1	RRP15
	rs3795367	1	229052855	C/T	2.42E-5	SIPA1L2
	rs6709502	2	30922575	C/T	3.11E-4	CAPN13
	rs7593405	2	78240187	C/T	1.00E-1	
	rs1511199	2	167729549	G/T	5.91E-3	XIRP2

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs4667173	2	188258463	A/G	4.87E-2	TFP1
	rs6782432	3	58786097	A/G	4.09E-1	C3orf67
	rs2968671	4	5092742	A/C	6.58E-3	CYTL1
	rs6447091	4	41392741	A/G	4.44E-2	LIMCH1
	rs2331548	4	170337334	A/G	2.38E-1	SH3RF1
	rs10520847	5	16956718	A/C	9.02E-3	MYO10
	rs6458287	6	12580242	C/T	3.59E-2	RPL15P3
	rs1782449	6	91206909	C/T	2.47E-1	MAP3K7
	rs9487754	6	112343990	C/T	2.22E-3	FYN
	rs4398772	6	115869717	A/G	8.17E-1	
	rs7803071	7	83104559	C/T	6.83E-2	SEMA3A
	rs7796348	7	157098551	C/T	7.00E-3	PTPRN2
	rs10105819	8	19272477	C/T	1.33E-1	SH2D4A
	rs7863972	9	3334203	C/T	1.19E-1	RFX3
	rs875295	12	26232155	A/C	5.09E-4	SSPN
	rs17631397	13	29351522	A/G	1.51E-1	UBL3
	rs1929753	13	75709051	C/T	2.32E-1	LMO7
	rs171603	16	9375312	A/G	1.16E-3	RPL21P119
	rs4455065	18	5064049	A/G	2.03E-2	C18orf42
	rs10483243	22	47390986	A/C	7.84E-4	FAM19A5
Leucine	rs3784191	14	31993041	C/T	5.71E-4	AKAP6
	rs3784759	15	87184048	C/T	5.81E-4	ACAN
Phenylalanine	rs2731672	5	176775080	A/G	4.59E-4	GRK6
Pyruvate	rs360102	1	222374597	A/G	1.07E-4	TMEM63A
	rs10802397	1	242860493	A/G	3.00E-5	SYD3
	rs1260326	2	27642591	C/T	2.85E-5	GCKR
	rs1356609	3	34278287	A/G	2.54E-4	PDCD6IP
	rs6899062	5	25646520	A/C	1.42E-2	
	rs1039693	6	54705257	A/G	2.35E-2	FAM83B
	rs2798976	6	78314890	C/T	2.46E-2	HTR1B
	rs7091933	10	9477937	A/G	9.46E-5	
	rs12367514	12	68311754	C/T	4.56E-3	BEST3
	rs2717417	12	69315698	C/T	6.40E-3	PTPRR
	rs189247	15	95387634	A/G	3.04E-3	SATA8
	rs1061273	19	5093817	C/T	1.83E-4	KDM4B
	rs4805339	19	34365730	G/T	5.97E-4	
	rs2423967	20	15529484	A/G	4.73E-3	MACROD2
Tyrosine	rs6664203	1	104328358	A/G	6.42E-4	AMY1C
	rs4696687	4	8057265	C/T	2.94E-3	AFAP1
	rs1947372	4	124781096	A/G	2.02E-4	SPRY1
	rs7451051	6	111851511	A/G	4.07E-8	REV3L
	rs517684	11	78721866	A/G	4.21E-5	TENM4
	rs4346460	20	23517400	C/T	1.27E-5	CST9
Urea	rs7545292	1	208667453	G/T	7.32E-3	DTL
	rs901666	1	241418873	A/G	1.91E-2	EFCAB2

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs4420731	2	22910802	A/C	4.56E-2	
	rs3732191	2	47951936	A/C	4.16E-2	FBXO11
	rs6781602	3	81497660	C/T	3.44E-1	GBE1
	rs4234851	4	81429574	C/T	2.65E-1	
	rs7655290	4	141543207	A/G	3.04E-1	SCOC
	rs17476583	5	85958923	A/C	1.10E-2	
	rs10040945	5	96442155	C/T	5.26E-2	LIX1
	rs11738868	5	169880846	A/G	7.32E-3	KCNIP1
	rs4718709	7	67011579	A/G	1.37E-2	
	rs7830040	8	83079508	C/T	1.90E-2	SNX16
	rs1180628	8	116500037	A/C	8.59E-2	TRPS1
	rs1132743	9	131401844	A/G	4.59E-2	PRRC2B
	rs1926559	10	114042907	C/T	1.79E-4	TECTB
	rs2238023	12	2072798	C/T	3.77E-1	CACNA1C
	rs2046932	12	38866707	A/G	7.21E-2	
	rs17538030	14	32639991	C/T	1.01E-1	NPAS3
	rs7497239	15	22459149	A/G	2.24E-1	NPAP1
	rs6502246	17	13124626	A/C	3.58E-2	HS3ST3A1
	rs2614996	18	40534091	A/G	2.30E-1	SETBP1
	rs1469580	18	64802946	A/G	2.65E-1	CCDC102B
	rs4471960	20	8398247	A/G	7.20E-2	PLCB1
	rs6031225	20	41981244	A/C	5.48E-2	TOX2
Valine	rs958325	4	89566860	A/G	2.92E-6	RP11-10L7.1
	rs454282	19	35291623	A/G	2.23E-3	URI1
EstC	rs646776	1	109530572	C/T	2.81E-6	CELSR2
	rs693	2	21143847	A/G	4.05E-6	APOB
	rs4384906	3	33215318	G/T	4.20E-4	SUSD5
	rs7655585	4	181174294	A/G	1.74E-4	
	rs2474152	6	144054752	C/T	6.88E-4	PHACTR2
	rs1904683	10	52619168	G/T	2.22E-4	PRKG1
	rs7186194	16	82183721	G/T	4.86E-4	CDH13
	rs878971	18	31300672	A/G	8.14E-3	INO80C
FreeC	rs693	2	21143847	A/G	2.35E-8	APOB
	rs4852500	2	70461429	A/C	2.07E-2	DFAM136A
	rs1904683	10	52619168	G/T	2.90E-4	PRKG1
	rs7159503	14	88458634	A/G	2.46E-4	TTC8
	rs1532085	15	56470658	A/G	8.95E-9	LIPC
	rs897587	17	74907220	C/T	2.55E-3	RBFOX3
	rs878971	18	31300672	A/G	1.10E-2	INO80C
FAw3	rs3934285	1	104310962	A/G	6.09E-3	
	rs2143109	1	167504689	A/G	7.37E-4	PRRX1
	rs11582261	1	178141143	G/T	2.29E-4	CACNA1E
	rs7603546	2	240455954	C/T	7.90E-4	
	rs6868240	5	4991915	A/C	5.39E-4	ADAMTS16
	rs2474152	6	144054752	C/T	2.63E-4	PHACTR2

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs11592435	10	67332087	A/G	9.52E-7	CTNNA3
	rs444323	10	109904338	A/G	3.41E-3	
	rs1898354	10	118818194	C/T	2.25E-4	KIAA1598
	rs102275	11	61314379	C/T	2.48E-7	FADS1-3
	rs4936947	11	124003995	C/T	8.74E-4	TBRG1
	rs12146774	12	118136223	C/T	6.46E-4	HSPB8
	rs3786394	18	31313063	A/C	3.43E-3	INO80C
	rs2304130	19	19650528	A/G	5.57E-6	ZNF101
FAw6	rs693	2	21143847	A/G	7.63E-6	APOB
	rs10185680	2	24186957	A/G	6.88E-5	APOB
	rs4616468	2	64015783	C/T	1.62E-4	UGP2
	rs1900741	2	135836232	C/T	1.57E-6	ZRANB3
	rs4384906	3	33215318	G/T	3.27E-4	SUSD5
	rs7653056	3	177272705	A/G	1.73E-2	NAALADL2
	rs6818812	4	164995577	C/T	4.11E-3	MARCH
	rs7655585	4	181174294	A/G	2.33E-4	
	rs17492381	5	152429608	C/T	4.73E-3	GRIA1
	rs10807218	6	40057564	C/T	4.44E-4	MOCS1
	rs2474152	6	144054752	C/T	1.45E-3	PHACTR2
	rs6969660	7	7619666	A/C	4.17E-5	AC006465.3
	rs1904683	10	52619168	G/T	4.03E-4	PRKG1
	rs9595132	13	43946832	A/G	2.65E-4	TSC22D1
	rs9634510	13	97572905	C/T	2.51E-5	FARP1
	rs1927384	13	101943751	A/G	1.37E-2	TPP2
	rs7193689	16	7474240	A/G	2.97E-3	RBFOX1
	rs7186194	16	82183721	G/T	5.23E-4	CDH13
	rs133903	22	24490565	A/C	1.02E-3	MYO18B
FAw79S	rs6728178	2	21105581	A/G	4.70E-5	APOB
	rs1260326	2	27642591	C/T	2.25E-8	GCKR
	rs10495936	2	46819850	C/T	7.41E-5	SOCS5
	rs2474152	6	144054752	C/T	1.69E-3	PHACTR2
	rs17152855	7	25692400	A/C	1.65E-3	NFE2L3
	rs9634510	13	97572905	C/T	1.49E-4	FARP1
	rs878971	18	31300672	A/G	2.26E-3	INO80C
TotFA	rs6728178	2	21105581	A/G	1.96E-5	APOB
	rs10185680	2	24186957	A/G	2.79E-4	APOB
	rs1260326	2	27642591	C/T	2.05E-7	GCKR
	rs2474152	6	144054752	C/T	7.09E-4	PHACTR2
	rs4722537	7	25700068	G/T	2.51E-4	NFE2L3
	rs2339732	10	52597248	G/T	4.29E-4	PRKG1
	rs9634510	13	97572905	C/T	5.00E-5	FARP1
	rs878971	18	31300672	A/G	2.28E-3	INO80C
LA	rs956796	1	146999164	A/G	2.85E-3	ANP32E
	rs693	2	21143847	A/G	8.01E-7	APOB
	rs10185680	2	24186957	A/G	3.52E-4	APOB

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs2464698	2	70394102	C/T	9.64E-2	PCYOX1
	rs968051	2	103911743	A/G	4.99E-4	
	rs1900741	2	135836232	C/T	2.08E-6	
	rs4691101	4	165010531	C/T	3.37E-3	MARCH
	rs9368699	6	31910520	C/T	2.19E-2	SNORD52
	rs10947842	6	40023327	C/T	6.67E-4	MOCS1
	rs2474152	6	144054752	C/T	3.67E-3	PHACTR2
	rs6969660	7	7619666	A/C	7.87E-6	AC006465.3
	rs1904683	10	52619168	G/T	4.30E-4	PRKG1
	rs174546	11	61326406	C/T	**2.46E-10**	FADS1-3
	rs9634510	13	97572905	C/T	5.59E-5	FARP1
	rs7186194	16	82183721	G/T	2.93E-3	CDH13
	rs2410062	21	39150790	A/G	1.70E-3	ETS2
TotPUFA	rs2031373	1	62795310	A/C	1.00E-3	DOCK7
	rs3934285	1	104310962	A/G	1.58E-3	
	rs10924989	1	243826691	A/G	7.54E-4	ZNF496
	rs11682343	2	85580646	A/G	1.02E-2	SH2D6
	rs4384906	3	33215318	G/T	9.91E-4	SUSD5
	rs1866929	3	73997737	C/T	1.57E-2	
	rs10511248	3	107260259	G/T	6.20E-4	CBLB
	rs6830724	4	90550905	C/T	4.99E-5	GPRIN3
	rs2389925	4	121237670	A/G	3.93E-4	MAD2L1
	rs12659587	5	50305167	C/T	4.91E-3	PARP8
	rs9790855	5	77634610	C/T	8.53E-3	AP3B1
	rs11746854	5	78864910	A/G	3.75E-4	HOMER1
	rs17492381	5	152429608	C/T	2.01E-2	KCTD16
	rs2474152	6	144054752	C/T	9.94E-4	PHACTR2
	rs7780859	7	20540830	C/T	2.05E-2	ABCB5
	rs1881726	7	147371266	C/T	3.81E-3	CNTNAP2
	rs2978521	8	62279233	A/G	1.28E-3	CLVS1
	rs6982636	8	126548497	A/G	5.88E-4	TRIB1
	rs10808546	8	126565000	C/T	2.41E-4	TRIB1
	rs2150197	9	116782172	A/G	2.00E-3	
	rs12685378	9	128827265	A/G	3.48E-4	NUP188
	rs790661	10	106779870	A/G	3.31E-4	SORCS3
	rs1898354	10	118818194	C/T	4.06E-3	KIAA1598
	rs174556	11	61337211	C/T	**1.01E-51**	FADS1
	rs174570	11	61353788	C/T	**6.41E-34**	FADS2-3
	rs1149593	11	76141588	A/G	4.92E-4	TSKU
	rs4936947	11	124003995	C/T	1.04E-3	
	rs11046207	12	21887312	C/T	1.64E-3	ABCC9
	rs10843712	12	30339528	C/T	5.16E-4	IPO8
	rs7159503	14	88458634	A/G	3.65E-3	TTC8
	rs3786394	18	31313063	A/C	1.53E-3	INO80C
	rs7247764	19	50367713	C/T	1.39E-2	MARK4
	rs363452	21	29901558	A/G	9.76E-4	GRIK1

Continued on next page

Table B.1 – continued from previous page

Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs412519	21	42916410	A/G	7.35E-3	AP0011626.1
	rs761336	22	36333741	A/C	2.59E-5	GGA1
DHA	rs4970405	1	1088878	A/G	1.16E-4	C1orf159
	rs4573267	8	64168117	A/G	7.93E-3	TTPA
	rs1898354	10	118818194	C/T	1.18E-3	KIAA1598
	rs174537	11	61309256	G/T	7.88E-6	PURAQTL1
	rs3786394	18	31313063	A/C	5.48E-3	INO80C
	rs2304130	19	19650528	A/G	2.93E-5	ZNF101
	rs761336	22	36333741	A/C	8.43E-6	GGA1
MUFA	rs6728178	2	21105581	A/G	1.14E-6	APOB
	rs1260326	2	27642591	C/T	**1.41E-9**	GCKR
	rs10495936	2	46819850	C/T	2.38E-4	SOCS5
	rs9683662	4	62577150	C/T	2.65E-2	LPHN3
	rs17152855	7	25692400	A/C	4.11E-3	NFE2L3
	rs9634510	13	97572905	C/T	2.73E-4	FARP1
	rs4243298	18	62688709	G/T	6.09E-4	CDH19
TotPG	rs10185680	2	24186957	A/G	2.93E-5	APOB
	rs2474152	6	144054752	C/T	8.90E-4	PHACTR2
	rs6969660	7	7619666	A/C	7.91E-5	AC006465.3
	rs1904683	10	52619168	G/T	4.50E-4	PRKG1
	rs9634510	13	97572905	C/T	3.09E-4	FARP1
	rs1532085	15	56470658	A/G	5.58E-8	LIPC
	rs878971	18	31300672	A/G	2.24E-3	INO80C
	rs1948557	19	36410329	A/G	2.27E-3	TSHZ3
PC	rs10185680	2	24186957	A/G	7.18E-4	APOB
	rs1866852	2	191066692	C/T	7.50E-4	
	rs425974	4	189752867	A/C	5.21E-4	RP11-219G10
	rs16871677	5	3448061	C/T	2.24E-3	IRX1
	rs2474152	6	144054752	C/T	2.85E-4	PHACTR2
	rs6969660	7	7619666	A/C	6.98E-5	AC006465.3
	rs2270214	7	140276302	A/G	1.34E-2	MRPS33
	rs1904683	10	52619168	G/T	8.30E-4	PRKG1
	rs7122481	11	132467431	A/G	1.62E-4	OPCML
	rs9634510	13	97572905	C/T	1.04E-3	FARP1
	rs3649	15	43696596	C/T	8.78E-3	BLOC1S6
	rs1532085	15	56470658	A/G	8.69E-9	LIPC
	rs878971	18	31300672	A/G	1.76E-3	INO80C
	rs1948557	19	36410329	A/G	1.03E-3	TSHZ3
	rs3804019	21	38568806	A/C	2.15E-3	CBLB
SM	rs991694	3	30677144	A/G	2.10E-3	TGFBR2
	rs10511248	3	107260259	G/T	1.07E-3	CBLB
	rs4723497	7	36046214	C/T	8.99E-5	EEPD1
	rs878971	18	31300672	A/G	2.50E-3	INO80C
FAw3toFA	rs2056387	1	234250153	A/C	6.79E-5	RYR2
	rs4555301	2	29551330	A/G	6.01E-3	ALK

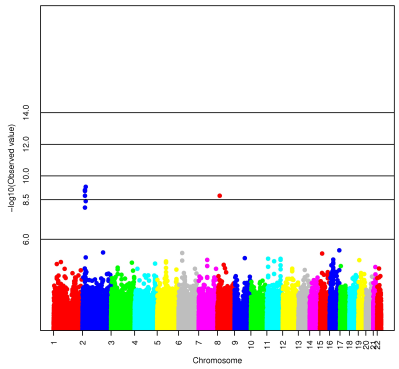
Continued on next page

Table B.1 – continued from previous page

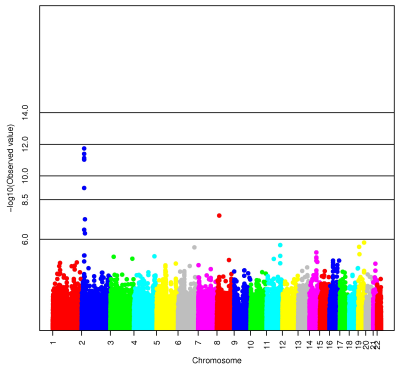
Metabolic T	Markers	Chr	Pos	Al1/Al2	P	Gene
	rs25440	2	34413857	C/T	6.38E-3	
	rs614281	2	45778107	C/T	9.37E-5	PRKCE
	rs2203586	2	164707177	A/G	4.85E-4	FIGN
	rs1532191	3	164888628	G/T	1.43E-4	
	rs7628387	3	178046264	A/C	1.70E-3	TBL1XR1
	rs1373947	5	13677894	A/C	4.59E-3	DNAH5
	rs319003	5	41399645	C/T	1.18E-4	PLCXD3
	rs3816024	5	108463913	A/G	1.49E-3	FER
	rs4583918	5	144279029	A/C	5.72E-4	
	rs9295553	6	22650885	A/G	3.83E-5	HDGFL1
	rs9373409	6	144328911	C/T	2.33E-3	PLAGL1
	rs2072184	7	19522123	A/G	1.55E-5	TMEM196
	rs2289496	8	98357340	A/C	3.65E-4	TSPYL5
	rs11166643	8	135723522	C/T	2.14E-4	ZAAT
	rs7856857	9	101529148	A/C	1.77E-2	GRIN3A
	rs174546	11	61326406	C/T	**1.18E-17**	FADS1-3
	rs10771468	12	29089014	A/G	8.62E-3	FAR2
	rs1679802	14	55563670	A/G	2.64E-5	PEL12
	rs3743398	15	87199411	C/T	8.16E-5	ACAN
	rs30988	16	2214301	C/T	1.08E-3	E4F1
	rs485018	17	72626322	A/G	1.20E-2	SEC14L1
	rs2000778	18	56271952	A/G	8.63E-4	MC4R
	rs736950	22	26043329	A/G	5.02E-6	MN1
FAw79StoFA	rs6662607	1	67733523	C/T	1.04E-4	GADD45A
	rs1260326	2	27642591	C/T	2.01E-6	GCKR
	rs319003	5	41399645	C/T	1.69E-4	PLCXD3
	rs2072184	7	19522123	A/G	4.74E-5	TMEM196
	rs6094373	20	35402488	C/T	5.51E-4	SRC
TGtoPG	rs673548	2	21149196	A/G	8.45E-1	APOB
	rs1260326	2	27642591	C/T	2.00E-2	GCKR
	rs4724206	7	43263377	A/G	6.73E-1	HECW1
	rs10096633	8	19875201	C/T	4.61E-1	LPL
	rs1044531	9	114178204	C/T	6.63E-2	CRM1
	rs4573621	10	113977538	A/G	9.36E-1	GPAM
	rs2543356	14	75658039	C/T	2.59E-1	GPATCH2L
	rs7149330	14	89705959	A/G	1.13E-1	KCNK13

Metabolic T, each of the fifty one serum metabolites included in the analysis, **Markers**, SNPs included in the best-fitting models found for each of the NMR serum metabolites for type I error $\alpha = 10^{-5}$ and shape parameter $\lambda = 1$; **Chr**, Chromosome; **Pos**, SNPs position at the corresponding chromosome; **A1/A2**, listed are 1^{st} major allele, 2^{nd} minor allele; **P**, p-value obtained in the single-SNP analysis for each trait; **Gene**, the plausible candidate gene in the locus or the nearest annotated gene to the lead SNP.

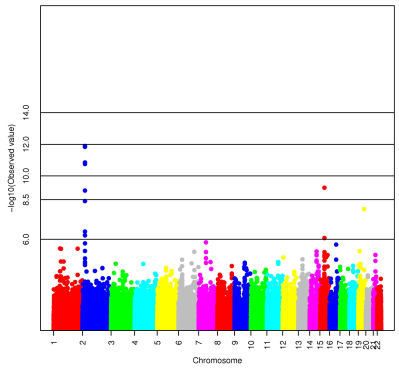
Figure B.1: The Manhattan plots of P-values for the leading traits obtained in the single-SNP association analysis.



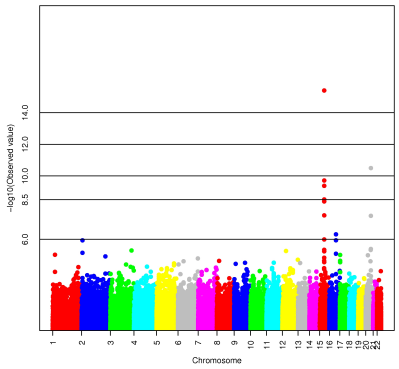
(a) Total lipids in medium VLDL



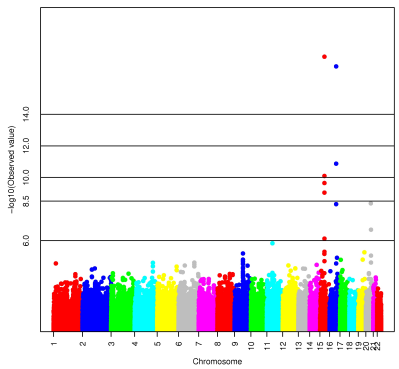
(b) Total lipids in small VLDL



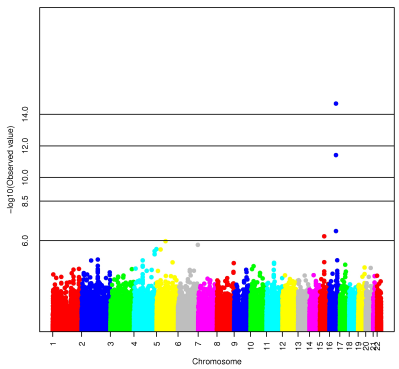
(c) Total lipids in very small VLDL



(d) Total lipids in very large HDL

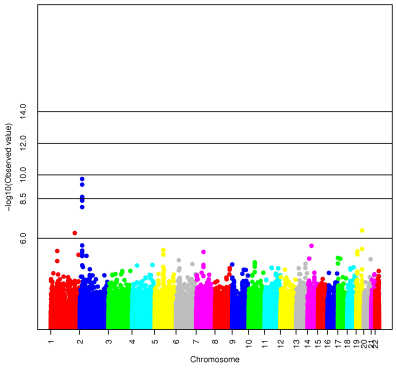


(e) Total lipids in large HDL

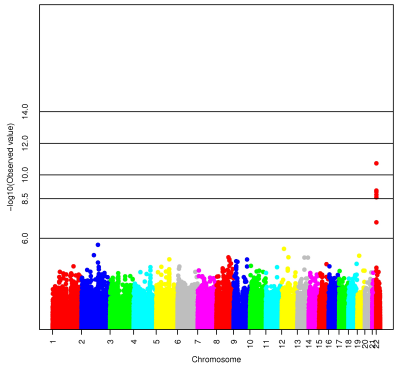


(f) Apolipoprotein A1

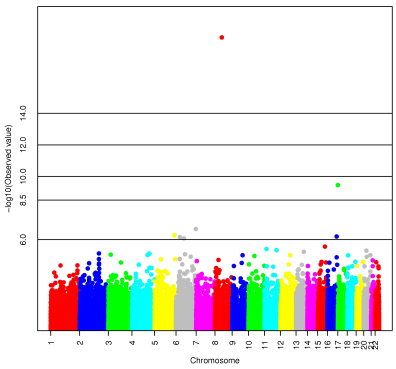
Figure B.1: The Manhattan plots of P-values for the leading traits obtained in the single-SNP association analysis.



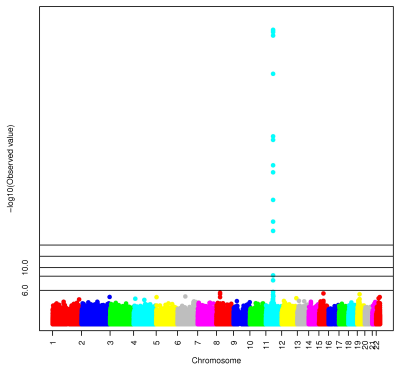
(g) Apolipoprotein B



(h) Citrate

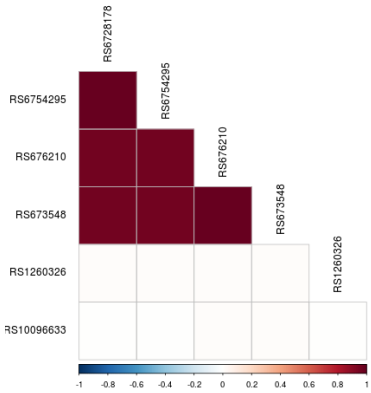


(i) Creatinine

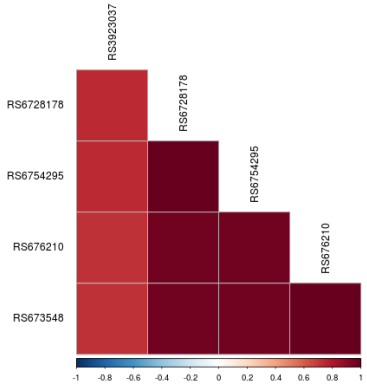


(j) Total PUFA

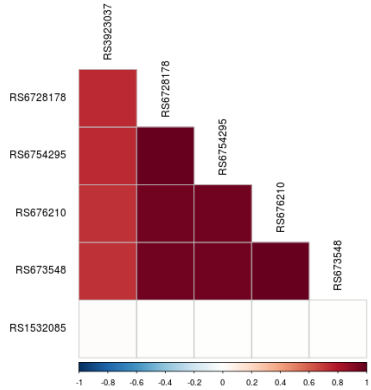
Figure B.2: Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the single-SNP analysis.



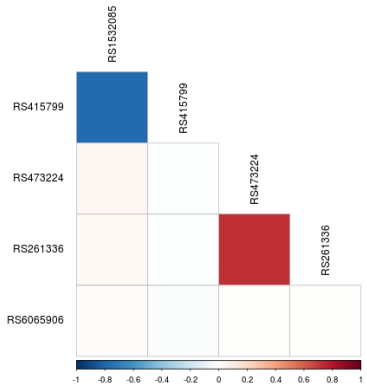
a) Total lipids in medium VLDL



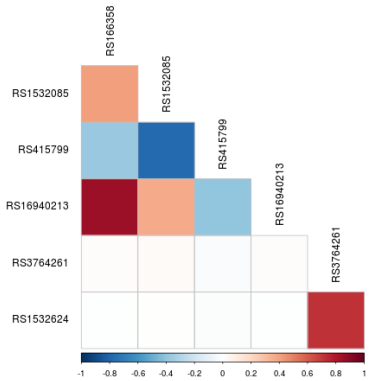
b) Total lipids in small VLDL



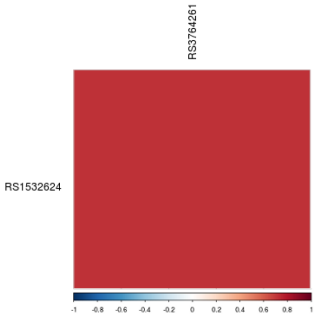
c) Total lipids in very small VLDL



d) Total lipids in very large HDL



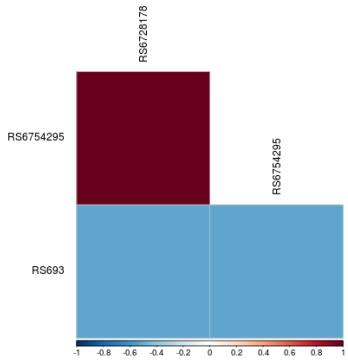
e) Total lipids in large HDL



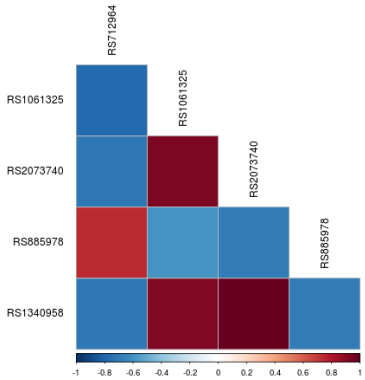
f) Apolipoprotein A1

The colour intensity reflects the amount of LD. Results are only shown for the leading traits.

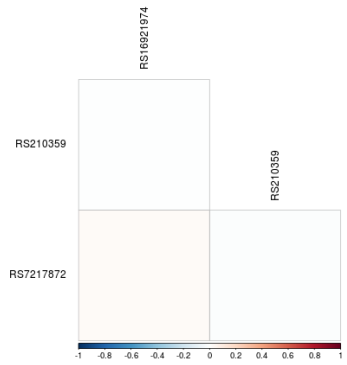
Figure B.2: Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the single-SNP analysis.



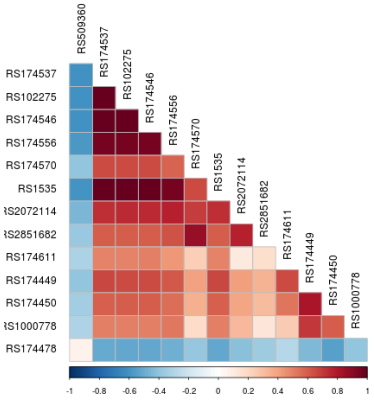
g) Apolipoprotein B



h) Citrate



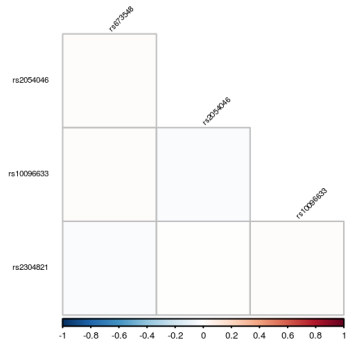
i) Creatinine



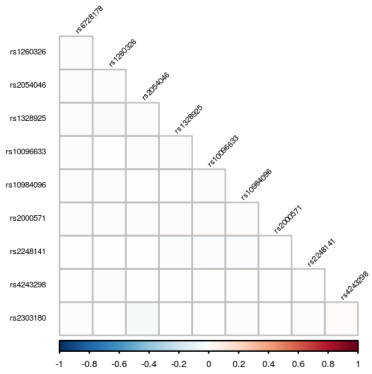
j) Total PUFA

The colour intensity reflects the amount of LD. Results are only shown for the leading traits.

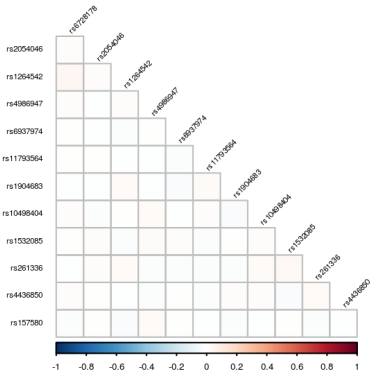
Figure B.3: Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the HyperLasso regression analysis.



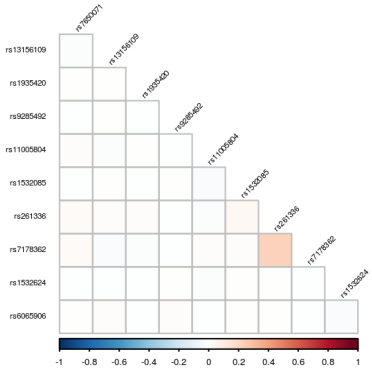
a) Total lipids in medium VLDL



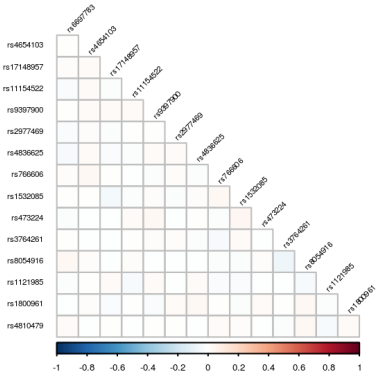
b) Total lipids in small VLDL



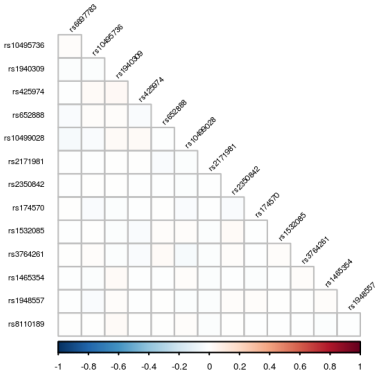
c) Total lipids in very small VLDL



d) Total lipids in very large HDL

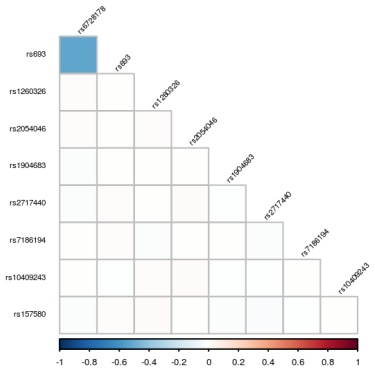


e) Total lipids in large HDL

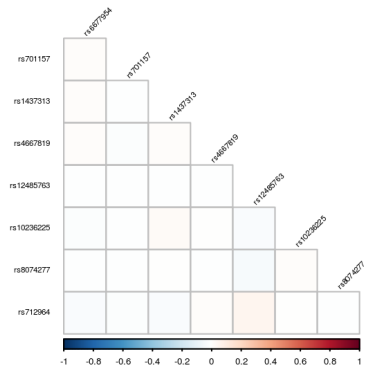


f) Apolipoprotein A1

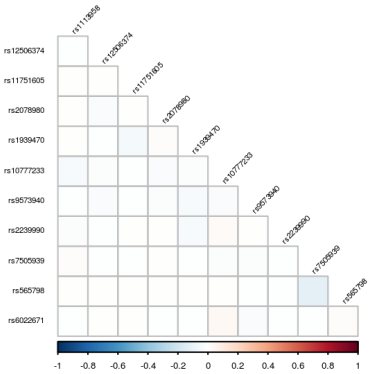
Figure B.3: Heatmaps displaying linkage disequilibrium (r) between SNPs identified by the HyperLasso regression analysis.



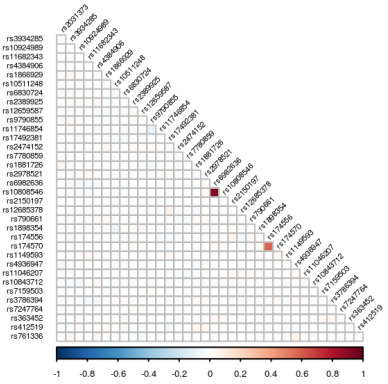
g) Apolipoprotein B



h) Citrate



i) Creatinine



j) Total PUFA

The colour intensity reflects the amount of LD. Results are only shown for the leading traits.

Table B.2: Results of the SUR analysis of VLDL system

	rs3923037	**rs6728178**	**rs6754295**	**rs676210**	**rs673548**	**rs1260326**	**rs10096633**	**rs1532085**
XXL-VLDL-L	0.8	0.2	0.1	0.9	1	**7.50E-6**	**6.50E-4**	0.8
XL-VLDL-L	0.4	0.2	0.2	0.8	0.7	**2.50E-8**	**6.50E-5**	0.2
L-VLDL-L	0.5	0.2	0.1	0.8	0.6	**3.50E-10**	**1.50E-5**	0.1
M-VLDL-L	0.5	0.09	0.1	0.7	0.5	**2.50E-10**	**6.50E-7**	0.2
S-VLDL-L	0.5	0.06	0.2	0.2	0.2	**1.50E-6**	**9.50E-7**	0.9
XS-VLDL-L	0.9	0.2	0.7	0.6	0.7	0.03	0.01	**7.50E-8**

A P-value given in bold indicates a statistically significant association between a SNP and a metabolite measure with a threshold $P < 0.05/k$ where $k = 8$ is the total n umber of tested SNPs

Table B.3: Results of the SUR analysis of HDL system

	rs166358	**rs1532085**	**rs415799**	**rs16940213**	**rs261336**	**rs3764261**	**rs1532624**	**rs6065906**
XL-HDL-L	0.3	6.50E-3	0.7	0.1	**1.50E-6**	0.5	0.01	**3.50E-7**
L-HDL-L	0.4	**1.50E-4**	0.4	0.09	7.50E-3	**6.50E-4**	0.1	**2.50E-3**
M-HDL-L	0.3	0.2	0.3	0.05	0.8	**4.50E-5**	0.7	**4.00E-3**
S-HDL-L	0.1	0.7	0.2	0.02	0.02	8.50E-3	0.5	**9.50E-8**

A P-value given in bold indicates a statistically significant association between a SNP and a metabolite measure with a threshold $P < 0.05/k$ where $k = 8$ is the total n umber of tested SNPs

Table B.4: Results of the SUR analysis of metabolites involved in the citric acid cycle

	rs712964	**rs1061325**	**rs2073740**	**rs885978**	**rs1340958**
Cit	0.07	0.4	0.03	0.2	0.03
bOHBut	1	0.6	0.2	0.8	0.1
Ace	0.7	0.5	0.5	0.7	0.5
AcAce	0.8	0.4	0.2	1	0.1

A P-value given in bold indicates a statistically significant association between a SNP and a metabolite measure with a threshold $P < 0.05/k$ where $k = 5$ is the total n umber of tested SNPs

Table B.5: Results of the SUR analysis of metabolites involved in the urea cycle

	rs16921974	**rs7217872**
Crea	0.2	0.06
Urea	1	0.04

A P-value given in bold indicates a statistically significant association between a SNP and a metabolite measure with a threshold $P < 0.05/k$ where $k = 2$ is the total n umber of tested SNPs

Appendix C

Differential Networks

Table C.1: Clinical characteristic of the NFBC1966 study participants.

Clinical Traits		**Male**	
	[1]**NFG**	[1]**IFG**	[2]**P-value**
Fasting glucose, mmol/L	5.01 (0.31)	5.84 (0.27)	< 2.20E-16
Fasting insulin, IU/L	8.20 (3.41)	10.89 (6.71)	< 2.20E-16
Body mass index, kg/m2	24.96 (3.28)	26.09 (3.92)	3.73E-7
Waist circumference, cm	88.21 (9.06)	91.33 (10.75)	4.60E-7
Hip circumference, cm	97.08 (6.17)	99.02 (6.99)	7.58E-7
Waist-to-hip ratio	0.91 (0.06)	0.92 (0.06)	1.06E-4
Systolic blood pressure, mm Hg	129.40 (12.16)	133.16 (13.55)	1.65E-6
Diastolic blood pressure, mm Hg	79.47 (11.09)	83.01 (12.18)	6.27E-8
Total cholesterol, mmol/l	5.34 (1.05)	5.52 (1.13)	8.03E-3
Triglycerides, mmol/l	1.01 (0.47)	1.09 (0.52)	9.77E-3
Low density lipoprotein cholesterol, mmol/l	2.21 (0.63)	2.30 (0.64)	8.03E-3
High density lipoprotein cholesterol, mmol/l	1.50 (0.33)	1.52 (0.33)	0.43
Clinical Traits		**Female**	
	[1]**NFG**	[1]**IFG**	[2]**P-value**
Fasting glucose, mmol/L	4.83 (0.36)	5.83 (0.28)	< 2.20E-16
Fasting insulin, IU/L	8.00 (3.35)	11.38 (6.73)	9.57E-14
Body mass index, kg/m2	24.07 (4.44)	27.09 (6.53)	3.29E-9
Waist circumference, cm	78.57 (11.48)	86.31 (15.53)	4.91E-11
Hip circumference, cm	97.05 (8.50)	102.08 (11.61)	7.39E-9
Waist-to-hip ratio	0.81 (0.08)	0.84 (0.10)	6.13E-8
Systolic blood pressure, mm Hg	119.27 (11.88)	125.27 (13.59)	1.22E-7
Diastolic blood pressure, mm Hg	74.44 (10.54)	78.76 (11.87)	2.13E-6
Total cholesterol, mmol/l	5.10 (1.03)	5.15 (1.10)	0.64
Triglycerides, mmol/l	0.87 (0.39)	0.97 (0.45)	0.012
Low density lipoprotein cholesterol, mmol/l	1.91 (0.57)	1.99 (0.64)	0.19
High density lipoprotein cholesterol, mmol/l	1.74 (0.36)	1.66 (0.36)	0.026

[1] Data are means ± SD; [2] P-value obtained from Mann-Whitney test.

Table C.2: Lipoprotein subclass measures and their mean concentrations for normal (NFG) and impaired fasting glucose (IFG).

Lipoproteins subclass fractions	Abbreviation	[1]NFG	[2]IFG	[3]P
Phospholipids in extremely large VLDL	XXL-VLDL-PL	2.2E-3	2.7E-3	**5.5E-4**
Concentration in extremely large VLDL particles	XXL-VLDL-P	7.3E-11	8.7E-11	0.011
Phospholipids in very large VLDL	XL-VLDL-PL	0.011	0.013	**2.1E-3**
Triglycerides in very large VLDL	XL-VLDL-TG	0.042	0.046	0.021
Concentration of very large VLDL particles	XL-VLDL-P	6.6E-10	7.3E-10	**9.2E-3**
Cholesterol in large VLDL	L-VLDL-C	0.063	0.067	0.06
Free cholesterol in large	L-VLDL-FC	0.03	0.033	0.048
Phospholipids in large VLDL	L-VLDL-PL	0.047	0.05	0.072
Triglycerides in large VLDL	L-VLDL-TG	0.16	0.16	0.19
Cholesterol esters in large VLDL	L-VLDL-CE	0.032	0.034	0.089
Concentration of large VLDL particles	L-VLDL-P	4.3E-9	4.6E-9	0.12
Cholesterol in medium VLDL	M-VLDL-C	0.18	0.18	0.098
Free cholesterol in medium VLDL	M-VLDL-FC	0.074	0.077	0.2
Phospholipids in medium VLDL	M-VLDL-PL	0.12	0.12	0.23
Triglycerides in medium VLDL	M-VLDL-TG	0.28	0.29	0.52
Cholesterol esters in medium VLDL	M-VLDL-CE	0.1	0.11	0.074
Concentration of medium VLDL particles	M-VLDL-P	1.6E-8	1.6E-8	0.36
Cholesterol in small VLDL	S-VLDL-C	0.3	0.31	0.023
Free cholesterol in small VLDL	S-VLDL-FC	0.11	0.11	0.055
Phospholipids in small VLDL	S-VLDL-PL	0.16	0.17	0.059
Triglycerides in small VLDL	S-VLDL-TG	0.26	0.27	0.15
Concentration of small VLDL particles	S-VLDL-P	3.2E-8	3.3E-8	0.072
Phospholipids in very small VLDL	XS-VLDL-PL	0.15	0.16	0.072
Triglycerides in very small VLDL	XS-VLDL-TG	0.12	0.12	0.044
Concentration of very small VLDL particles	XS-VLDL-P	3.6E-8	3.7E-8	0.034
Free cholesterol in IDL	IDL-FC	0.22	0.23	0.056
Phospholipids in IDL	IDL-PL	0.32	0.33	0.034
Concentration of IDL particles	IDL-P	9.7E-8	1.0E-7	0.017
Cholesterol in large LDL	L-LDL-C	1.12	1.17	0.012
Free cholesterol in large LDL	L-LDL-FC	0.29	0.3	0.017
Phospholipids in large LDL	L-LDL-PL	0.37	0.39	**9.4E-3**
Cholesterol esters in large LDL	L-LDL-CE	0.83	0.86	0.012
Concentration of large LDL particles	L-LDL-P	1.7E-7	1.8E-7	**7.3E-3**
Cholesterol in medium LDL	M-LDL-C	0.68	0.71	**7.4E-3**
Phospholipids in medium LDL	M-LDL-PL	0.24	0.25	**3.3E-3**
Cholesterol esters in medium LDL	M-LDL-CE	0.5	0.55	0.011
Concentration of medium LDL particles	M-LDL-P	1.4E-7	1.5E-7	**3.3E-3**
Cholesterol in small LDL	S-LDL-C	0.41	0.43	**3.4E-3**
Concentration of small LDL particles	S-LDL-P	1.6E-7	1.7E-7	**1.4E-3**
Cholesterol in very large HDL	XL-HDL-C	0.2	0.21	0.22

Continued on next page

Table C.2 – continued from previous page

Lipoproteins subclass fractions	**Abbreviation**	[1]**NFG**	[2]**IFG**	[3]**P**
Free cholesterol in very large HDL	XL-HDL-FC	0.051	0.052	0.04
Phospholipids in very large HDL	XL-HDL-PL	0.17	0.17	0.82
Triglycerides in very large HDL	XL-HDL-TG	0.011	0.012	0.46
Cholesterol esters in very large HDL	XL-HDL-CE	0.15	0.16	0.21
Concentration of very large HDL particles	XL-HDL-P	3.0E-7	3.1E-7	0.5
Cholesterol in large HDL	L-HDL-C	0.34	0.33	0.66
Free cholesterol in large HDL	L-HDL-FC	0.066	0.066	0.77
Phospholipids in large HDL	L-HDL-PL	0.3	0.31	0.76
Cholesterol esters in large HDL	L-HDL-CE	0.27	0.27	0.63
Concentration of large HDL particles	L-HDL-P	8.2E-7	8.2E-7	0.67
Cholesterol in medium HDL	M-HDL-C	0.44	0.45	0.66
Free cholesterol in medium HDL	M-HDL-FC	0.08	0.081	0.52
Phospholipids in medium HDL	M-HDL-PL	0.37	0.38	0.48
Cholesterol esters in medium HDL	M-HDL-CE	0.37	0.37	0.69
Concentration of medium HDL particles	M-HDL-P	1.5E-6	1.6E-6	0.55
Triglycerides in small HDL	S-HDL-TG	0.04	0.041	0.19
Concentration of small HDL particle	S-HDL-P	4.4E-6	4.5E-6	0.093
Apolipoprotein A-I	ApoA.I.	1.64	1.67	0.1
Apolipoprotein B	ApoB.	1.04	1.08	**9.8E-3**
Apolipoprotein B / Apolipoprotein A-I	ApoB.ApoA.I	0.64	0.65	0.12

[1] Mean concentration for each [1]H lipoprotein measure in males with normal fasting glucose.
[2] Mean concentration for each [1]H lipoprotein measure in males with impaired fasting glucose.
[3] P-values obtained from Mann-Whitney test. Bold indicates significant differences in mean concentration between the two groups for the non-corrected threshold ($P < 0.01$).

Correlation maps using Pearson's correlation

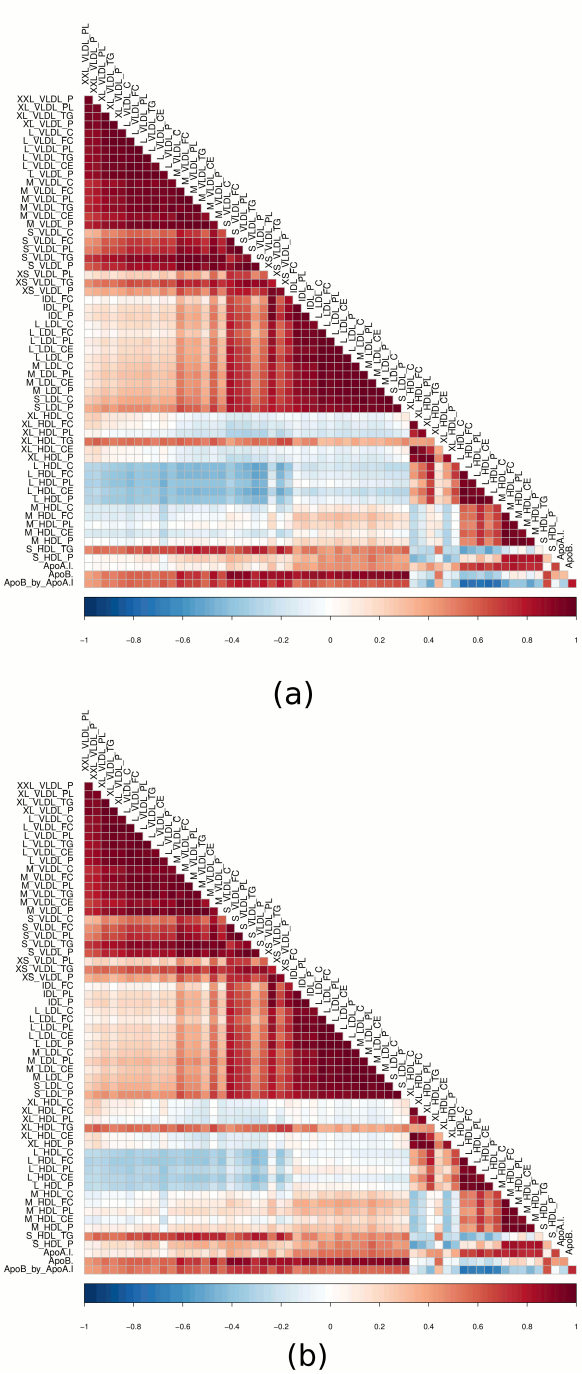
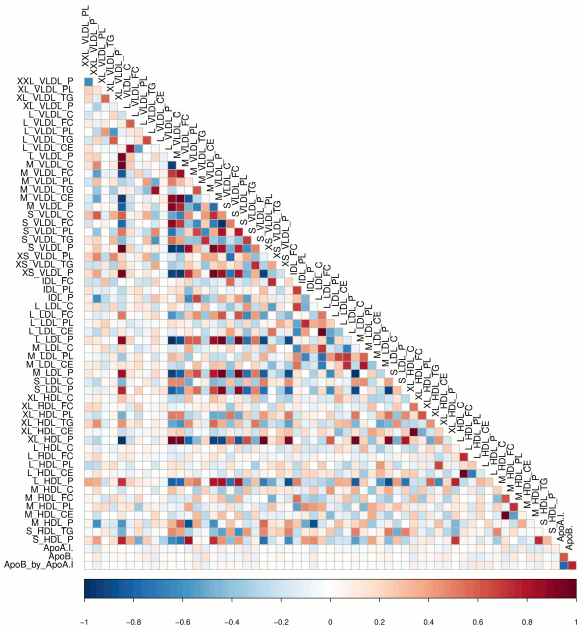
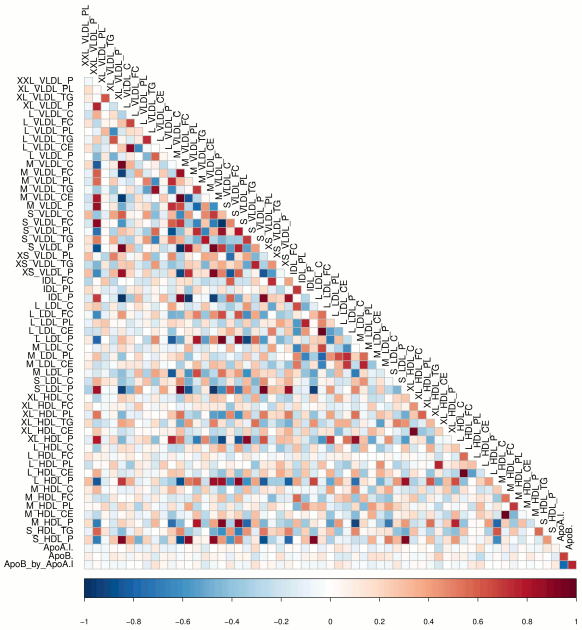


Figure C.1: The colour key gives the R value for the correlation between the lipoproteins subclass components calculated for males with (a) normal fasting glucose (NFG) and males with (b) impaired fasting glucose (IFG).

Correlation maps using standard partial correlation



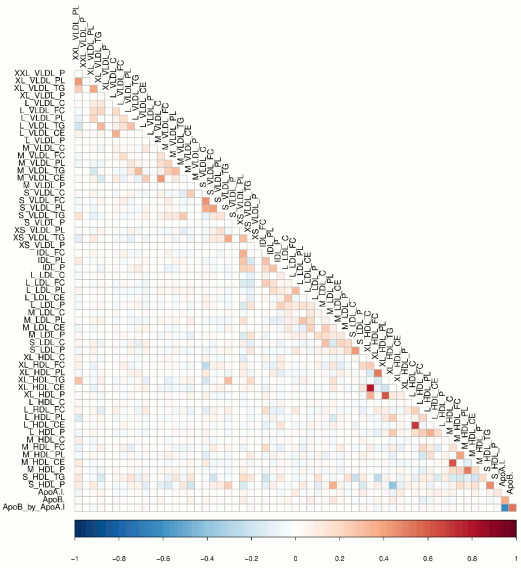
(a)



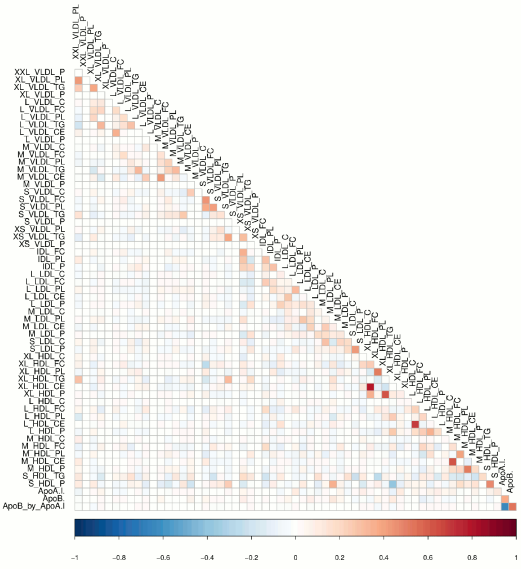
(b)

Figure C.2: The colour key gives the R value for the correlation between the lipoproteins subclass components calculated for males with (a) normal fasting glucose (NFG) and males with (b) impaired fasting glucose (IFG).

Correlation maps using shrinkage partial correlation



(a)



(b)

Figure C.3: The colour key gives the R value for the correlation between the lipoproteins subclass components calculated for males with (a) normal fasting glucose (NFG) and males with (b) impaired fasting glucose (IFG).

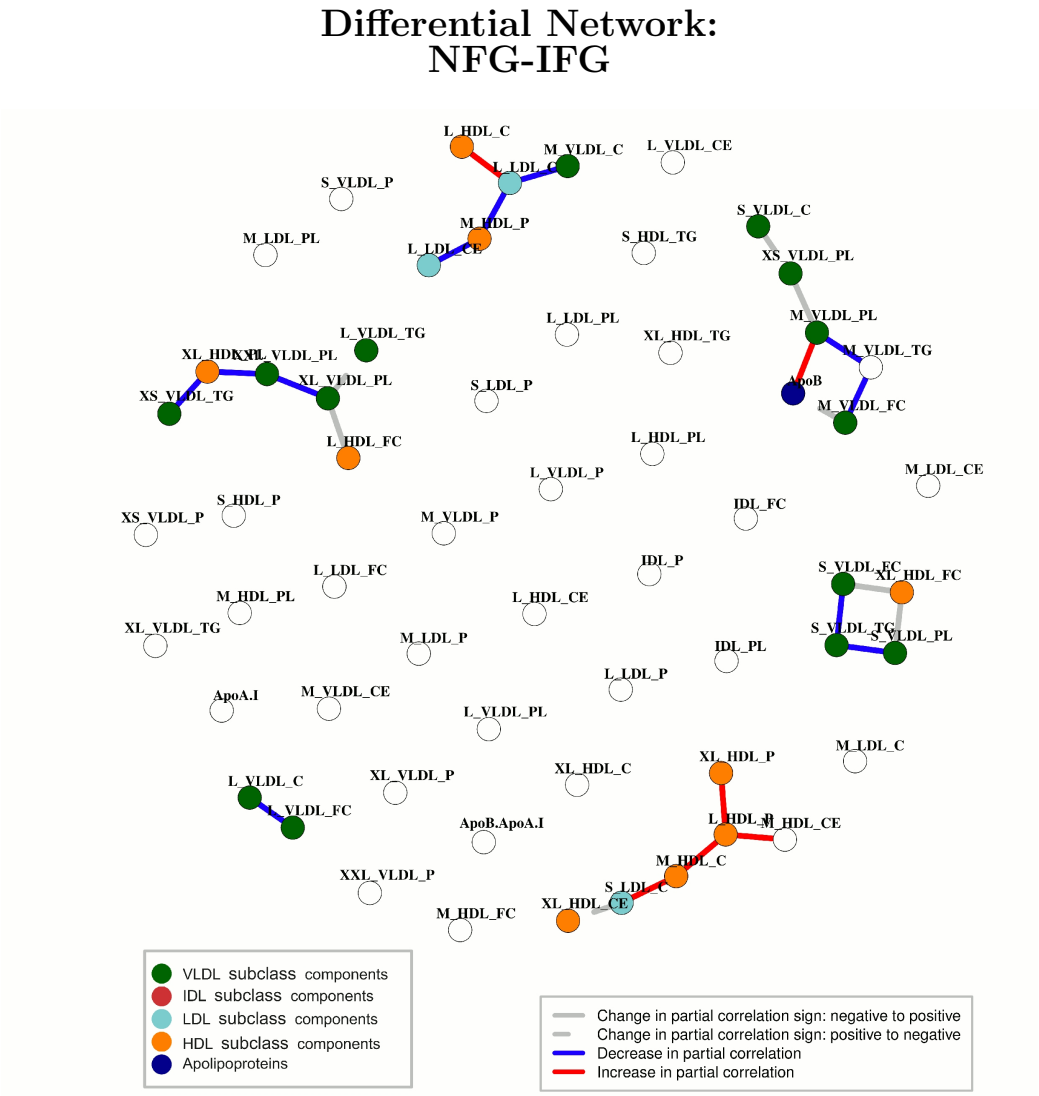


Figure C.4: The differential network inferred from females with normal fasting glucose (NFG) and impaired fasting glucose (IFG). Each of the connections indicates a significant change in partial correlation between two lipoprotein measures across the physiological conditions. Edge colours represent how the partial correlation between two measures has changed across the two groups.

Appendix D

GEMINi

Table D.1: Linear regression analysis in the NFBC1966 between body mass index (BMI) and the serum metabolite measures

Metabolic measure	Abbreviation	Estimate (β)	P
Total lipids in chylomicrons	XXL-VLDL-L	38.1	**3.99E-50**
Total lipids in very largeVLDL	XL-VLDL-L	15.4	**6.39E-50**
Total lipids in large VLDL	L-VLDL-L	4.7	**4.01E-55**
Total lipids in medium VLDL	M-VLDL-L	3.6	**1.02E-62**
Total lipids in small VLDL	S-VLDL-L	5.2	**3.08E-73**
Total lipids in very small VLDL	XS-VLDL-L	6.8	**3.38E-50**
Total lipids in IDL	IDL-L	2.7	**1.28E-37**
Total lipids in large LDL	L-LDL-L	2.2	**3.74E-41**
Total lipids in medium LDL	M-LDL-L	3.8	**2.24E-49**
Total lipids in small LDL	S-LDL-L	6.7	**4.80E-64**
Total lipids in very large HDL	XL-HDL-L	-2.6	**5.37E-14**
Total lipids in large HDL	L-HDL-L	-2.9	**5.80E-53**
Total lipids in medium HDL	M-HDL-L	-1.3	**3.03E-7**
Total lipids in small HDL	S-HDL-L	1.6	**3.26E-5**
Apolipoprotein A-I	ApoA1	-1	6.83E-4
Apolipoprotein B	ApoB	5.4	**1.04E-63**
Mean diameter for VLDL particles	VLDL-D	0.8	**4.13E-47**
Mean diameter for LDL particles	LDL-D	-3.2	**7.92E-14**
Mean diameter for HDL particles	HDL-D	-5.2	**1.04E-58**
3-hydroxybutyrate	bOHBut	-1	0.1
Acetate	Ace	-7.1	0.03
Acetoacetate	AcAce	-0.8	0.62
Alanine	Ala	8.5	**9.29E-21**
Citrate	Cit	-11.1	**2.84E-3**
Creatinine	Crea	-6.6	0.16
Glucose	Glc	0.5	**7.12E-10**
Glutamine	Gln	-2.9	**8.67E-5**
Glycerol	Glol	30.8	**4.70E-52**
Glycoprotein acetyls	Gp	4.5	**2.19E-60**
Histidine	His	18.9	5.99E-4

Continued on next page

Table D.1 – continued from previous page

Metabolic measure	Abbreviation	Estimate (β)	P
Isoleucine	Ile	89	**2.61E-82**
Lactate	Lac	1	**4.30E-12**
Leucine	Leu	64.9	**2.76E-67**
Phenylalanine	Phe	96.5	**5.27E-93**
Pyruvate	Pyr	17.4	**5.57E-11**
Tyrosine	Tyr	90.4	**6.87E-54**
Urea	Urea	1.8	0.5
Valine	Val	24.7	**6.38E-54**

A P-value given in bold indicates a statistically significant association between BMI and a metabolite measure with a Bonferroni corrected threshold, $p < 0.01/M$, where $M = 38$ is the total number of metabolites included in the analyses. Estimate (β), estimate BMI effect size.

Table D.2: Linear regression analysis in the NFBC1986 between body mass index (BMI) and the serum metabolite measures

Metabolic measure	Abbreviation	Estimate (β)	P
Total lipids in chylomicrons	XXL-VLDL-L	51.22	**1.58E-79**
Total lipids in very large VLDL	XL-VLDL-L	20.21	**4.58E-88**
Total lipids in large VLDL	L-VLDL-L	6.02	**2.58E-93**
Total lipids in medium VLDL	M-VLDL-L	4.44	**1.72E-100**
Total lipids in small VLDL	S-VLDL-L	6.09	**1.89E-102**
Total lipids in very small VLDL	XS-VLDL-L	7.01	**2.01E-53**
Total lipids in IDL	IDL-L	2.78	**9.22E-38**
Total lipids in large LDL	L-LDL-L	2.36	**4.28E-44**
Total lipids in medium LDL	M-LDL-L	4.16	**1.36E-52**
Total lipids in small LDL	S-LDL-L	6.96	**1.99E-63**
Total lipids in very large HDL	XL-HDL-L	-4.46	**3.73E-57**
Total lipids in large HDL	L-HDL-L	-3.13	**5.21E-69**
Total lipids in medium HDL	M-HDL-L	0.55	0.046
Total lipids in small HDL	S-HDL-L	4.42	**5.13E-30**
Apolipoprotein A-I	ApoA1	-0.91	3.16E-4
Apolipoprotein B	ApoB	5.88	**2.15E-87**
Mean diameter for VLDL particles	VLDL-D	0.56	**1.22E-45**
Mean diameter for LDL particles	LDL-D	-1.66	**5.74E-9**
Mean diameter for HDL particles	HDL-D	-4.84	**1.41E-94**
3-hydroxybutyrate	bOHBut	-7.30E-6	**1.05E-6**
Acetate	Ace	-6.86E-5	**1.29E-5**
Acetoacetate	AcAce	-1.17E-5	1.31E-3
Alanine	Ala	1.79E-5	**2.68E-12**
Citrate	Cit	-6.18E-5	**1.15E-23**
Creatinine	Crea	7.11E-5	**2.62E-5**
Glucose	Glc	1.69E-6	0.024
Glutamine	Gln	-1.26E-5	**8.36E-8**
Glycerol	Glol	4.09E-5	**8.24E-13**
Glycoprotein acetyls	Gp	1.46E-5	**9.63E-82**

Continued on next page

Table D.2 – continued from previous page

Metabolic measure	Abbreviation	Estimate (β)	P
Histidine	His	1.16E-4	**1.02E-11**
Isoleucine	Ile	2.16E-4	**9.12E-65**
Lactate	Lac	-2.17E-6	8.56E-4
Leucine	Leu	1.67E-4	**1.74E-60**
Phenylalanine	Phe	2.88E-4	**2.43E-85**
Pyruvate	Pyr	3.44E-5	**5.38E-5**
Tyrosine	Tyr	2.24E-4	**5.24E-51**
Urea	Urea	-9.39E-6	0.18
Valine	Val	5.17E-5	**5.84E-32**

A P-value given in bold indicates a statistically significant association between BMI and a metabolite measure with a Bonferroni corrected threshold, $p < 0.01/M$, where $M = 38$ is the total number of metabolites included in the analyses. Estimate (β), estimate BMI effect size.

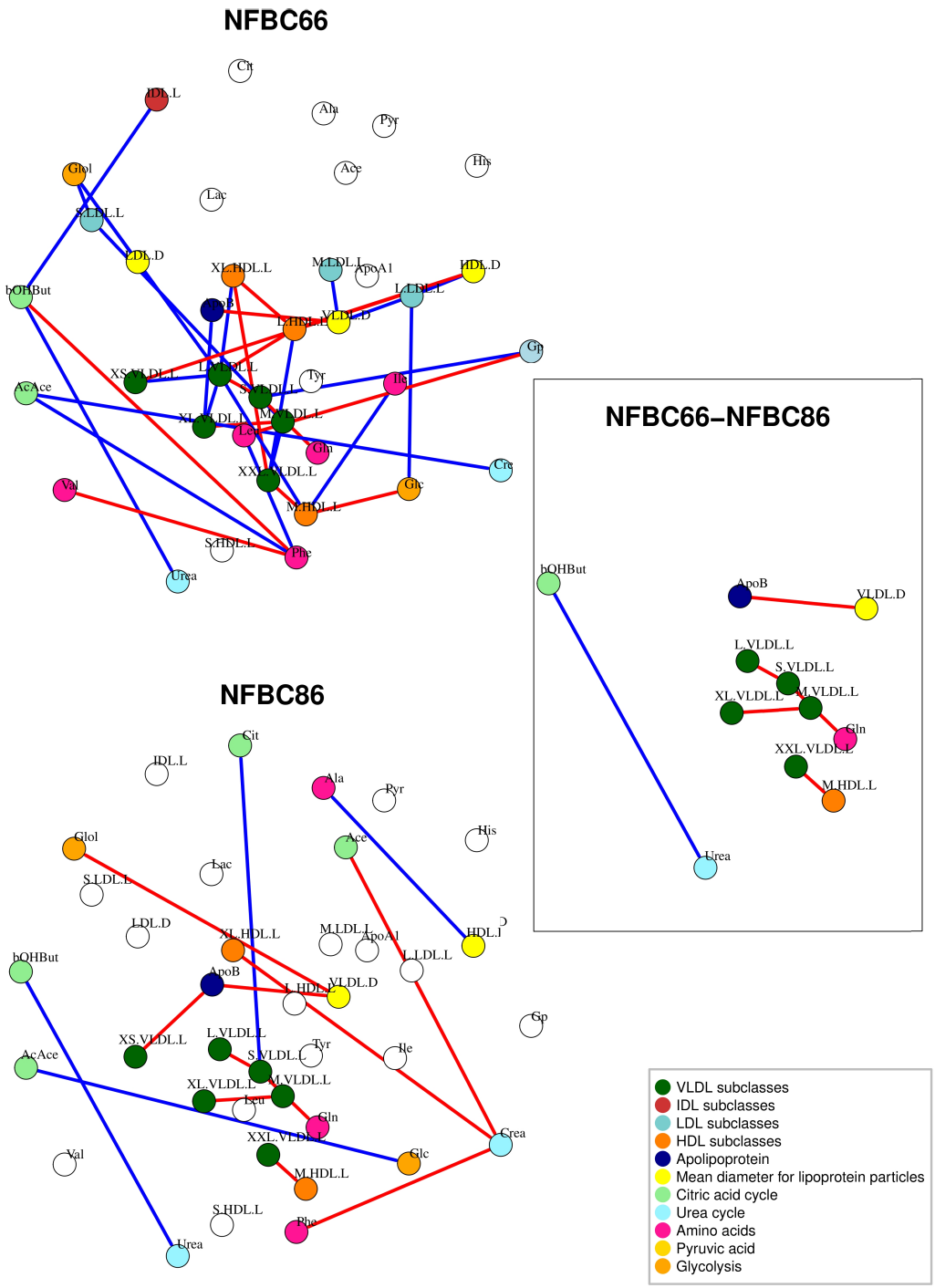


Figure D.1: Differential network analysis for serum metabolites inferred from normal and obese individuals. (a) Differential Network inferred from NFBC1966; (b) differential network inferred from NFBC1986; (c) Replicated differential connection between NFBC1966 and NFBC1986 differential networks. Individuals are classified using body mass index (BMI) definitions.

Appendix E

Variance and Covariance Regression

Table E.1: Connectivity of the correlation networks.

Metabolites	$k_{0.2}$	$k_{0.3}$	$k_{0.4}$	$k_{0.5}$	$k_{0.6}$	$k_{0.7}$	$k_{0.8}$
4-deoxyerythronic.acid	3	5	5	5	5	5	4
3-hydroxyisovalerate	6	9	8	7	6	6	6
lactate	3	7	7	6	5	4	4
alanine	4	6	6	8	10	10	9
acetate	3	5	5	4	2	0	0
pyruvate	4	4	4	3	3	3	3
succinate	5	5	3	3	3	2	2
citrate	5	4	4	3	4	4	3
dimethylamine	6	8	8	8	8	6	5
trimethylamine	1	4	4	3	3	3	3
dimethylgycine	7	7	6	3	3	3	2
creatine	1	0	0	0	1	1	1
creatinine	5	4	4	4	4	4	5
Proline.betaine	3	2	3	6	6	6	6
P.cresol.sulphate	5	4	3	4	4	3	3
phenylacetylglutamine	4	5	5	5	5	4	4
hippurate	6	7	8	3	1	2	2
N.methyl.nicotinic.acid	4	4	2	3	3	3	2
TMAO	3	6	6	5	5	5	4
glycine	6	8	7	5	5	4	4
urea	5	7	7	5	5	1	1
formate	5	3	1	3	3	3	3
Total Connectivity	94	114	106	96	94	82	76

k; connectivity for each of correlation networks under seven different concentrations of urinary cadmium (U-Cd (nmol/mmol creatinine)= 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)

Table E.2: Results from the variance regression analysis

Met	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
4DEA	0.07(0.04-0.09)	0.06(0.05-0.08)	0.06(0.05-0.09)	0.07(0.04-0.10)	0.07(0.03-0.12)	0.07(0.03-0.15)	0.07(0.02-0.19)
3HV	0.05(0.04-0.07)	0.05(0.04-0.06)	0.04(0.03-0.06)	0.04(0.03-0.07)	0.04(0.02-0.07)	0.04(0.02-0.08)	0.04(0.02-0.09)
Lac	0.13(0.09-0.19)	0.14(0.11-0.18)	0.15(0.11-0.20)	0.16(0.10-0.24)	0.17(0.09-0.31)	0.18(0.07-0.40)	0.20(0.06-0.53)
Ala	0.05(0.03-0.07)	0.05(0.04-0.06)	0.05(0.04-0.07)	0.05(0.03-0.08)	0.06(0.03-0.11)	0.06(0.02-0.14)	0.07(0.02-0.18)
AcO	0.38(0.25-0.54)	0.35(0.27-0.45)	0.34(0.25-0.46)	0.32(0.20-0.53)	0.32(0.15-0.62)	0.32(0.11-0.75)	0.32(0.09-0.90)
Pyr	0.15(0.10-0.21)	0.15(0.11-0.19)	0.15(0.11-0.20)	0.15(0.10-0.24)	0.16(0.08-0.29)	0.17(0.07-0.37)	0.18(0.05-0.46)
Suc	0.06(0.04-0.09)	0.06(0.05-0.08)	0.06(0.04-0.07)	0.05(0.03-0.08)	0.05(0.03-0.09)	0.05(0.02-0.10)	0.05(0.02-0.12)
Cit	0.12(0.09-0.17)	0.13(0.10-0.17)	0.14(0.10-0.18)	0.15(0.10-0.23)	0.16(0.09-0.29)	0.18(0.08-0.37)	0.20(0.07-0.47)
DMA	0.05(0.04-0.08)	0.05(0.04-0.07)	0.05(0.04-0.06)	0.04(0.03-0.07)	0.04(0.02-0.08)	0.04(0.02-0.09)	0.04(0.01-0.10)
TMA	0.10(0.07-0.14)	0.10(0.07-0.13)	0.10(0.07-0.13)	0.10(0.07-0.16)	0.10(0.06-0.19)	0.11(0.05-0.23)	0.12(0.04-0.29)
DMG	0.07(0.05-0.10)	0.05(0.04-0.06)	0.04(0.03-0.05)	0.03(0.02-0.04)	0.02(0.01-0.04)	0.02(0.01-0.04)	0.01(0.00-0.03)
Crea	0.08(0.05-0.11)	0.08(0.07-0.11)	0.10(0.07-0.13)	0.11(0.07-0.17)	0.13(0.07-0.24)	0.16(0.06-0.34)	0.19(0.06-0.48)
Creat	0.05(0.03-0.07)	0.05(0.04-0.06)	0.05(0.03-0.06)	0.05(0.03-0.07)	0.05(0.02-0.08)	0.05(0.02-0.10)	0.05(0.02-0.12)
PB	0.32(0.22-0.44)	0.29(0.23-0.37)	0.27(0.20-0.36)	0.26(0.16-0.40)	0.24(0.13-0.45)	0.24(0.10-0.52)	0.23(0.07-0.59)
PCS	0.14(0.10-0.20)	0.14(0.11-0.18)	0.14(0.11-0.19)	0.15(0.10-0.23)	0.16(0.08-0.29)	0.17(0.07-0.35)	0.18(0.06-0.44)
PAG	0.16(0.11-0.23)	0.17(0.13-0.22)	0.18(0.13-0.24)	0.19(0.12-0.30)	0.20(0.11-0.38)	0.22(0.09-0.49)	0.25(0.08-0.63)
Hip	0.24(0.16-0.34)	0.23(0.18-0.30)	0.23(0.17-0.31)	0.23(0.14-0.35)	0.23(0.12-0.42)	0.23(0.10-0.51)	0.24(0.08-0.61)
NMNA	0.55(0.38-0.79)	0.59(0.45-0.76)	0.63(0.47-0.85)	0.68(0.43-1.08)	0.75(0.38-1.42)	0.84(0.33-1.89)	0.95(0.29-2.51)
TMAO	0.36(0.25-0.50)	0.40(0.31-0.51)	0.44(0.33-0.59)	0.50(0.33-0.75)	0.56(0.32-0.98)	0.64(0.30-1.31)	0.74(0.29-1.74)
Gly	0.13(0.09-0.18)	0.13(0.10-0.17)	0.13(0.10-0.18)	0.13(0.09-0.21)	0.14(0.08-0.25)	0.15(0.06-0.31)	0.16(0.06-0.38)
Urea	0.05(0.04-0.07)	0.05(0.04-0.06)	0.04(0.03-0.06)	0.04(0.03-0.07)	0.04(0.02-0.08)	0.04(0.02-0.09)	0.04(0.01-0.10)
For	0.17(0.11-0.24)	0.18(0.14-0.24)	0.20(0.15-0.27)	0.22(0.14-0.35)	0.25(0.13-0.47)	0.29(0.12-0.64)	0.34(0.10-0.87)

Variance for each of the 22 urinary metabolites estimated from the variance regression along with the 95% credible interval for seven different concentrations of urinary cadmium ($^{1-7}$U-Cd (nmol/mmol creatinine)= 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)

Table E.3: Results from the covariance regression analysis

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
4DEA	3HV	0.29(0.10,0.45)	0.36(0.21,0.52)	0.42(0.25,0.59)	0.46(0.26,0.63)	0.48(0.26,0.68)	0.50(0.25,0.71)	0.51(0.24,0.74)
4DEA	Lac	-0.06(-0.27,0.15)	0.04(-0.14,0.22)	0.12(-0.06,0.32)	0.20(-0.01,0.41)	0.25(0.01,0.49)	0.30(0.02,0.56)	0.34(0.01,0.62)
4DEA	Ala	0.04(-0.16,0.24)	0.11(-0.07,0.28)	0.16(-0.02,0.34)	0.21(0.01,0.41)	0.25(0.02,0.49)	0.29(0.01,0.55)	0.32(0.00,0.60)
4DEA	AcO	-0.03(-0.22,0.17)	-0.06(-0.22,0.12)	-0.08(-0.27,0.10)	-0.11(-0.31,0.11)	-0.13(-0.38,0.13)	-0.15(-0.44,0.16)	-0.17(-0.50,0.19)
4DEA	Pyr	-0.15(-0.32,0.05)	-0.17(-0.33,0.00)	-0.18(-0.36,0.00)	-0.19(-0.40,0.04)	-0.20(-0.45,0.07)	-0.21(-0.49,0.11)	-0.21(-0.52,0.14)
4DEA	Suc	0.22(0.02,0.40)	0.19(0.00,0.37)	0.15(-0.04,0.34)	0.12(-0.11,0.34)	0.08(-0.20,0.34)	0.05(-0.28,0.35)	0.01(-0.35,0.37)
4DEA	Cit	0.08(-0.12,0.27)	0.06(-0.11,0.25)	0.05(-0.15,0.25)	0.03(-0.19,0.25)	0.01(-0.25,0.28)	0.00(-0.30,0.30)	-0.02(-0.36,0.32)
4DEA	DMA	0.13(-0.08,0.33)	0.18(0.00,0.34)	0.22(0.03,0.41)	0.26(0.02,0.48)	0.29(0.02,0.55)	0.31(0.01,0.61)	0.33(-0.01,0.65)
4DEA	TMA	0.08(-0.11,0.28)	0.10(-0.08,0.26)	0.11(-0.07,0.30)	0.12(-0.09,0.35)	0.13(-0.12,0.40)	0.14(-0.15,0.44)	0.15(-0.17,0.48)
4DEA	DMG	0.05(-0.15,0.27)	0.09(-0.09,0.28)	0.14(-0.06,0.34)	0.18(-0.05,0.40)	0.23(-0.05,0.47)	0.27(-0.05,0.54)	0.31(-0.07,0.61)
4DEA	Crea	0.18(-0.03,0.38)	0.11(-0.08,0.28)	0.04(-0.16,0.24)	-0.02(-0.25,0.22)	-0.07(-0.33,0.23)	-0.11(-0.41,0.26)	-0.14(-0.48,0.30)
4DEA	Creat	0.10(-0.12,0.30)	0.17(-0.03,0.35)	0.25(0.06,0.45)	0.33(0.10,0.56)	0.39(0.13,0.64)	0.44(0.15,0.69)	0.47(0.14,0.72)
4DEA	PB	0.35(0.16,0.52)	0.31(0.15,0.45)	0.27(0.08,0.43)	0.25(0.00,0.44)	0.22(-0.06,0.46)	0.21(-0.11,0.49)	0.20(-0.17,0.52)
4DEA	PCS	0.03(-0.16,0.24)	-0.02(-0.20,0.15)	-0.07(-0.25,0.11)	-0.12(-0.32,0.10)	-0.16(-0.41,0.10)	-0.20(-0.47,0.11)	-0.23(-0.53,0.12)
4DEA	PAG	-0.11(-0.31,0.09)	-0.20(-0.36,-0.04)	-0.27(-0.43,-0.09)	-0.33(-0.51,-0.12)	-0.38(-0.59,-0.14)	-0.42(-0.65,-0.15)	-0.46(-0.70,-0.15)
4DEA	Hip	-0.07(-0.28,0.13)	-0.12(-0.29,0.06)	-0.16(-0.34,0.03)	-0.19(-0.40,0.03)	-0.22(-0.46,0.03)	-0.25(-0.52,0.05)	-0.27(-0.57,0.09)
4DEA	NMNA	0.09(-0.10,0.28)	0.05(-0.13,0.23)	0.01(-0.18,0.18)	-0.03(-0.25,0.18)	-0.06(-0.32,0.20)	-0.09(-0.39,0.21)	-0.12(-0.45,0.23)
4DEA	TMAO	-0.08(-0.27,0.12)	-0.03(-0.20,0.16)	0.02(-0.17,0.23)	0.06(-0.17,0.30)	0.09(-0.16,0.36)	0.11(-0.19,0.42)	0.13(-0.20,0.48)
4DEA	Gly	0.13(-0.06,0.34)	0.12(-0.06,0.32)	0.10(-0.10,0.32)	0.07(-0.16,0.30)	0.04(-0.22,0.29)	0.01(-0.30,0.30)	-0.02(-0.38,0.32)
4DEA	Urea	-0.09(-0.29,0.10)	-0.11(-0.30,0.07)	-0.11(-0.31,0.08)	-0.11(-0.33,0.11)	-0.10(-0.36,0.15)	-0.09(-0.37,0.21)	-0.08(-0.39,0.29)
4DEA	For	0.05(-0.15,0.24)	0.05(-0.13,0.23)	0.03(-0.15,0.23)	0.02(-0.20,0.26)	0.00(-0.27,0.29)	-0.01(-0.35,0.32)	-0.03(-0.42,0.36)
3HV	Lac	0.12(-0.10,0.32)	0.13(-0.06,0.31)	0.15(-0.04,0.34)	0.18(-0.04,0.38)	0.20(-0.05,0.45)	0.23(-0.07,0.51)	0.26(-0.09,0.57)
3HV	Ala	0.25(0.06,0.44)	0.27(0.11,0.44)	0.29(0.12,0.46)	0.31(0.09,0.50)	0.32(0.07,0.54)	0.33(0.04,0.59)	0.35(0.00,0.63)
3HV	AcO	-0.14(-0.34,0.08)	-0.18(-0.34,0.01)	-0.20(-0.38,-0.01)	-0.22(-0.43,0.00)	-0.24(-0.48,0.03)	-0.25(-0.54,0.05)	-0.26(-0.58,0.10)
3HV	Pyr	-0.10(-0.28,0.10)	-0.11(-0.28,0.07)	-0.12(-0.31,0.07)	-0.14(-0.34,0.11)	-0.15(-0.40,0.14)	-0.16(-0.46,0.17)	-0.18(-0.51,0.20)
3HV	Suc	0.08(-0.12,0.27)	0.08(-0.10,0.25)	0.08(-0.11,0.29)	0.07(-0.15,0.31)	0.07(-0.20,0.34)	0.06(-0.25,0.36)	0.05(-0.30,0.40)
3HV	Cit	0.17(-0.02,0.36)	0.15(-0.02,0.32)	0.13(-0.06,0.33)	0.12(-0.12,0.34)	0.10(-0.19,0.36)	0.09(-0.24,0.38)	0.07(-0.30,0.40)
3HV	DMA	0.11(-0.10,0.30)	0.25(0.07,0.41)	0.36(0.17,0.54)	0.45(0.23,0.65)	0.53(0.28,0.72)	0.58(0.31,0.78)	0.61(0.33,0.82)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
3HV	TMA	0.19(-0.01,0.38)	0.25(0.07,0.41)	0.30(0.11,0.47)	0.34(0.14,0.53)	0.38(0.15,0.60)	0.41(0.14,0.64)	0.43(0.14,0.69)
3HV	DMG	0.27(0.07,0.46)	0.23(0.06,0.39)	0.19(-0.01,0.39)	0.16(-0.10,0.41)	0.14(-0.16,0.43)	0.12(-0.22,0.46)	0.11(-0.26,0.48)
3HV	Crea	-0.13(-0.32,0.09)	-0.08(-0.27,0.11)	-0.05(-0.25,0.16)	-0.03(-0.26,0.22)	-0.01(-0.29,0.29)	0.00(-0.34,0.36)	0.01(-0.38,0.42)
3HV	Creat	0.07(-0.14,0.27)	0.17(-0.01,0.35)	0.26(0.08,0.46)	0.34(0.12,0.55)	0.39(0.13,0.62)	0.43(0.13,0.67)	0.46(0.11,0.72)
3HV	PB	0.17(-0.03,0.36)	0.17(0.00,0.34)	0.17(-0.03,0.36)	0.17(-0.07,0.41)	0.16(-0.14,0.44)	0.15(-0.18,0.45)	0.14(-0.23,0.47)
3HV	PCS	-0.08(-0.26,0.12)	-0.09(-0.25,0.09)	-0.10(-0.28,0.10)	-0.11(-0.33,0.12)	-0.12(-0.37,0.16)	-0.13(-0.41,0.19)	-0.14(-0.45,0.22)
3HV	PAG	-0.13(-0.31,0.07)	-0.15(-0.31,0.02)	-0.16(-0.35,0.03)	-0.18(-0.38,0.05)	-0.19(-0.43,0.08)	-0.20(-0.46,0.12)	-0.21(-0.51,0.15)
3HV	Hip	-0.22(-0.40,-0.02)	-0.23(-0.38,-0.05)	-0.23(-0.40,-0.03)	-0.24(-0.43,0.00)	-0.25(-0.47,0.04)	-0.25(-0.52,0.07)	-0.26(-0.56,0.10)
3HV	NMNA	0.08(-0.12,0.29)	0.10(-0.08,0.29)	0.09(-0.12,0.30)	0.08(-0.15,0.30)	0.05(-0.21,0.31)	0.03(-0.28,0.33)	0.00(-0.35,0.36)
3HV	TMAO	0.24(0.04,0.42)	0.31(0.13,0.46)	0.36(0.19,0.54)	0.41(0.22,0.60)	0.46(0.22,0.66)	0.49(0.23,0.71)	0.52(0.23,0.75)
3HV	Gly	0.21(0.00,0.39)	0.23(0.04,0.43)	0.22(0.01,0.45)	0.19(-0.05,0.44)	0.16(-0.13,0.41)	0.13(-0.20,0.41)	0.10(-0.28,0.43)
3HV	Urea	-0.05(-0.26,0.17)	-0.10(-0.30,0.08)	-0.14(-0.35,0.06)	-0.17(-0.39,0.07)	-0.18(-0.43,0.10)	-0.19(-0.48,0.12)	-0.20(-0.52,0.17)
3HV	For	-0.10(-0.30,0.09)	-0.09(-0.27,0.09)	-0.08(-0.26,0.12)	-0.06(-0.28,0.17)	-0.05(-0.32,0.22)	-0.04(-0.35,0.29)	-0.03(-0.39,0.34)
Lac	Ala	0.16(-0.04,0.36)	0.25(0.08,0.41)	0.33(0.17,0.49)	0.39(0.21,0.56)	0.44(0.22,0.63)	0.48(0.24,0.68)	0.52(0.25,0.72)
Lac	AcO	0.47(0.30,0.62)	0.40(0.24,0.54)	0.33(0.16,0.49)	0.28(0.06,0.48)	0.23(-0.04,0.46)	0.18(-0.12,0.47)	0.15(-0.20,0.48)
Lac	Pyr	-0.18(-0.38,0.02)	-0.20(-0.37,-0.02)	-0.21(-0.39,-0.03)	-0.22(-0.44,-0.01)	-0.23(-0.48,0.02)	-0.24(-0.53,0.06)	-0.24(-0.58,0.11)
Lac	Suc	0.04(-0.15,0.24)	0.07(-0.11,0.25)	0.10(-0.09,0.30)	0.12(-0.10,0.35)	0.13(-0.12,0.40)	0.14(-0.14,0.44)	0.15(-0.19,0.48)
Lac	Cit	-0.04(-0.25,0.18)	-0.01(-0.21,0.17)	0.01(-0.18,0.20)	0.03(-0.20,0.25)	0.05(-0.23,0.33)	0.07(-0.27,0.38)	0.08(-0.30,0.43)
Lac	DMA	-0.18(-0.38,0.03)	-0.08(-0.25,0.11)	0.01(-0.18,0.22)	0.09(-0.14,0.34)	0.15(-0.11,0.43)	0.20(-0.10,0.50)	0.25(-0.09,0.56)
Lac	TMA	0.08(-0.12,0.29)	0.05(-0.12,0.23)	0.02(-0.17,0.21)	-0.01(-0.23,0.22)	-0.03(-0.29,0.25)	-0.05(-0.35,0.28)	-0.07(-0.41,0.32)
Lac	DMG	0.07(-0.14,0.27)	0.08(-0.11,0.27)	0.10(-0.11,0.30)	0.12(-0.10,0.36)	0.15(-0.13,0.42)	0.17(-0.15,0.48)	0.19(-0.18,0.53)
Lac	Crea	-0.10(-0.32,0.12)	-0.03(-0.23,0.15)	0.04(-0.14,0.23)	0.10(-0.14,0.32)	0.16(-0.12,0.41)	0.20(-0.14,0.50)	0.24(-0.16,0.57)
Lac	Creat	-0.08(-0.29,0.15)	-0.07(-0.28,0.14)	-0.03(-0.29,0.19)	0.02(-0.26,0.25)	0.09(-0.18,0.34)	0.15(-0.15,0.42)	0.20(-0.14,0.50)
Lac	PB	0.06(-0.14,0.26)	0.09(-0.09,0.27)	0.11(-0.08,0.31)	0.11(-0.10,0.34)	0.12(-0.14,0.36)	0.12(-0.18,0.41)	0.12(-0.24,0.45)
Lac	PCS	-0.21(-0.41,-0.01)	-0.23(-0.39,-0.05)	-0.24(-0.41,-0.06)	-0.25(-0.46,-0.04)	-0.26(-0.51,-0.02)	-0.27(-0.56,0.02)	-0.28(-0.60,0.05)
Lac	PAG	-0.23(-0.41,-0.02)	-0.24(-0.39,-0.07)	-0.25(-0.43,-0.07)	-0.27(-0.47,-0.07)	-0.28(-0.52,-0.03)	-0.29(-0.57,0.01)	-0.30(-0.61,0.04)
Lac	Hip	-0.04(-0.24,0.19)	-0.08(-0.25,0.10)	-0.14(-0.32,0.05)	-0.20(-0.39,0.02)	-0.25(-0.47,0.01)	-0.29(-0.55,0.02)	-0.33(-0.62,0.03)
Lac		-0.11(-0.32,0.10)	-0.12(-0.28,0.06)	-0.13(-0.31,0.07)	-0.15(-0.36,0.09)	-0.16(-0.41,0.13)	-0.17(-0.45,0.18)	-0.17(-0.50,0.23)
NMNA								
Lac	TMAO	0.04(-0.17,0.25)	0.07(-0.11,0.24)	0.10(-0.09,0.29)	0.13(-0.10,0.34)	0.15(-0.11,0.40)	0.18(-0.13,0.46)	0.20(-0.14,0.51)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
Lac	Gly	0.11(-0.11,0.32)	0.19(0.01,0.36)	0.25(0.08,0.42)	0.31(0.10,0.50)	0.35(0.10,0.57)	0.39(0.10,0.62)	0.41(0.08,0.68)
Lac	Urea	-0.20(-0.39,0.00)	-0.24(-0.41,-0.07)	-0.28(-0.45,-0.10)	-0.30(-0.49,-0.10)	-0.32(-0.55,-0.10)	-0.34(-0.61,-0.06)	-0.35(-0.65,-0.03)
Lac	For	0.01(-0.20,0.22)	0.04(-0.14,0.22)	0.06(-0.12,0.26)	0.08(-0.14,0.31)	0.10(-0.18,0.37)	0.11(-0.22,0.43)	0.12(-0.27,0.49)
Ala	AcO	-0.15(-0.34,0.06)	-0.17(-0.33,0.02)	-0.18(-0.36,0.02)	-0.20(-0.40,0.04)	-0.21(-0.44,0.07)	-0.22(-0.48,0.12)	-0.23(-0.52,0.16)
Ala	Pyr	0.06(-0.15,0.27)	0.11(-0.06,0.30)	0.14(-0.06,0.34)	0.16(-0.09,0.40)	0.17(-0.13,0.45)	0.18(-0.19,0.50)	0.18(-0.26,0.54)
Ala	Suc	0.19(-0.01,0.38)	0.16(-0.02,0.33)	0.13(-0.07,0.32)	0.12(-0.12,0.34)	0.11(-0.15,0.37)	0.11(-0.21,0.41)	0.10(-0.26,0.46)
Ala	Cit	0.36(0.18,0.52)	0.35(0.19,0.50)	0.34(0.15,0.49)	0.33(0.10,0.52)	0.33(0.05,0.56)	0.32(0.00,0.59)	0.31(-0.06,0.62)
Ala	DMA	-0.09(-0.31,0.14)	0.04(-0.15,0.23)	0.16(-0.04,0.36)	0.25(0.01,0.49)	0.32(0.03,0.59)	0.38(0.06,0.66)	0.42(0.07,0.71)
Ala	TMA	0.15(-0.04,0.34)	0.12(-0.05,0.29)	0.10(-0.09,0.28)	0.08(-0.14,0.30)	0.06(-0.20,0.33)	0.04(-0.27,0.36)	0.02(-0.32,0.39)
Ala	DMG	0.26(0.07,0.45)	0.35(0.17,0.53)	0.42(0.22,0.61)	0.45(0.23,0.66)	0.47(0.22,0.69)	0.48(0.19,0.69)	0.46(0.10,0.71)
Ala	Crea	-0.09(-0.30,0.12)	-0.02(-0.22,0.17)	0.05(-0.14,0.25)	0.12(-0.11,0.34)	0.17(-0.11,0.42)	0.22(-0.12,0.50)	0.26(-0.15,0.56)
Ala	Creat	-0.10(-0.33,0.12)	-0.04(-0.24,0.17)	0.06(-0.18,0.27)	0.16(-0.09,0.40)	0.26(-0.01,0.52)	0.34(0.03,0.62)	0.40(0.06,0.69)
Ala	PB	-0.10(-0.30,0.10)	-0.09(-0.27,0.08)	-0.08(-0.28,0.11)	-0.08(-0.31,0.15)	-0.08(-0.34,0.19)	-0.08(-0.38,0.23)	-0.08(-0.42,0.27)
Ala	PCS	0.01(-0.21,0.21)	0.05(-0.14,0.26)	0.07(-0.13,0.29)	0.08(-0.16,0.32)	0.08(-0.21,0.38)	0.08(-0.27,0.43)	0.08(-0.35,0.48)
Ala	PAG	-0.08(-0.29,0.14)	-0.04(-0.23,0.19)	-0.02(-0.22,0.19)	-0.02(-0.25,0.22)	-0.03(-0.32,0.28)	-0.04(-0.40,0.33)	-0.05(-0.48,0.38)
Ala	Hip	-0.19(-0.38,0.03)	-0.23(-0.40,-0.06)	-0.27(-0.44,-0.09)	-0.31(-0.50,-0.11)	-0.35(-0.57,-0.11)	-0.38(-0.62,-0.09)	-0.41(-0.67,-0.09)
Ala	NMNA	0.00(-0.20,0.19)	-0.10(-0.27,0.07)	-0.18(-0.34,0.00)	-0.24(-0.43,-0.04)	-0.30(-0.51,-0.07)	-0.35(-0.58,-0.08)	-0.39(-0.63,-0.08)
Ala	TMAO	0.03(-0.18,0.23)	0.12(-0.07,0.29)	0.19(0.00,0.38)	0.25(0.02,0.47)	0.30(0.04,0.54)	0.34(0.03,0.61)	0.37(0.00,0.65)
Ala	Gly	0.57(0.40,0.70)	0.55(0.41,0.66)	0.53(0.39,0.66)	0.52(0.34,0.67)	0.50(0.28,0.69)	0.49(0.20,0.73)	0.48(0.12,0.76)
Ala	Urea	-0.06(-0.25,0.15)	-0.08(-0.26,0.10)	-0.10(-0.30,0.09)	-0.12(-0.32,0.10)	-0.13(-0.37,0.12)	-0.14(-0.43,0.17)	-0.15(-0.48,0.21)
Ala	For	0.21(-0.01,0.40)	0.17(0.00,0.37)	0.13(-0.05,0.33)	0.08(-0.13,0.31)	0.04(-0.23,0.32)	0.00(-0.33,0.35)	-0.03(-0.41,0.38)
AcO	Pyr	-0.01(-0.21,0.20)	0.00(-0.18,0.18)	0.00(-0.18,0.19)	0.01(-0.21,0.22)	0.01(-0.24,0.26)	0.02(-0.28,0.31)	0.02(-0.32,0.36)
AcO	Suc	0.15(-0.07,0.36)	0.15(-0.04,0.34)	0.13(-0.05,0.33)	0.12(-0.09,0.33)	0.10(-0.17,0.35)	0.08(-0.24,0.39)	0.06(-0.34,0.42)
AcO	Cit	-0.15(-0.35,0.06)	-0.15(-0.33,0.04)	-0.15(-0.33,0.04)	-0.16(-0.36,0.06)	-0.16(-0.40,0.08)	-0.17(-0.45,0.12)	-0.17(-0.50,0.16)
AcO	DMA	-0.36(-0.52,-0.18)	-0.34(-0.49,-0.17)	-0.32(-0.48,-0.13)	-0.30(-0.50,-0.09)	-0.29(-0.52,-0.03)	-0.27(-0.54,0.03)	-0.26(-0.56,0.09)
AcO	TMA	-0.05(-0.25,0.17)	-0.08(-0.25,0.10)	-0.11(-0.29,0.07)	-0.14(-0.35,0.07)	-0.16(-0.41,0.09)	-0.18(-0.48,0.12)	-0.20(-0.53,0.14)
AcO	DMG	-0.01(-0.22,0.23)	-0.07(-0.25,0.12)	-0.13(-0.33,0.07)	-0.18(-0.41,0.06)	-0.22(-0.49,0.07)	-0.25(-0.56,0.09)	-0.27(-0.61,0.13)
AcO	Crea	0.12(-0.09,0.34)	0.15(-0.04,0.33)	0.19(0.00,0.37)	0.22(0.01,0.43)	0.25(-0.01,0.50)	0.28(-0.05,0.56)	0.30(-0.07,0.62)
AcO	Creat	-0.19(-0.40,0.03)	-0.24(-0.41,-0.05)	-0.27(-0.45,-0.06)	-0.29(-0.50,-0.05)	-0.29(-0.55,-0.03)	-0.28(-0.58,0.02)	-0.28(-0.61,0.06)
AcO	PB	-0.11(-0.31,0.10)	-0.06(-0.24,0.13)	-0.01(-0.20,0.18)	0.03(-0.20,0.26)	0.05(-0.21,0.32)	0.07(-0.23,0.39)	0.09(-0.26,0.45)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
AcO	PCS	0.07(-0.13,0.27)	0.05(-0.12,0.22)	0.04(-0.14,0.22)	0.02(-0.18,0.23)	0.01(-0.24,0.26)	0.00(-0.29,0.30)	-0.01(-0.35,0.33)
AcO	PAG	0.00(-0.21,0.20)	-0.01(-0.18,0.18)	-0.01(-0.19,0.19)	-0.01(-0.22,0.21)	-0.01(-0.27,0.26)	0.00(-0.30,0.30)	0.00(-0.34,0.35)
AcO	Hip	0.00(-0.21,0.20)	-0.01(-0.18,0.17)	-0.01(-0.19,0.18)	-0.01(-0.23,0.22)	-0.01(-0.27,0.25)	-0.01(-0.30,0.29)	-0.01(-0.33,0.34)
AcO	NMNA	-0.10(-0.29,0.10)	-0.09(-0.26,0.09)	-0.07(-0.25,0.12)	-0.06(-0.27,0.17)	-0.04(-0.30,0.22)	-0.03(-0.34,0.27)	-0.01(-0.39,0.32)
AcO	TMAO	-0.21(-0.38,-0.01)	-0.21(-0.37,-0.03)	-0.21(-0.37,-0.01)	-0.21(-0.40,0.03)	-0.20(-0.44,0.07)	-0.19(-0.47,0.11)	-0.19(-0.50,0.16)
AcO	Gly	-0.17(-0.37,0.05)	-0.15(-0.32,0.03)	-0.13(-0.32,0.06)	-0.12(-0.33,0.11)	-0.11(-0.36,0.15)	-0.09(-0.39,0.21)	-0.08(-0.43,0.27)
AcO	Urea	0.15(-0.07,0.35)	0.15(-0.03,0.33)	0.14(-0.04,0.33)	0.13(-0.08,0.34)	0.11(-0.14,0.36)	0.09(-0.20,0.37)	0.08(-0.27,0.40)
AcO	For	0.11(-0.11,0.36)	0.04(-0.15,0.26)	-0.03(-0.22,0.17)	-0.08(-0.31,0.16)	-0.13(-0.44,0.16)	-0.17(-0.53,0.18)	-0.20(-0.60,0.20)
Pyr	Suc	-0.08(-0.28,0.13)	-0.03(-0.19,0.16)	0.02(-0.15,0.22)	0.06(-0.14,0.28)	0.10(-0.15,0.36)	0.13(-0.16,0.43)	0.16(-0.18,0.49)
Pyr	Cit	0.04(-0.17,0.25)	0.08(-0.10,0.25)	0.10(-0.09,0.29)	0.13(-0.11,0.35)	0.14(-0.14,0.41)	0.16(-0.15,0.46)	0.18(-0.18,0.51)
Pyr	DMA	0.01(-0.20,0.21)	-0.01(-0.17,0.17)	-0.02(-0.21,0.18)	-0.02(-0.25,0.21)	-0.03(-0.31,0.25)	-0.03(-0.35,0.30)	-0.04(-0.39,0.32)
Pyr	TMA	-0.07(-0.26,0.13)	-0.05(-0.21,0.12)	-0.04(-0.22,0.15)	-0.03(-0.24,0.20)	-0.02(-0.28,0.25)	-0.02(-0.32,0.29)	-0.02(-0.36,0.33)
Pyr	DMG	-0.22(-0.40,-0.01)	-0.18(-0.35,0.01)	-0.16(-0.34,0.05)	-0.16(-0.37,0.08)	-0.17(-0.42,0.11)	-0.18(-0.50,0.14)	-0.21(-0.56,0.18)
Pyr	Crea	-0.17(-0.38,0.06)	-0.11(-0.29,0.06)	-0.06(-0.25,0.13)	-0.02(-0.24,0.21)	0.02(-0.26,0.30)	0.05(-0.28,0.37)	0.08(-0.31,0.43)
Pyr	Creat	-0.13(-0.33,0.09)	-0.11(-0.28,0.08)	-0.09(-0.27,0.10)	-0.08(-0.31,0.15)	-0.08(-0.34,0.20)	-0.07(-0.37,0.24)	-0.07(-0.42,0.29)
Pyr	PB	0.00(-0.21,0.21)	-0.09(-0.26,0.08)	-0.17(-0.35,0.01)	-0.24(-0.44,-0.01)	-0.29(-0.53,-0.03)	-0.34(-0.61,-0.04)	-0.38(-0.67,-0.06)
Pyr	PCS	0.77(0.67,0.85)	0.81(0.75,0.87)	0.85(0.78,0.90)	0.87(0.79,0.94)	0.89(0.80,0.95)	0.90(0.81,0.96)	0.91(0.81,0.97)
Pyr	PAG	0.74(0.63,0.83)	0.73(0.64,0.80)	0.71(0.62,0.80)	0.70(0.58,0.80)	0.69(0.54,0.81)	0.68(0.49,0.83)	0.68(0.44,0.84)
Pyr	Hip	-0.03(-0.24,0.18)	0.02(-0.15,0.20)	0.06(-0.13,0.26)	0.09(-0.12,0.32)	0.12(-0.13,0.40)	0.15(-0.16,0.47)	0.17(-0.20,0.53)
Pyr	NMNA	-0.37(-0.52,-0.18)	-0.29(-0.44,-0.12)	-0.22(-0.39,-0.03)	-0.16(-0.38,0.08)	-0.11(-0.37,0.18)	-0.06(-0.37,0.26)	-0.02(-0.38,0.33)
Pyr	TMAO	0.08(-0.11,0.29)	0.05(-0.12,0.23)	0.02(-0.16,0.21)	0.00(-0.21,0.22)	-0.02(-0.27,0.24)	-0.04(-0.33,0.25)	-0.06(-0.38,0.27)
Pyr	Gly	-0.05(-0.24,0.15)	-0.01(-0.17,0.17)	0.03(-0.14,0.22)	0.06(-0.16,0.29)	0.08(-0.20,0.37)	0.10(-0.24,0.43)	0.11(-0.28,0.49)
Pyr	Urea	0.14(-0.07,0.35)	0.16(-0.02,0.32)	0.17(-0.01,0.35)	0.18(-0.03,0.37)	0.19(-0.06,0.42)	0.20(-0.09,0.47)	0.21(-0.14,0.52)
Pyr	For	-0.02(-0.23,0.19)	-0.07(-0.24,0.11)	-0.11(-0.30,0.07)	-0.14(-0.36,0.07)	-0.17(-0.43,0.09)	-0.19(-0.49,0.11)	-0.21(-0.55,0.13)
Suc	Cit	0.45(0.27,0.60)	0.45(0.31,0.58)	0.46(0.32,0.60)	0.47(0.29,0.64)	0.49(0.27,0.67)	0.50(0.24,0.70)	0.51(0.22,0.73)
Suc	DMA	-0.03(-0.24,0.17)	-0.01(-0.18,0.18)	0.02(-0.17,0.22)	0.04(-0.19,0.27)	0.07(-0.22,0.33)	0.09(-0.24,0.39)	0.10(-0.27,0.44)
Suc	TMA	-0.01(-0.21,0.18)	-0.03(-0.21,0.16)	-0.04(-0.23,0.15)	-0.06(-0.27,0.16)	-0.07(-0.32,0.19)	-0.09(-0.36,0.22)	-0.10(-0.41,0.24)
Suc	DMG	-0.04(-0.25,0.18)	-0.04(-0.22,0.15)	-0.02(-0.23,0.19)	0.01(-0.25,0.27)	0.04(-0.27,0.35)	0.07(-0.29,0.41)	0.10(-0.29,0.48)
Suc	Crea	0.16(-0.05,0.35)	0.09(-0.11,0.28)	0.06(-0.16,0.27)	0.04(-0.21,0.29)	0.05(-0.24,0.33)	0.05(-0.30,0.40)	0.06(-0.33,0.48)
Suc	Creat	-0.03(-0.23,0.19)	0.04(-0.13,0.24)	0.10(-0.10,0.34)	0.15(-0.10,0.42)	0.19(-0.11,0.47)	0.21(-0.12,0.51)	0.22(-0.14,0.55)

Table E.3 – continued from previous page

Met1	Met2	[1]U-Cd	[2]U-Cd	[3]U-Cd	[4]U-Cd	[5]U-Cd	[6]U-Cd	[7]U-Cd
Suc	PB	0.16(-0.06,0.35)	0.13(-0.06,0.31)	0.11(-0.08,0.31)	0.09(-0.15,0.32)	0.07(-0.21,0.33)	0.05(-0.26,0.35)	0.04(-0.31,0.37)
Suc	PCS	0.10(-0.10,0.29)	0.12(-0.05,0.30)	0.14(-0.05,0.32)	0.16(-0.05,0.38)	0.17(-0.07,0.42)	0.18(-0.11,0.46)	0.19(-0.16,0.50)
Suc	PAG	-0.01(-0.22,0.20)	0.04(-0.13,0.22)	0.07(-0.11,0.26)	0.10(-0.11,0.32)	0.12(-0.13,0.37)	0.14(-0.15,0.41)	0.15(-0.18,0.46)
Suc	Hip	0.01(-0.20,0.20)	0.03(-0.15,0.21)	0.04(-0.15,0.23)	0.05(-0.16,0.27)	0.05(-0.19,0.31)	0.06(-0.23,0.35)	0.06(-0.27,0.38)
Suc	NMNA	-0.02(-0.22,0.18)	-0.04(-0.20,0.16)	-0.05(-0.23,0.15)	-0.05(-0.26,0.17)	-0.05(-0.30,0.19)	-0.06(-0.35,0.23)	-0.06(-0.39,0.28)
Suc	TMAO	-0.08(-0.28,0.12)	-0.06(-0.23,0.14)	-0.04(-0.23,0.16)	-0.03(-0.24,0.21)	-0.02(-0.27,0.24)	-0.01(-0.30,0.28)	-0.01(-0.34,0.32)
Suc	Gly	0.27(0.08,0.46)	0.27(0.07,0.42)	0.28(0.05,0.45)	0.30(0.08,0.49)	0.32(0.08,0.54)	0.35(0.07,0.59)	0.37(0.06,0.64)
Suc	Urea	-0.21(-0.39,-0.02)	-0.23(-0.38,-0.06)	-0.25(-0.41,-0.06)	-0.26(-0.44,-0.05)	-0.27(-0.50,-0.03)	-0.28(-0.54,-0.01)	-0.29(-0.57,0.02)
Suc	For	0.24(0.04,0.44)	0.19(0.01,0.38)	0.16(-0.04,0.35)	0.15(-0.08,0.37)	0.14(-0.12,0.39)	0.14(-0.16,0.44)	0.14(-0.21,0.49)
Cit	DMA	0.04(-0.17,0.25)	0.09(-0.09,0.26)	0.12(-0.07,0.32)	0.16(-0.08,0.40)	0.19(-0.10,0.46)	0.21(-0.12,0.52)	0.23(-0.15,0.57)
Cit	TMA	-0.13(-0.32,0.07)	-0.15(-0.31,0.04)	-0.16(-0.32,0.03)	-0.17(-0.37,0.05)	-0.18(-0.42,0.08)	-0.19(-0.46,0.12)	-0.20(-0.50,0.16)
Cit	DMG	-0.04(-0.26,0.17)	0.08(-0.11,0.26)	0.18(-0.06,0.39)	0.24(-0.06,0.49)	0.26(-0.07,0.54)	0.27(-0.09,0.56)	0.25(-0.13,0.57)
Cit	Crea	-0.06(-0.26,0.16)	-0.04(-0.24,0.15)	-0.01(-0.21,0.19)	0.02(-0.22,0.26)	0.06(-0.25,0.32)	0.09(-0.27,0.39)	0.12(-0.32,0.46)
Cit	Creat	-0.10(-0.29,0.10)	0.00(-0.17,0.20)	0.09(-0.11,0.33)	0.16(-0.09,0.43)	0.22(-0.06,0.48)	0.26(-0.05,0.54)	0.29(-0.08,0.60)
Cit	PB	0.10(-0.11,0.30)	-0.03(-0.22,0.14)	-0.15(-0.34,0.05)	-0.23(-0.46,0.00)	-0.30(-0.56,-0.04)	-0.36(-0.63,-0.05)	-0.39(-0.67,-0.07)
Cit	PCS	0.10(-0.11,0.30)	0.12(-0.05,0.30)	0.14(-0.03,0.33)	0.16(-0.05,0.37)	0.18(-0.08,0.41)	0.19(-0.12,0.46)	0.21(-0.15,0.50)
Cit	PAG	0.05(-0.16,0.25)	0.09(-0.08,0.28)	0.12(-0.05,0.32)	0.16(-0.07,0.38)	0.19(-0.08,0.44)	0.21(-0.10,0.49)	0.24(-0.13,0.55)
Cit	Hip	0.09(-0.12,0.30)	0.04(-0.14,0.23)	0.01(-0.19,0.21)	-0.01(-0.24,0.21)	-0.01(-0.26,0.24)	-0.02(-0.31,0.28)	-0.01(-0.35,0.33)
Cit	NMNA	0.08(-0.12,0.28)	0.05(-0.12,0.23)	0.02(-0.16,0.22)	-0.01(-0.23,0.21)	-0.03(-0.29,0.23)	-0.06(-0.36,0.26)	-0.08(-0.41,0.29)
Cit	TMAO	0.05(-0.14,0.25)	0.06(-0.12,0.24)	0.07(-0.12,0.27)	0.08(-0.15,0.32)	0.08(-0.19,0.36)	0.09(-0.24,0.39)	0.09(-0.28,0.44)
Cit	Gly	0.43(0.25,0.58)	0.47(0.33,0.61)	0.50(0.35,0.65)	0.53(0.34,0.69)	0.55(0.34,0.72)	0.56(0.33,0.75)	0.58(0.31,0.78)
Cit	Urea	-0.29(-0.47,-0.10)	-0.24(-0.40,-0.07)	-0.20(-0.38,0.00)	-0.17(-0.38,0.06)	-0.15(-0.40,0.12)	-0.13(-0.42,0.19)	-0.11(-0.47,0.24)
Cit	For	0.22(0.01,0.41)	0.14(-0.04,0.31)	0.07(-0.12,0.25)	0.01(-0.21,0.23)	-0.04(-0.28,0.24)	-0.08(-0.37,0.25)	-0.12(-0.42,0.29)
DMA	TMA	0.17(-0.03,0.36)	0.22(0.05,0.38)	0.25(0.07,0.43)	0.28(0.07,0.50)	0.30(0.05,0.55)	0.33(0.05,0.60)	0.34(0.04,0.64)
DMA	DMG	0.06(-0.15,0.28)	0.14(-0.03,0.32)	0.21(0.01,0.41)	0.27(0.03,0.51)	0.32(0.00,0.58)	0.35(-0.02,0.64)	0.37(-0.04,0.68)
DMA	Crea	0.01(-0.20,0.20)	0.06(-0.11,0.24)	0.10(-0.09,0.32)	0.14(-0.11,0.38)	0.17(-0.12,0.45)	0.19(-0.14,0.50)	0.22(-0.15,0.55)
DMA	Creat	0.52(0.34,0.69)	0.54(0.41,0.67)	0.56(0.40,0.69)	0.57(0.34,0.73)	0.59(0.29,0.77)	0.60(0.27,0.79)	0.61(0.28,0.82)
DMA	PB	0.24(0.04,0.43)	0.17(0.00,0.34)	0.11(-0.07,0.30)	0.05(-0.18,0.28)	0.00(-0.26,0.28)	-0.04(-0.35,0.27)	-0.08(-0.42,0.28)
DMA	PCS	-0.03(-0.24,0.18)	-0.03(-0.20,0.16)	-0.02(-0.20,0.19)	-0.01(-0.23,0.22)	0.00(-0.27,0.27)	0.00(-0.29,0.32)	0.01(-0.33,0.36)
DMA	PAG	-0.03(-0.22,0.17)	-0.02(-0.19,0.16)	0.00(-0.18,0.21)	0.01(-0.19,0.25)	0.02(-0.23,0.29)	0.02(-0.27,0.34)	0.03(-0.31,0.39)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
DMA	Hip	-0.25(-0.44,-0.04)	-0.25(-0.41,-0.08)	-0.25(-0.42,-0.05)	-0.25(-0.45,0.00)	-0.25(-0.48,0.05)	-0.26(-0.53,0.08)	-0.26(-0.56,0.11)
DMA	NMNA	-0.14(-0.34,0.07)	-0.09(-0.26,0.09)	-0.06(-0.25,0.15)	-0.03(-0.26,0.23)	-0.01(-0.27,0.30)	0.00(-0.30,0.36)	0.02(-0.33,0.42)
DMA	TMAO	0.54(0.38,0.67)	0.57(0.44,0.68)	0.59(0.45,0.71)	0.61(0.45,0.75)	0.62(0.43,0.78)	0.63(0.42,0.80)	0.64(0.41,0.82)
DMA	Gly	0.06(-0.16,0.28)	0.13(-0.05,0.31)	0.18(-0.03,0.40)	0.22(-0.03,0.47)	0.24(-0.04,0.53)	0.25(-0.06,0.58)	0.26(-0.11,0.61)
DMA	Urea	-0.29(-0.46,-0.10)	-0.30(-0.45,-0.13)	-0.30(-0.46,-0.10)	-0.29(-0.48,-0.07)	-0.29(-0.52,-0.03)	-0.28(-0.54,0.02)	-0.28(-0.57,0.07)
DMA	For	-0.10(-0.28,0.10)	-0.10(-0.27,0.07)	-0.11(-0.29,0.08)	-0.11(-0.33,0.13)	-0.11(-0.37,0.17)	-0.11(-0.41,0.21)	-0.11(-0.46,0.25)
TMA	DMG	0.22(0.02,0.43)	0.23(0.06,0.41)	0.23(0.05,0.44)	0.22(0.00,0.46)	0.20(-0.07,0.48)	0.18(-0.16,0.49)	0.14(-0.23,0.50)
TMA	Crea	-0.20(-0.41,0.01)	-0.17(-0.35,0.02)	-0.14(-0.33,0.07)	-0.10(-0.34,0.14)	-0.08(-0.37,0.23)	-0.05(-0.40,0.32)	-0.03(-0.43,0.39)
TMA	Creat	0.08(-0.14,0.29)	0.09(-0.10,0.27)	0.10(-0.10,0.29)	0.12(-0.11,0.35)	0.13(-0.15,0.41)	0.15(-0.18,0.46)	0.16(-0.21,0.52)
TMA	PB	0.05(-0.14,0.26)	0.08(-0.09,0.25)	0.11(-0.08,0.31)	0.12(-0.10,0.37)	0.13(-0.12,0.41)	0.14(-0.16,0.43)	0.13(-0.20,0.46)
TMA	PCS	-0.02(-0.22,0.18)	0.02(-0.15,0.21)	0.04(-0.15,0.24)	0.05(-0.18,0.28)	0.05(-0.21,0.33)	0.05(-0.25,0.38)	0.05(-0.31,0.41)
TMA	PAG	-0.14(-0.32,0.06)	-0.10(-0.27,0.08)	-0.07(-0.26,0.13)	-0.04(-0.28,0.21)	-0.03(-0.31,0.28)	-0.01(-0.34,0.33)	0.00(-0.37,0.37)
TMA	Hip	-0.15(-0.34,0.06)	-0.11(-0.28,0.07)	-0.08(-0.26,0.13)	-0.05(-0.26,0.19)	-0.03(-0.29,0.25)	-0.01(-0.31,0.31)	0.00(-0.33,0.36)
TMA	NMNA	0.04(-0.16,0.24)	0.07(-0.10,0.25)	0.10(-0.09,0.28)	0.12(-0.10,0.34)	0.13(-0.12,0.39)	0.15(-0.15,0.45)	0.16(-0.18,0.50)
TMA	TMAO	0.19(0.01,0.37)	0.26(0.09,0.40)	0.31(0.14,0.47)	0.36(0.17,0.54)	0.40(0.17,0.59)	0.43(0.17,0.64)	0.46(0.17,0.69)
TMA	Gly	0.10(-0.09,0.30)	0.05(-0.12,0.24)	0.01(-0.18,0.20)	-0.04(-0.25,0.19)	-0.08(-0.32,0.20)	-0.11(-0.40,0.20)	-0.14(-0.47,0.21)
TMA	Urea	0.09(-0.11,0.28)	0.02(-0.18,0.19)	-0.04(-0.25,0.16)	-0.08(-0.30,0.15)	-0.11(-0.36,0.17)	-0.13(-0.44,0.20)	-0.15(-0.48,0.25)
TMA	For	0.09(-0.12,0.28)	0.11(-0.06,0.29)	0.12(-0.07,0.32)	0.13(-0.10,0.36)	0.12(-0.16,0.40)	0.12(-0.20,0.44)	0.11(-0.25,0.48)
DMG	Crea	0.09(-0.11,0.30)	0.08(-0.13,0.29)	0.06(-0.17,0.29)	0.04(-0.20,0.28)	0.02(-0.24,0.27)	0.00(-0.31,0.33)	-0.02(-0.41,0.41)
DMG	Creat	0.31(0.10,0.50)	0.19(0.01,0.38)	0.10(-0.14,0.35)	0.05(-0.23,0.36)	0.04(-0.27,0.39)	0.06(-0.27,0.42)	0.10(-0.27,0.48)
DMG	PB	-0.01(-0.20,0.19)	0.15(-0.03,0.32)	0.28(0.07,0.46)	0.39(0.14,0.61)	0.47(0.19,0.71)	0.52(0.23,0.76)	0.55(0.24,0.77)
DMG	PCS	-0.23(-0.43,-0.02)	-0.16(-0.33,0.03)	-0.11(-0.33,0.11)	-0.10(-0.35,0.17)	-0.11(-0.39,0.19)	-0.14(-0.45,0.19)	-0.18(-0.54,0.19)
DMG	PAG	-0.21(-0.41,0.01)	-0.17(-0.36,0.02)	-0.16(-0.39,0.05)	-0.17(-0.41,0.06)	-0.20(-0.46,0.07)	-0.24(-0.53,0.08)	-0.29(-0.61,0.08)
DMG	Hip	-0.22(-0.42,-0.01)	-0.23(-0.39,-0.04)	-0.24(-0.44,-0.02)	-0.25(-0.50,0.02)	-0.26(-0.54,0.04)	-0.27(-0.57,0.04)	-0.27(-0.60,0.07)
DMG	NMNA	-0.01(-0.21,0.20)	-0.01(-0.19,0.20)	-0.02(-0.24,0.19)	-0.05(-0.28,0.20)	-0.09(-0.35,0.19)	-0.13(-0.42,0.18)	-0.17(-0.51,0.20)
DMG	TMAO	0.16(-0.06,0.37)	0.28(0.12,0.45)	0.37(0.17,0.55)	0.42(0.16,0.62)	0.43(0.13,0.68)	0.42(0.08,0.69)	0.39(-0.01,0.68)
DMG	Gly	0.27(0.06,0.46)	0.26(0.07,0.43)	0.25(0.03,0.46)	0.25(-0.02,0.49)	0.25(-0.04,0.52)	0.26(-0.07,0.55)	0.27(-0.09,0.59)
DMG	Urea	0.10(-0.14,0.30)	0.03(-0.15,0.21)	-0.02(-0.23,0.20)	-0.05(-0.29,0.19)	-0.07(-0.36,0.21)	-0.08(-0.40,0.26)	-0.09(-0.43,0.32)
DMG	For	-0.04(-0.26,0.18)	0.10(-0.10,0.29)	0.19(-0.05,0.41)	0.24(-0.01,0.46)	0.26(-0.01,0.51)	0.25(-0.09,0.53)	0.24(-0.20,0.58)
Crea	Creat	0.06(-0.16,0.29)	-0.01(-0.24,0.20)	-0.06(-0.29,0.16)	-0.08(-0.31,0.16)	-0.08(-0.36,0.21)	-0.08(-0.43,0.29)	-0.08(-0.50,0.38)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
Crea	PB	0.00(-0.21,0.23)	0.01(-0.19,0.21)	0.01(-0.21,0.23)	0.01(-0.22,0.24)	0.00(-0.28,0.28)	0.00(-0.32,0.33)	-0.01(-0.37,0.38)
Crea	PCS	-0.07(-0.29,0.16)	-0.07(-0.25,0.14)	-0.06(-0.24,0.14)	-0.05(-0.27,0.19)	-0.05(-0.33,0.24)	-0.04(-0.37,0.32)	-0.03(-0.42,0.37)
Crea	PAG	-0.05(-0.28,0.18)	-0.05(-0.23,0.15)	-0.03(-0.21,0.16)	-0.02(-0.24,0.21)	-0.01(-0.29,0.28)	0.00(-0.34,0.37)	0.01(-0.38,0.44)
Crea	Hip	0.04(-0.19,0.28)	-0.02(-0.22,0.21)	-0.07(-0.28,0.13)	-0.12(-0.34,0.12)	-0.15(-0.44,0.13)	-0.18(-0.53,0.18)	-0.21(-0.60,0.24)
Crea	NMNA	0.03(-0.19,0.24)	-0.06(-0.24,0.12)	-0.13(-0.32,0.06)	-0.19(-0.41,0.04)	-0.24(-0.48,0.05)	-0.27(-0.54,0.06)	-0.31(-0.59,0.09)
Crea	TMAO	-0.09(-0.30,0.12)	-0.01(-0.19,0.18)	0.05(-0.12,0.25)	0.11(-0.12,0.33)	0.15(-0.15,0.41)	0.18(-0.18,0.48)	0.21(-0.21,0.55)
Crea	Gly	-0.12(-0.34,0.09)	-0.11(-0.34,0.11)	-0.05(-0.26,0.18)	0.03(-0.19,0.27)	0.10(-0.18,0.38)	0.16(-0.19,0.49)	0.22(-0.21,0.58)
Crea	Urea	-0.04(-0.27,0.18)	0.08(-0.13,0.31)	0.16(-0.05,0.39)	0.21(-0.02,0.44)	0.24(-0.05,0.50)	0.26(-0.09,0.56)	0.28(-0.19,0.61)
Crea	For	0.29(0.02,0.52)	0.18(-0.05,0.40)	-0.03(-0.22,0.18)	-0.21(-0.44,0.03)	-0.35(-0.59,-0.05)	-0.45(-0.69,-0.10)	-0.52(-0.75,-0.13)
Creat	PB	0.09(-0.14,0.31)	0.03(-0.18,0.22)	-0.03(-0.29,0.18)	-0.07(-0.35,0.19)	-0.10(-0.40,0.20)	-0.11(-0.44,0.22)	-0.11(-0.47,0.25)
Creat	PCS	-0.18(-0.38,0.04)	-0.13(-0.30,0.07)	-0.09(-0.28,0.14)	-0.05(-0.28,0.22)	-0.03(-0.31,0.28)	-0.01(-0.35,0.33)	0.00(-0.40,0.38)
Creat	PAG	-0.14(-0.34,0.07)	-0.13(-0.30,0.07)	-0.13(-0.32,0.08)	-0.13(-0.36,0.11)	-0.13(-0.40,0.13)	-0.13(-0.44,0.18)	-0.13(-0.50,0.22)
Creat	Hip	-0.27(-0.47,-0.05)	-0.19(-0.39,0.03)	-0.15(-0.39,0.17)	-0.17(-0.43,0.17)	-0.22(-0.49,0.08)	-0.28(-0.55,0.02)	-0.34(-0.62,0.01)
Creat	NMNA	-0.23(-0.44,-0.03)	-0.16(-0.35,0.06)	-0.12(-0.34,0.21)	-0.10(-0.36,0.23)	-0.12(-0.40,0.19)	-0.14(-0.42,0.18)	-0.17(-0.49,0.23)
Creat	TMAO	0.03(-0.17,0.25)	0.05(-0.14,0.22)	0.07(-0.16,0.28)	0.09(-0.20,0.34)	0.11(-0.22,0.40)	0.14(-0.23,0.45)	0.16(-0.22,0.50)
Creat	Gly	0.05(-0.17,0.26)	0.10(-0.10,0.30)	0.14(-0.10,0.47)	0.17(-0.10,0.50)	0.18(-0.10,0.45)	0.18(-0.13,0.46)	0.17(-0.19,0.52)
Creat	Urea	-0.09(-0.30,0.13)	-0.25(-0.45,0.01)	-0.30(-0.53,0.00)	-0.25(-0.51,0.03)	-0.14(-0.41,0.16)	-0.03(-0.36,0.31)	0.08(-0.35,0.44)
Creat	For	-0.29(-0.49,-0.10)	-0.25(-0.45,-0.06)	-0.20(-0.39,0.01)	-0.13(-0.34,0.09)	-0.06(-0.32,0.21)	0.00(-0.32,0.32)	0.05(-0.34,0.40)
PB	PCS	0.02(-0.18,0.24)	-0.09(-0.25,0.09)	-0.18(-0.36,0.01)	-0.26(-0.47,-0.03)	-0.33(-0.56,-0.07)	-0.38(-0.63,-0.08)	-0.42(-0.67,-0.09)
PB	PAG	0.02(-0.18,0.23)	-0.10(-0.27,0.07)	-0.21(-0.39,-0.02)	-0.29(-0.50,-0.06)	-0.36(-0.59,-0.10)	-0.42(-0.65,-0.12)	-0.46(-0.71,-0.14)
PB	Hip	-0.08(-0.28,0.14)	-0.09(-0.26,0.09)	-0.09(-0.29,0.11)	-0.09(-0.32,0.15)	-0.09(-0.36,0.19)	-0.09(-0.38,0.22)	-0.08(-0.41,0.25)
PB	NMNA	-0.06(-0.26,0.13)	0.00(-0.18,0.18)	0.05(-0.15,0.27)	0.09(-0.15,0.34)	0.12(-0.15,0.39)	0.14(-0.17,0.44)	0.15(-0.20,0.48)
PB	TMAO	0.19(-0.02,0.38)	0.19(0.02,0.37)	0.19(-0.03,0.40)	0.17(-0.08,0.42)	0.14(-0.15,0.43)	0.11(-0.22,0.43)	0.08(-0.28,0.43)
PB	Gly	0.08(-0.15,0.30)	0.15(-0.04,0.35)	0.17(-0.07,0.43)	0.16(-0.12,0.44)	0.12(-0.18,0.42)	0.08(-0.24,0.38)	0.03(-0.32,0.38)
PB	Urea	-0.26(-0.46,-0.05)	-0.18(-0.35,0.01)	-0.10(-0.30,0.13)	-0.03(-0.27,0.23)	0.02(-0.27,0.34)	0.05(-0.27,0.41)	0.08(-0.29,0.48)
PB	For	0.04(-0.18,0.25)	0.11(-0.08,0.29)	0.18(-0.02,0.38)	0.24(0.01,0.47)	0.29(0.03,0.55)	0.33(0.04,0.60)	0.37(0.03,0.66)
PCS	PAG	0.80(0.72,0.87)	0.80(0.73,0.86)	0.80(0.72,0.86)	0.80(0.70,0.87)	0.80(0.68,0.88)	0.80(0.65,0.90)	0.80(0.63,0.91)
PCS	Hip	0.04(-0.17,0.23)	0.06(-0.12,0.25)	0.09(-0.10,0.28)	0.11(-0.11,0.33)	0.13(-0.14,0.39)	0.14(-0.18,0.45)	0.15(-0.21,0.50)
PCS	NMNA	-0.33(-0.50,-0.13)	-0.25(-0.41,-0.08)	-0.18(-0.36,0.01)	-0.12(-0.34,0.11)	-0.07(-0.33,0.20)	-0.03(-0.33,0.28)	0.01(-0.33,0.36)
PCS	TMAO	0.01(-0.17,0.22)	0.01(-0.15,0.19)	0.01(-0.18,0.21)	0.01(-0.21,0.24)	0.00(-0.26,0.27)	0.00(-0.30,0.31)	-0.01(-0.34,0.34)

Table E.3 – continued from previous page

Met1	Met2	^{1}U-Cd	^{2}U-Cd	^{3}U-Cd	^{4}U-Cd	^{5}U-Cd	^{6}U-Cd	^{7}U-Cd
PCS	Gly	-0.14(-0.33,0.07)	-0.10(-0.28,0.08)	-0.06(-0.24,0.14)	-0.03(-0.23,0.20)	0.00(-0.24,0.27)	0.03(-0.27,0.34)	0.05(-0.30,0.39)
PCS	Urea	0.08(-0.13,0.28)	0.11(-0.07,0.28)	0.13(-0.06,0.32)	0.15(-0.06,0.35)	0.16(-0.08,0.40)	0.17(-0.12,0.44)	0.18(-0.16,0.48)
PCS	For	0.08(-0.14,0.29)	0.04(-0.15,0.23)	-0.01(-0.20,0.18)	-0.05(-0.28,0.17)	-0.09(-0.38,0.19)	-0.12(-0.46,0.21)	-0.15(-0.54,0.23)
PAG	Hip	0.08(-0.14,0.27)	0.09(-0.09,0.27)	0.12(-0.06,0.31)	0.14(-0.07,0.36)	0.16(-0.10,0.42)	0.18(-0.14,0.47)	0.20(-0.18,0.53)
PAG	NMNA	-0.30(-0.46,-0.10)	-0.24(-0.40,-0.06)	-0.19(-0.36,0.02)	-0.14(-0.34,0.11)	-0.09(-0.35,0.20)	-0.05(-0.35,0.27)	-0.02(-0.37,0.33)
PAG	TMAO	0.09(-0.10,0.30)	0.12(-0.05,0.31)	0.13(-0.05,0.33)	0.14(-0.08,0.38)	0.15(-0.12,0.42)	0.15(-0.17,0.47)	0.15(-0.21,0.51)
PAG	Gly	-0.12(-0.31,0.08)	-0.09(-0.26,0.10)	-0.07(-0.24,0.12)	-0.05(-0.26,0.18)	-0.04(-0.30,0.23)	-0.02(-0.36,0.30)	-0.01(-0.42,0.36)
PAG	Urea	0.07(-0.13,0.28)	0.08(-0.10,0.27)	0.08(-0.10,0.27)	0.08(-0.12,0.29)	0.08(-0.16,0.32)	0.08(-0.23,0.37)	0.07(-0.29,0.42)
PAG	For	0.05(-0.15,0.27)	0.00(-0.17,0.19)	-0.04(-0.23,0.14)	-0.08(-0.31,0.13)	-0.12(-0.39,0.14)	-0.15(-0.45,0.16)	-0.17(-0.51,0.19)
Hip	NMNA	0.18(-0.03,0.36)	0.17(-0.04,0.36)	0.19(-0.06,0.40)	0.23(-0.03,0.45)	0.27(0.01,0.52)	0.32(0.01,0.59)	0.36(0.01,0.65)
Hip	TMAO	-0.15(-0.33,0.05)	-0.17(-0.33,0.00)	-0.19(-0.36,0.00)	-0.20(-0.40,0.01)	-0.22(-0.45,0.03)	-0.23(-0.50,0.05)	-0.24(-0.54,0.07)
Hip	Gly	-0.25(-0.42,-0.06)	-0.24(-0.42,-0.05)	-0.23(-0.42,0.00)	-0.21(-0.41,0.04)	-0.19(-0.42,0.12)	-0.17(-0.46,0.17)	-0.15(-0.49,0.21)
Hip	Urea	-0.25(-0.43,-0.05)	-0.23(-0.40,-0.04)	-0.22(-0.40,-0.02)	-0.22(-0.42,0.01)	-0.22(-0.45,0.05)	-0.21(-0.50,0.09)	-0.21(-0.55,0.14)
Hip	For	0.29(0.08,0.47)	0.27(0.10,0.43)	0.25(0.08,0.43)	0.24(0.04,0.45)	0.24(-0.02,0.49)	0.23(-0.09,0.53)	0.23(-0.15,0.58)
NMNA	TMAO	-0.02(-0.21,0.19)	-0.06(-0.23,0.13)	-0.10(-0.27,0.11)	-0.13(-0.33,0.11)	-0.16(-0.40,0.11)	-0.19(-0.45,0.14)	-0.21(-0.50,0.16)
NMNA	Gly	0.06(-0.14,0.27)	0.06(-0.13,0.29)	0.03(-0.16,0.28)	-0.01(-0.23,0.25)	-0.05(-0.31,0.21)	-0.09(-0.40,0.24)	-0.13(-0.48,0.25)
NMNA	Urea	-0.16(-0.34,0.04)	-0.19(-0.36,-0.01)	-0.22(-0.39,-0.03)	-0.24(-0.43,-0.02)	-0.25(-0.48,0.00)	-0.26(-0.53,0.03)	-0.27(-0.57,0.05)
NMNA	For	0.09(-0.12,0.30)	0.14(-0.06,0.32)	0.19(-0.01,0.38)	0.24(0.02,0.45)	0.29(0.01,0.51)	0.32(0.00,0.58)	0.36(0.00,0.63)
TMAO	Gly	0.03(-0.18,0.24)	0.13(-0.07,0.35)	0.19(-0.04,0.42)	0.21(-0.05,0.46)	0.20(-0.08,0.45)	0.18(-0.17,0.47)	0.15(-0.27,0.50)
TMAO	Urea	-0.07(-0.27,0.14)	-0.09(-0.26,0.09)	-0.11(-0.29,0.08)	-0.13(-0.34,0.10)	-0.14(-0.38,0.12)	-0.15(-0.44,0.15)	-0.16(-0.48,0.18)
TMAO	For	0.11(-0.10,0.31)	0.04(-0.14,0.22)	-0.01(-0.21,0.17)	-0.06(-0.30,0.17)	-0.11(-0.38,0.19)	-0.14(-0.44,0.19)	-0.17(-0.49,0.21)
Gly	Urea	-0.17(-0.35,0.03)	-0.22(-0.40,-0.03)	-0.25(-0.44,-0.05)	-0.26(-0.47,-0.04)	-0.26(-0.48,-0.01)	-0.26(-0.52,0.05)	-0.25(-0.55,0.11)
Gly	For	0.11(-0.10,0.32)	0.08(-0.12,0.27)	0.06(-0.13,0.26)	0.05(-0.16,0.29)	0.05(-0.22,0.33)	0.04(-0.29,0.39)	0.04(-0.34,0.45)
Urea	For	-0.16(-0.34,0.06)	-0.16(-0.33,0.04)	-0.17(-0.35,0.01)	-0.19(-0.40,0.03)	-0.21(-0.47,0.05)	-0.23(-0.54,0.09)	-0.25(-0.60,0.12)

The pair-wise correlations calculated from the covariance regression fits for each pair of urinary metabolites along with the 95% credible interval for seven different concentrations of urinary cadmium ($^{1-7}$U-Cd (nmol/mmol creatinine)= 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8)

Appendix F

Permission for third party copyright

Page Number	Type of work	Souce work	Copyright holder& year	Work out of copyright	Permission to re-use
Page 29	Figure 1.4	Neuropsychopharmacology (2008), 34(1)173-186	Macmillan Publishers Ltd copyright (2009)	yes	yes
Page 41	Figure 2.1	Biochemical and Biophysical Research Communications, 375/3	Elsevier (2008)	yes	yes
Page 45	Figure 2.2	Wikimedia Commons, free media repository	Free Software Foundation	yes	
Page 48	Figure 2.3	Nature Reviews Genetics (2009), 10, 109121	Nature Publishing Group (2009)	yes	yes
Page 64	Figure 3.2	PLoS genetics (2008) 4(7), e1000130	open-access article	yes	yes
Page 77-100	Research Article	PLoS ONE 6(9), e24702	open-access article	yes	yes

This is a License Agreement between Beatriz Valcarcel ("You") and Nature Publishing Group ("Nature Publishing Group") The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

License number	Reference confirmation email for license number
License date	Sep 10, 2013
Licensed content publisher	Nature Publishing Group
Licensed content publication	Neuropsychopharmacology
Licensed content title	Metabolomics: A Global Biochemical Approach to the Study of Central Nervous System Diseases
Licensed content author	Rima Kaddurah-Daouk and K Ranga Rama Krishnan
Licensed content date	Oct 8, 2008
Type of Use	Thesis / Dissertation
Volume number	34
Issue number	1
Your reference number	2805
Title of your thesis / dissertation	Statistical association networks as complex phenotypes: new methods and applications
Expected completion date	
Estimated size (number of pages)	204
Billing Type	Invoice
Billing address	5 tyssen street
	London, E8 2LY
	United Kingdom
Total	0.00 USD

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 10, 2013

This is a License Agreement between Beatriz Valcarcel ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Beatriz Valcarcel
Customer address	5 tyssen street London, E8 2LY
License number	3225370757033
License date	Sep 10, 2013
Licensed content publisher	Elsevier
Licensed content publication	Biochemical and Biophysical Research Communications
Licensed content title	A multi-metabolite analysis of serum byH NMR spectroscopy: Early systemic signs of Alzheimer's disease
Licensed content author	Taru Tukiainen,Tuulia Tynkkynen,Ville-Petteri Mäkinen,Pasi Jylänki,Antti Kangas,Johanna Hokkanen,Aki Vehtari,Olli Gröhn,Merja Hallikainen,Hilkka Soininen,Miia Kivipelto,Per-Henrik Groop,Kimmo Kaski,Reino Laatikainen,Pasi Soininen,Tuula Pirttilä, et al.
Licensed content date	24 October 2008
Licensed content volume number	375
Licensed content issue number	3
Number of pages	6
Start Page	356
End Page	361
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Sep 10, 2013

This is a License Agreement between Beatriz Valcarcel ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3225371191879
License date	Sep 10, 2013
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Plasma lipoproteins: genetic influences and clinical implications
Licensed content author	Robert A. Hegele
Licensed content date	Feb 1, 2009
Volume number	10
Issue number	2
Type of Use	reuse in a thesis/dissertation
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
High-res required	no
Figures	Figure 1 . Overview of lipoprotein metabolism.
Author of this NPG article	no
Your reference number	2805
Title of your thesis / dissertation	Statistical association networks as complex phenotypes: new methods and applications
Expected completion date	Oct 2013
Estimated size (number of pages)	204
Total	0.00 GBP
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to