Abstract of thesis entitled

"Connecting the dots: Integrative analysis of genomic, metabolomic and phenotypic data from a population cohort"

Submitted by

Yiming Li

for the degree of Doctor of Philosophy
at The University of Hong Kong
in January 2019

Lower back pain (LBP) is one of the most prevalent global health issues and a main cause of disability. Lumbar disc degeneration (LDD) is one of the major reasons for LBP, which could be evaluated by radiographic observations through magnetic resonance imaging (MRI). Nevertheless, these MRI observations, as diagnostics for LBP/LDD, are prone to human error and may be insufficient in detecting real-time variations of complex biological systems. This thesis aims to identify novel LDD biomarkers related to altered metabolism, which could aid personalized diagnosis and treatment of LBP.

This study is based on a population cohort of $3,584$ southern Chinese. Over $1,000$ individuals were followed with MRI scans, which were read by experienced physicians specialized in LDD. Statistical analyses showed that the severity of LDD is significantly greater in the lower lumbar region, the lower disc levels forming a cluster more related to age. Accordingly, a systematic way to quantify the degree of LDD from raw MRI reads is proposed in this thesis. Apart from phenotyping, $2,482$ samples in the cohort were genotyped, and the serum samples of 757 individuals were acquired for proton nuclear magnetic resonance spectroscopy, resulting in 130 metabolomic measurements over three molecular windows.

In order to discover genetic variants associated with different metabolomic measurements, genome-wide association studies (GWAS) were conducted on 571 individuals for each of the 130 metabolomic measurements. In total, 123 unique single nucleotide polymorphisms were found to be significantly associated with one or more metabolomic measurements; among them, intergenic variants were underrepresented, whereas exonic, intronic and UTR3 variants were enriched. My results suggest significant associations between 42 different metabolomic traits and a number of genetic loci. Polyunsaturated fatty acids were found to be significantly

associated with the FADS1/FADS2 loci, and CTTNBP2 was identified as a potential risk locus for a cluster of lipid / fatty acid related metabolites.

The human metabolome was next estimated based on the summary statistics from previous GWAS and genomic data via meta-analysis and polygenic scoring. The associations between (estimated) metabolomic data and various phenotypes (anthropometric, behavioral, clinical, and LDD-related) were tested using different regression methods, ranging from simple linear models to Lasso. Potential metabolomic biomarkers for LDD were identified, including blood lipid levels, the mean diameter for very low density lipoprotein particles, sphingomyelins and tyrosine.

Through GWAS, polygenic scoring and association analysis, this study pinpoints metabolomic biomarkers for LDD with a purely data-driven approach. It also proposes a new way to analyze genomic, metabolomic and phenotypic data in an integrative manner, utilizing metabolome prediction models. This process of the integration of big omics data could help us discover known and novel metabolomic biomarkers associated with complex traits and gain a better understanding of the mechanisms of these associations.

(452 words)

Connecting the dots:

Integrative analysis of genomic, metabolomic and phenotypic data from a population cohort

by

Yiming Li

B. Sc. *H.K.*

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at The University of Hong Kong.

January 2019

Dedicated to my beloved parents, Wenling Sun and Yan Li.

Declaration

I declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed _____

Yiming Li

Acknowledgements

First of all, I would like to extend my deepest gratitude to my primary supervisor Prof. Pak Sham. Back when I was a freshman in statistics and practically knew nothing about statistical genetics, Prof. Sham generously accepted me as a summer student, introducing me to academic research and in particular, this fascinating field. Without him, my life would be totally different, and I would be always grateful to him for his continuous guidance and invaluable advice over all these years. My sincere appreciation also goes to my co-supervisors Dr. Miaoxin Li and Dr. Stacey Cherny for their kind support and all the enlightening discussions. Many a time their knowledgeable comments have inspired me with novel ideas and helped me avoid potential pitfalls.

Furthermore, I would like to thank all the volunteers in the Hong Kong disc degeneration cohort and my colleagues from the degenerative disc disease group in the School of Biomedical Sciences. Without their dedication, my study would not have been possible. Especially, I would like to thank Dr. Jaro Karppinen and Dr. Dino Samartzis for performing phenotypic assessments. Also, the metabolomic data is only available thanks to the efforts of Dr. Dino Samartzis – thank you so much for making this happen. As experts in human degenerative skeletal disorders, Prof. Kathryn Cheah and Prof. Danny Chan have given me many insightful suggestions – as a student from a computational background, I really needed hearing a biologist's perspective and deeply appreciate their advice and help.

I wish to gratefully acknowledge Prof. Francis Chin and Prof. Guosheng Yin, my former supervisors (on top of that, two of my favorite professors) during my undergraduate career, for broadening my horizons and stimulating my interest in computer science and statistics. Whenever I became lost and desolated, their kind encouragements have always boosted my confidence and determination in research.

I would like to express my heartfelt gratitude to The University of Hong Kong Foundation for Educational Development and Research ("HKU Foundation") for its generous support for my research studies at the University of Hong Kong.

My appreciation also goes to Johnny Kwan, Desmond Campbell, Clara Tang, Xiaowei Zhang, Yan Li, Peikai Chen, Xueya Zhou, Timothy Mak, Sam Choi, Jacob Hsu, Robert Porsch, Yiming Qin, Jill Ding, the IT team at the Centre for Genomic Sciences, all my other lovely

friends and colleagues in Sham lab, as well as my old friends Brian He and Daiwei He – I feel lucky to have met you all. Thank you so much for all the moments we have shared and your friendship.

Finally and most importantly, I would like to thank my mom and dad for everything. Words cannot describe my gratitude and to you, I dedicate this thesis – I love you guys.

Table of contents

List of figures

List of tables

Nomenclature

Symbols

^{1}H	Proton (hydrogen-1)
α	Significance threshold for hypothesis testing, or the acceptable probability for type I error (rejecting the null hypothesis when it is true)
δ	The chemical shift
F	Inbreeding coefficient
H_0	The null hypothesis
μ	The mean of a random variable
m/z	Mass-to-charge ratio
n	Number of observations, or the sample size
p	Number of features (covariates, predictors, etc.)
$\hat{\pi}$	The kinship coefficient
R^2	Pairwise correlation (e.g. SNP-SNP correlation)
r_{pb}	Point biserial correlation
σ^2	The variance of a random variable
τ	Kendall's tau (rank distance measure)
u	The unified atomic mass unit (or dalton)

Abbreviations and acronyms

AF	Annulus fibrosus
AI	Artificial intelligence
AIC	Akaike information criterion
ANOVA	Analysis of variance
AUC	Area under the curve
B-H	Benjamini-Hochberg (procedure for multiple testing control)
BLUP	Best linear unbiased predictor

BMI	Body mass index
BMU	Best matching unit (in self-organizing map analysis)
bp	Base pair
CDCV	Common disease / common variant (hypothesis)
CDRV	Common disease / rare variant (hypothesis)
CE	Capillary electrophoresis
CF	Cystic fibrosis
CLC	Chondrocyte-like cell
CNV	Copy number variation
CT	Computed tomography
DALY	Disability-adjusted life year
DB	Disc bulging
dbSNP	The single nucleotide polymorphism database
DHA	Docosahexaenoic acid
DNA	Deoxyribonucleic acid
EBAM	Empirical Bayesian analysis of microarrays (metabolites)
EC	Esterified cholesterol
EP	Vertebral endplates
eQTL	Expression quantitative trait locus
FADS	Fatty acid desaturase (gene family)
FA	Fatty acid
FC	Free cholesterol
FDR	False discovery rate
GC	Gas chromatography
GIGO	Garbage in, garbage out
GPU	Graphics processing unit
GRC	The Genome Reference Consortium
GRIP	Genetic risk prediction
GRN	Gene regulatory network
GWAS	Genome-wide association studies
HDL	High density lipoprotein
hg19	Human reference genome (UCSC) version 19, i.e. the GRCh37 reference assembly
HGP	Human genome project
HIZ	High intensity zone

HPLC	High-performance liquid chromatography
HWE	Hardy-Weinberg equilibrium
IBD	Identical/Identity by descent
IBS	Identical/Identity by state
IDD	Intervertebral disc degeneration
IDL	Intermediate density lipoprotein
Indel	Insertion or deletion (in the genome)
IQR	Interquantile range
IVD	Intervertebral disc
LA	Linoleic acid
LBP	Lower back pain
LDD	Lumbar disc degeneration
LD	Linkage disequilibrium
LDL	Low density lipoprotein
LIPID	Lipid extracts (molecular window in ^{1}H NMR)
LIPO	Lipoprotein lipids (molecular window in ^{1}H NMR)
LMWM	Low molecular weight metabolites (molecular window in ^{1}H NMR)
MAF	Minor allele frequency
MC	Modic change
MDS	Multidimensional scaling
MLE	Maximum likelihood estimator
ML	Machine learning
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
NC	Notochordal cell
ncRNA	Non-coding ribonucleic acid
NMR	Nuclear magnetic resonance
NN	Neural network
NP	Nucleus pulposus
ODI	Oswestry disability index
PCA	Principal component analysis
PC	Phosphatidylcholine
PDF	Probability density function
PGLY	Phosphoglycerides

pH	Potential of hydrogen
PLS-DA	Partial least squares discriminant analysis
ppm	Parts per million
PRS	Polygenic risk score
PUFA	Polyunsaturated fatty acid
QC	Quality control
QQ plot	Quantile-quantile plot
RF	Random forest
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
SAM	Significance analysis of microarrays (metabolites)
Sat FA	Saturated fatty acid
SD	Standard deviation
SIL	Signal intensity loss
SM	Sphingomyelin
SNP	Single nucleotide polymorphism
SNR	Signal-to-noise ratio
SN	Schmorl's node
SNV	Single nucleotide variant
SOM	Self-organizing map
SS	Schneiderman's score
SVM	Support vector machine
TAG	Triacylglycerol
TC	Total serum cholesterol
TG	Total serum triglycerides
TLR	Toll-like receptor
truncnorm	Truncated normal (distribution)
UTR	Untranslated region
VAS	Visual analog scale
VLDL	Very low density lipoprotein

1

Introduction

1.1 Statistical analysis of big omics data

1.1.1 Genomic data

It is, indeed, a "small"[1] world, if we take into consideration how many people are living on Earth. At the time of writing, we share our planet with over 7.6 billion human beings [Worldometers, 2008] – every one of us being so similar, yet so different. Like all the other living creatures, we grow, adapt, respond to whatever today has in store for us and reproduce. But at the same time, just as no two leaves are exactly alike, all human beings are different from one another. How do the trillions of cells in our body know how to function properly? What makes each of us unique? The partial[2] answer lies in our genome.

Let us start by examining the building blocks of the human genome. Deoxyribonucleic acid (DNA) is a chain of nucleotides containing genetic instructions for the activities of almost all living organisms; it consists of two paired strands, coiled around one another to form a double helix. Each nucleotide (one of the "rungs" of the double helix) is composed of one of four nitrogen-containing nucleobases – adenine (A), thymine (T), cytosine (C) and guanine (G); we call a unit of two nucleobases bound to each other (A-T or G-C) a base pair (bp). As

[1] Approximately $63,819$ billion square meters of land is habitable on Earth [Pianka, 2007], giving us less than $8,400$ square meters per person. This is actually extremely sparse since we need public space as well.

[2] The other part of the answer is blowin' in the wind (pardon me for being cheesy). Environmental factors are also important.

illustrated in Figure 1.1, DNA is stored in thread-like structures called chromosomes in the nucleus of each cell.

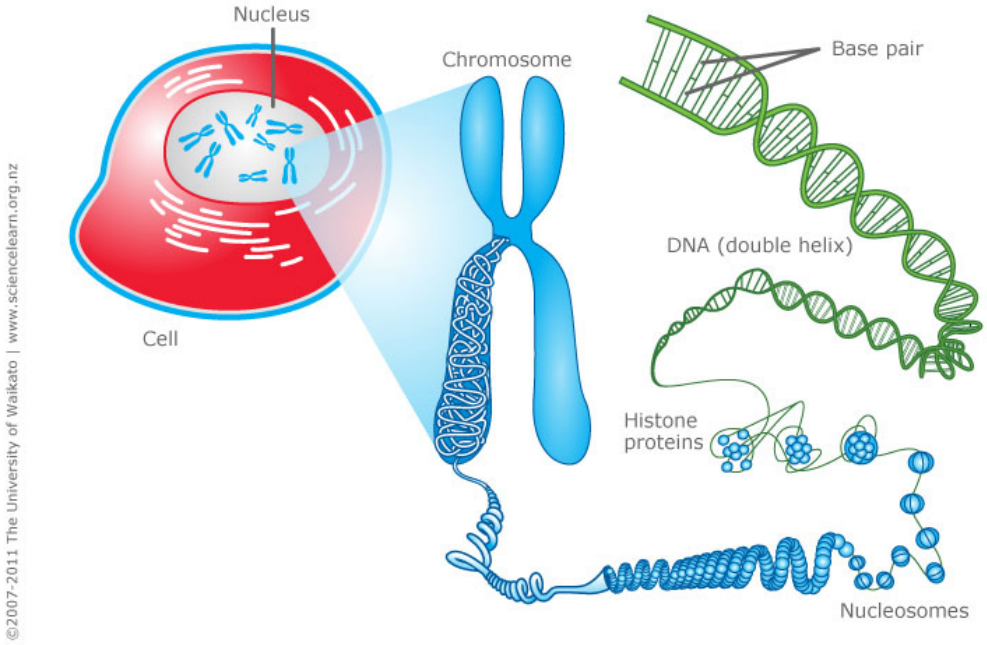


Fig. 1.1 Cell, chromosome and DNA [Science Learn Hub, 2011].

Based on DNA strands, ribonucleic acid (RNA) strands are created via transcription, which are next translated to specify the sequence of amino acids within proteins (c.f. Figure 1.2). This process is summarized in the central dogma, which was first proposed by [Crick, 1970].

A gene is a sequence of DNA that encodes function. The human genome project (HGP) has estimated that human beings have between $20,000$ and $25,000$ genes, whose length ranges from several hundred to over 2 million bases [Venter, M. D. Adams, et al., 2001]. Genes could include both coding regions (exons) and non-coding segments (introns) between exons. Certain non-coding sequences also have biological functions, for instance, regulate neighboring coding regions [Carey, 2015].

The human genome is simply a human's complete set of DNA – it has approximately three billion base pairs of DNA arranged into 46 chromosomes [National Human Genome Research Institute, 2007]. The term "genomics" was originally coined by Dr. Roderick[3]. Unlike genetics, which studies inheritance mainly in terms of single genes, genomics aims at characterizing and quantifying genes collectively [Klug, Cummings, et al., 2003].

[3] In 1986, over plenty of beer. The name was intended for a new scientific journal [Kuska, 1998].

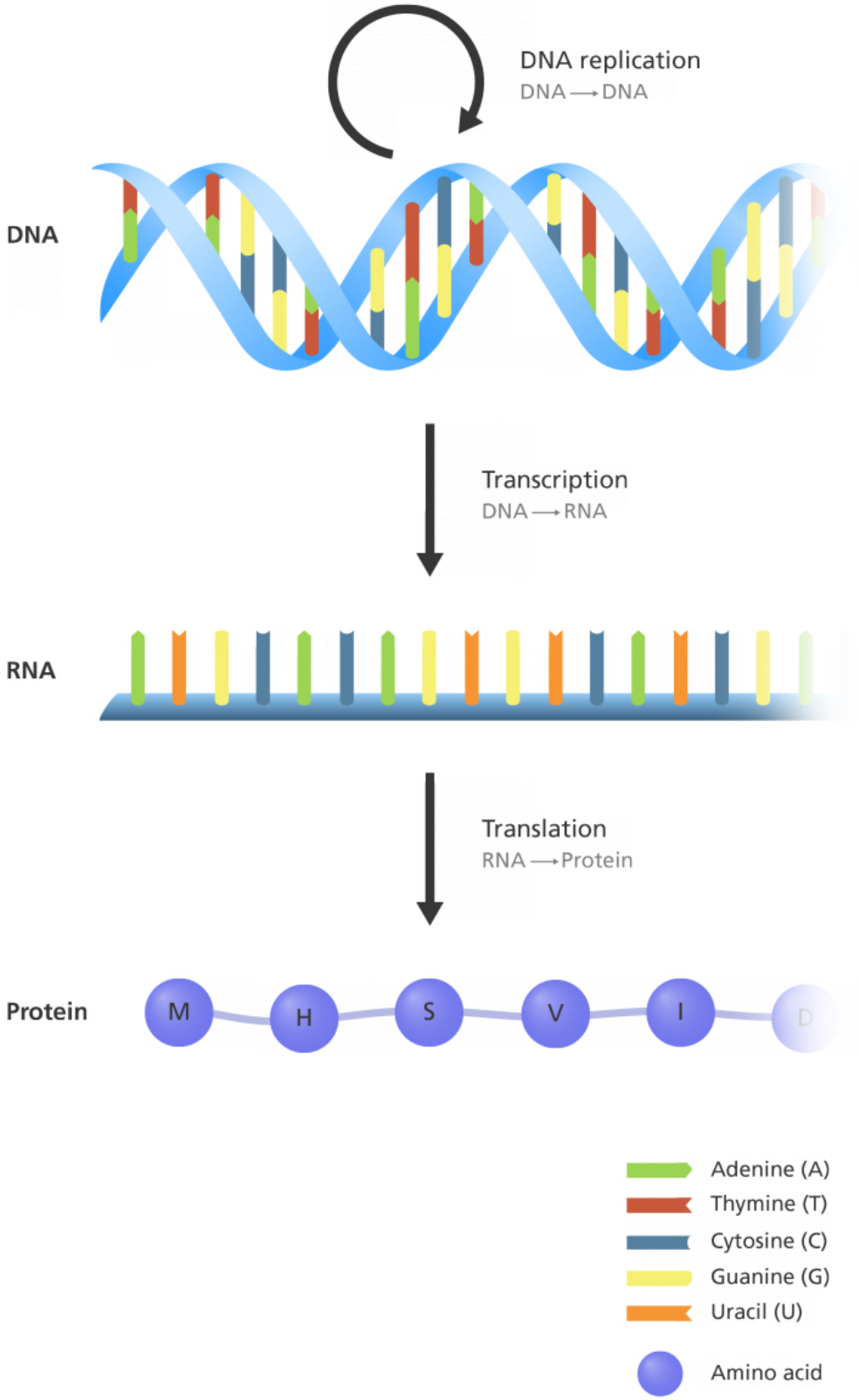


Fig. 1.2 The central dogma [Genome Research Limited, 2016].

1.1.1.1 Genetic variation

DNA sequences are prone to variation. Essentially, this simple fact is why you, my reader, are different from me, and also why we now stand as human beings instead of little globs. In a population, there could be different DNA sequences at the same locus (a specific position on a chromosome). This is known as genetic polymorphism, and the alternative sequences at a locus are called alleles. Genetic variation could be categorized into three classes.

Single nucleotide polymorphisms (SNPs) and single-nucleotide variants (SNVs) refer to variation in a single nucleotide at a locus. This substitution could be a transition (interchange of purine or pyrimidine nucleic acids[4]) or a transversion (interchange of a purine and pyrimidine nucleic acid[4]). If the variation is common within a population (> 1%), it is called a SNP; otherwise, it is a SNV. SNPs could fall in coding/non-coding regions of genes or intergenic regions. Those in coding regions may or may not influence the protein sequence – synonymous SNPs do not affect the sequence of amino acids in proteins, whereas nonsynonymous SNPs do change it. SNPs not in coding regions could still affect transcription factor binding, messenger RNA (mRNA) degradation, gene splicing, or the sequence of non-coding RNA. They may also alter gene expression, denoted as expression quantitative trait loci (eQTLs). [Kruglyak and Nickerson, 2001] estimated that there are approximately 10 million common SNPs, which could account for 90% of the variation in the world's human population. Hence, the study of SNPs is crucial in understanding the underlying genetics of a wide range of human diseases. The HapMap project aims to describe the common patterns of these variations, also taking into consideration linkage disequilibrium (LD; the non-random association of alleles at two or more loci) [International HapMap Consortium, 2003].

The second type of genetic variation are indels[5], which refer to insertion or deletion of short DNA sequences. They could range from 1 to $10,000$ bps in length [Mills et al., 2006]. 16% to 25% of all sequence polymorphisms in humans exist in the form of indels – this frequency is markedly lower than that of SNPs [Mills et al., 2006]. Still, many diseases are associated with indels, examples including cystic fibrosis [F. S. Collins et al., 1987] and fragile X syndrome [Warren et al., 1987].

[4] Adenine (A) and guanine (G) are purine nucleic acids, whereas cytosine (C) and thymine (T) are pyrimidine nucleic acids.

[5] The functional consequences described for SNPs/SNVs also apply to indels and to a certain extend, to structural variations.

Finally, genetic variations occurring over a larger DNA sequence are called structural variations[6]. These include copy number variations (CNVs) and chromosomal rearrangement events (deletion, insertion, inversion, duplication, etc.). Structural variations in the human genome can affect gene dosage (the number of copies of a certain gene present in a genome) and therefore, diseases and other phenotypic variations [Feuk et al., 2006].

1.1.1.2 Genome-wide association studies

How do genetic variations affect human phenotypes? Genetic association studies seek to answer this question by testing if the allele of a genetic variant is found more frequently than expected in individuals with the phenotype of interest [Risch and Merikangas, 1996].

With the emergence of big biobanks (human genetic repositories) and the development of methods for genotyping (e.g. DNA microarrays), genome-wide association studies (GWAS) were introduced. GWAS could scan the entire genome for SNPs and other variants significantly associated with traits of interest. This approach is completely data-driven and could detect associations, but not causality [Pearson and T. A. Manolio, 2008].

The common disease / common variant (CDCV) hypothesis[7] states that common disorders are likely affected by genetic variation that is also common in the population [Reich and Lander, 2001]. This hypothesis is quite intuitive. Assuming that common genetic variants could influence diseases, the effect size of any particular variant must be small relative to that identified for rare disorders [Bush and Moore, 2012]. If common alleles have small effects yet still common disorders show heritability (proportion of variation of a trait in a population due to genetic variation), then multiple common alleles must influence disease susceptibility at the same time [Bush and Moore, 2012]. GWAS could typically identify these common variants with small effect sizes (lower right corner of Figure 1.3), making them ideal for detecting genetic effects on human traits.

The statistical power of a GWAS is determined by the effect size of the susceptibility loci and the study's sample size. The success of a GWAS also depends on a thoughtful study design,

[6] Please see footnote 5.

[7] A contender to the CDCV hypothesis is the common disease / rare variant (CDRV) hypothesis, which argues that the genetic susceptibility to common diseases is mainly contributed by rare variants with relatively high penetrance [N. J. Schork et al., 2009]. Several common diseases (e.g. breast cancer) do have rare, Mendelian forms (upper left corner of Figure 1.3), but their many other forms follow the CDCV hypothesis and hence still could be studied using GWAS [N. J. Schork et al., 2009]. Indeed, in practice, both CDCV and CDRV hypotheses hold, each to a different extent. The question regarding the "correctness" of each hypothesis is an empirical one [N. J. Schork et al., 2009; G. Gibson, 2012].

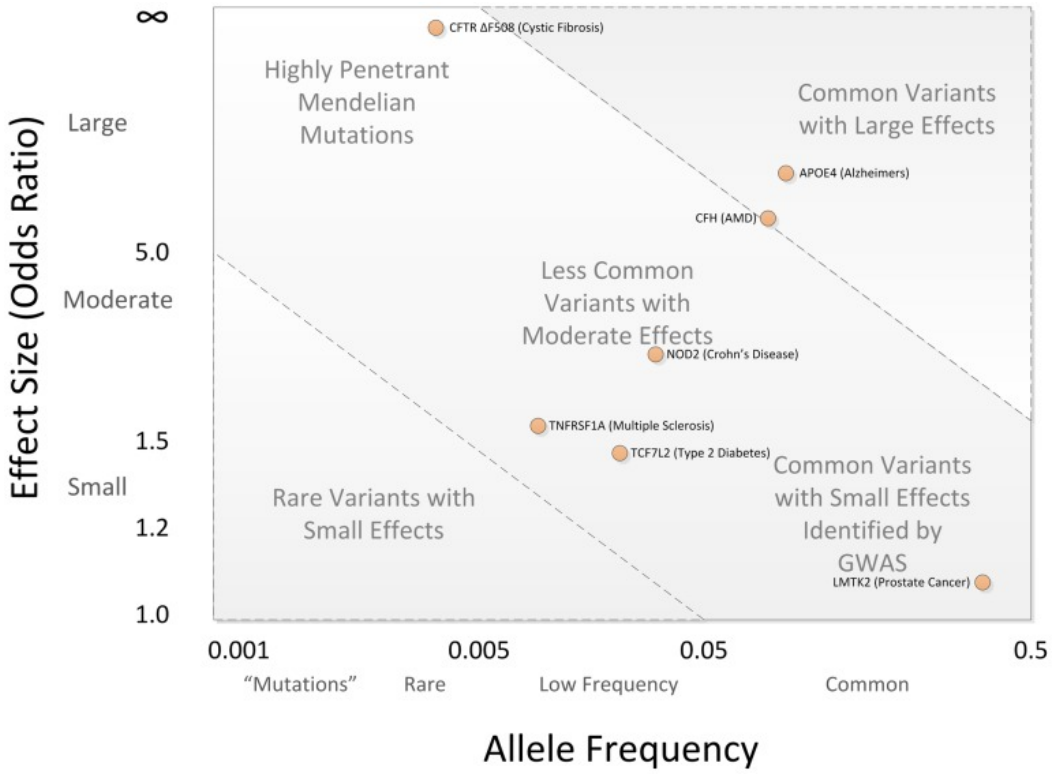


Fig. 1.3 Spectrum of disease allele effects [Bush and Moore, 2012].

careful quality control and genotype imputation[8], as well as proper control for population stratification and other confounding factors.

Since we are performing a lot of regressions at the same time, the probability of false positives is now quite large. This is called the multiple testing problem. To account for multiple testing, as a rule of thumb, a p-value of 5×10^{-8} is the threshold for genome-wide significance [Barsh et al., 2012]. This number is equivalent to a threshold of $\alpha = 0.05$ Bonferroni-corrected[9] for 1 million independent variants (approximately the number of independent SNPs estimated using the HapMap Phase II data set [Consortium et al., 2007]) [Kanai et al., 2016].

1.1.1.3 Genetic risk prediction

Complex disorders such as schizophrenia, diabetes, and multiple sclerosis are fundamentally determined by genetic and environmental factors [N. J. Schork, 1997]. Due to the seriousness

[8]To impute is to replace missing data with substituted values.

[9]The Bonferroni correction sets the significance cut-off at $\frac{\alpha}{n}$, where α is the significance level and n is the number of tests.

and prevalence of these diseases, there has been plenty of research effort in developing methods for predicting their risk. Genetic risk prediction (GRIP) could open the door for personalized treatment for complex diseases, which is crucial in both medical research and providing high-quality, affordable health care.

Polygenic risk scores

Numerous complex disease susceptibility loci have been successfully identified by GWAS in recent years [McCarthy and Hirschhorn, 2008]. One of the first approaches to utilize these GWAS results in GRIP is constructing polygenic risk scores (PRS) by calculating a weighted sum of the known susceptibility loci [Speliotes, Willer, et al., 2010].

One issue with only using known susceptibility loci is that, even though the variants at these loci affect the risk for the corresponding disorders, they could only explain a small faction of the genetic risk variance in the general population [McCarthy and Hirschhorn, 2008]. Because of this drawback, the PRS approach fails to perform satisfactorily when applied to certain diseases and conditions. For instance, it merely produced an area under the curve (AUC) of 0.515 with the atherosclerosis risk in communities dataset [Speliotes, Willer, et al., 2010], indicating that the prediction is only slightly better than chance[10].

Therefore, rather than solely estimating the disease risk based on the known susceptibility loci, we need to select a wider range of risk alleles reaching genome-wide significance for constructing PRS [Z. Wei et al., 2009]. More generally speaking, designing a precise and efficient GRIP method is two-fold. We need to first select which genetic variants to include in the predictive model, and next develop a metric for GRIP integrating the selected markers (e.g. PRS) [Wimmer et al., 2013; Schrodi et al., 2014].

Nowadays, instead of taking a weighted sum of only the known susceptibility loci, PRS sums a number of trait-associated alleles across many genetic loci, typically weighted by effect sizes estimated from a GWAS [Dudbridge, 2013].

$$PRS = \sum_i (x_i \log(\beta_i)) \tag{1.1}$$

[10] When the prediction for a binary classification problem is made by tossing a fair coin, we expect an AUC of 0.5.

Equation 1.1 is a general formula of PRS. Here x_i is the number of risk alleles (0, 1, 2) at SNP i, whereas $\log(\beta_i)$ are the log odds ratios from per-SNP logistic regressions.

There are various SNP pre-selection methods for constructing PRS. SNP thresholding selects a subset of SNPs that are more associated with the trait of interest [Euesden et al., 2014]. Thresholding is usually followed by LD-based methods like pruning and clumping, which try to choose a subset of SNPs that are not highly correlated [Euesden et al., 2014].

Besides being utilized for GRIP, PRS can be very powerful in detecting genetic effects when no single SNPs are genome-wide significant, as well as establishing a common genetic basis for related diseases [Dudbridge, 2013].

Regression methods

Regression methods are also intuitive ways to perform GRIP for complex diseases. A general logistic regression model has been proposed by [Q. Yang et al., 2003] to estimate both the risk and the standard error of the risk estimates. Furthermore, its predictive power could be improved by utilizing best linear unbiased predictors (BLUPs) to account for random effects [Campos et al., 2013]. At present, regression methods are quite widely adopted in the field of GRIP for complex diseases ranging from cerebrovascular disease [Tsai et al., 2013] to prostate cancer [Mondul et al., 2013].

Nevertheless, these seemingly well-established techniques still suffer from various issues. Recall that there exists LD between different genetic loci. One potentially serious problem is the high correlation between pairs of genetic markers that are close to one another. This multicollinearity violates the basic assumptions of regression analysis, which could make the regression results suspicious at best [Schrodi et al., 2014]. Moreover, by nature, regression analysis is more focused on estimating the effect of various parameters (the genetic markers) on the dependent variable (the disease status) than performing classification (whether the individual suffers from the disease or not) itself. Some have argued that in light of this fact, regression analysis is not very pertinent in a clinical setting [Z. Wei et al., 2009].

Machine learning approaches

To tackle the previously mentioned problems, machine learning (ML) approaches have been introduced into the field of GRIP. Unlike the traditional approaches (e.g. PRS) obeying

explicit, rule-based algorithms, ML methods are intelligent in that they could directly learn from data and build a purely data-driven model for risk prediction [Dietterich, 1997].

As discussed previously, potential interactions between genetic variants may incur problems in regression analysis. When we learn from genomic data, though, these interactions could actually become advantageous instead. During the learning process, ML methods could make inferences about the interactions between variant pairs and allocate weights to the variants in the prediction model accordingly. In this way, the optimal binary classification power could be achieved [Z. Wei et al., 2009]. For example, support vector machines (SVMs) are supervised ML algorithms that attempt to recognize patterns in the data and find a maximum-margin hyperplane that separates the data into different classes [Cortes and Vapnik, 1995]. [Z. Wei et al., 2009] has suggested that by taking into consideration interactions between a large group of genetic variants through SVMs, the GRIP performance could be significantly improved. Some other ML methods commonly used in GRIP include multifactor dimensionality reduction [R. L. Collins et al., 2013], naïve Bayes classifier [Malovini et al., 2012], and random forest (RF) [Khalilia et al., 2011].

ML methods are generally more accurate and efficient than traditional methods in terms of GRIP, but could be overly optimistic when over-fitting is present [Okser et al., 2014]. Apart from this, ML methods are often highly dependent on the available sample size and also require fine-tuning hyperparameters to achieve the optimal prediction performance, which could be time-consuming. Therefore, traditional methods like PRS is still more widely used for GRIP when the sample size is limited.

1.1.2 Transcriptomic data

All the RNA molecules in one cell or a population of cells form the transcriptome. First initiated in the early 1990s, transcriptomics (the study of whole transcriptomes) could provide us with insights into the mechanisms of gene expression, the functional roles of different genes and the gene regulatory network (GRN) [Brazma and Vilo, 2000].

1.1.2.1 Gene expression

Gene expression is the process of "interpreting" the information stored in a gene. The field of transcriptomics has hugely transformed with key techniques like microarrays and RNA

sequencing (RNA-seq) [Lowe et al., 2017], which could measure gene expression levels quickly and robustly.

Microarray data analysis

A DNA microarray (DNA chip) is a collection of microscopic DNA spots attached to a solid surface, which can be used to simultaneously measure the expression levels of numerous genes [Schena et al., 1995]. The raw data is first transformed into a gene expression matrix. After data normalization, we could perform clustering and visualize the results through heatmaps and dendrograms (c.f. Figure 1.4). Clustering proves to be useful in detecting gene expression profiles for distinguishing between cases and controls [Brazma and Vilo, 2000].

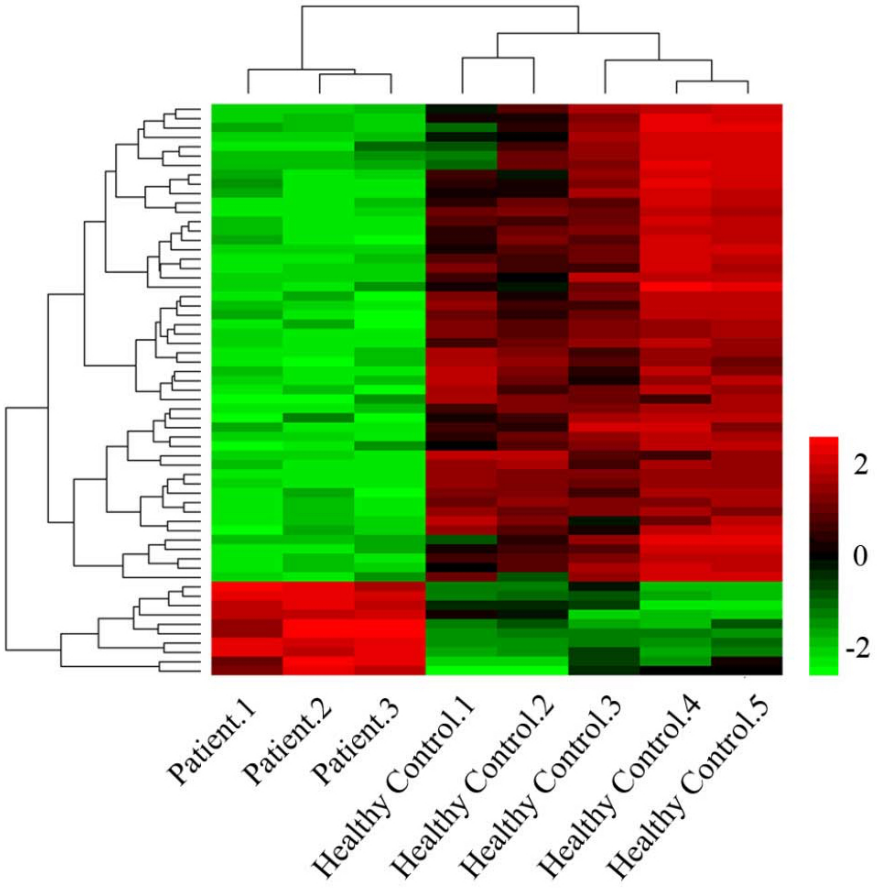


Fig. 1.4 Visualization of the gene expression matrix [Liu et al., 2014]. Up-regulated genes are colored red, and down-regulated genes are colored green. It could be seen that neurofibromatosis type I cases and controls have different gene expression profiles.

Additionally, significance analysis of microarrays (SAM) [Tusher et al., 2008] and empirical Bayesian analysis of microarrays (EBAM) [Efron et al., 2001] are two well-established statistical methods for detecting differentially expressed genes in microarray data. In both procedures, a modified version of the t-statistic is used to pinpoint genes whose expression levels significantly differ between two groups. [Schwender et al., 2003] has shown that of the two, SAM is better with simulated data but performs worse when applied to real data sets.

RNA-seq data analysis

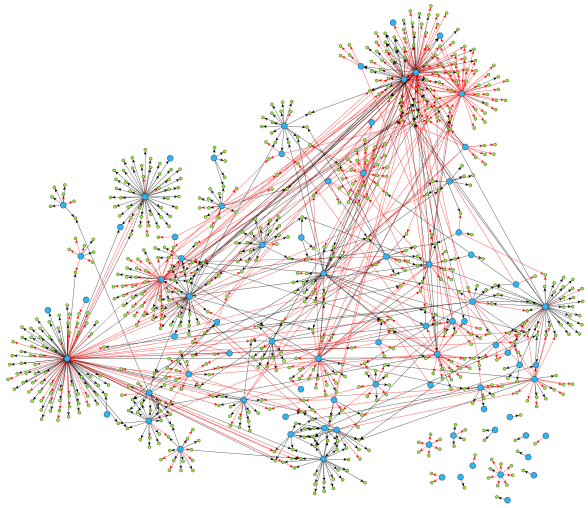
RNA-seq utilizes high-throughput sequencing to record all RNAs in a biological sample at a given moment [Chu and Corey, 2012]. It is used more and more widely for measuring the transcriptome since it is able to quantify a large range of expression levels with absolute values [Kukurba and Montgomery, 2015].

A generic roadmap for RNA-seq data analysis has been proposed by [Conesa et al., 2016]. After quality control, we could perform transcriptome profiling on the RNA-seq data – steps including read alignment, transcript discovery and quantification [Conesa et al., 2016]. The key part of RNA-seq data analysis utilizes differential gene expression methods, which aim to estimate the degree of differential expression between two or more conditions based on read counts from replicated samples [Dündar et al., 2015; Conesa et al., 2016]. Finally, gene set enrichment and pathway analyses could characterize the molecular functions or pathways involving differentially expressed genes [Conesa et al., 2016].

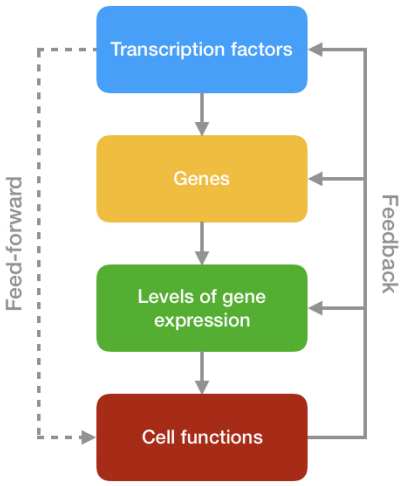
1.1.2.2 Gene regulatory networks

Cells could be considered as containers of various chemicals interacting with each other to control the gene expression levels of mRNAs and proteins. These molecular regulators and their interactions comprise a gene regulatory network (GRN).

A typical GRN is shown in Figure 1.5a. Genes can be viewed as nodes in the network, with inputs being proteins (e.g. transcription factors), and outputs being the levels of gene expression (c.f. Figure 1.5b). A GRN is an abstraction of the cell's molecular dynamics – modeling GRNs could help us predict novel regulations and gain biological insights into the cell's functional organization [Barabasi and Oltvai, 2004].



(a) Example of a GRN [Ma et al., 2014]. Large blue circles denote transcription factors, whereas small green circles denote other genes. Edges represent direct regulatory interactions (inhibiting: red; excitatory: black).



(b) Simplified control process of a GRN. Changed cell behaviors and structures could in return, influence genes, mRNAs and proteins; therefore the feedback circuits.

Fig. 1.5 Illustration of GRNs.

There are numerous mathematical models for GRNs, including coupled ordinary differential equations, Boolean networks [Kauffman, 1969], continuous networks [Vohradsky, 2001] and stochastic gene networks [A. Arkin et al., 1998]. Based on these models, we can try to predict gene expression levels in a GRN, which could aid exploration of how drugs affect a group of genes [Barabasi and Oltvai, 2004].

1.1.2.3 Gene-based association

Recall that GWAS is capable of searching the entire genome for genetic variants significantly associated with complex traits (c.f. Section 1.1.1.2). Nevertheless, it is hard to understand the mechanisms underlying the significant associations from GWAS. Could we conduct a modified version of GWAS, also incorporating transcriptomic data? This is what gene-based association methods like PrediXcan [Gamazon et al., 2015] aim at.

The PrediXcan [Gamazon et al., 2015] / MetaXcan [Barbeira et al., 2016] methods are illustrated in Figure 1.6. To recap, GWAS run a set of regressions $Y = X_l b + \varepsilon$, where X_l is the individual dosage and Y is the phenotype of interest. The regression coefficients b describe SNP-based associations. In contrast, PrediXcan first calculates imputed transcriptomes

(T_g) with X_l based on a database of transcriptome prediction models (PredictDB, available at http://predictdb.org/) and next regresses Y on the predicted expression levels T_g, i.e. $Y = T_g\gamma + \varepsilon$ [Gamazon et al., 2015]. The regression coefficients γ are our gene-based results. MetaXcan could compute the gene-level association results directly using the summary statistics from GWAS, which is useful when we do not have access to the raw genomic data [Barbeira et al., 2016].

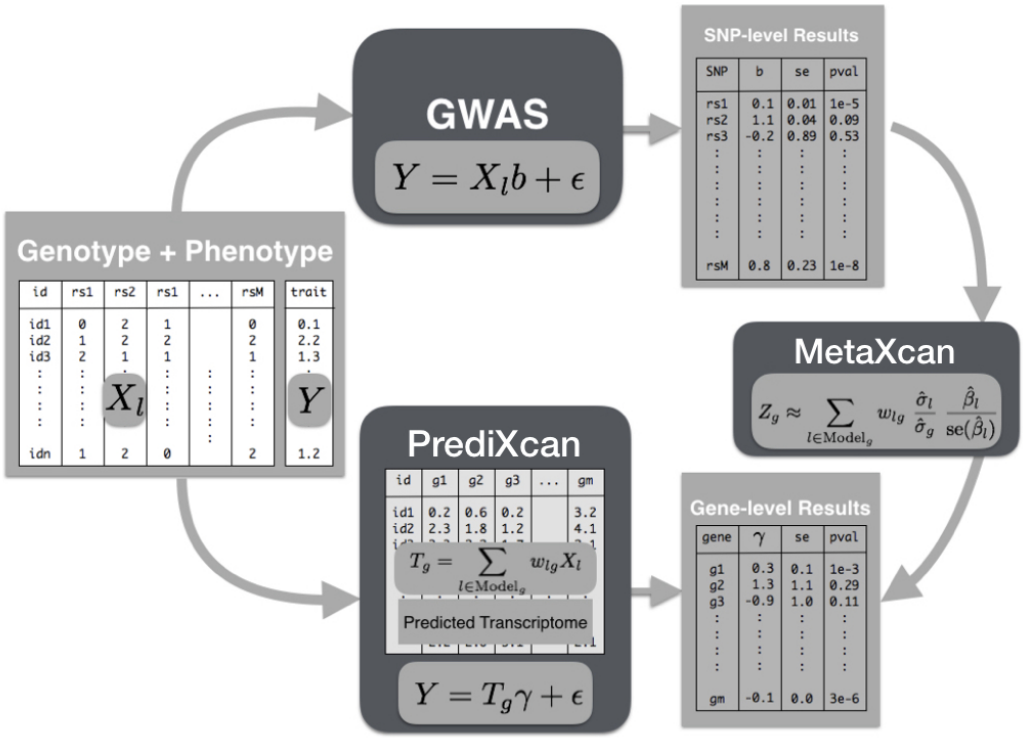


Fig. 1.6 Flow of analysis of GWAS, PrediXcan and MetaXcan [Barbeira et al., 2016].

Gene-based association methods benefit from reduced multiple-testing burden and more interpretable results [Gamazon et al., 2015; Barbeira et al., 2016]. It has been shown that PrediXcan can discover known and novel genes associated with complex traits and help us understand the mechanism of these associations [Gamazon et al., 2015].

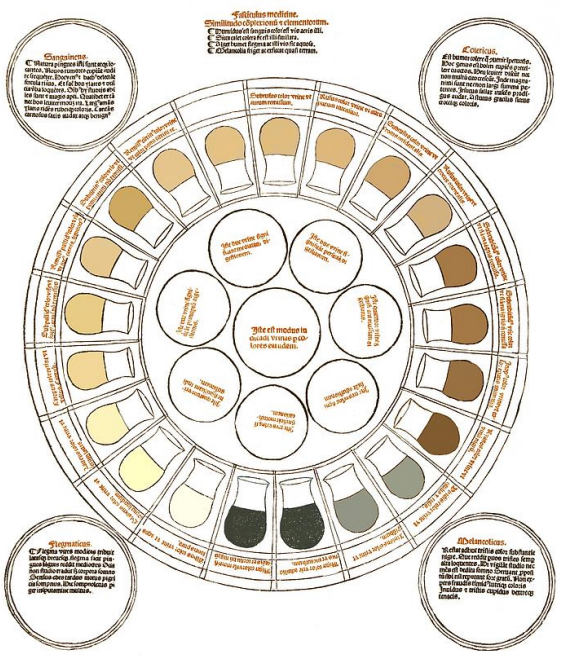
1.1.3 Metabolomic data

One of the defining features of a living creature is metabolism, the set of life-sustaining chemical reactions within the organism's cells. Metabolism converts food to energy and building blocks for nucleic acids, lipids, proteins and some carbohydrates. It also eliminates

nitrogenous wastes. The intermediates and products of metabolism are called metabolites. Normally, we restrict the term metabolite to small biological molecules ($<$ 900 daltons[11] [M. R. Arkin and Wells, 2004]). The metabolome refers to the complete set of small-molecule metabolites found at a given time within a biological sample.

Indeed, an organism could not function properly if its metabolism is out of order. The idea that we could infer about a person's health from his or her biological fluids dates back to the medieval period. In the Middle Ages, people brought flasks containing their urine to the doctor (Figure 1.7a), and the doctor would diagnose their conditions based on the urine wheel (Figure 1.7b). Visual inspection of urine is actually still used as a first-hand reference for self-diagnosis nowadays – for instance, if the urine is red, the patient may be suffering from kidney disease or various other conditions [U.S. National Library of Medicine, 2018].

(a) People giving samples of urine to Constantine the African, a physician in the 11^{th} century [Newton, 1994].



(b) A medieval urine wheel mapping color of the urine to diseases [Jungersen, 2004].

Fig. 1.7 Uroscopy in the Middle Ages.

Fast forwarding to the contemporary era, Roger Williams introduced the concept that individuals might have a "metabolic profile" that could be reflected in the composition of their biological fluids [R. J. Williams, 1956]. [Boulton et al., 1967] used paper chromatography

[11] One dalton, often denoted as u, is $\frac{1}{12}$ of the mass of a carbon-12 atom [McNaught, 1997].

and suggested a possible association between metabolic patterns in urine and diseases such as schizophrenia. With technological advancements, we could now quantitatively measure metabolic profiles much more accurately and meticulously. This paves the way for the emerging field of metabolomics, which aims to study the metabolome as well as all the chemical processes involving metabolites [Shulaev, 2006; Patti et al., 2012].

1.1.3.1 Measuring the metabolome

To measure the metabolome, we first need to prepare the biological sample. This could be serum, urine, saliva or cultured cells. Biological samples are often pulverized into smaller particles in order to increase their surface area exposed to the extraction buffer chosen based on the chemical characteristics of the samples [Cambiaghi et al., 2016]. The prepared sample, by nature, is a highly complex mixture. Hence, prior to applying certain detection methods, we could simplify it by separating some compounds from the others. Common compound separation methods include gas chromatography (GC) [James and Martin, 1952], high-performance liquid chromatography (HPLC) [Knox et al., 1978] and capillary electrophoresis (CE) [Manz et al., 1992].

There is a variety of compound detection techniques to quantify the metabolome from the prepared biological sample. Two examples are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy.

Mass spectrometry

The basic idea behind MS is to ionize chemical species and plot a mass spectrum (like Figure 1.8) of the ion signals sorted according to their mass-to-charge ratio (m/z) [Fenn et al., 1989]. Each peak in a mass spectrum shows a signal of unique m/z in the sample, and heights of the peaks correspond to the relative abundance of the various signals in the sample [Broad Institute, 2018].

MS is one of the most widely applied techniques, as it identifies metabolites reliably and rapidly (the analysis time is between 5 and 140 minutes) [Cambiaghi et al., 2016]. One of the disadvantages of MS is that it often requires separation by GC, HPLC or CE beforehand [Cambiaghi et al., 2016].

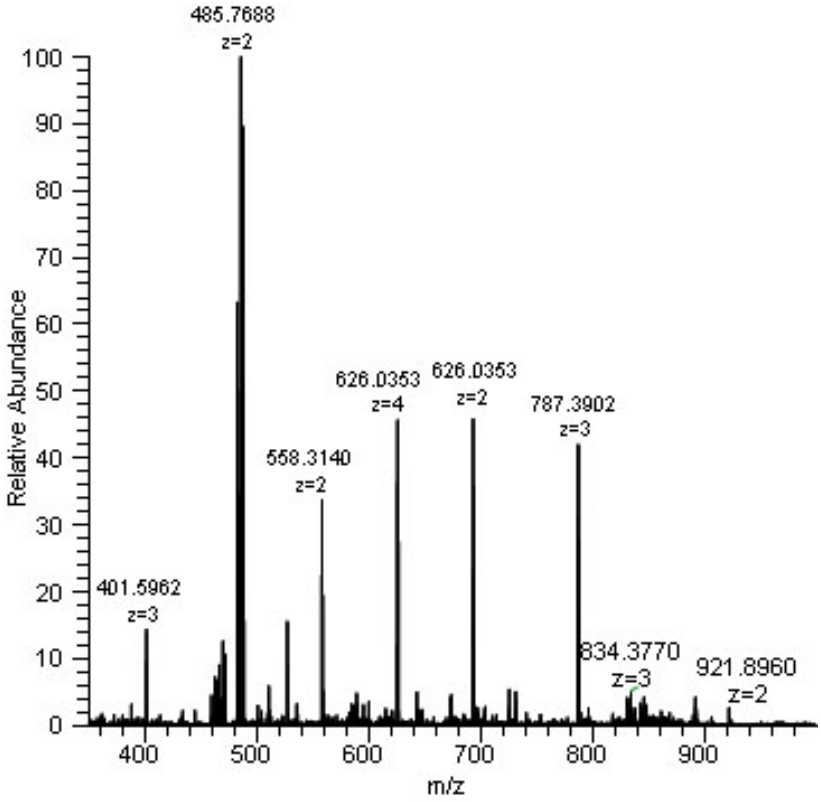


Fig. 1.8 Example of a mass spectrum [Broad Institute, 2018].

NMR spectroscopy

First described and measured by [Rabi et al., 1939], NMR is a physical phenomenon in which nuclei in a magnetic field absorb and re-emit electromagnetic radiation. The electron distribution and the local magnetic field of the same type of nucleus are usually dependent on the local geometry (e.g. bond lengths and binding partners), which is reflected in resonance frequencies [Rabi et al., 1939].

The resonance frequency of a nucleus relative to a standard in a magnetic field is called the chemical shift (δ). We could infer about a molecule's structure based on the position and number of chemical shifts. NMR spectroscopy takes advantage of this observation to detect and measure metabolites in a biological sample simultaneously using NMR spectrometers, which spin the biological sample of interest inside a very strong magnet and detect the NMR signals produced by radio-frequency receivers.

NMR spectrometers must be tuned to a specific nucleus, for example, the proton (^{1}H). Figure 1.9 is a proton NMR spectrum for ethyl acetate. We could see that different types of ^{1}H have

different chemical shifts, measured in parts per million (ppm). In the spectrum, the height of peaks displays the intensity of resonance signals.

NMR spectroscopy is the only metabolite detection method that does not require prior separation [Beckonert et al., 2007]. Therefore, it is non-destructive and the biological sample could be retained for further analyses. However, it is less sensitive than MS-based techniques and needs larger amounts of sample [Cambiaghi et al., 2016].

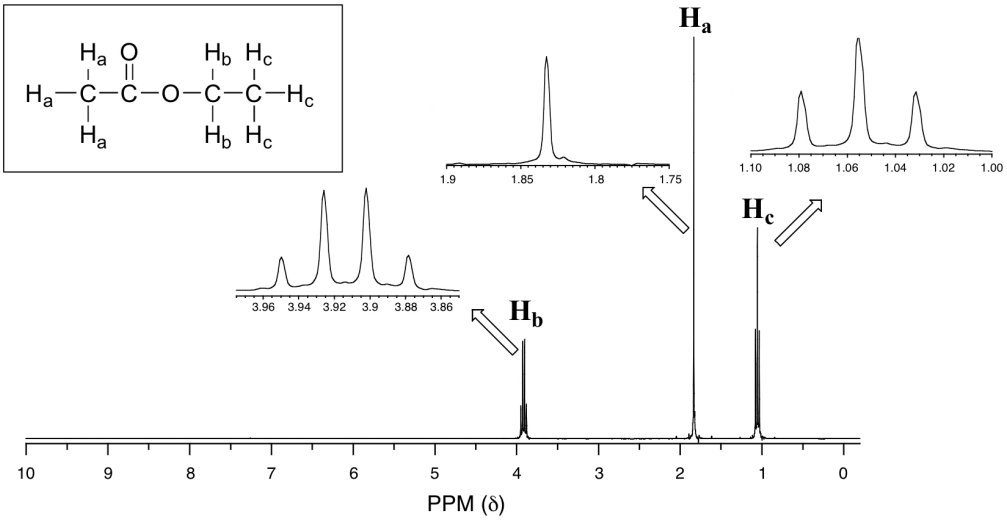


Fig. 1.9 A proton NMR spectrum for ethyl acetate [Soderberg, 2016].

Different detection and separation techniques have different sensitivity, resolution, and limitations in identifying different metabolites; therefore, when choosing a method for measuring the metabolome, we need to consider the characteristic of the biological sample and what type of analysis we aim to conduct [Castle et al., 2006; A. Zhang et al., 2012].

1.1.3.2 Analysis of metabolomic data

Data pre-processing and cleaning

The raw signals detected by scientific instruments (e.g. mass spectra and NMR data) first undergo data pre-processing, including noise reduction, time correction, peak detection and so on, to quantify different metabolites [Cambiaghi et al., 2016].

Data cleaning is crucial in the analysis of the now quantified metabolomic data. Usually, we identify and filter out variables that are of near-constant or close to zero value, which

are unlikely to be of use for subsequent analysis [Xia et al., 2012]. Since metabolomic data is often of different orders of magnitudes among samples and features, it is important to perform data normalization to reduce systematic biases and failure in identifying significant associations [Cambiaghi et al., 2016; Xia et al., 2012].

Feature selection and clustering

Typically, metabolomic data takes the form of large and feature-rich data matrices, which is quite similar with microarray data. Both metabolomic and microarray data analysis seeks to identify features significantly associated with certain conditions (biomarker discovery) or for disease diagnosis (classification) [Xia et al., 2012]. Additionally, both kinds of studies are challenged with the large p, small n problem (high-dimensional feature space with limited sample size). Therefore, feature selection methods widely adopted in microarray analysis such as SAM and EBAM (c.f. Section 1.1.2.1) could be adapted to metabolomic studies.

Principal component analysis (PCA) projects the metabolomic data to a lower-dimensional space capturing variation in the data as much as possible. Clustering of samples with similar metabolomic profiles could be detected when we conduct data analysis in the lower-dimensional PCA space. This could assist us in finding novel disease biomarkers.

We could also identify groups of functionally related metabolites with clustering analysis on the metabolites [Sugimoto et al., 2012]. Hierarchical clustering, k-means clustering [Hartigan, 1975] and self-organizing maps (SOMs) [Kohonen, 1998] are three commonly used methods for clustering.

Univariate data analysis

In practice, metabolomic data analysis usually starts with applying univariate methods (i.e. one variable at a time) like t-tests, one-way analysis of variance (ANOVA) and correlation analysis, aiming to identify the metabolites that show significant changes under the studied conditions [Saccenti et al., 2014; Cambiaghi et al., 2016].

Since when dealing with metabolomic data, we need to perform individual statistical tests for tenths to hundreds of metabolites, just as in GWAS (c.f. Section 1.1.1.2), the issue of multiple testing arises. There are many approaches dedicated to this issue (e.g. [Hochberg and Benjamini, 1990] and [Storey, 2002]); these methods are also applicable in the context of metabolomics [Broadhurst and Kell, 2006].

Multivariate data analysis and classification

Univariate data analysis could provide us with a general sense of the metabolomic data. Their preliminary findings could be reinforced (or rejected) via multivariate methods, for example, partial least squares discriminant analysis (PLS-DA) [Barker and Rayens, 2003].

Like PCA, PLS regression tries to reduce the dimensionality of metabolomic data while aiming to maintain a large proportion of the observed variation. What's better, it could also account for the relationship between the independent and dependent variables. PLS-DA refers to the PLS regression variant in which the dependent variable is categorical. Since compared with PCA, PLS-DA is highly prone to over-fitting [Westerhuis et al., 2010], we need to validate the results by permutation testing or cross-validation [Szymańska et al., 2012]. Successful applications of PLS-DA in metabolomics include identification of metabolomic markers for distinguishing subgroups of motor neuron diseases [Rozen et al., 2005] and predicting preeclampsia in early pregnancy [Kenny et al., 2010].

ML methods are also useful in the analysis of metabolomic data, especially for the classification of disease status based on metabolomic fingerprints. Unlike PCA or PLS-DA, SVMs can be extended to nonlinear cases with the help of kernels. [Mahadevan et al., 2008] has generated a more accurate predictive model than PLS-DA for pneumococcal disease. RF is an ensemble learning algorithm that consists of many decision trees. By averaging over several trees, both overfitting and the variance of the algorithm's performance are reduced. [T. Chen et al., 2013] has shown that RF outperforms PLS-DA and SVM for predicting colorectal cancer based on metabolomic profiles.

Pathway analysis

Metabolomic pathway analysis tries to identify the pathways with significant impact on a given biological process by studying the interactions among genes and metabolites within a sample [Xia and Wishart, 2010a]. Enrichment analysis and topological analysis are two commonly used ways to perform pathway analysis on metabolites [Xia and Wishart, 2010a].

In enrichment analysis, we check if there exist significant expression changes (i.e. enrichment or underrepresentation) among metabolite sets [Cambiaghi et al., 2016]. The set of enriched or underrepresented metabolites could be mapped to biological pathways or disease conditions, paving way for further investigation [Xia and Wishart, 2010b].

Topological analysis, on the other hand, is conducted based on metabolic networks (e.g. the one shown in Figure 1.10). The importance of individual metabolites in the network could be measured by their centrality [Aittokallio and Schwikowski, 2006], and the impact of a certain pathway could be evaluated by summing the importance of all the metabolites in the pathway and next dividing by the sum of the importance of all metabolites in each considered pathway [Xia and Wishart, 2010a].

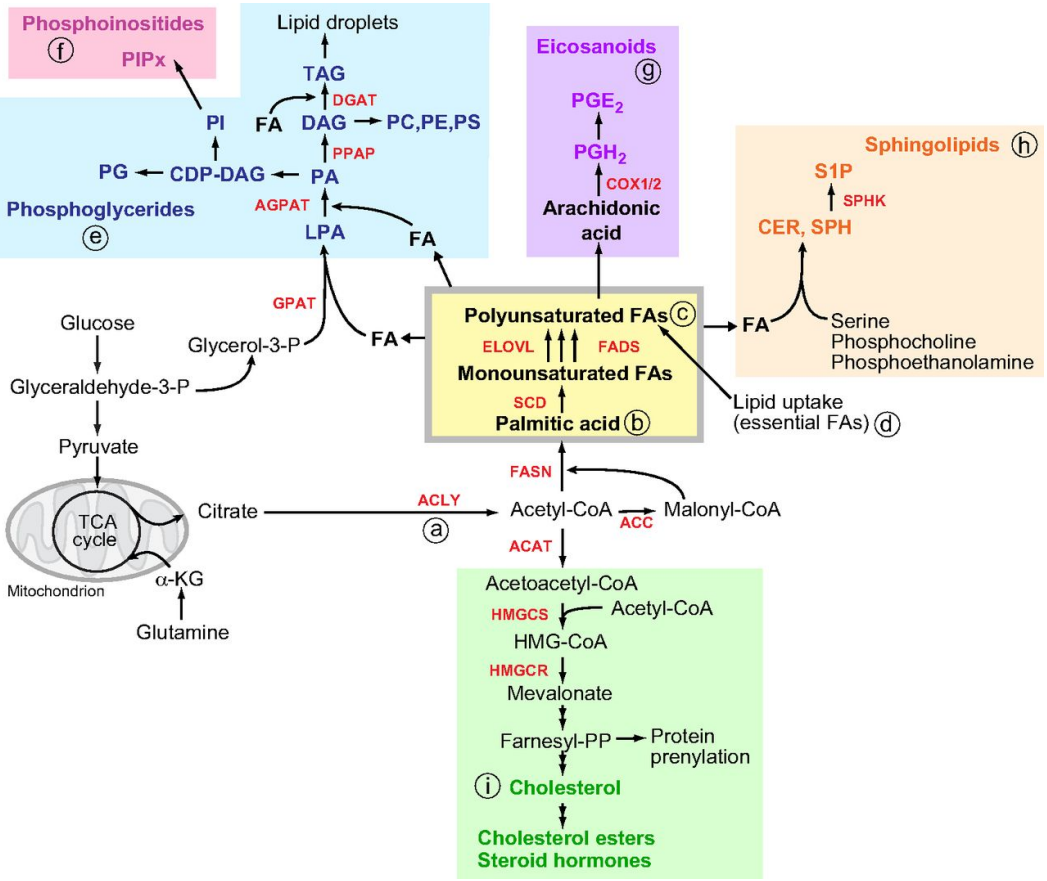


Fig. 1.10 Metabolic network illustrating lipid biosynthesis [Baenke et al., 2013].

Through pathway analysis, the findings (e.g. metabolites selected as biomarkers) from the previous steps could be linked back to the biological context. Furthermore, the identified pathways could be integrated with transcriptomic data to gain a whole picture of all relevant mechanisms [Cavill et al., 2015].

1.1.4 Integrative analysis of big omics data

In previous sections, we have given an overview of genomic (Section 1.1.1), transcriptomic (Section 1.1.2) and metabolomic (Section 1.1.3) data, as well as their corresponding analytic methods. As we have seen, generally speaking, the study of omics data refers to comprehensive, or global, assessment of a complete set of molecules[12] [Hasin et al., 2017].

Figure 1.11 is a simplified illustration of the relationship between big omics data. The reality is much more complex. For instance, as shown in Figure 1.5b, proteins could in return, influence the expression of genes (the transcriptome); modified cell structures could also have feedback effects on genes, mRNAs, and proteins.

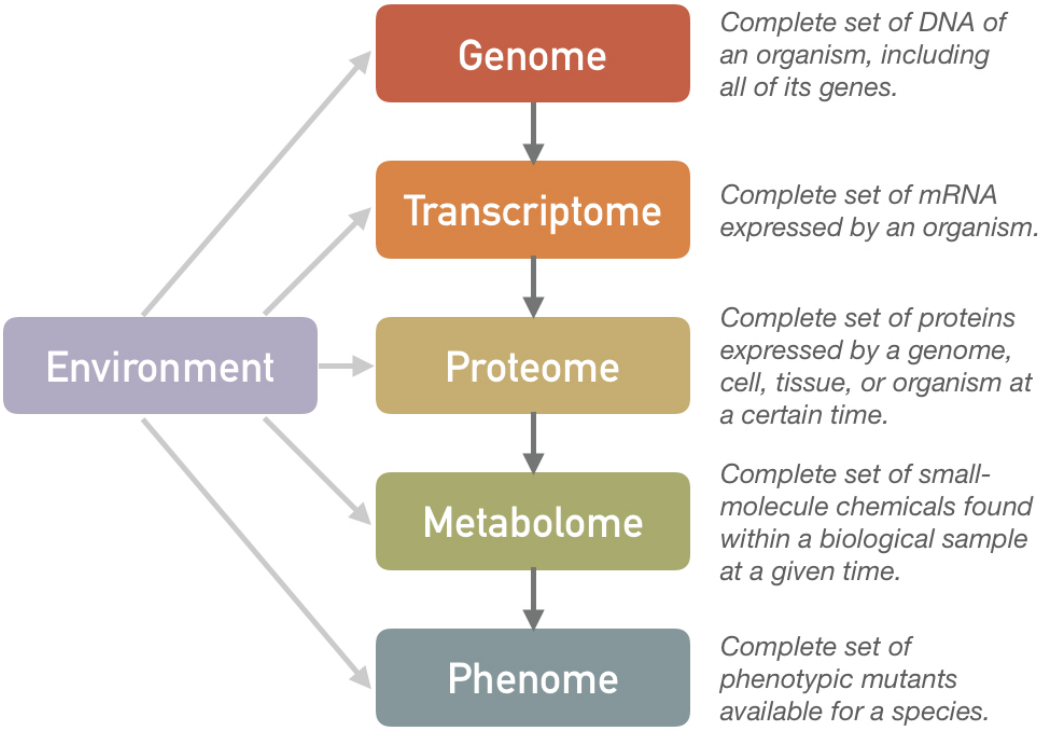


Fig. 1.11 The relationship between big omics data.

The analysis of each type of omics data could provide us with useful biomarkers of traits of interest and help us understand the mechanisms underlying different complex diseases [Hasin et al., 2017]. However, relying on only one data type has certain limitations.

[12]Extract from the Oxford English Dictionary [Dictionary, 2004]: "There are three different fields of application for the *-ome* suffix: (1) in medicine, forming nouns with the sense 'swelling, tumor'; (2) in botany or zoology, forming nouns in the sense 'a part of an animal or plant with a specified structure'; and (3) in cellular and molecular biology, forming nouns with the sense 'all constituents considered collectively'." When we refer to omics data, the third application is used.

Take genomic data analysis for example. For many traits, all the known risk loci identified from past GWAS could only explain a relatively small amount of the heritable component [T. A. Manolio et al., 2009]. Additionally, common diseases usually result from variations in gene regulation and not the coding regions of genes [Hasin et al., 2017]. To account for these problems, integrative analysis with transcriptomic data could be helpful. Different expression levels between the maternal and paternal allele can be used for the investigation of effects of rare variants [Rivas et al., 2015], which may resolve part of the missing heritability problem [T. A. Manolio et al., 2009]. On the other hand, the GRN could help us model changes in gene regulation (c.f. Section 1.1.2.2).

Integrative analysis of different types of omics data could overcome the disadvantages of using only one data type (e.g. inaccurate depiction of the truth due to considering limited types of effects), potentially increase the study's statistical power and give us a more comprehensive understanding of the flow of information underlying complex traits [Hasin et al., 2017; Manzoni et al., 2016]. Based on the initial focus of investigation, there are three approaches in multi-omics data analysis, namely genome-first (focusing on the mechanisms through which GWAS loci affect traits), phenotype-first (studying the pathways contributing to diseases without focusing on one specific locus) and environment-first (investigating how the environment interacts with genes or perturbs pathways) [Hasin et al., 2017].

One challenge in the integrative analysis of big omics data is differentiating causality from correlation [Hasin et al., 2017]. This is particularly difficult due to the correlative nature of omics datasets. Moreover, all analyses of high-dimensional data suffer from the "large p, small n" problem. This is exaggerated for integrated omics data (compared with a single type of omics data), since both the complexity and dimensionality of the data elevate. Lastly, analyzing integrated omics data could be very computationally expensive. Optimization methods and better hardware (e.g. graphics processing units, or GPUs) need to be utilized in implementing and carrying out the analysis.

1.2 Lumbar disc degeneration

1.2.1 Elements of the human spine

A human's spinal column could protect the spinal cord and support his or her head. The ribs, as well as back and neck muscles, are attached to the spinal column. A child is born

with approximately 33 vertebrae, but as he or she grows, several vertebrae fuse together. As shown in Figure 1.12, the adult vertebrae consist of 26 bones – 7 cervical vertebrae, 12 thoracic vertebrae, 5 lumbar vertebrae, the sacrum, and the coccyx.

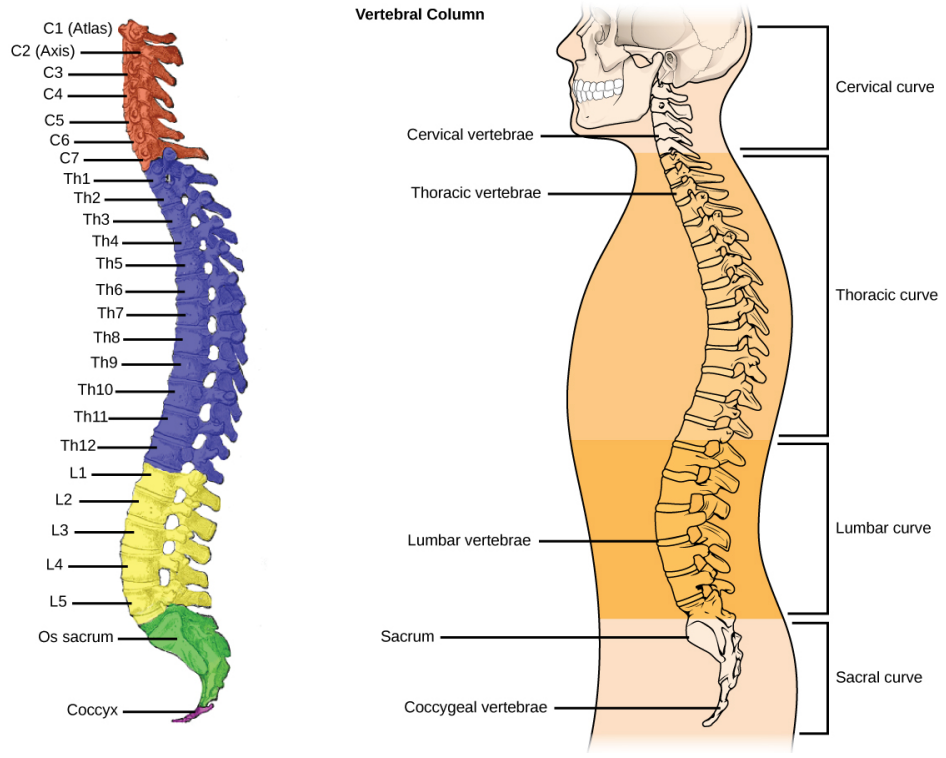


Fig. 1.12 Sections of the human spine [Lumen Learning, 2007].

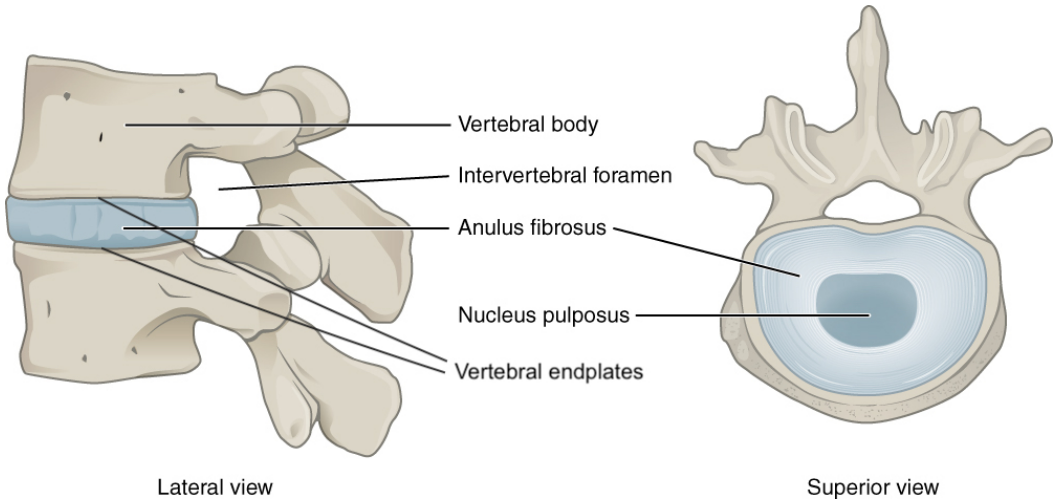


Fig. 1.13 An IVD consists of a NP, a peripheral AF and two VEPs [OpenStax, 2013].

Intervertebral discs (IVDs) are round, rubbery pads placed between adjacent vertebrae. They could absorb shock and cushion the vertebrae as the body moves. Figure 1.13 is an illustration of an IVD. The core of an IVD is called nucleus pulposus (NP), which consists of gel-like matter and a loose network of collagen fibers. The NP allows discs to withstand torsion and compression, and is surrounded by a tough exterior named annulus fibrosus (AF). The AF contains a ring of ligament fibers which protects the NP and connects the adjacent vertebrae. Between the vertebral body and the IVD, there are vertebral endplates (EPs). EPs are thin layers of cartilage covering the entire NP but not AF. They could prevent NP from bulging into the vertebral body and absorb pressure.

1.2.2 Lumbar disc degeneration and lower back pain

1.2.2.1 Lower back pain as a common global health problem

Lower back pain (LBP) (Figure 1.14a) is one of the most common global health issues and a major cause of disability [Kaplan et al., 2013]. [Vos et al., 2012] estimated that LBP is one of the top ten conditions accounting for the highest number of disability-adjusted life years[13] (DALYs) worldwide. Some patients with severe LBP could develop sciatica, i.e. pain spreading down the leg from the lower back (Figure 1.14b). Sciatica also has a great public health burden due to its high incidence and major socioeconomic costs [Younes et al., 2006].

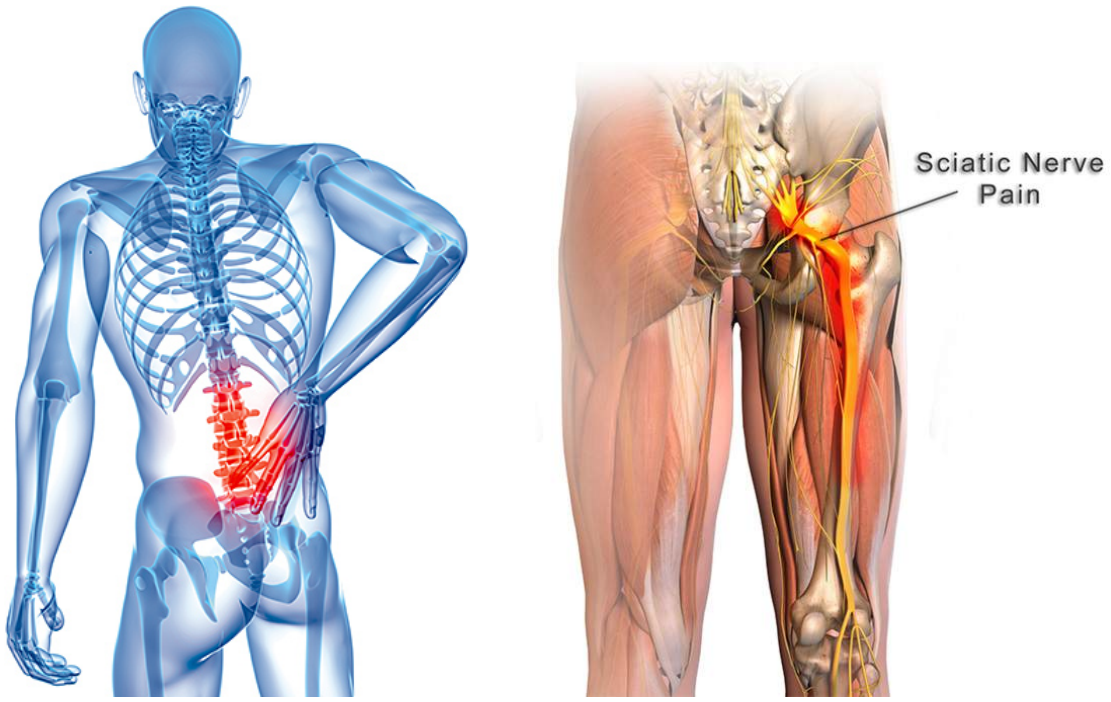
The lifetime prevalence of common LBP is estimated to be 60% to 70% in industrialized countries [Kaplan et al., 2013]. Additionally, [Hoy et al., 2012] has shown that the global 1-month prevalence of activity-limiting LBP is $23.2\% \pm 2.9\%$, i.e. during one month, out of 100 people, on average 20 to 26 suffer from activity-limiting LBP in the global population.

Since LBP is such a common condition that could affect a patient's work or study performance as well as general well-being, identifying potential biomarkers and risk factors for LBP is of great importance [Kaplan et al., 2013].

1.2.2.2 Lumbar disc degeneration as a cause for lower back pain

One of the most common causes for back pain is intervertebral disc degeneration (IDD), which refers to the deterioration of IVDs (e.g. loss of cushioning ability) over time [National

[13] The disability-adjusted life year is the number of years lost due to bad health, disability or early death. It is a measure of disease burden, the impact of a health problem measured by mortality, financial cost, etc.



(a) Location of LBP [Leotraining, 2016]. (b) Location of sciatica [Diagrams for all, 2018].

Fig. 1.14 Illustration of LBP and sciatica.

Institute of Neurological Disorders and Stroke, 2017]. The process of disc degeneration is an aberrant response to progressive structural defect mediated by cells [M. A. Adams and Roughley, 2006]. A degenerate disc has a structural defect and its aging, in terms of biochemical, histologic, metabolic and functional changes, is accelerated [M. A. Adams and Roughley, 2006].

If we refer to Figures 1.12 and 1.14a, we could see that LBP happens in the lumbar region of the spine (vertebrae L1 to L5), which supports much of the weight of the upper body. Therefore, lumbar disc degeneration (LDD), in specific, is relatively more relevant to LBP. LDD could be evaluated by magnetic resonance imaging (MRI), a medical imaging technique which forms detailed pictures of the body's anatomy using a magnetic field and radio waves. Typical radiographic observations related to LDD include disc space narrowing, disc bulging, disc herniation, disc dehydration, modic changes, EP damages and annular tears [M. A. Adams and Roughley, 2006].

Disc space narrowing

The height of an IVD is influenced by several factors, ranging from dehydration induced by increased load (potentially reversible) to the reduction of disc tissue volume due to structural failure (often irreversible) [Berlemann et al., 1998].

Prior to the wide application of MRI, disc space narrowing was probably the medical imaging finding most frequently used to indicate LDD [Battié et al., 2004]. However, with the prevalence of MRI, we could attain a more detailed image of the IVD, and it is found that the existence of other types of disc changes, e.g. bulging, render it less clear what disc narrowing reflect [Battié et al., 2004; Berlemann et al., 1998]. Hence, changes in disc volume may be more relevant to LDD than changes in disc height [Battié et al., 2004].

Still, severe disc narrowing is a strong indicator of LDD, though severe narrowing on a single level is more likely to reflect a traumatic, rather than a systemic origin [Frobin et al., 2001].

Disc bulging and herniation

Disc bulging (DB) and herniation both happen when radial fissures allow migration of NP relative to AF such that the IVD periphery is affected [M. A. Adams and Roughley, 2006].

If there exists protrusion but no tear in the outer layer of the disc, we observe DB. If the AF breaks and the NP has leaked out of it, disc herniation is present. Therefore, disc herniation is also referred to as a "slipped" disc, though this term is inaccurate since the IVDs are firmly attached between the vertebrae and actually cannot slip out of place [OpenStax, 2013].

Potentially resulting from intensive repetitive loading, DB and disc herniation are suggested to be associated with LDD and LBP [Videman et al., 2003].

Disc dehydration

In the laboratory, we usually do not observe disc herniation in severely degenerated discs, presumably because the NP is so dehydrated that it could no longer exert hydrostatic pressure to the AF [M. A. Adams and Roughley, 2006]. Dehydration of the NP reduces the flexibility and often the height of the disc, which could reflect the degree of LDD [Luoma et al., 2001].

Disc dehydration could be measured in terms of signal intensity loss (SIL) on MRI. [Schneiderman et al., 1987] proposed a classification scheme for SIL – grade 0 indicates a disc with

a normal height and signal intensity, grade 1 marks a disc with a speckled pattern of SIL, grade 2 refers to a disc with diffuse SIL, and grade 3 corresponds to a black disc with disc narrowing. [Luoma et al., 2001] shows that SIL may be a more sensitive measure of LDD compared to disc narrowing itself.

Modic changes

Modic changes (MCs) are vertebral body marrow changes adjacent to the EPs, which could be observed through MRI [Modic et al., 1988].

Typically, MCs could be categorized into three types [Modic et al., 1988]. Type 1 MC refers to marrow edema and disruption of the EPs, type 2 MC indicates fatty degeneration of the adjacent vertebral marrow, whereas type 3 MC represents the presence of bone sclerosis (hardening) and relative loss of bone marrow [De Roos et al., 1987; Modic et al., 1988]. When different types (mainly 1 and 2 or 2 and 3) are observed simultaneously at the adjacent vertebral body, we declare MC of mixed types [Määttä et al., 2016].

In previous studies, MCs are found to be highly associated with LDD [Modic et al., 1988] and LBP in population-based [Kjaer et al., 2005; Mok et al., 2016] as well as clinical cohorts [Toyone et al., 1994; O. K. Jensen et al., 2014].

Endplate damages

As could be seen from Figure 1.13, upon compression, EPs are the spine's "weak spot". Indeed, damage of the EP may decompress the adjacent NP and transfer the load to the AF, causing it to bulge into the NP cavity [M. A. Adams et al., 2000; M. A. Adams and Roughley, 2006]. If NP flows through a damaged EP into the adjacent vertebra, a Schmorl's node (SN) could be created [Hamanishi et al., 1994].

It has been shown that under experimental conditions, the disrupt of endplate could lead to IVD [Holm et al., 2004]. In particular, SNs are shown to be highly heritable and related to LDD [F. Williams et al., 2007].

Annular tears

An annular tear occurs when the AF rips resulting from too much stress on the IVD – it is often classified into three categories, according to its shape and location [Osti et al., 1992].

First of all, typically as a result of natural aging, radial tears extend from the center of the IVD to the outer layer and could cause disc herniation [Osti et al., 1992]. Secondly, peripheral tears occur in the outer fibers and are usually due to bone outgrowth or injury – these could contribute to LDD [Osti et al., 1992]. Finally, circumferential tears are concentric tears between the outer layers, which normally result from injury or compressive stress concentrations in older discs [Osti et al., 1992; Goel et al., 1995].

The annular tears could be seen on MRI as areas with a brighter signal, which are called high intensity zones (HIZs). [Peng et al., 2006] shows that the HIZs of patients with LBP could be considered as reliable indicators for painful AF damage.

1.2.3 Prevalence of lumbar disc degeneration

The epidemiology[14] of LDD is difficult to discuss due to the lack of a standard definition (or a precise measurement) of disc degeneration [Battié et al., 2004]. To attain better consistency and reliability, we usually study LDD-related MRI findings (c.f. Section 1.2.2.2) instead.

The prevalences of different MRI features regarding LDD have been estimated from population-based cohorts in many previous studies. The results are shown in Table 1.1. As could be seen, the reported prevalences vary vastly across studies. This is probably because the cohorts recruited in different studies have different age distributions and exposure to risk factors [Battié et al., 2004]. Another possible reason is the variation in the definitions and readings of the MRI features [Battié et al., 2004]. Future research may benefit from devising a scheme or algorithm (e.g. using computer vision) for standardized definitions of MRI findings related to LDD [Battié et al., 2004].

1.2.4 Etiology of lumbar disc degeneration

The etiology of LDD is highly multifactorial – genetic and environmental risk factors, as well as their interactions, could contribute to LDD [Battié et al., 2004].

[14] Epidemiology is the study of the incidence, distribution, and possible determinants of health and disease conditions [Last et al., 2001].

Table 1.1 Prevalence of different MRI features regarding LDD in the general population.

MRI feature	Prevalence
Disc narrowing	15% – 53%
Disc bulging	22% – 48%
Signal intensity loss	9% – 86%
Modic change	19% – 56%
Schmorl's node	6% – 79%
High intensity zone	15% – 28%

Results are from [Battié et al., 2004; T. S. Jensen et al., 2010; Y. Wang et al., 2012; Teraguchi et al., 2016].

1.2.4.1 Age, sex and environmental risk factors

As a progressive disorder, LDD is heavily influenced by age. Previous research has revealed that signs of LDD could be identified as early as in childhood and across age groups, great variability in LDD-related MRI features exists [Battié et al., 2004].

Regarding sex, it has been shown that male discs tend to be more degenerated than female discs at most ages [J. A. Miller et al., 1988]. However, this gender influence is still controversial, since some later studies have failed to find a significant association between gender and LDD [Teraguchi et al., 2014].

Apart from age and sex, it has been found that LDD could be worsened by various risk factors, for instance, obesity [Samartzis et al., 2012; Teraguchi et al., 2014] and smoking [Battié et al., 1991]. It is also noteworthy that contrary to popular belief, [Battié et al., 2009] found that occupational and leisure physical loading conditions throughout adulthood do not contribute to LDD – quite the opposite, routine loading may actually benefit the IVD. The association between physical loading and LDD is still controversial and under investigation [Battié et al., 2004].

1.2.4.2 Genetic risk factors

After adjusting for age, weight, smoking, occupation and physical activity, the heritability of LDD[15] is estimated to be 74% with a 95% confidence interval of $(64\%, 81\%)$ by twin study methods [Sambrook et al., 1999].

[15] As a summary score taking into consideration disc height, signal intensity, bulging, and anterior osteophyte formation [Sambrook et al., 1999].

In light of this relatively high heritability, researchers have conducted GWAS aiming to identify novel loci associated with LDD and gain a better understanding of LDD's underlying genetic factors. However, unfortunately, most of the genes found to be significantly associated with LDD from various GWAS have a weak level of cumulative association evidence [Eskola et al., 2012]. According to [Eskola et al., 2012], the only previously reported genes with moderate levels of evidence are *ASPN (D-repeat)*, *COL11A1 (rs1676486)*, *GDF5 (rs143383)*, *SKT (rs16924573)*, *THBS2 (rs9406328)* and *MMP9 (rs17576)*.

This lack of credibility of most reported genetic associations is, again, partly due to the often ambiguous definitions of LDD phenotypes [Eskola et al., 2012]. Additionally, large population-based cohorts are needed for future research [Eskola et al., 2012].

1.2.4.3 Metabolomic risk factors

In an adult's body, the IVDs are avascular (i.e. with few or no blood vessels). Metabolites are transported into the disc by diffusion (small molecules) or bulk fluid flow (large molecules) via the EP [M. A. Adams and Roughley, 2006], as illustrated in Figure 1.15. In recent years, the role of altered metabolism in the development and progression of LDD has gained more and more interest [Samartzis et al., 2013a].

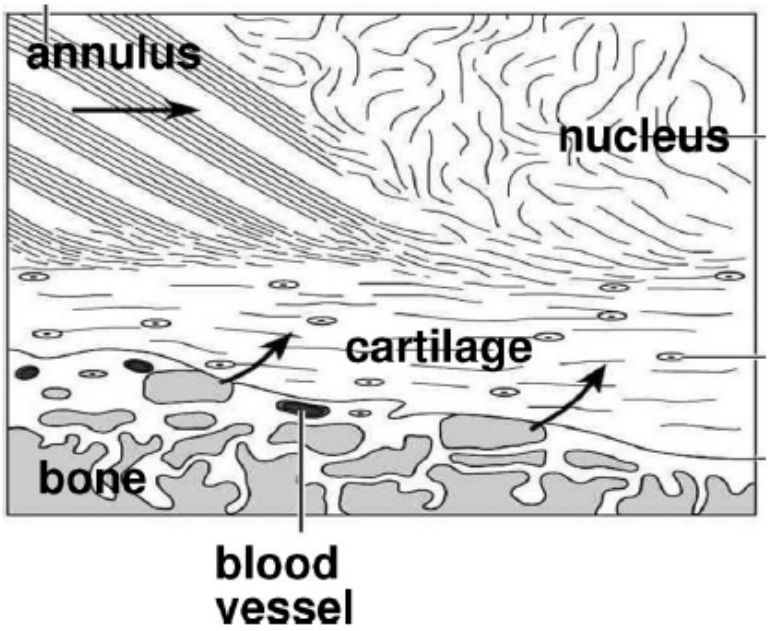


Fig. 1.15 Routes of metabolite transport from surrounding blood vessels into the center of an IVD via the EP [M. A. Adams and Roughley, 2006].

One of the examples illustrating how metabolism could potentially influence LDD is related to metabolite transport. Experiments have shown that a chronic lack of oxygen renders disc cells quiescent, and a chronic deficiency of glucose can kill them [Horner and Urban, 2001]. Deficiencies in nutrient supply reduce the number of viable disc cells and limit the metabolic activity of the cells that are still alive [Urban et al., 2004], which leads to degeneration and limited ability to recover from any injury. By studying diffusion, [Rajasekaran et al., 2004] has demonstrated that EP permeability (and hence disc metabolite transport) normally reduces during aging and increase when degeneration and EP damages are present. Therefore, studying disk metabolism could probably help us distinguish aging from degeneration.

Another example is relevant to the presence of anaerobic metabolism when the center of an IVD suffers from low oxygen tension. Anaerobic metabolism results in higher levels of lactic acid and a lower pH (potential of hydrogen, a scale of acidity/basicity) value [Urban et al., 2004], and lactic acid has been shown to be a metabolic marker for discogenic back pain [Keshari et al., 2008].

The current literature about the influence of altered metabolism on LDD is still quite limited. The analysis of the relationship between metabolomics and LDD could help us gain insight into the underlying biological mechanism of the degeneration process, as well as identify novel biomarkers for LDD, which could potentially aid diagnosis and treatment of LBP.

1.3 Aims and organization of this thesis

1.3.1 Research objectives

LBP is one of the most prevalent global health problems and a main cause of disability [Kaplan et al., 2013]. Since LDD is one of the major reasons for LBP and could be measured more accurately utilizing MRI techniques, in this thesis, I choose to focus on studying LDD through a set of integrative analyses of genomic, metabolomic and phenotypic data from a large population cohort.

The first objective of this thesis is to identify novel genetic variants associated with various metabolomic measurements via GWAS. Through polygenic scoring, this thesis aims to estimate the human metabolome based on genomic data. Secondly, the association between (estimated) metabolomic data and LDD-related phenotypes is tested. I hypothesize that the

(estimated) metabolomic measurements could be used as potential biomarkers for LDD and seek to discover them, if any, in this step.

The ultimate aim of this thesis is (1) to propose a new way of analyzing big omics data in an integrative manner, utilizing metabolome prediction models; and (2) to gain a better understanding of the underlying biological mechanisms of LDD (especially the ones related to altered metabolism) with a data-driven approach.

1.3.2 Thesis organization and flow of data analysis

This thesis is divided into six chapters, organized as follows.

To begin with, Chapter 2 provides details of the studied cohort and describes how the data analyzed in this thesis is collected and pre-processed. Findings of exploratory analysis of the serum metabolome and its phenotypic associations are presented in Chapter 3, and results of the genome-wide association studies of metabolomic measurements are shown in Chapter 4. In Chapter 5, basic and LDD-related phenotypes are associated with metabolomic measurements estimated via polygenic scoring. This thesis concludes by summarizing and discussing its findings, as well as outlining potential areas for future research.

Figure 1.16 demonstrates the overall flow of analysis in this thesis.

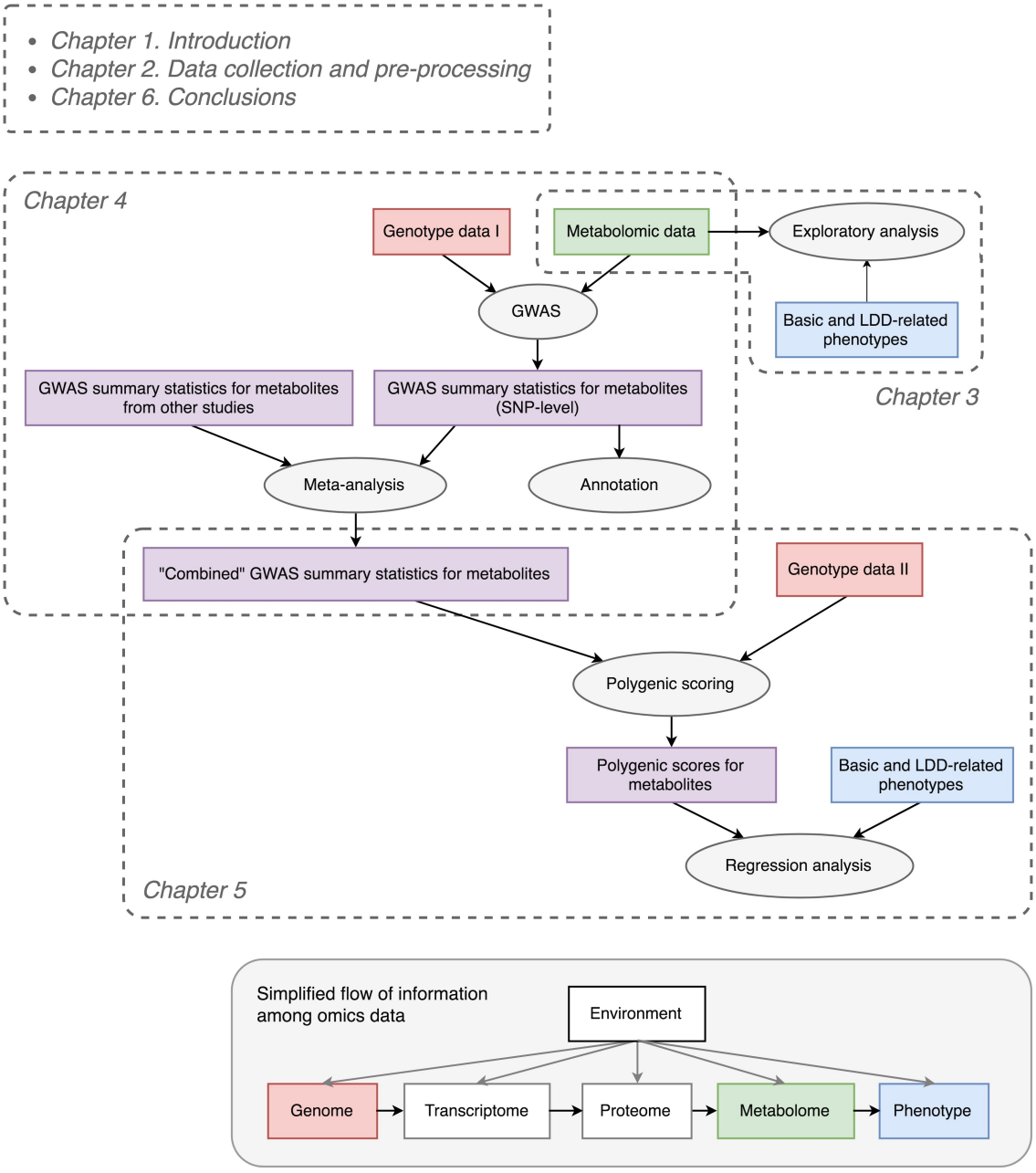


Fig. 1.16 Organization and flow of analysis of this thesis.

2

Data collection and pre-processing

2.1 Sample recruitment

The population cohort studied in this thesis consists of $3,584$ volunteers recruited from 1999 to 2011 by open invitation, following approval from the institutional ethics board [K. M. Cheung et al., 2009; Samartzis et al., 2011; Y. Li et al., 2016]. Of all the individuals, $2,139$ (59.68%) were female.

The volunteers who are of southern Chinese ancestry, living in Hong Kong and between 15 and 55 years old were selected for participation in the study. Additionally, I excluded the subjects with a known history of spinal tumor, spinal infection or spinal deformities [K. M. Cheung et al., 2009; Samartzis et al., 2011; Y. Li et al., 2016].

2.2 Data collection

2.2.1 Questionnaire data

After obtaining informed consent from the volunteers, we asked them to fill out questionnaires collecting basic personal information, including their age (in years) and cigarette smoking status (in pack-years[1]).

[1]Pack-year is calculated via $(\textit{Packs smoked per day}) \times (\textit{Years as a smoker})$. One pack-year is simply smoking 20 cigarettes a day for one year. If someone has smoked 10 cigarettes a day for 6 years, he or she would have a 3 pack-year history. If someone is a non-smoker, he or she would have a 0 pack-year history.

Clinical assessments were also included in the questionnaire. To start with, the volunteers were asked to specify whether they have ever experienced lower back pain or sciatica. Moreover, the visual analog scale (VAS; c.f. Figure 2.1) enabled the individuals to indicate their pain intensity in a continuous manner. We asked them to report both the VAS score on the test day and the severest VAS score ever experienced.

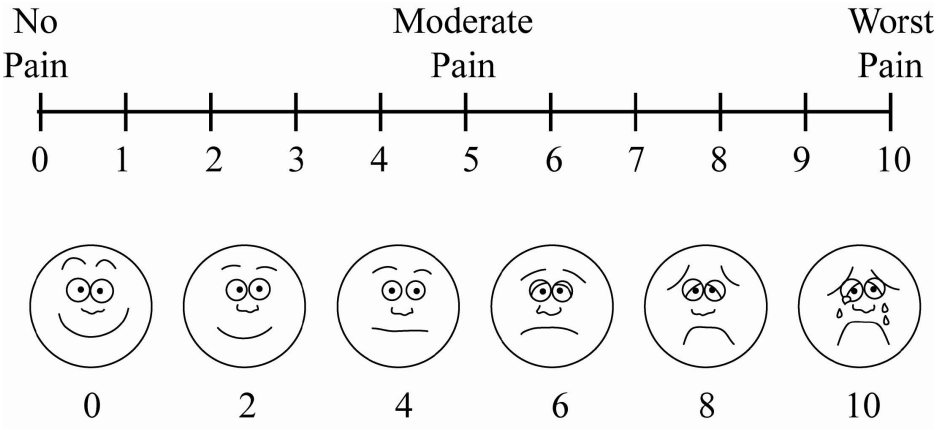


Fig. 2.1 Example of the visual analog scale [Yale University, 2018]. The volunteer was asked to draw a cross anywhere on the scale, and we measured the length from "0" to the cross.

To quantify the disability incurred by LBP, we invited the individuals to fill out the Oswestry LBP questionnaire, which contained ten topics regarding the intensity of pain, lifting, sexual function, social life, sleep quality, as well as ability to care for oneself, walk, sit, stand and travel [Fairbank and Pynsent, 2000].

Each of the topics was reflected by six statements describing possible scenarios in the volunteer's life, and the individual was asked to select the statement closest to his or her situation. The six statements each corresponded to a score of 0 (least amount of disability) to 5 (most severe disability); the ten scores (from the selected statements for the ten topics) were summed and multiplied by two to obtain the Oswestry Disability Index (ODI), which ranges from 0 (no disability) to 100 (maximum disability possible) [Fairbank and Pynsent, 2000].

Tables 2.1 and 2.2 summarize the collected questionnaire data. The cohort mainly consisted of people from 45 to 55 years old, and the majority was non-smokers. Most of them have experienced LBP at some point in their life, but less than half have sciatica. Most volunteers (72.21%) in the cohort had minimal disability (ODI < 20%), though a small proportion of people (3.49%) were severely disabled by LBP, crippled, or even bed-bound (ODI > 41%).

Table 2.1 Summary statistics of binary questionnaire data.

Phenotype	Proportion of trues	Proportion of falses	# of non-NAs
Smoking	0.1286	0.8714	3429
LBP	0.8018	0.1982	3300
Sciatica	0.4299	0.5701	3296

Table 2.2 Summary statistics of continuous questionnaire data.

Phenotype	Minimum	First quartile	Median	Mean	Third quartile	Maximum	# of non-NAs
Age	17.36	45.83	52.22	50.03	56.33	86.33	1823
Smoking (pack-years)	0.000	0.000	0.000	1.652	0.000	95.000	3395
VAS (test day)	0.00	0.00	5.00	15.53	26.00	100.00	3186
VAS (severest)	0.00	14.00	52.00	48.65	78.00	100.00	3186
ODI	0.000	0.000	4.444	9.906	14.000	95.560	3179

2.2.2 Anthropometric measurements

Anthropometric measurements were taken for a majority (over 88%) of volunteers. Height (in meters) was measured without shoes, and weight (in kilograms) was measured in light clothing. The mean height of the measured individuals was 1.60 with a 0.09 standard deviation (SD), whereas the weight was of mean 61.07 with SD $= 11.40$.

The body mass index (BMI) was then calculated via $\frac{Weight}{Height^2}$. BMI is a commonly used way to determine whether a person is overweight, underweight or of normal weight. The mean BMI was 22.93, with 3.30 SD.

2.2.3 Magnetic resonance imaging scan and evaluation

From the whole cohort, $1,416$ probands (the first recruited subject within a family) were followed longitudinally (two time points) with MRI scans. On each volunteer, MRI of the whole spine was performed using sagittal T2-weighted fast spin-echo sequences (repetition time = $3,000$ milliseconds; echo time = 92 milliseconds; slice thickness = 5 millimeters) [K. M. Cheung et al., 2009; Samartzis et al., 2011; Y. Li et al., 2016]. In this thesis, I only consider the second time point[2] since it used MRI machines with a higher resolution, and its time was more akin to metabolomic measurements.

[2]The questionnaire data and anthropometric measurements described in Sections 2.2.1 and 2.2.2 were also measured over two time points; I only use those at the second time point.

This thesis is focused on the lumbar region of the spine, and we asked Dr. Jaro Karppinen (denoted as JK), an experienced physician to assess the L1 to L5 region (c.f. Figure 1.12) of the MRI scans, blinded to the clinical data of the subjects. From JK's reads, I selected five MRI features associated with LDD for analysis in my study, which are disc bulging (DB), signal intensity loss (SIL), high intensity zone (HIZ), modic change (MC) and Schmorl's node (SN). Description of these features could be found in Section 1.2.2.2.

Regarding DB, JK graded a disc with no displacement as 0, a disc with protrusion as 1, and a disc with extrusion as 2. The counts and distribution of DB $= 0, 1, 2$ at each disc level for all the studied volunteers are shown in Table 2.3 and Figure 2.2. On lower disc levels, there tend to be more discs that have developed DB, and the DB condition generally becomes more severe (e.g. extrusion instead of protrusion).

Table 2.3 Contingency table of the counts of each DB status at different disc levels.

Disc level	DB=0	DB=1	DB=2	Unknown
L1	1385	28	3	0
L2	1304	111	1	0
L3	1158	256	2	0
L4	867	534	9	6
L5	829	567	18	2

SIL was measured according to Schneiderman's scoring (SS) scheme (on a scale of $0, 1, 2, 3$), whose details are described in Section 1.2.2.2. Table 2.4 is a contingency table of the counts of each SS value at different disc levels (also c.f. Figure 2.2 for the distribution), demonstrating that disc dehydration, or SIL, is greater at lower disc levels.

Table 2.4 Contingency table of the counts of each SS value at different disc levels.

Disc level	SS=0	SS=1	SS=2	SS=3	Unknown
L1	1206	144	39	27	0
L2	936	283	143	54	0
L3	632	424	280	79	1
L4	383	361	406	260	6
L5	431	292	432	258	3

For HIZ, MC and SN, JK marked the presence (or absence) of each phenotype on each disc level for all the individuals. As could be seen in Table 2.5 and Figure 2.2, the occurrence

of HIZ, MC and SN are quite low in the studied population. HIZ and MC generally worsen when the disc level is lower, but the trend is opposite for SN. This demonstrates the possibly developmental nature of SN.

Table 2.5 Contingency table of the counts of HIZ, MC and SN status at different disc levels.

Phenotype	Disc level	Absence of phenotype	Presence of phenotype	Unknown
HIZ	L1	1374	1	41
HIZ	L2	1361	14	41
HIZ	L3	1354	21	41
HIZ	L4	1229	145	42
HIZ	L5	1184	190	42
MC	L1	1372	3	41
MC	L2	1359	16	41
MC	L3	1350	25	41
MC	L4	1297	77	42
MC	L5	1244	130	42
SN	L1	1328	88	0
SN	L2	1314	102	0
SN	L3	1349	67	0
SN	L4	1373	42	1
SN	L5	1406	9	1

Additional to JK's reads, Dr. Dino Samartzis (denoted as DS), another expert in LDD, has read the MC types in the lumbar region for another subset of $1,713$ probands in the cohort (c.f. Table 2.6). In the table, a subject with "mixed types" can have a combination of any of types from 1 to 3. Only a relatively small proportion of individuals suffered from any type of MC, which agreed with JK's reads. Within the people with MC, type 2 MC was the most prevalent, followed by mixed types.

Table 2.6 Counts of different MC types in the cohort (read by Dr. Samartzis).

Type	Count	Notation
None	$1,363$	0
Type 1	37	1
Type 2	235	2
Type 3	0	3
Mixed types	78	4
Sum	$1,713$	/

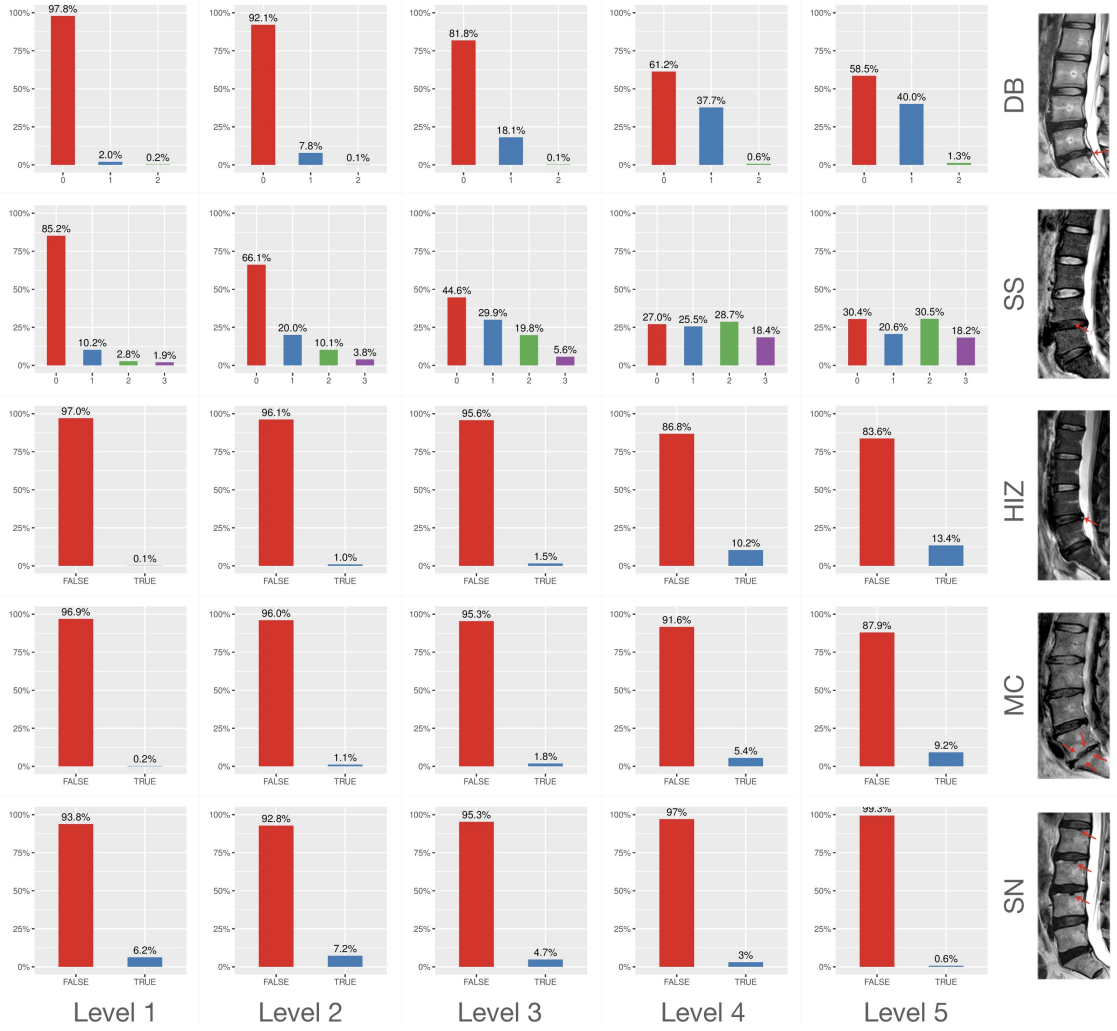


Fig. 2.2 Bar plots showing the distributions of DB, SS, HIZ, MC and SN at different disc levels. The missing values are not plotted, so whenever there is missing data for a phenotype on a certain disc level, in that specific bar plot, the labeled percentages of plotted bars do not add to 100%. MRI images courtesy of [Y. Li, 2016].

This thesis studies (1) the 5 MRI phenotypes read by JK over 5 disc levels ($5 \times 5 = 25$ in total); and (2) DS's MC type reads in the lumbar region. In JK's reads, $1,293$ individuals had complete MRI phenotypic measurements, whereas for DS, $1,713$ individuals had no missing data.

2.2.4 Genotyping

The blood samples of $2,482$ volunteers were obtained for genotyping. DNA was extracted from the blood samples and underwent concentration quality control. The DNA samples were next genotyped using Illumina's OmniZhongHua-8 BeadChip, which is a population-specific whole genome array covering 77% of common variation (minor allele frequency, or MAF $> 5\%$), 73% of intermediate variation (MAF $> 2.5\%$) as well as 65% of low frequency variation (MAF $> 1\%$) in the Chinese population [Illumina, 2016]. Therefore, OmniZhongHua-8 BeadChip is ideal for Chinese population GWAS and hence my study, which is based on a population cohort of southern Chinese.

In total, $900,015$ SNPs (on chromosomes 1 to 22 and chromosome X) were genotyped. The raw data was converted to PLINK format [Purcell et al., 2007] using Illumina's GenomeStudio.

2.2.5 Metabolomic measurements

The serum samples of 814 individuals were acquired for metabolomic measurements. After extracting lipids from the serum samples utilizing a standard protocol described in [Folch et al., 1957; Adosraku et al., 1994], we performed ^{1}H NMR measurements over three molecular windows for all serum samples on a Bruker AVANCE 500 DRX spectrometer operating at 500.13 MHz, following a procedure presented in [Tukiainen et al., 2008].

As illustrated in Figure 2.3, the three molecular windows are lipoprotein lipids (LIPO), low molecular weight metabolites (LMWM) and lipid extracts (LIPID). The LIPO window mainly consists of broad signals of macromolecules, e.g. albumin and lipoprotein lipids [Tukiainen et al., 2008]. On the other hand, the LMWM window applies a pulse sequence suppressing the macromolecule signals, hence enabling the detection of small molecules [Tukiainen et al., 2008], e.g. amino acids, lactate, and glucose. When we extracted lipids from the samples, the lipoprotein particles were broken down, yielding useful information regarding the individual

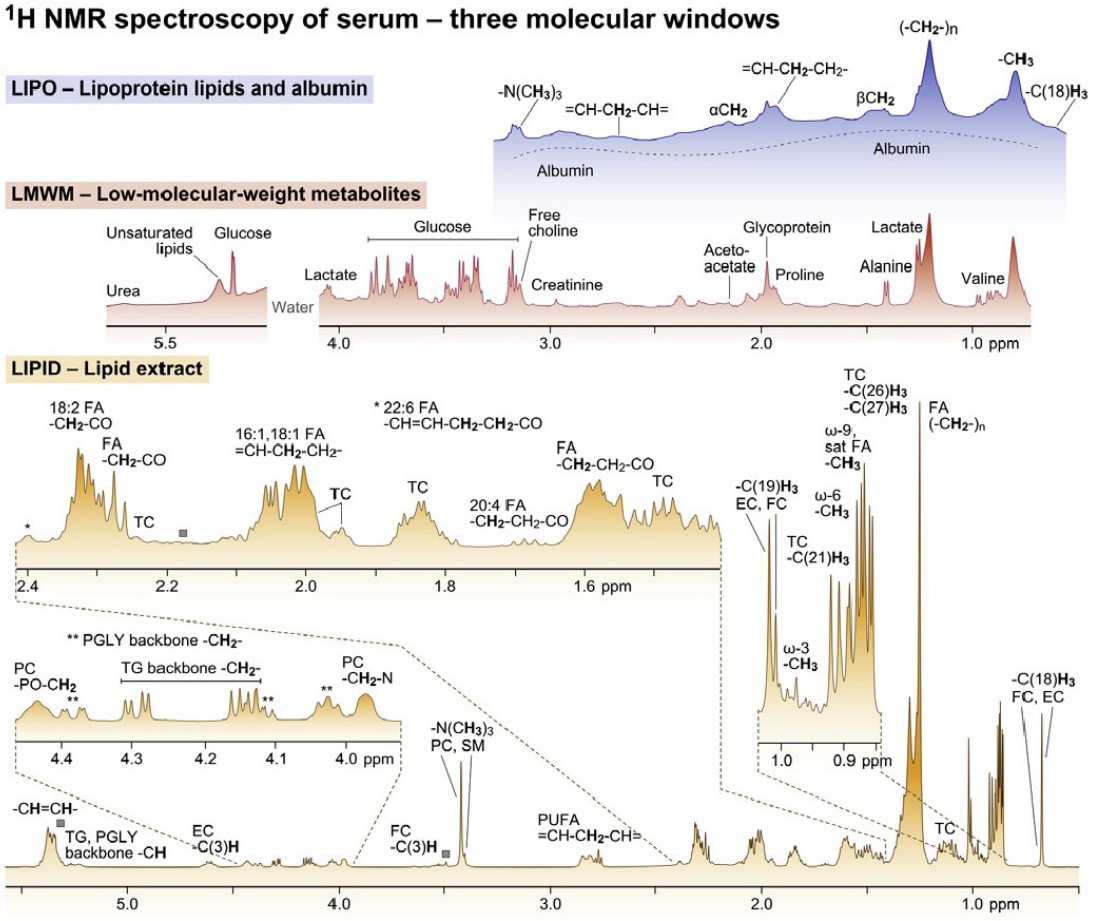


Fig. 2.3 Three molecular windows of ^{1}H NMR measurements [Tukiainen et al., 2008]. Acronyms used in the figure: EC, esterified cholesterol; FA, fatty acid; FC, free cholesterol; PC, phosphatidylcholine; PGLY, phosphoglycerides; PUFA, polyunsaturated fatty acid; sat, saturated; SM, sphingomyelin; TC, total serum cholesterol; TG, total serum triglycerides.

lipid species inside lipoprotein particles – this information is reflected in the LIPID window [Tukiainen et al., 2008].

The water region in the LMWM window, as well as the narrow spectral regions (marked with grey squares in Figure 2.3) with potential residual solvent peaks in the LIPID window were excluded from subsequent analysis [Tukiainen et al., 2008]. The raw metabolomic data was then quantified using a protocol described in [Tukiainen et al., 2008; Larmo et al., 2013], which includes data pre-processing (scaling, transforming, corrections), fitting cross-validated regression models for the LIPO window, as well as lineshape fitting of the LMWM and LIPID windows.

The protocol resulted in 137 quantified metabolomic measurements, falling into the three molecular windows (c.f. Table 2.7). Among the 814 volunteers with serum samples collected, 757 had complete metabolomic measurements (i.e. no missing data).

Table 2.7 Types of quantified metabolomic measurements.

Window	No. of measurements	Description
LIPO window	91	Lipoprotein lipids
LMWM window	23	Low molecular weight metabolites
LIPID window	23	Lipid extracts

2.3 Data pre-processing

2.3.1 Quantifying lumbar disc degeneration

For ease of subsequent analysis, I devise a scheme for quantifying LDD, i.e. defining composite MRI phenotypes based on the reads by the two clinicians.

2.3.1.1 Truncated normal conversion of MRI reads

The truncated normal (*truncnorm*) distribution is a probability distribution derived from that of a normally distributed random variable by truncating the random variable from either below or above (or both).

Suppose X is a random variable following $N(\mu, \sigma^2)$. If we condition X on $a < X < b, -\infty \leq a < b \leq \infty$, it would follow a truncated normal distribution.

The probability density function (PDF) of a truncated normal distribution is:

$$f(x;\mu,\sigma,a,b) = \begin{cases} \frac{\phi(\frac{x-\mu}{\sigma})}{\sigma\left(\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})\right)} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\xi^2\right) \tag{2.2}$$

and

$$\Phi(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-t^2/2} dt \tag{2.3}$$

$\phi(\xi)$ and $\Phi(\xi)$ are, respectively, the PDF and cumulative distribution function of the standard normal distribution.

From Equation 2.1, we could observe the following.

- If $a = -\infty$, by definition, $\Phi\left(\frac{a-\mu}{\sigma}\right) = 0$. The distribution is either upper truncated (if $b < \infty$) or non-truncated (if $b = \infty$).
- If $-\infty < a < b < \infty$, the distribution is doubly truncated.
- If $b = \infty$, by definition, $\Phi\left(\frac{b-\mu}{\sigma}\right) = 1$. The distribution is either lower truncated (if $a > -\infty$) or non-truncated (if $a = -\infty$).

The truncated normal distribution could be used to transform the MRI reads (either boolean or ordinal) into continuous *truncnorm* scores.

Take SS (ordinal; possible values: $0, 1, 2, 3$) for instance. To start with, for each disc level, the standard normal distribution $N(0,1)$ was cut into four parts so that their areas were, respectively, the proportion of people with SS $= 0, 1, 2, 3$ for that disc level. Now there were altogether four truncated normal distributions, corresponding to the four areas. The mean of each truncated normal distribution[3] was then calculated, and SS was directly converted to these means (SS $= 0$ was converted to the mean of the truncated normal distribution corresponding to the leftmost area and so on). Figure 2.4 is an illustration of this procedure.

2.3.1.2 The relationship between MRI reads and disc levels

As could be seen in Figure 2.2, the severity of LDD-related conditions was highly dependent on disc levels. This association is further visualized in Figure 2.5 using conditional density plots [Meyer et al., 2017], which are generalized from spine plots[4] [Hummel, 1996].

[3]The mean of $truncnorm(\mu, \sigma, a, b)$ is $\mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$.

[4]Spine plots, or spinograms, are derived from stacked bar plots, where the widths of the bars reflect the relative frequencies of x and the heights of the bars correspond to the conditional relative frequencies of y in every x category [Hummel, 1996]. Spine plots discretize x, whereas conditional density plots perform smoothing on the explanatory variable [Meyer et al., 2017].

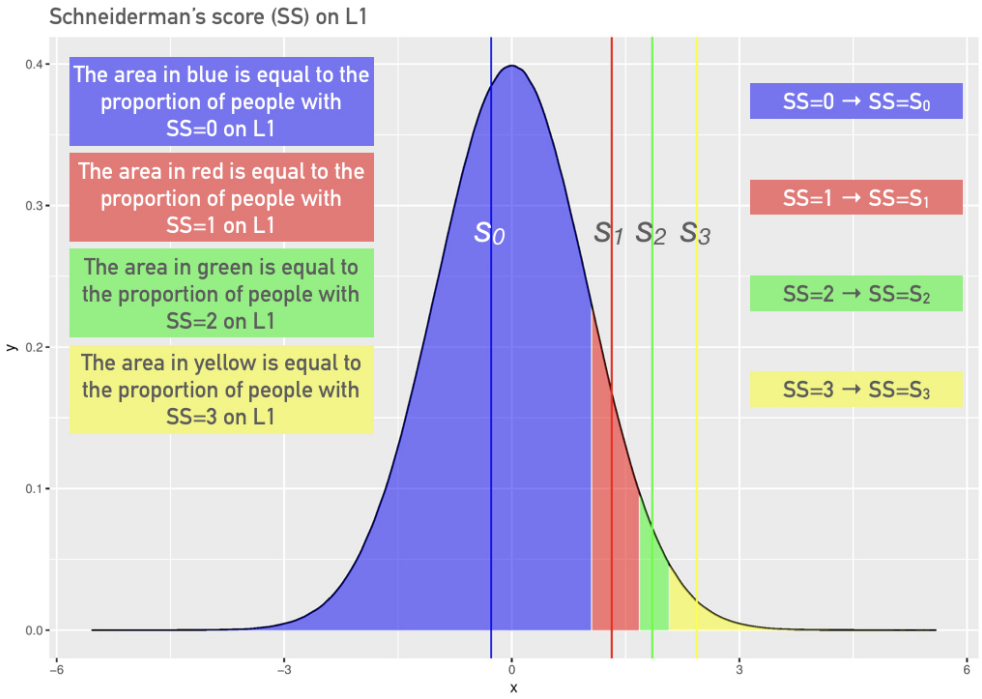


Fig. 2.4 Converting SS on L1 into *truncnorm* score. The truncated normal distributions would be different for each MRI measurement on each disc level.

Again we could observe from Figure 2.5 that (1) there were more discs suffering from DB, SIL, HIZ, and MC at lower disc levels; and (2) there were fewer discs with SN at lower disc levels. The validity of these observations could be examined through log-linear models.

Take SS for instance. We would like to test for complete independence between SIL score SS and disc level L using the data in Table 2.4, a 5×4 contingency table (here the column for missing data is omitted). Denote the probability of an observation falling into cell (i, j) as π_{ij} $(i = 1, \cdots, 5; j = 1, \cdots, 4)$, i.e. $SS = j - 1$ is observed on disc level $L = i$. The joint distribution of SS and L is then defined by π_{ij}. Furthermore, denote the observed count in cell (i, j) as y_{ij}, which is a realization of a random variable Y_{ij} following $Poisson(n\pi_{ij})$, where n is the total number of observations.

The null hypothesis of our model assumes complete independence between SS and L:

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \tag{2.4}$$

where $\pi_{i.}$ $(i = 1, \cdots, 5)$ is the marginal probability that an observation falls in row i of Table 2.4 and $\pi_{.j}$ $(j = 1, \cdots, 4)$ is the corresponding column margin.

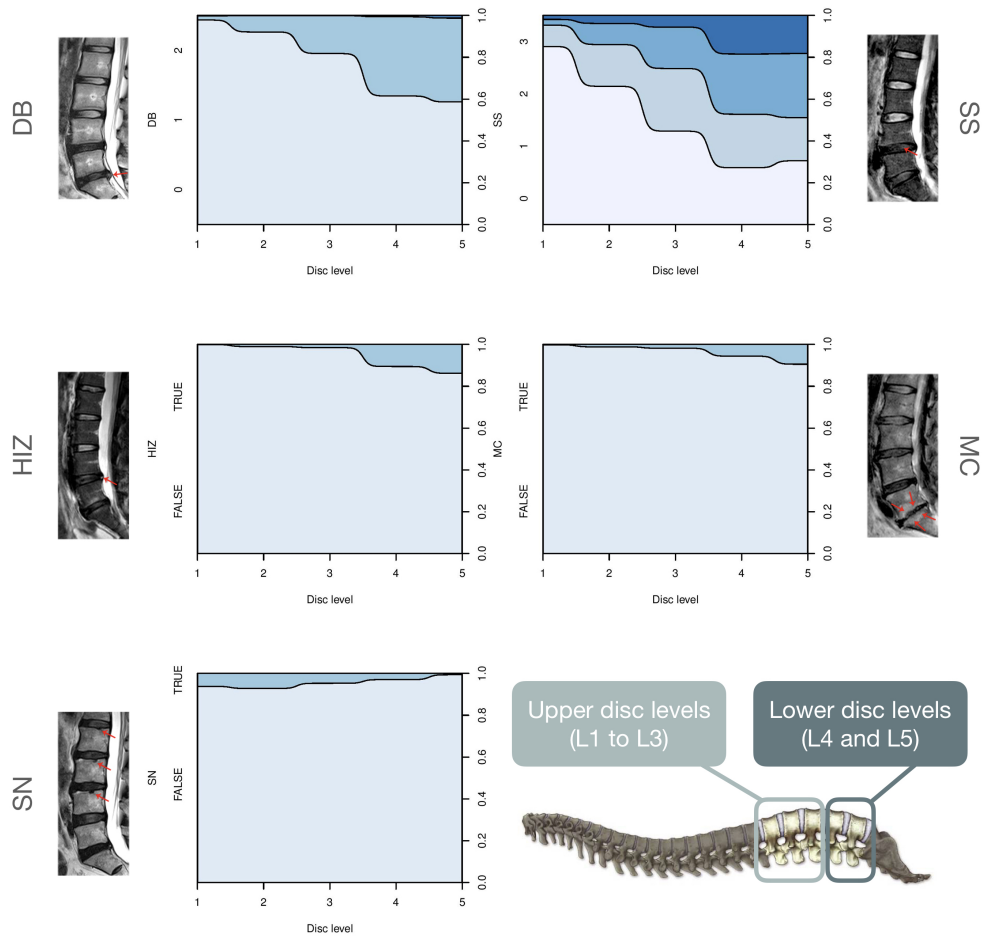


Fig. 2.5 Conditional density plots of DB, SS, HIZ, MC and SN over L1 to L5. MRI images courtesy of [Y. Li, 2016]. Spine image courtesy of [Cedars-Sinai, 2018].

The expected counts μ_{ij} under H_0 satisfy:

$$\log \mu_{ij} = \log n\pi_{ij} = \log n + \log \pi_{i.} + \log \pi_{.j} \tag{2.5}$$

The maximum likelihood estimators (MLEs) of μ_{ij} under H_0 then derived:

$$\hat{\mu}_{ij} = \frac{y_{i.} y_{.j}}{n^2} \tag{2.6}$$

To test for complete independence, I performed a likelihood ratio test comparing H_0 and the saturated model (where each cell in Table 2.4 has its own distribution).

Results of the fitted log-linear models are visualized in Figure 2.6. Generally speaking, there were (1) significantly more discs suffering from DB, SIL, HIZ, and MC at lower disc levels; and (2) significantly fewer discs with SN at lower disc levels. Additionally, The five disc levels seemed to form two clusters – {L1, L2, L3} and {L4, L5}. These are in line with past studies [Y. Li et al., 2016], and support that SN is more developmental versus the other four MRI phenotypes[5].

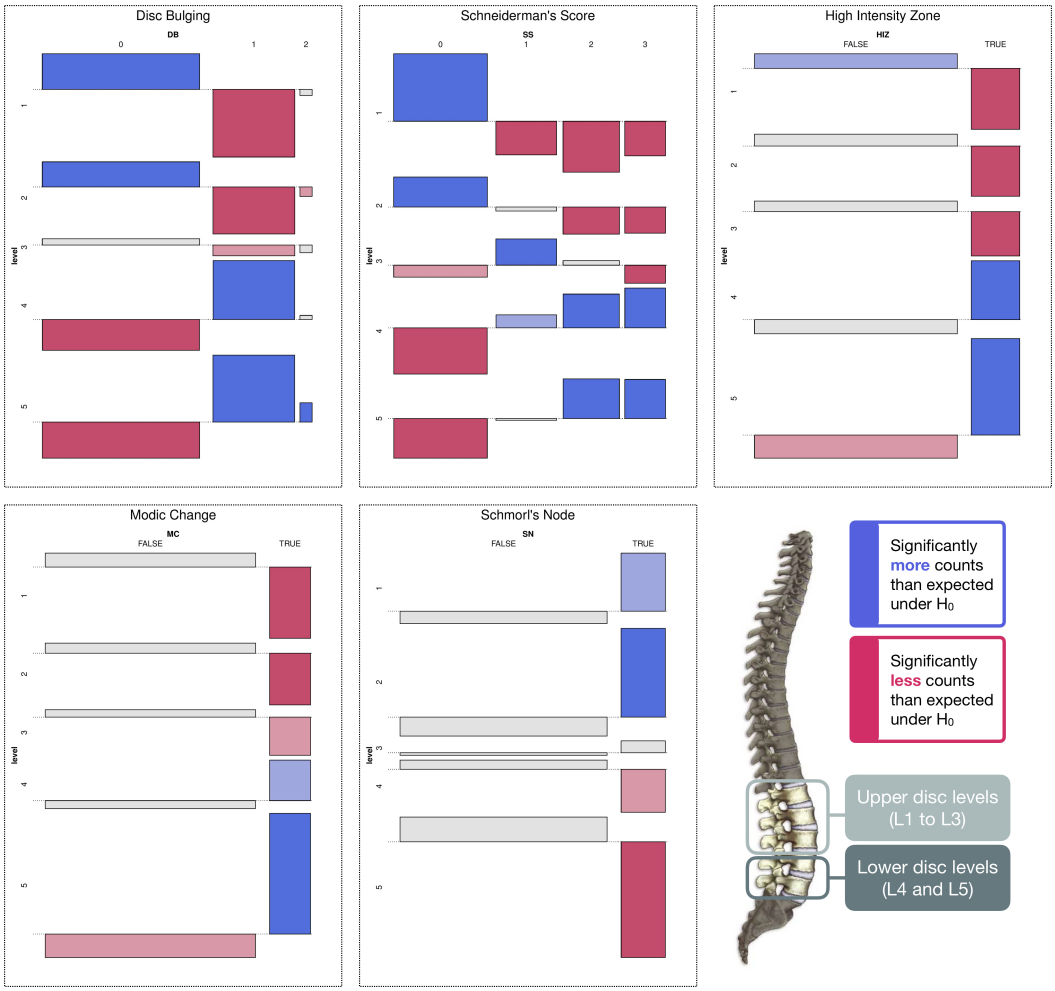


Fig. 2.6 Association plots indicating our data's deviation from log-linear models assuming complete independence between MRI phenotypes and disc levels. Spine image courtesy of [Cedars-Sinai, 2018].

[5]Note that in human beings, more weight is exerted on the lower disc levels on average since our posture is often upright. Therefore, if a spinal condition is more degenerative in nature, it would occur more frequently/severely at lower disc levels.

2.3.1.3 Defining composite MRI phenotypes

Since it is more meaningful to take into account an overall picture of the disc degeneration status, composite scores were defined as sums of certain *truncnorm*-converted single phenotypes, grouped according to disc level (L1 to L3 as upper, L4 and L5 as lower [Y. Li et al., 2016]), type of MRI phenotype, or the hypothesized fundamental cause of the condition (degenerative versus developmental).

The 25 continuous composite scores are listed below. Note that all the MRI measurements were first *truncnorm*-converted.

- Considering all types of MRI measurements (10 in total)
  - Degenerative score: SS L3, 4, 5 + DB L3, 4, 5 + HIZ L3, 4, 5 [Y. Li et al., 2016]
  - Developmental score: SN L1, 2, 3, 4, 5 + SS L1, 2 + DB L1, 2 [Y. Li et al., 2016]
  - Overall LDD severity: Add up all MRI measurements on all disc levels
  - Upper LDD severity: Add up all the MRI measurements on L1 to L3
  - Lower LDD severity: Add up all the MRI measurements on L4 and L5
  - Li LDD severity: Add up all the MRI measurements on Li ($i = 1, \cdots, 5$)
- Considering one type of MRI measurement ($15 = 3 \times 5$ in total)
  - MRI measurement *MRIM* (*MRIM* = DB, SN, SS, HIZ, MC)
    - Overall *MRIM*: Add up *MRIM* L1 to L5
    - Upper *MRIM*: Add up *MRIM* L1 to L3
    - Lower *MRIM*: Add up *MRIM* L4 and L5

The summary statistics of the continuous composite phenotypes could be found in Table 2.8. For the sake of comparison, I first divided the composite scores by the number of single MRI phenotypes used to calculate the score before calculating the summary statistics. For example, all the degenerative scores were divided by 9 prior to summary statistics calculation. In subsequent analysis, though, I would use the raw composite scores instead of the normalized (by division) ones.

Table 2.8 Summary statistics of the continuous composite MRI phenotypes.

Composite phenotype	Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum	# of non-NAs
Degenerative score	-0.5960	-0.3218	-0.0140	0.0034	0.2850	1.3540	1366
Developmental score	-0.1676	-0.1676	-0.1676	-0.0004	0.0756	1.2690	1414
Overall LDD severity	-0.2729	-0.1569	-0.0256	0.0012	0.1318	0.8320	1366
Upper LDD severity	-0.1600	-0.1600	-0.0842	0.0012	0.0975	0.9808	1374
Lower LDD severity	-0.4422	-0.2744	-0.0173	0.0023	0.1980	1.0360	1367
L1 LDD severity	-0.0669	-0.0669	-0.0669	-0.0003	-0.0669	1.5210	1375
L2 LDD severity	-0.1544	-0.1544	-0.1544	0.0013	0.1008	1.3220	1375
L3 LDD severity	-0.2586	-0.2586	-0.0314	0.0024	0.1283	1.4920	1374
L4 LDD severity	-0.4333	-0.4333	-0.0977	0.0033	0.2191	1.2300	1369
L5 LDD severity	-0.4511	-0.4511	-0.1340	0.0021	0.3791	1.2280	1373
Overall DB	-0.3646	-0.3646	-0.0538	-0.0011	0.2629	1.4890	1408
Upper DB	-0.1789	-0.1789	-0.1789	0.0000	-0.1789	2.1460	1416
Lower DB	-0.6431	-0.6431	0.1339	-0.0009	0.1487	2.6980	1408
Overall SS	-0.8161	-0.4410	-0.0789	-0.0013	0.3821	1.9040	1407
Upper SS	-0.5702	-0.5702	-0.1916	0.0002	0.3969	2.2090	1415
Lower SS	-1.1850	-0.7040	0.0889	0.0000	0.4470	2.7220	1408
Overall HIZ	-0.1058	-0.1058	-0.1058	0.0001	-0.1058	1.3130	1373
Upper HIZ	-0.023	-0.023	-0.023	0.000	-0.023	1.725	1375
Lower HIZ	-0.2301	-0.2301	-0.2301	0.0002	-0.2301	1.6630	1373
Overall MC	-0.0778	-0.0778	-0.0778	0.0001	-0.0778	1.7720	1373
Upper MC	-0.0277	-0.0277	-0.0277	0.0000	-0.0277	1.9040	1375
Lower MC	-0.1529	-0.1529	-0.1529	0.0001	-0.1529	1.8980	1373
Overall SN	-0.0939	-0.0939	-0.0939	0.0001	-0.0939	2.2100	1414
Upper SN	-0.1274	-0.1274	-0.1274	0.0000	-0.1274	1.9870	1416
Lower SN	-0.0438	-0.0438	-0.0438	0.0000	-0.0438	2.5440	1414

Since to clinicians studying LDD, MC is a phenotype with particular interest, I also define 6 binary composite phenotypes related to the existence and type of MC as below. Again, all the phenotypes were first *truncnorm*-converted.

- From JK's reads (3 in total)
  - Overall MC (binary): Whether there exists MC on any of the disc levels
  - Upper MC (binary): Whether there exists MC on L1 to L3
  - Lower MC (binary): Whether there exists MC on L4 or L5
- From DS's reads (3 in total)
  - Whether there exists MC
  - Whether there exists type 1 MC
  - Whether there exists type 2 MC

The summary statistics of the 6 binary composite MRI phenotypes are shown in Table 2.8. We could observe that MC generally worsened at lower disc levels, and type 2 MC was slightly more prevalent than type 1 MC. Only a small proportion ($< 21\%$) of individuals suffered from MC in the studied population. Note that the composite phenotypes "Overall MC (binary; JK)" and "Existence of MC (DS)" are almost identical to each other. However, based on the data set of $1,228$ samples read both by JK and MC, the phi coefficient[6] between the two phenotypes was estimated to be 0.6330, indicating the existence of a strong, yet not perfect, correlation. The different focus and habits of the two clinicians when reading MRI may be a confounding factor here.

Table 2.9 Summary statistics of the binary composite MRI phenotypes.

Composite phenotype	Proportion of trues	Proportion of falses	# of non-NAs
Overall MC (binary; JK)	0.1529	0.8471	1373
Upper MC (binary; JK)	0.0291	0.9709	1375
Lower MC (binary; JK)	0.1347	0.8653	1373
Existence of MC (DS)	0.2043	0.7957	1713
Existence of type 1 MC (DS)	0.0226	0.9773	1635
Existence of type 2 MC (DS)	0.1437	0.8563	1635

[6]The phi coefficient measures the association between two binary variables x and y, which could be obtained via estimating the Pearson correlation coefficient for x and y.

To sum up, to portray the degree of LDD of a certain individual, I have defined altogether 31 composite phenotypes based on the raw MRI reads of clinicians. Among the 31, there are 25 continuous and 3 binary ones calculated using JK's reads, as well as 3 binary ones related to MC type from DS's reads.

2.3.2 Pre-processing of metabolomic data

2.3.2.1 Data filtering

The purpose of conducting filtering on the metabolomic data (137 metabolomic measurements for 814 individuals, as described in Section 2.2.5) is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information was used in the filtering process; hence the result can be used with any downstream analysis.

From all the metabolomic measurements, 5% (7) with very small values (close to baseline or detection limit) were detected using the sample median. Additionally, 5% of the metabolomic measurements (7) with near-constant values throughout the experiment conditions (housekeeping or homeostasis) were identified using the interquartile range (IQR).

The same 7 measurements (listed below) were found via the two different approaches, which were filtered out and not used in the following analysis.

- Concentration of chylomicrons and extremely large very low density lipoprotein (VLDL) particles
- Concentration of very large VLDL particles
- Concentration of large VLDL particles
- Concentration of medium VLDL particles
- Concentration of small VLDL particles
- Concentration of very small VLDL particles
- Concentration of intermediate density lipoprotein (IDL) particles

Table 2.10 lists all the 130 metabolomic measurements studied in this thesis, as well as their molecular windows and abbreviations.

Table 2.10 Metabolomic measurements studied in this thesis.

Abbreviation	Molecular window	Full name
Bis.DB	LIPID	Ratio of bisallylic groups to double bonds
Bis.FA	LIPID	Ratio of bisallylic groups to total fatty acids
CH2.DB	LIPID	Average number of methylene groups per a double bond
CH2.in.FA	LIPID	Average number of methylene groups in a fatty acid chain
DB.in.FA	LIPID	Average number of double bonds in a fatty acid chain
DHA	LIPID	22:6, docosahexaenoic acid (DHA)
Est.C	LIPID	Esterified cholesterol
FALen	LIPID	Description of average fatty acid chain length (not actual carbon number)
FAw3	LIPID	Omega-3 fatty acids
FAw3.FA	LIPID	Ratio of omega-3 fatty acids to total fatty acids
FAw6	LIPID	Omega-6 and -7 fatty acids
FAw6.FA	LIPID	Ratio of omega-6/7 fatty acids to total fatty acids
FAw79S	LIPID	Omega-9 and saturated fatty acids
FAw79S.FA	LIPID	Ratio of omega-9 and saturated fatty acids to total fatty acids
Free.C	LIPID	Free cholesterol
LA	LIPID	18:2, linoleic acid (LA)
MUFA	LIPID	Monounsaturated fatty acids
otPUFA	LIPID	Other polyunsaturated fatty acids than 18:2
PC	LIPID	Phosphatidylcholine (and other cholines)
SM	LIPID	Sphingomyelins
TG.PG	LIPID	Ratio of triglycerides to phosphoglycerides
Tot.FA	LIPID	Total fatty acids
TotPG	LIPID	Total phosphoglycerides
Alb	LIPO	Albumin
ApoA1	LIPO	Apolipoprotein A-I (Lipido)

(Continued on next page)

Table 2.10 Metabolomic measurements studied in this thesis (cont'd).

Abbreviation	Molecular window	Full name
ApoB	LIPO	Apolipoprotein B (Lipido)
ApoB.ApoA1	LIPO	Apolipoprotein B by apolipoprotein A-I (Lipido)
HDL.C	LIPO	Total cholesterol in HDL
HDL.D	LIPO	Mean diameter for HDL particles
HDL2.C	LIPO	Total cholesterol in HDL2 (Lipido)
HDL3.C	LIPO	Total cholesterol in HDL3 (Lipido)
IDL.C	LIPO	Total cholesterol in IDL
IDL.C.eFR	LIPO	Total cholesterol in IDL (Lipido)
IDL.FC	LIPO	Free cholesterol in IDL
IDL.L	LIPO	Total lipids in IDL
IDL.PL	LIPO	Phospholipids in IDL
IDL.TG	LIPO	Triglycerides in IDL
L.HDL.C	LIPO	Total cholesterol in large HDL
L.HDL.CE	LIPO	Cholesterol esters in large HDL
L.HDL.FC	LIPO	Free cholesterol in large HDL
L.HDL.L	LIPO	Total lipids in large HDL
L.HDL.P	LIPO	Concentration of large HDL particles
L.HDL.PL	LIPO	Phospholipids in large HDL
L.LDL.C	LIPO	Total cholesterol in large LDL
L.LDL.CE	LIPO	Cholesterol esters in large LDL
L.LDL.FC	LIPO	Free cholesterol in large LDL
L.LDL.L	LIPO	Total lipids in large LDL
L.LDL.P	LIPO	Concentration of large LDL particles
L.LDL.PL	LIPO	Phospholipids in large LDL
L.VLDL.C	LIPO	Total cholesterol in large VLDL
L.VLDL.CE	LIPO	Cholesterol esters in large VLDL
L.VLDL.FC	LIPO	Free cholesterol in large VLDL
L.VLDL.L	LIPO	Total lipids in large VLDL
L.VLDL.PL	LIPO	Phospholipids in large VLDL
L.VLDL.TG	LIPO	Triglycerides in large VLDL
LDL.C	LIPO	Total cholesterol in LDL

(Continued on next page)

Table 2.10 Metabolomic measurements studied in this thesis (cont'd).

Abbreviation	Molecular window	Full name
LDL.C.eFR	LIPO	Total cholesterol in LDL (Lipido)
LDL.D	LIPO	Mean diameter for LDL particles (includes IDL particles)
M.HDL.C	LIPO	Total cholesterol in medium HDL
M.HDL.CE	LIPO	Cholesterol esters in medium HDL
M.HDL.FC	LIPO	Free cholesterol in medium HDL
M.HDL.L	LIPO	Total lipids in medium HDL
M.HDL.P	LIPO	Concentration of medium HDL particles
M.HDL.PL	LIPO	Phospholipids in medium HDL
M.LDL.C	LIPO	Total cholesterol in medium LDL
M.LDL.CE	LIPO	Cholesterol esters in medium LDL
M.LDL.L	LIPO	Total lipids in medium LDL
M.LDL.P	LIPO	Concentration of medium LDL particles
M.LDL.PL	LIPO	Phospholipids in medium LDL
M.VLDL.C	LIPO	Total cholesterol in medium VLDL
M.VLDL.CE	LIPO	Cholesterol esters in medium VLDL
M.VLDL.FC	LIPO	Free cholesterol in medium VLDL
M.VLDL.L	LIPO	Total lipids in medium VLDL
M.VLDL.PL	LIPO	Phospholipids in medium VLDL
M.VLDL.TG	LIPO	Triglycerides in medium VLDL
S.HDL.L	LIPO	Total lipids in small HDL
S.HDL.P	LIPO	Concentration of small HDL particles
S.HDL.TG	LIPO	Triglycerides in small HDL
S.LDL.C	LIPO	Total cholesterol in small LDL
S.LDL.L	LIPO	Total lipids in small LDL
S.LDL.P	LIPO	Concentration of small LDL particles
S.VLDL.C	LIPO	Total cholesterol in small VLDL
S.VLDL.FC	LIPO	Free cholesterol in small VLDL
S.VLDL.L	LIPO	Total lipids in small VLDL
S.VLDL.PL	LIPO	Phospholipids in small VLDL
S.VLDL.TG	LIPO	Triglycerides in small VLDL
Serum.C	LIPO	Serum total cholesterol

(Continued on next page)

Table 2.10 Metabolomic measurements studied in this thesis (cont'd).

Abbreviation	Molecular window	Full name
Serum.TG	LIPO	Serum total triglycerides
VLDL.D	LIPO	Mean diameter for VLDL particles
VLDL.TG	LIPO	Triglycerides in VLDL
VLDL.TG.eFR	LIPO	Triglycerides in VLDL (Lipido)
XL.HDL.C	LIPO	Total cholesterol in very large HDL
XL.HDL.CE	LIPO	Cholesterol esters in very large HDL
XL.HDL.FC	LIPO	Free cholesterol in very large HDL
XL.HDL.L	LIPO	Total lipids in very large HDL
XL.HDL.P	LIPO	Concentration of very large HDL particles
XL.HDL.PL	LIPO	Phospholipids in very large HDL
XL.HDL.TG	LIPO	Triglycerides in very large HDL
XL.VLDL.L	LIPO	Total lipids in very large VLDL
XL.VLDL.PL	LIPO	Phospholipids in very large VLDL
XL.VLDL.TG	LIPO	Triglycerides in very large VLDL
XS.VLDL.L	LIPO	Total lipids in very small VLDL
XS.VLDL.PL	LIPO	Phospholipids in very small VLDL
XS.VLDL.TG	LIPO	Triglycerides in very small VLDL
XXL.VLDL.L	LIPO	Total lipids in chylomicrons and extremely large VLDL
XXL.VLDL.PL	LIPO	Phospholipids in chylomicrons and extremely large VLDL
XXL.VLDL.TG	LIPO	Triglycerides in chylomicrons and extremely large VLDL
AcAce	LMWM	Acetoacetate
Ace	LMWM	Acetate
Ala	LMWM	Alanine
bOHBut	LMWM	3-hydroxybutyrate
Cit	LMWM	Citrate
Crea	LMWM	Creatinine
Glc	LMWM	Glucose
Gln	LMWM	Glutamine
Glol	LMWM	Glycerol

(Continued on next page)

Table 2.10 Metabolomic measurements studied in this thesis (cont'd).

Abbreviation	Molecular window	Full name
Gly	LMWM	Glycine
Gp	LMWM	Glycoproteins
His	LMWM	Histidine
Ile	LMWM	Isoleucine
Lac	LMWM	Lactate
Leu	LMWM	Leucine
MobCH	LMWM	Double bond protons of mobile lipids
MobCH2	LMWM	Mobile lipids -CH_2-
MobCH3	LMWM	Mobile lipids -CH_3
Phe	LMWM	Phenylalanine
Pyr	LMWM	Pyruvate
Tyr	LMWM	Tyrosine
Urea	LMWM	Urea
Val	LMWM	Valine

2.3.2.2 Data normalization

After data filtering, data normalization was performed on the metabolomic data using MetaboAnalyst [Xia et al., 2012], an online server for metabolomic data analysis.

Sample-wise, the data was normalized by sum to adjust for differences among the samples. Measurement-wise, auto-scaling[7] was performed to make metabolomic measurements, which are vastly different in terms of their magnitudes by nature, more comparable to one another. The normalization results could be seen in Figures 2.7 and 2.8.

[7]Mean-centered and divided by the standard deviation. After auto-scaling, the data would approximately have mean 0 and standard deviation 1.

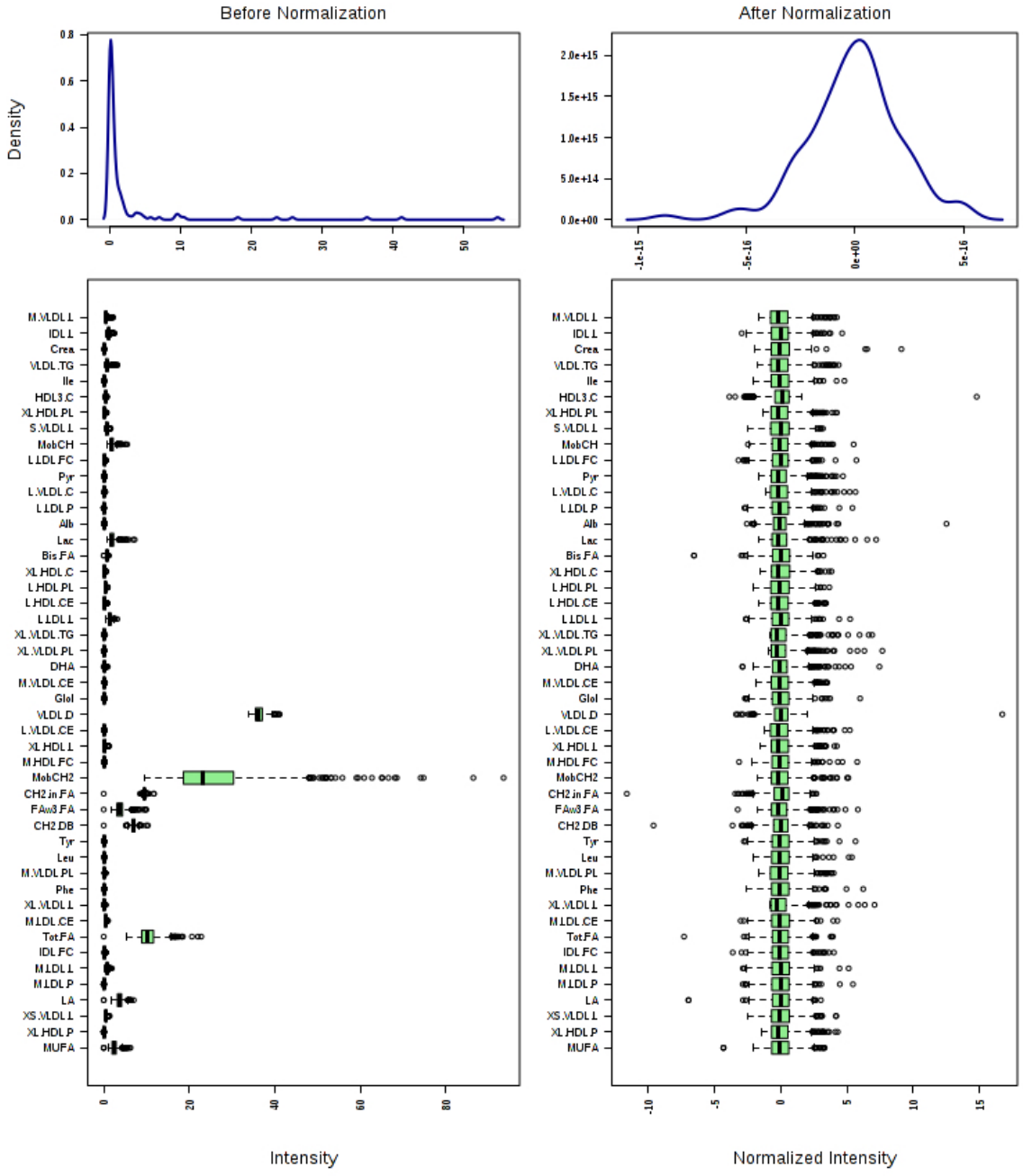


Fig. 2.7 Metabolomic data normalization results (measurement view). The boxplots only show 50 measurements due to space limitation; the density plots are based on all data.

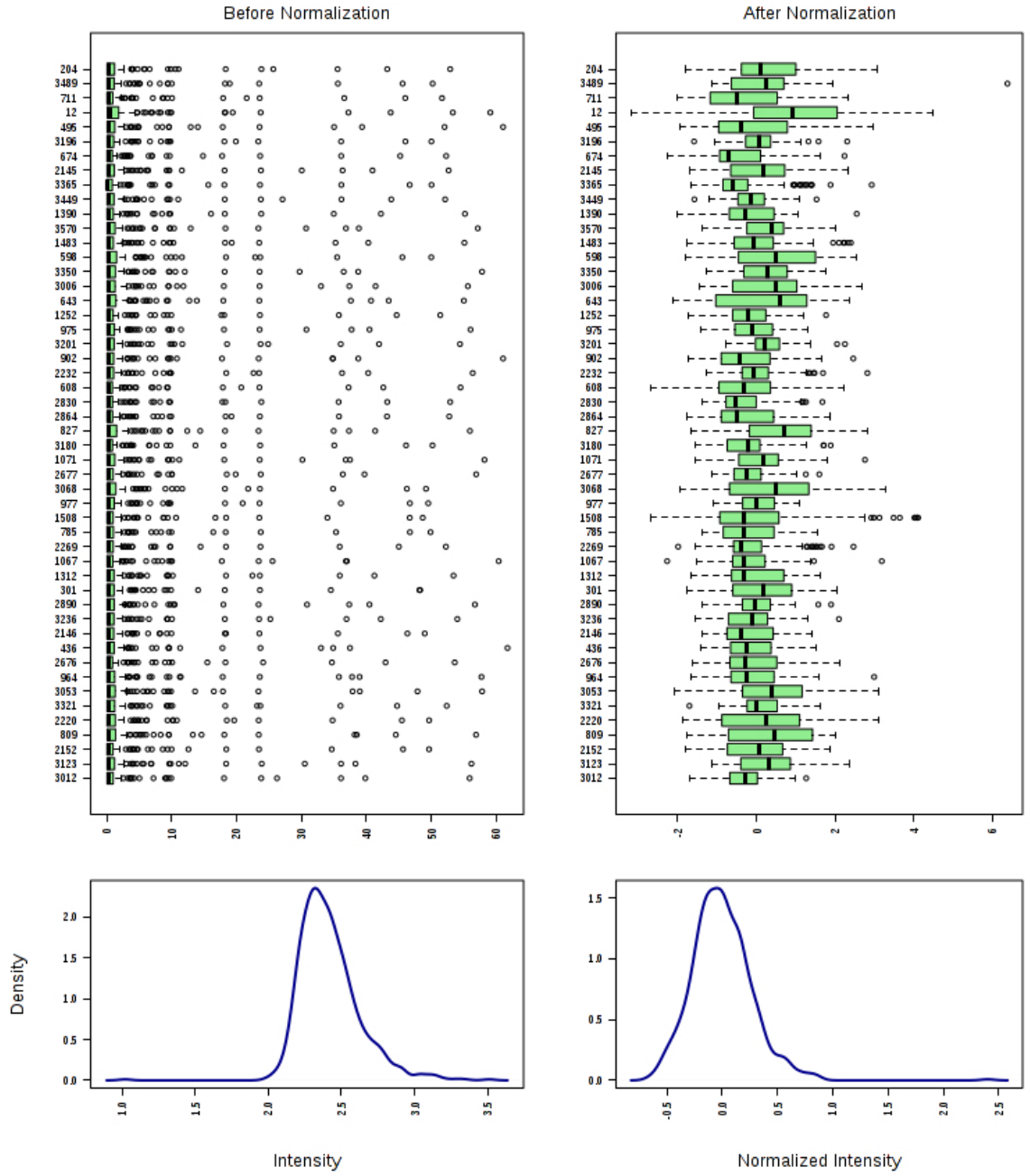


Fig. 2.8 Metabolomic data normalization results (sample view). The boxplots only show 50 measurements due to space limitation; whereas the density plots are based on all data.

2.3.2.3 Dimensionality reduction on metabolites

In data analysis, as the number of variables grows, so does the volume of the space and the available data would become "sparser". Consequently, the amount of data needed to support the statistical reliability of a result would increase[8]. Indeed, some problems will become intractable when the dimensionality of the data is too high – this is generally referred to as the curse of dimensionality.

Since there are 130 metabolomic measurements in the studied data set, any further analysis could suffer from this problem. Performing dimensionality reduction on metabolites could both help resolve this issue and identify new meaningful underlying variables.

In order to reduce the dimensionality of our data set, hierarchical clustering of metabolomic measurements was performed. Complete linkage was used as the agglomeration method to avoid chaining [Wilks, 2011] and the distance measure was set to be Kendall's τ [Kendall, 1938]. For two vectors x and y, Kendall's τ is defined as:

$$\tau = \sum_{i,j} K_{i,j}(x,y) \tag{2.7}$$

where

$$K_{i,j}(x,y) = \begin{cases} 0 & \text{if } x_i, x_j \text{ in same order as } y_i, y_j \\ 1 & \text{otherwise} \end{cases}$$

The reason for selecting Kendall's τ as the distance measure is two-fold. Firstly, the Pearson correlation coefficient[9] only depicts a linear relationship between two variables, whereas rank-based measures could capture all types of monotonic relationship. Secondly, when compared with other rank-based distances like Spearman's ρ, Kendall's τ benefits from having more interpretable and reliable confidence intervals [Newson, 2002].

As shown in Figure 2.9, the resulting dendrogram was cut into subtrees via dynamic tree cutting [Langfelder et al., 2007] with a cut-off of $\tau = 0.2$. Compared with cutting the tree at a fixed height, the dynamic tree cut algorithm benefits from being able to identify nested clusters [Langfelder et al., 2007]. In this study, the clusters were built bottom-up. The minimum cluster size was set to be 2 and the relative sensitivity to cluster splitting was controlled at a high level, producing a large number of clusters separated by small gaps.

[8] As a matter of fact, this amount increases exponentially with dimensionality [Donoho et al., 2000].

[9] Distance measures could be based on correlations, e.g. $1 - |corr(x,y)|$.

Each resulting subtree corresponded to one new metabolomic feature – either itself (if the subtree consisted of one metabolomic measurement) or the average of all the measurements in that subtree. In total, 66 metabolomic features (listed in Table 2.11) were defined, among them 35 composite and 31 single. The dimensionality of our metabolomic data was thus vastly decreased (by almost 50%). These metabolomic features defined via hierarchical clustering would be used in Section 5.2.4.

Table 2.11 Metabolomic features defined via hierarchical clustering.

Metabolomic feature	Metabolomic measurement(s)
gr1	Total lipids / Free cholesterol / Phospholipids in large LDL; Concentration of large LDL particles; Total cholesterol in LDL (Lipido)
gr2	Total lipids / Free cholesterol / Phospholipids / Triglycerides in medium VLDL; Triglycerides in VLDL
gr3	Total lipids / Total cholesterol / Free cholesterol / Phospholipids in small VLDL; Triglycerides in very small VLDL
gr4	Total lipids / Phospholipids in medium LDL; Concentration of medium LDL particles; Total cholesterol in LDL
gr5	Total lipids / Free cholesterol / Phospholipids in medium HDL; Concentration of medium HDL particles
gr6	Total lipids / Free cholesterol / Phospholipids in very large HDL; Concentration of very large HDL particles
gr7	Total lipids / Phospholipids / Triglycerides in very large VLDL; Free cholesterol in large VLDL
gr8	Mean diameter for LDL particles (includes IDL particles); Mean diameter for HDL particles; Average number of methylene groups in a fatty acid chain; Description of average fatty acid chain length (not actual carbon number)
gr9	Total cholesterol in HDL / HDL2 (Lipido); Apolipoprotein A-I (Lipido)
gr10	Phospholipids in large HDL, Total lipids in large HDL, Concentration of large HDL particles
gr11	Total cholesterol / Free cholesterol / Cholesterol esters in large HDL

(Continued on next page)

Table 2.11 Metabolomic features defined via hierarchical clustering (cont'd).

Metabolomic feature	Metabolomic measurement(s)
gr12	Triglycerides in small VLDL; Serum total triglycerides; Triglycerides in VLDL (Lipido)
gr13	Total lipids / Phospholipids / Triglycerides in large VLDL
gr14	Total lipids / Phospholipids / Triglycerides in chylomicrons and extremely large VLDL
gr15	Total lipids / Total cholesterol in small LDL; Concentration of small LDL particles
gr16	Average number of double bonds in a fatty acid chain; Ratio of bisallylic groups to double bonds; Ratio of bisallylic groups to total fatty acids
gr17	Total lipids / Phospholipids in very small VLDL; Triglycerides in IDL
gr18	Omega-3 fatty acids; Other polyunsaturated fatty acids than 18:2, 22:6; docosahexaenoic acid (DHA)
gr19	Total cholesterol / Cholesterol esters in medium HDL
gr20	Total cholesterol / Cholesterol esters in large LDL
gr21	Total cholesterol / Cholesterol esters in medium LDL
gr22	Total cholesterol / Cholesterol esters in very large HDL
gr23	Total cholesterol / Cholesterol esters in large VLDL
gr24	Serum total cholesterol; Esterified cholesterol
gr25	Total cholesterol / Cholesterol esters in medium VLDL
gr26	Total lipids / Total cholesterol in IDL
gr27	Total lipids in small HDL; Concentration of small HDL particles
gr28	Free cholesterol / Phospholipids in IDL
gr29	Total cholesterol in IDL (Lipido); Apolipoprotein B (Lipido)
gr30	Total phosphoglycerides; Phosphatidylcholine (and other cholines)
gr31	Omega-9 and saturated fatty acids; Total fatty acids
gr32	Omega-6 and -7 fatty acids; 18:2, linoleic acid (LA)
gr33	Mobile lipids -CH_3; Double bond protons of mobile lipids
gr34	Mobile lipids -CH_2-; Monounsaturated fatty acids
gr35	Leucine; Valine

(Continued on next page)

Table 2.11 Metabolomic features defined via hierarchical clustering (cont'd).

Metabolomic feature	Metabolomic measurement(s)
AcAce	Acetoacetate
Ace	Acetate
Ala	Alanine
Alb	Albumin
ApoB.ApoA1	Apolipoprotein B by apolipoprotein A-I (Lipido)
bOHBut	3-hydroxybutyrate
CH2.DB	Ave no. of CH2 per a double bond
Cit	Citrate
Crea	Creatinine
FAw3.FA	Ratio of omega-3 fatty acids to total fatty acids
FAw6.FA	Ratio of omega-6/7 fatty acids to total fatty acids
FAw79S.FA	Ratio of omega-9 and saturated fatty acids to total fatty acids
Free.C	Free cholesterol
Glc	Glucose
Gln	Glutamine
Glol	Glycerol
Gly	Glycine
Gp	Glycoproteins
HDL3.C	Total cholesterol in HDL3 (Lipido)
His	Histidine
Ile	Isoleucine
Lac	Lactate
Phe	Phenylalanine
Pyr	Pyruvate
S.HDL.TG	Triglycerides in small HDL
SM	Sphingomyelins
TG.PG	Ratio of triglycerides to phosphoglycerides
Tyr	Tyrosine
Urea	Urea
VLDL.D	Mean diameter for VLDL particles
XL.HDL.TG	Triglycerides in very large HDL

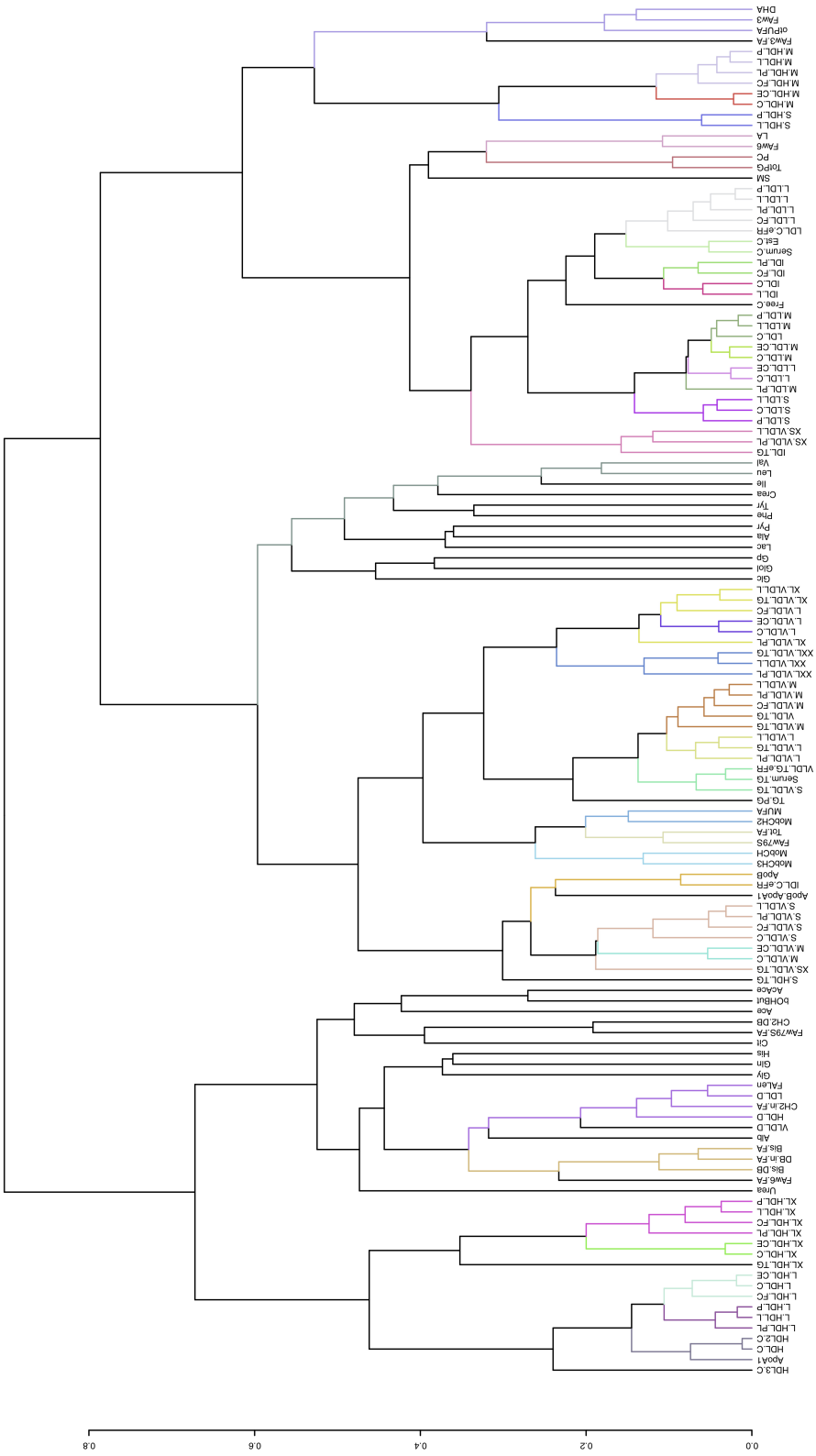


Fig. 2.9 Dendrogram cut via dynamic tree cutting. The black branches are "subtrees" containing only one metabolomic measurement, whereas the colored branches indicate subtrees clustering ≥ 2 metabolomic measurements.

2.4 Summary of integrated data

2.4.1 Basic description and sample sizes

The sample sizes of various types of data of the cohort are shown in Table 2.12. In this thesis, I consider altogether 40 phenotypes (c.f. Table 2.13) and only use the complete observations (e.g. for exploratory analysis of metabolomic data, I consider the 757) for analysis.

Table 2.12 Sample sizes of different types of data in the population cohort.

Data type	No. of observations	No. of complete observations
GWAS (before quality control)	2,482	/
Metabolomic	814	757
Phenotypic (composite scores – JK)	1,416	1,366 to 1,416
Phenotypic (composite scores – DS)	1,713	1,635 to 1,713

Table 2.13 Categories of phenotypes studied in this thesis.

Type	Count	Details
Basic	4	Height, weight, BMI, amount of cigarette smoking
Clinical	5	Lower back pain (binary), sciatica (binary), Oswestry disbility total score (continuous), VAS score on the test day (continuous) and severest VAS score ever experienced (continuous)
MRI	31	31 composite phenotypes (25 continuous and 3 binary from JK's reads; 3 binary related to modic change types from DS's reads)

The number of available matched samples upon data integration is demonstrated in Figure 2.10. Here data set I and data set II correspond to "Genotype data I" and "Genotype data II" in Figure 1.16. The studies deliberately use non-overlapping data sets to avoid possible over-fitting incurred by recycling data in "training" and "testing" phases.

The size of data set II depends on the studied phenotype. When taking the intersect between GWAS data and different phenotypes, the size of data set II is, respectively – basic phenotypes: 1,214; clinical phenotypes: 795; composite MRI phenotypes (JK): 750 to 769; composite MRI phenotypes (DS): 632.

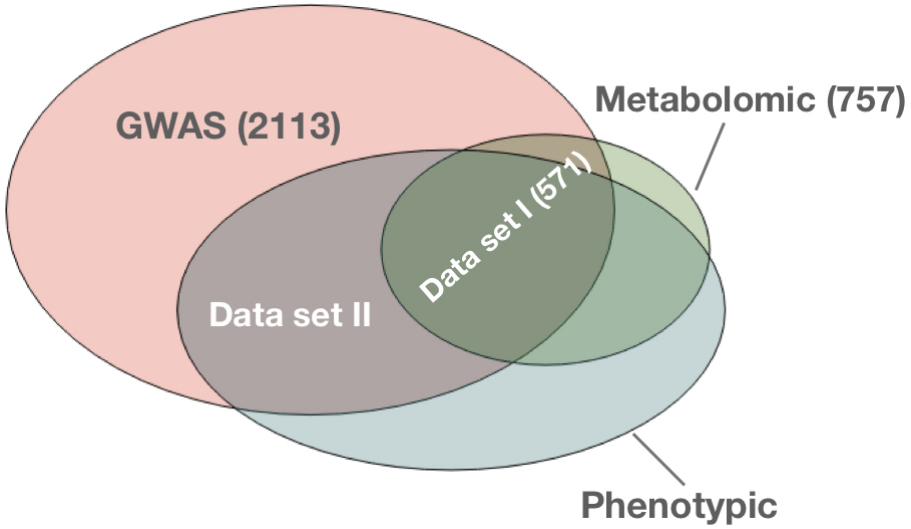
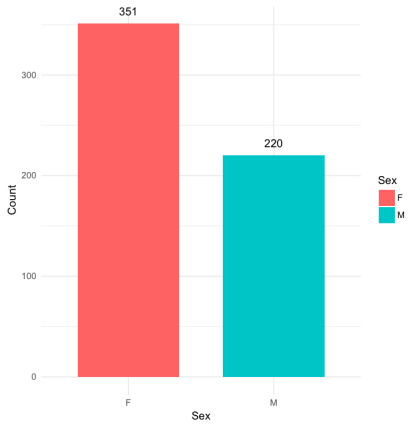


Fig. 2.10 Venn diagram showing the number of matched samples in the integrated data. Note that the areas of the ovals are not proportional to the corresponding sample sizes.

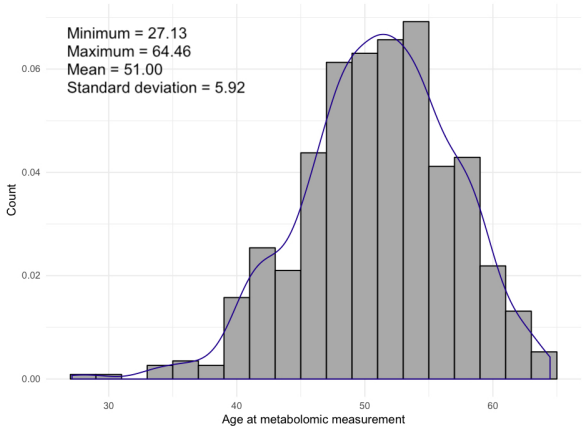
2.4.2 Descriptive statistics

The age and sex distributions of data set I are visualized in Figure 2.11. It could be observed that there were more females (61.47%) in the data set. Moreover, the data set mainly consisted of middle-aged and elderly people.

Since data set II varies with the phenotype of interest, we select one particular phenotype, namely overall LDD severity, for illustrating the age and sex distributions. The sample size of data set II for overall LDD severity was 751, and its age and sex distributions are visualized in Figure 2.12. It could be seen that again, there were relatively more females (60.59%) and people aged from 45 to 55 in the data set. The distributions were quite similar to Figure 2.11, except that the age distribution of data set II (overall LDD severity) was more skewed to the left, as well as had a slightly larger range and variance.

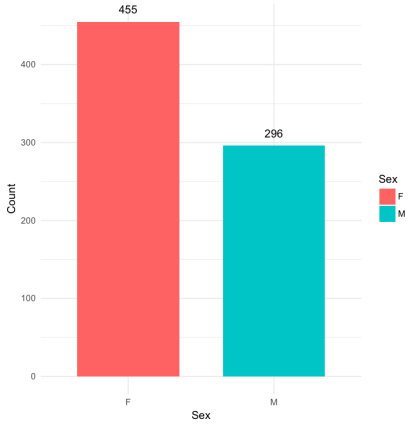


(a) Sex distribution of data set I.

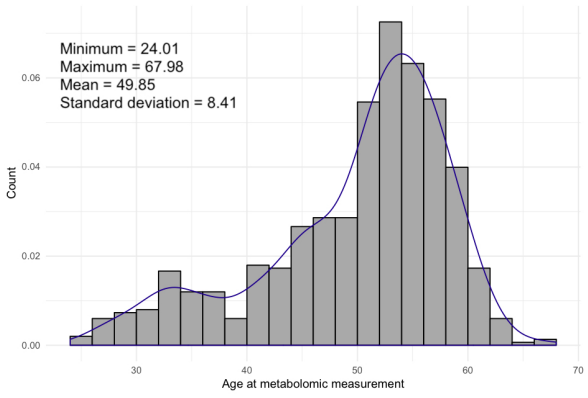


(b) Age distribution of data set I.

Fig. 2.11 Age and sex distributions of data set I.



(a) Sex distribution of data set II (overall LDD severity).



(b) Age distribution of data set II (overall LDD severity).

Fig. 2.12 Age and sex distributions of data set II (overall LDD severity).

3

Exploratory analysis of the serum metabolome and its phenotypic associations

3.1 Introduction

Low back pain (LBP), as a major cause of disability, is one of the most common global health problems [Kaplan et al., 2013]. LBP is often caused by lumbar disc degeneration (LDD) [Luoma et al., 2000], which could be evaluated by radiographic observations through magnetic resonance imaging (MRI) [M. A. Adams and Roughley, 2006]. Nevertheless, these MRI observations, as diagnostics for LBP/LDD, are prone to human error and may be insufficient in detecting variations of biological systems [Zenker et al., 2007]. Therefore, one of the urgent needs in LBP research is the identification of novel biomarkers for LDD, which could aid personalized diagnosis and treatment of LBP.

Previous genome-wide association studies (GWAS) have identified various genes associated with LDD [Eskola et al., 2012]. However, the genetic approach does not take into consideration the complex dynamics of the patient's biological environment, which limits its usage in personalized medicine [Nicholson and Wilson, 2003]. Metabolomics, on the other hand, studies an individual's metabolome at a given time and proves to be more useful in real-time diagnosis [Nicholson and Wilson, 2003].

Proton NMR metabolomics is an efficient technique to systematically quantify an individual's metabolome and gain molecular information regarding a variety of metabolites in different biofluids, for instance, serum [Mäkinen et al., 2008; Beckonert et al., 2003; Tang et al., 2004]. This chapter focuses on studying the direct association between serum ^{1}H NMR spectroscopy data and LDD related phenotypes. Additionally, self-organizing map (SOM) analysis is carried out to gain insight into the metabolomic continuum underlying LDD and other phenotypes.

3.2 Materials and methods

3.2.1 Study sample

As described in Section 2.2.5, the serum samples of 814 individuals were obtained for the application of ^{1}H NMR spectroscopy. For each individual, 137 metabolomic measurements were recorded, which belonged to one of the three molecular windows – lipoprotein lipids (LIPO), low-molecular-weight metabolites (LMWM) and lipid extracts (LIPID).

The metabolomic data set next underwent data filtering and normalization (c.f. Section 2.3.2) to reduce noise and increase the robustness of consequent analyses. After data pre-processing, the data set included 130 metabolomic measurements (c.f. Table 2.10) for 757 individuals.

This study also utilizes the composite LDD phenotypes defined in Section 2.3.1.3 based on the MRI reads of two experienced physicians, Dr. Jaro Karppinen (JK) and Dr. Dino Samartzis (DS). Among the 757 samples with metabolomic measurements, the MRI scans of 427 individuals were read by JK and those of 526 people were read by DS.

3.2.2 Correlation analysis

To assess the strength of the relationship between metabolomic measurements and LDD, the correlation between every metabolomic measurement and each of the 31 composite LDD phenotypes defined in Section 2.3.1.3 was calculated.

For each of the 25 continuous composite LDD phenotypes, the correlation between every metabolomic measurement and the LDD phenotype was calculated using Pearson's formula.

A two-sided test was also conducted to determine whether the correlation is significantly different from 0 (i.e. there exists a significant linear correlation).

For each of the 6 binary composite LDD phenotypes, the point biserial correlation (r_{pb}) [Olsson et al., 1982] was calculated instead. r_{pb} is a special case of Pearson's correlation coefficient measuring the relationship between one continuous variable X and one binary variable Y. Assume that the binary variable Y has two values 0 and 1. The data set could then be divided into two groups, $Y = 0$ and $Y = 1$. The formula for r_{pb} is:

$$r_{pb} = \frac{\mu_1 - \mu_0}{S_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{3.1}$$

where μ_i is the mean of X in the group with $Y = i$, n_i is the number of samples in the group with $Y = i$, n is the overall sample size, and S_n is the standard deviation (Equation 3.2).

$$S_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{3.2}$$

Again, two-sided tests were performed to evaluate whether the calculated point biserial correlations significantly deviate from 0.

3.2.2.1 Controlling for multiple testing

Since this study has calculated $130 \times 31 = 4,030$ correlations and hence performed $4,030$ statistical tests, the issue of multiple testing arose. For example, if I choose p-value $= 0.05$ as the significance threshold, the probability of observing at least one significant result due to chance is $1 - (1 - 0.05)^{4,030} \approx 1$.

Bonferroni correction is probably the easiest way to circumvent the multiple testing problem. It sets the p-value significance cut-off at $\frac{\alpha}{n}$, where α is the significance level and n is the number of tests. This approach is widely adopted in GWAS studies, where as a rule of thumb, a p-value of 5×10^{-8} (equivalent to a threshold of $\alpha = 0.05$ Bonferroni-corrected for 1 million independent variants) is set as the threshold for genome-wide significance. However, Bonferroni correction may lead to a high probability of type II errors[1]. Therefore, as a less

[1] A type II error occurs if we do not reject the null hypothesis (H_0) when it is false.

conservative alternative to Bonferroni-type adjustments, false discovery rate (FDR) control[2] is often recommended in health studies [Glickman et al., 2014].

In my study, the p-values P_i ($i = 1,\ldots,4,030$) were adjusted using the FDR approach, and the FDR was controlled at level $\alpha = 0.1$ through the Benjamini-Hochberg (B-H) procedure [Hochberg and Benjamini, 1990]:

1. Find the largest k such that $P_{(k)} \leq 0.1\frac{k}{m}$, where $m = 4,030$ is the total number of hypotheses tested.
2. Reject the null hypothesis (i.e. correlation $= 0$) for all $H_{(k)}$ for $i = 1,\ldots,k$.

Note that the FDR threshold $\alpha = 0.1$ means if there are altogether n significant adjusted p-values (q-values) at 0.1 cut-off, the number of expected falses among these findings would be $0.1n$.

Since there existed heterogeneity among the MRI phenotypes, the stratified FDR approach would perform better than aggregating all the tests [L. Sun et al., 2006]. Therefore, instead of lumping all $4,030$ p-values, the B-H procedure was carried out stratified by phenotype (i.e. 31 strata, 130 tests in each stratum).

3.2.3 Self-organizing map analysis

3.2.3.1 A brief introduction to self-organizing maps

Self-organizing map (SOM) is a type of neural network (NN) trained using unsupervised learning to represent multidimensional data in much lower dimensional spaces (often two; hence called a map) [Kohonen, 1998]. Since metabolomic data is of high dimensions by nature, SOMs are widely used in its analysis and visualization [Mäkinen et al., 2008; Beckonert et al., 2003; Xia et al., 2012].

Figure 3.1 illustrates a typical two-dimensional SOM. It consists of an input layer and a two dimensional "lattice" of neurons, each of which fully connected to the input layer. Each neuron has a topological position in the lattice and contains a vector of weights of the same dimension as the input vectors.

[2] A type I error occurs if we reject the null hypothesis (H_0) when it is true. The FDR is defined as the expected proportion of "false discoveries" among all discoveries. It conceptualizes the rate of type I errors when conducting multiple hypothesis tests.

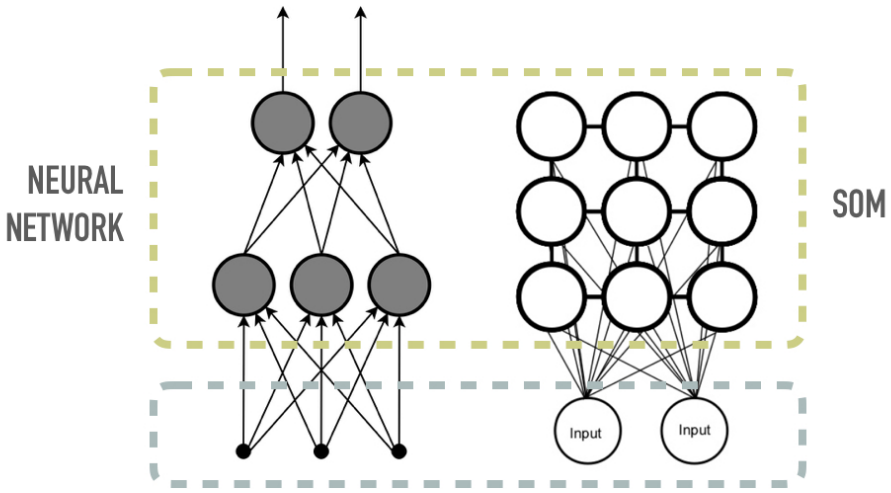


Fig. 3.1 Illustration of NNs and SOMs. Instead of having multiple hidden layers as in a typical NN, SOM feeds the input layer to a "lattice" of neurons. Note that the lines connecting the neurons in the lattice represent adjacency – there are no lateral connections between nodes within the lattice.

As opposed to other dimensional reduction methods (e.g. multidimensional scaling) which focus on maintaining the overall *dissimilarity* structure, SOMs emphasize localized *similarity*, i.e. in a trained SOM, on average, neighboring nodes have more similar weights to each other than those from the opposite sides of the map [Kohonen, 1998].

The training process of SOMs occurs in several steps and over many iterations:

1. Randomly initialize the weights.
2. Choose one training example at random and feed it to the lattice.
3. Find the best matching unit (BMU).
4. Determine the BMU's neighborhood.
5. Each neighboring node's weights are adjusted to be closer to the input vector. The closer a node is to the BMU, the more its weights get altered.
6. Repeat steps 2 to 5 n times. In each iteration, the size of the neighborhood used in steps 4 and 5 decreases.

3.2.3.2 Fitting a self-organizing map to metabolomic data

In this study, the training examples are the 757 samples in the metabolomic data set. A rule of thumb for choosing the SOM size is to have $5\sqrt{n}$ grid nodes in total, where n is the sample size [Vesanto and Alhoniemi, 2000]. In my case, the recommended SOM size would be $5 \times \sqrt{757} \approx 138$ nodes. It has also been found that non-symmetrical SOMs have fewer edge effects[3] [Kohonen, 1998]. Therefore, I chose a $9 \times 15 = 135$ hexagonal sheet of map units for the analysis (on average, 5.6 samples per node).

The R package "kohonen" [Wehrens and Buydens, 2007] was used to fit a SOM to the metabolomic data. To start with a globally optimal embedding, the SOM was first initialized with the space spanned by the first two eigenvectors attained from performing PCA on the metabolomic data [Wittek et al., 2013]. During training, the whole data set was presented to the map $1,000$ times, and the learning rate linearly decreased from 0.05 to 0.01. A Gaussian neighborhood was selected. The initial neighborhood[4] covered two-thirds of the map and decreased linearly in each iteration so that after 33 iterations, only the BMU was considered – we no longer adjusted the weights of neighboring neurons and the algorithm essentially became k-means.

Each neuron in the trained SOM represents a "metabolomic" neighborhood, and the individuals assigned to the same neuron share similar metabolomic characteristics. In the trained map, any two neighbors would, on average, have more similar metabolomic spectra than two randomly picked samples.

3.2.3.3 Quality evaluation of the fitted self-organizing map

The quality of the fitted SOM was evaluated using the map convergence index, which is defined to be the mean of the map's topographic accuracy and embedding accuracy [Hamel, 2016].

A fitted SOM's *topographic accuracy* measures how continuous it is. For training data $\{x_1, \ldots, x_n\}$, if the best matching and the second best matching units of x_i are adjacent, we declare local continuity; if not, there exists a local discontinuity (or a local topographic error) [Kiviluoto, 1996]. The entire map's topographic error is the total number of local topographic

[3] With edge effects, the training examples tend to be assigned to nodes clustered around edges of the map, leaving many empty neurons. This should be avoided [Vesanto and Alhoniemi, 2000].

[4] One of the parameters used in steps 4 and 5 of the SOM training algorithm, presented in Section 3.2.3.1.

errors divided by n [Kiviluoto, 1996]; the topographic accuracy is next defined as 1 minus the topographic error [Hamel, 2016].

When a SOM is completely *embedded*, there is no significant difference between the population of training samples and that of neurons in the map, i.e. the training data could be effectively represented by neurons in the SOM [Hamel and Ott, 2012; Hamel, 2016]. In my study, the embedding accuracy was calculated through the Kolmogorov–Smirnov test [Kolmogorov, 1933; Smirnov, 1948].

3.2.3.4 Coloring the fitted self-organizing map using phenotypic data

After fitting the SOM, the map was colored according to the phenotypic properties (i.e. the 40 phenotypes listed in Table 2.13) of different parts of the SOM. For each phenotype, the map was colored through the following steps:

1. Match phenotype and metabolomic data by sample ID.
2. For each node in the fitted SOM,
    - Gather all the samples assigned to it.
    - The "node statistic" $NodeStat$ is calculated.
        - If the phenotype is continuous, calculate the mean.
        - If the phenotype is binary (with/without a condition), calculate the proportion of samples with the phenotype.
3. Calculate k sample quantiles of the node statistics, where $k =$ number of colors in the sequential palette used.
4. Assign each node the color corresponding to the i^{th} sample quantile that is closest to $NodeStat$ of that node.

Recall that in the trained SOM, each node has a characteristic metabolomic profile. The phenotypic colorings could visualize the association between phenotypes and different underlying metabolomic patterns.

3.2.3.5 Statistical significance of self-organizing map colorings

To verify the statistical significance of one particular SOM coloring, we need to estimate the regional variability of the coloring and test if this "bumpiness" is significantly larger than that solely resulting from chance.

The bumpiness of the coloring pattern, or the "map statistic" *MapStat*, could be defined as:

$$MapStat = \sum_{i=1}^{135} [NodeStat_i - mean(NodeStat)]^2 \tag{3.3}$$

where 135 is the total number of nodes in the SOM.

The significance of the observed *MapStat* could be checked via permutation testing:

1. Shuffle the phenotype used for coloring for all the people with both phenotypic and metabolomic data. Now we have a pseudo-phenotype.
2. Match the pseudo-phenotype and metabolomic data by sample ID.
3. Calculate $PseudoNodeStat_i$ ($i = 1, \ldots, 135$) for the integrated pseudo-data. *NodeStat* is defined previously in Section 3.2.3.4.
4. Calculate *PseudoMapStat* for the integrated pseudo-data using Equation 3.3.
5. Repeat 1 to 4 n times. Now we have n *PseudoMapStat*.

The p-value from permutation testing is then asymptotically:

$$\text{p-value} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{ObservedMapStat \leq PseudoMapStat_j} \tag{3.4}$$

where $\mathbb{1}$ is the indicator function:

$$\mathbb{1}_{<condition>} = \begin{cases} 0 & \text{if <}condition\text{> is false} \\ 1 & \text{if <}condition\text{> is true} \end{cases}$$

Equation 3.4 is an approximation to the probability of attaining the observed *MapStat* (or a more extreme value) from the random distribution. As n increases, the approximation is closer to the truth. In this study, n was set to be $10,000$.

Next, to control for multiple testing, the 40 p-values (regarding the 40 SOM colorings – one for each phenotype) were adjusted using the B-H FDR procedure (c.f. Section 3.2.2.1). The p-values were stratified based on whether the phenotype is directly based on LDD/LBP[5]. If for a phenotype, the adjusted p-value of its coloring is less than 0.1, significance of the association between the phenotype and the underlying metabolomic continuum is declared.

3.3 Results

3.3.1 Correlation analysis

By performing $4,030$ correlation tests between 130 metabolomic measurements and 31 composite LDD phenotypes (defined in Section 2.3.1.3), 210 significantly correlated metabolomic-phenotypic pairs were identified[6].

Table 3.1 shows the significant results from the analysis. A vast majority of the significant correlations (203 in total) identified regards modic change (MC). The results are also visualized in Figure 3.2 utilizing the R package "corrplot" [T. Wei and Simko, 2017]. From the correlation plot, we could observe that type 1 MC tends to be negatively correlated with clusters of metabolomic measurements related to chylomicrons and large VLDL. It also correlates positively with clusters of large HDL related metabolomic measurements. On the contrary, upper MC and type 2 MC tend to have a negative correlation with clusters of large HDL related metabolomic measurements.

Additionally, the developmental score was found to be positively correlated with acetate with a q-value of 0.0279. The other two LDD phenotypes significantly correlated with certain metabolomic measurements (acetate and LDL related) both concerned the upper disc levels (L1, L2, and L3), which are also more developmental in nature compared with the lower ones. It has been found that acetate could function as an epigenetic metabolite to enhance lipid synthesis [Gao et al., 2016]. My results indicate that LDD and lipid metabolism possibly have shared genetic components.

[5]There were two strata – (1) weight, height, BMI and smoking (not directly based on LDD/LBP); and (2) the other 36 phenotypes.

[6]This study declares findings with q-value < 0.1 as significant. Therefore, among the 210, the number of expected false positives was $210 \times 0.1 = 21$.

Table 3.1 Significant correlations between LDD phenotypes and metabolomic measurements.

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
Developmental score	Ace	0.1768	0.0279
Upper LDD severity	Ace	0.1536	0.0621
Upper LDD severity	S.LDL.P	0.1615	0.0555
Upper LDD severity	S.LDL.C	0.1446	0.0879
Upper LDD severity	S.LDL.L	0.1606	0.0555
Upper SS	S.LDL.P	0.1593	0.0852
Upper SS	S.LDL.L	0.1540	0.0852
Overall MC (binary)	HDL2.C	-0.0553	0.0402
Overall MC (binary)	L.HDL.FC	-0.0630	0.0803
Overall MC (binary)	XL.HDL.PL	-0.0916	0.0115
Overall MC (binary)	XL.HDL.P	-0.0720	0.0803
Overall MC (binary)	Alb	-0.0118	0.0043
Overall MC (binary)	Ace	0.0239	0.0043
Overall MC (binary)	bOHBut	-0.1212	0.0000
Overall MC (binary)	AcAce	-0.0460	0.0000
Overall MC (binary)	MobCH2	0.0705	0.0264
Overall MC (binary)	L.VLDL.TG	0.0349	0.0203
Overall MC (binary)	XXL.VLDL.PL	0.1430	0.0005
Overall MC (binary)	XXL.VLDL.L	0.1219	0.0012
Overall MC (binary)	XXL.VLDL.TG	0.1047	0.0000
Overall MC (binary)	XL.VLDL.PL	0.1215	0.0055
Overall MC (binary)	L.VLDL.CE	0.0948	0.0851
Overall MC (binary)	XL.VLDL.TG	0.0775	0.0115
Overall MC (binary)	XL.VLDL.L	0.0935	0.0029
Overall MC (binary)	Glc	-0.0243	0.0000
Overall MC (binary)	Lac	-0.0154	0.0000
Overall MC (binary)	Pyr	0.1924	0.0115
Overall MC (binary)	Crea	-0.0714	0.0029
Overall MC (binary)	Leu	0.0868	0.0117
Overall MC (binary)	FAw3	0.0550	0.0180
Overall MC (binary)	DHA	-0.0043	0.0436

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
Upper MC (binary)	HDL2.C	-0.0752	0.0469
Upper MC (binary)	L.HDL.FC	-0.1178	0.0865
Upper MC (binary)	L.HDL.CE	-0.1822	0.0579
Upper MC (binary)	XL.HDL.PL	-0.1346	0.0445
Upper MC (binary)	XL.HDL.FC	-0.0853	0.0889
Upper MC (binary)	XL.HDL.P	-0.1106	0.0889
Upper MC (binary)	Alb	0.0135	0.0005
Upper MC (binary)	Ace	0.1608	0.0055
Upper MC (binary)	bOHBut	-0.3455	0.0000
Upper MC (binary)	AcAce	-0.2133	0.0000
Upper MC (binary)	MobCH2	-0.0427	0.0113
Upper MC (binary)	L.VLDL.TG	-0.0323	0.0092
Upper MC (binary)	L.VLDL.L	-0.0118	0.0640
Upper MC (binary)	M.VLDL.TG	-0.0533	0.0815
Upper MC (binary)	XXL.VLDL.PL	0.0890	0.0011
Upper MC (binary)	XXL.VLDL.L	0.0577	0.0005
Upper MC (binary)	XXL.VLDL.TG	0.0374	0.0000
Upper MC (binary)	XL.VLDL.PL	0.0590	0.0050
Upper MC (binary)	L.VLDL.C	0.0018	0.0640
Upper MC (binary)	L.VLDL.FC	0.0043	0.0257
Upper MC (binary)	XL.VLDL.TG	0.0065	0.0050
Upper MC (binary)	XL.VLDL.L	0.0215	0.0011
Upper MC (binary)	Glc	0.0243	0.0000
Upper MC (binary)	Lac	0.0255	0.0000
Upper MC (binary)	Pyr	0.1371	0.0257
Upper MC (binary)	Crea	0.1150	0.0046
Upper MC (binary)	Leu	0.2435	0.0243
Upper MC (binary)	L.LDL.P	0.1317	0.0889
Upper MC (binary)	M.HDL.C	-0.0336	0.0563
Upper MC (binary)	M.HDL.CE	-0.0476	0.0128
Upper MC (binary)	M.HDL.L	-0.0097	0.0445
Upper MC (binary)	M.HDL.P	-0.0005	0.0413

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
Upper MC (binary)	FAw3.FA	0.0489	0.0640
Upper MC (binary)	otPUFA	0.1543	0.0027
Upper MC (binary)	FAw3	0.0254	0.0055
Upper MC (binary)	DHA	0.0868	0.0221
Lower MC (binary)	HDL2.C	-0.0269	0.0328
Lower MC (binary)	L.HDL.FC	-0.0272	0.0513
Lower MC (binary)	L.HDL.C	-0.0214	0.0909
Lower MC (binary)	XL.HDL.PL	-0.0598	0.0110
Lower MC (binary)	XL.HDL.P	-0.0437	0.0705
Lower MC (binary)	Alb	-0.0263	0.0024
Lower MC (binary)	Ace	-0.0503	0.0047
Lower MC (binary)	bOHBut	-0.0744	0.0000
Lower MC (binary)	AcAce	-0.0113	0.0000
Lower MC (binary)	FAw79S	0.0680	0.0964
Lower MC (binary)	MobCH2	0.0811	0.0202
Lower MC (binary)	L.VLDL.TG	0.0386	0.0202
Lower MC (binary)	XXL.VLDL.PL	0.1273	0.0006
Lower MC (binary)	XXL.VLDL.L	0.1125	0.0014
Lower MC (binary)	XXL.VLDL.TG	0.0987	0.0000
Lower MC (binary)	XL.VLDL.PL	0.1180	0.0043
Lower MC (binary)	XL.VLDL.TG	0.0780	0.0135
Lower MC (binary)	XL.VLDL.L	0.0932	0.0031
Lower MC (binary)	Glc	-0.0456	0.0000
Lower MC (binary)	Lac	-0.0358	0.0000
Lower MC (binary)	Pyr	0.1741	0.0145
Lower MC (binary)	Crea	-0.1367	0.0043
Lower MC (binary)	Leu	0.0106	0.0241
Lower MC (binary)	S.LDL.P	0.0669	0.0328
Lower MC (binary)	FAw3	0.0598	0.0241
Lower MC (binary)	DHA	-0.0223	0.0769
MC exists (DS read)	HDL3.C	0.0031	0.0389
MC exists (DS read)	L.HDL.L	0.0317	0.0890

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
MC exists (DS read)	L.HDL.FC	0.0273	0.0213
MC exists (DS read)	L.HDL.C	0.0236	0.0308
MC exists (DS read)	L.HDL.CE	0.0223	0.0710
MC exists (DS read)	XL.HDL.C	-0.0358	0.0378
MC exists (DS read)	XL.HDL.PL	-0.0187	0.0036
MC exists (DS read)	XL.HDL.FC	-0.0209	0.0768
MC exists (DS read)	XL.HDL.L	-0.0334	0.0077
MC exists (DS read)	XL.HDL.P	-0.0306	0.0036
MC exists (DS read)	Alb	0.0341	0.0001
MC exists (DS read)	Ace	0.0236	0.0001
MC exists (DS read)	bOHBut	-0.0823	0.0000
MC exists (DS read)	AcAce	-0.0643	0.0000
MC exists (DS read)	MobCH	-0.0361	0.0706
MC exists (DS read)	FAw79S	-0.0914	0.0532
MC exists (DS read)	MobCH2	-0.0833	0.0028
MC exists (DS read)	L.VLDL.PL	-0.1265	0.0308
MC exists (DS read)	L.VLDL.TG	-0.1342	0.0026
MC exists (DS read)	L.VLDL.L	-0.1223	0.0659
MC exists (DS read)	M.VLDL.TG	-0.1147	0.0058
MC exists (DS read)	XXL.VLDL.PL	-0.0229	0.0000
MC exists (DS read)	XXL.VLDL.L	-0.0656	0.0000
MC exists (DS read)	XXL.VLDL.TG	-0.0901	0.0000
MC exists (DS read)	XL.VLDL.PL	-0.0889	0.0003
MC exists (DS read)	L.VLDL.FC	-0.0994	0.0061
MC exists (DS read)	XL.VLDL.TG	-0.1118	0.0005
MC exists (DS read)	XL.VLDL.L	-0.0990	0.0001
MC exists (DS read)	Glc	0.0773	0.0000
MC exists (DS read)	Lac	0.0034	0.0000
MC exists (DS read)	Pyr	0.0403	0.0010
MC exists (DS read)	Crea	-0.0273	0.0064
MC exists (DS read)	Leu	-0.0129	0.0092
MC exists (DS read)	FAw3.FA	0.0369	0.0035

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
MC exists (DS read)	otPUFA	0.0208	0.0233
MC exists (DS read)	FAw3	-0.0178	0.0092
MC exists (DS read)	DHA	-0.0466	0.0092
Type 1 MC (DS read)	HDL3.C	0.1679	0.0308
Type 1 MC (DS read)	L.HDL.FC	0.2434	0.0229
Type 1 MC (DS read)	L.HDL.C	0.2655	0.0138
Type 1 MC (DS read)	L.HDL.CE	0.2720	0.0168
Type 1 MC (DS read)	XL.HDL.C	0.1881	0.0091
Type 1 MC (DS read)	XL.HDL.CE	0.1921	0.0309
Type 1 MC (DS read)	XL.HDL.PL	0.2143	0.0014
Type 1 MC (DS read)	XL.HDL.FC	0.2064	0.0091
Type 1 MC (DS read)	XL.HDL.L	0.2097	0.0025
Type 1 MC (DS read)	XL.HDL.P	0.2104	0.0007
Type 1 MC (DS read)	Alb	0.0776	0.0007
Type 1 MC (DS read)	CH2.DB	0.0977	0.0073
Type 1 MC (DS read)	Ace	0.1741	0.0004
Type 1 MC (DS read)	bOHBut	0.0271	0.0000
Type 1 MC (DS read)	AcAce	0.0632	0.0000
Type 1 MC (DS read)	S.VLDL.PL	-0.2326	0.0361
Type 1 MC (DS read)	S.VLDL.L	-0.2435	0.0221
Type 1 MC (DS read)	MobCH	-0.1354	0.0495
Type 1 MC (DS read)	FAw79S	-0.1622	0.0702
Type 1 MC (DS read)	MobCH2	-0.1979	0.0025
Type 1 MC (DS read)	TG.PG	-0.2014	0.0710
Type 1 MC (DS read)	L.VLDL.PL	-0.3028	0.0144
Type 1 MC (DS read)	L.VLDL.TG	-0.2868	0.0007
Type 1 MC (DS read)	L.VLDL.L	-0.2747	0.0663
Type 1 MC (DS read)	M.VLDL.TG	-0.2902	0.0075
Type 1 MC (DS read)	M.VLDL.L	-0.2886	0.0431
Type 1 MC (DS read)	XXL.VLDL.PL	-0.0773	0.0000
Type 1 MC (DS read)	XXL.VLDL.L	-0.0964	0.0000
Type 1 MC (DS read)	XXL.VLDL.TG	-0.1306	0.0000

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
Type 1 MC (DS read)	XL.VLDL.PL	-0.1996	0.0008
Type 1 MC (DS read)	L.VLDL.FC	-0.2415	0.0075
Type 1 MC (DS read)	XL.VLDL.TG	-0.2162	0.0008
Type 1 MC (DS read)	XL.VLDL.L	-0.1959	0.0004
Type 1 MC (DS read)	Glc	-0.0156	0.0000
Type 1 MC (DS read)	Glol	0.0878	0.0309
Type 1 MC (DS read)	Lac	-0.1750	0.0000
Type 1 MC (DS read)	Pyr	-0.0080	0.0079
Type 1 MC (DS read)	Crea	-0.1298	0.0055
Type 1 MC (DS read)	Leu	-0.2259	0.0166
Type 1 MC (DS read)	FAw3.FA	-0.0036	0.0092
Type 1 MC (DS read)	otPUFA	-0.0527	0.0020
Type 1 MC (DS read)	FAw3	-0.1039	0.0091
Type 1 MC (DS read)	DHA	-0.0341	0.0037
Type 2 MC (DS read)	HDL3.C	-0.0598	0.0409
Type 2 MC (DS read)	L.HDL.FC	-0.0640	0.0542
Type 2 MC (DS read)	L.HDL.C	-0.0726	0.0606
Type 2 MC (DS read)	L.HDL.CE	-0.0752	0.0715
Type 2 MC (DS read)	XL.HDL.C	-0.1067	0.0567
Type 2 MC (DS read)	XL.HDL.PL	-0.1040	0.0052
Type 2 MC (DS read)	XL.HDL.FC	-0.0963	0.0542
Type 2 MC (DS read)	XL.HDL.L	-0.1150	0.0055
Type 2 MC (DS read)	XL.HDL.P	-0.1151	0.0024
Type 2 MC (DS read)	Alb	0.0313	0.0005
Type 2 MC (DS read)	CH2.DB	-0.0205	0.0052
Type 2 MC (DS read)	Ace	-0.0080	0.0003
Type 2 MC (DS read)	bOHBut	-0.0674	0.0000
Type 2 MC (DS read)	AcAce	-0.0423	0.0000
Type 2 MC (DS read)	MobCH	-0.0009	0.0886
Type 2 MC (DS read)	MobCH2	-0.0263	0.0032
Type 2 MC (DS read)	L.VLDL.PL	-0.0470	0.0660
Type 2 MC (DS read)	L.VLDL.TG	-0.0617	0.0010

(Continued on next page)

Table 3.1 (cont'd).

Phenotype	Metabolomic measurement	Correlation	Adjusted p-value
Type 2 MC (DS read)	M.VLDL.TG	-0.0461	0.0166
Type 2 MC (DS read)	XXL.VLDL.PL	0.0431	0.0001
Type 2 MC (DS read)	XXL.VLDL.L	-0.0022	0.0001
Type 2 MC (DS read)	XXL.VLDL.TG	-0.0219	0.0000
Type 2 MC (DS read)	XL.VLDL.PL	-0.0095	0.0006
Type 2 MC (DS read)	L.VLDL.FC	-0.0245	0.0094
Type 2 MC (DS read)	XL.VLDL.TG	-0.0374	0.0005
Type 2 MC (DS read)	XL.VLDL.L	-0.0254	0.0002
Type 2 MC (DS read)	Glc	0.1091	0.0000
Type 2 MC (DS read)	Glol	0.0625	0.0542
Type 2 MC (DS read)	Lac	0.0465	0.0000
Type 2 MC (DS read)	Pyr	0.0836	0.0023
Type 2 MC (DS read)	Crea	0.0168	0.0006
Type 2 MC (DS read)	Leu	0.0744	0.0176
Type 2 MC (DS read)	IDL.C	-0.0182	0.0660
Type 2 MC (DS read)	FAw3.FA	0.0361	0.0052
Type 2 MC (DS read)	otPUFA	0.0274	0.0711
Type 2 MC (DS read)	FAw3	0.0009	0.0059
Type 2 MC (DS read)	DHA	-0.0525	0.0020

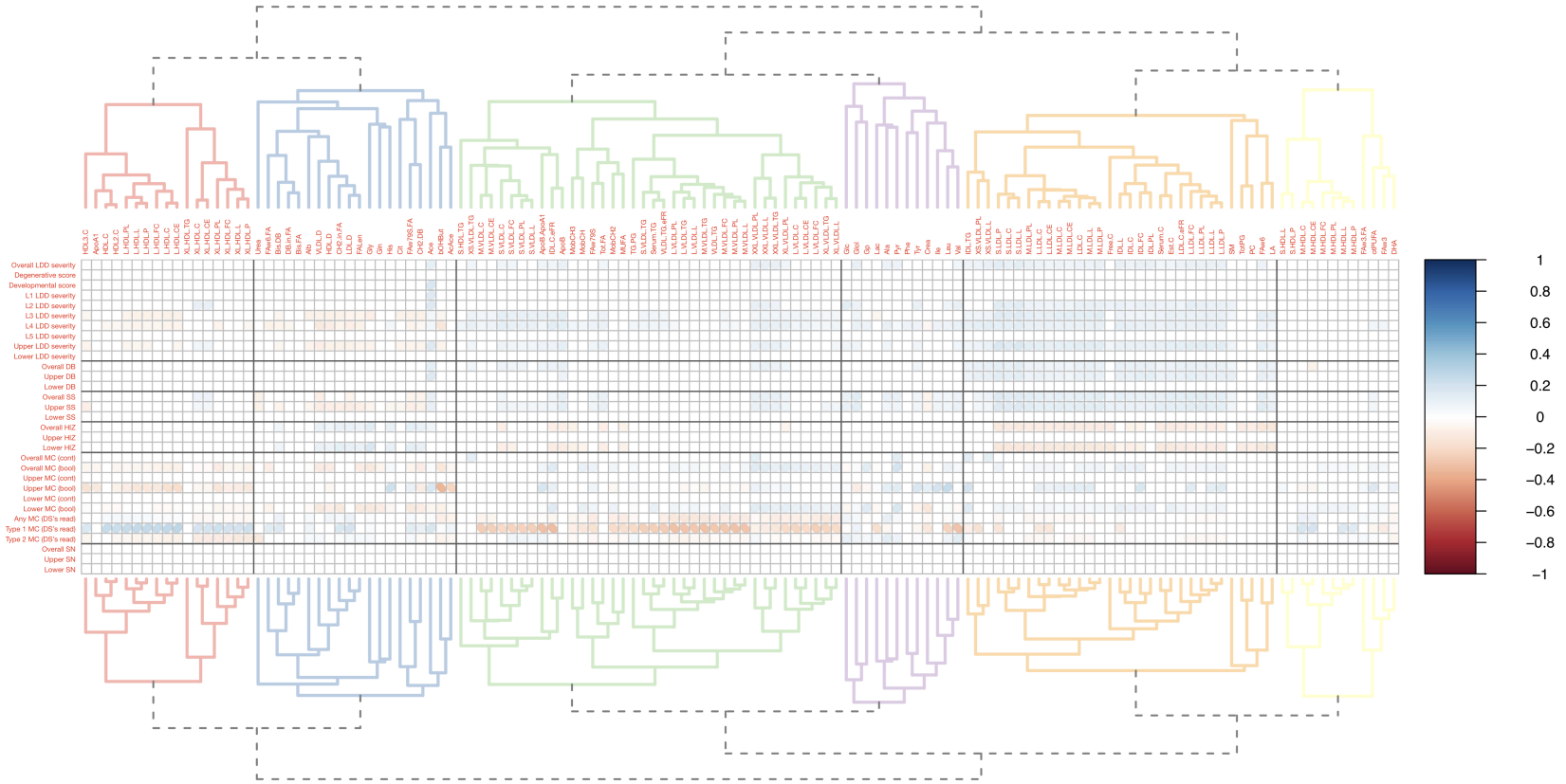


Fig. 3.2 Correlation plot between metabolomic measurements and LDD phenotypes. The metabolomic measurements are clustered using hierarchical clustering based on Kendall's correlation – the dendrogram is cut into six subtrees by height. The LDD phenotypes are categorized into (1) general LDD severity, (2) disc bulging, (3) signal intensity loss, (4) high intensity zone, (5) modic change and (6) Schmorl's node. Positive (Negative) correlations are colored blue (red), and the insignificant correlations are colored white (for visualization purpose, we set a more lenient significance threshold of q-value $= 0.5$ in this graph as opposed to 0.1 in Table 3.1). The color is darker when the correlation is of a larger magnitude.

3.3.2 Self-organizing map analysis

As described in Section 3.2.3.2, a SOM was constructed from the ^{1}H NMR data, reducing the 757 metabolomic spectra of different individuals into $9 \times 15 = 135$ representative spectral models. Each of the characteristic metabolomic spectra was assigned to a unique hexagonal unit in the SOM based on localized similarity, and the 757 samples were allocated to their best-matching cells rather uniformly (c.f. Figure 3.3).

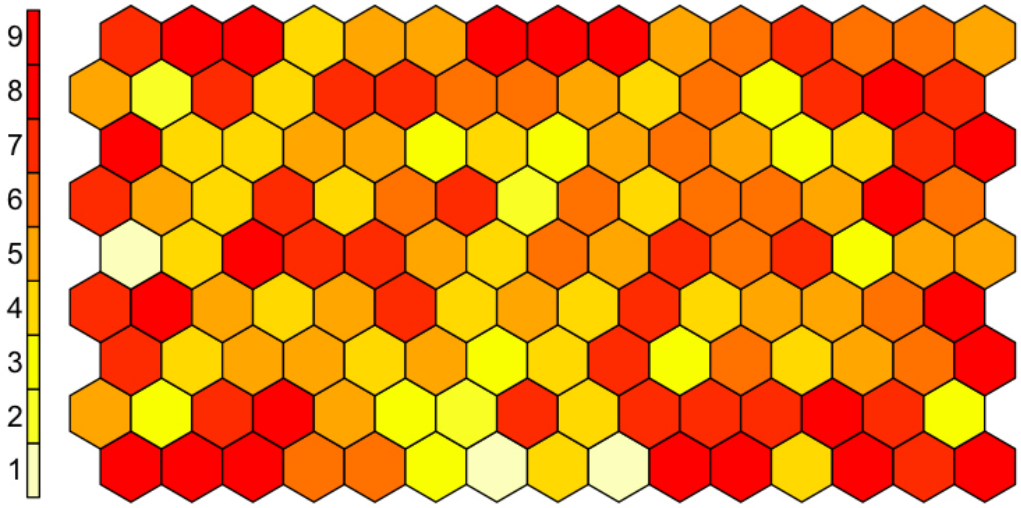
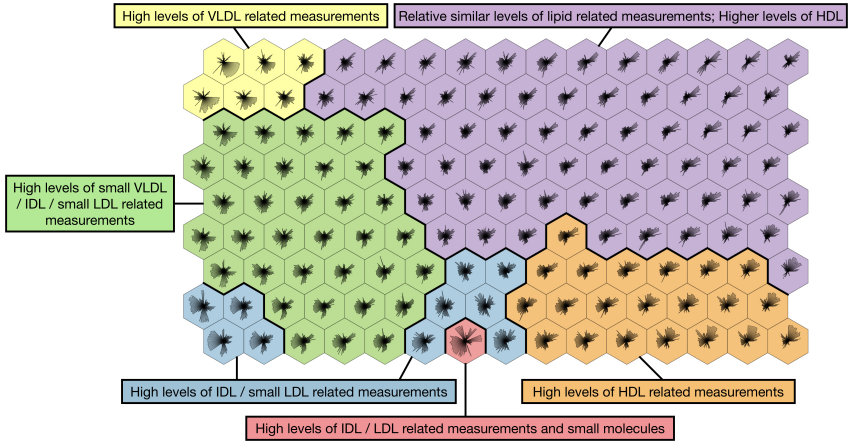


Fig. 3.3 Count plot of the fitted SOM. The number of samples falling into each cell ranges from 1 to 9, with an average of 5.61 and a median of 6. Each cell is colored according to the number of samples in it and the color scale on the left of the figure – 1 corresponds to light yellow and 9 corresponds to red.

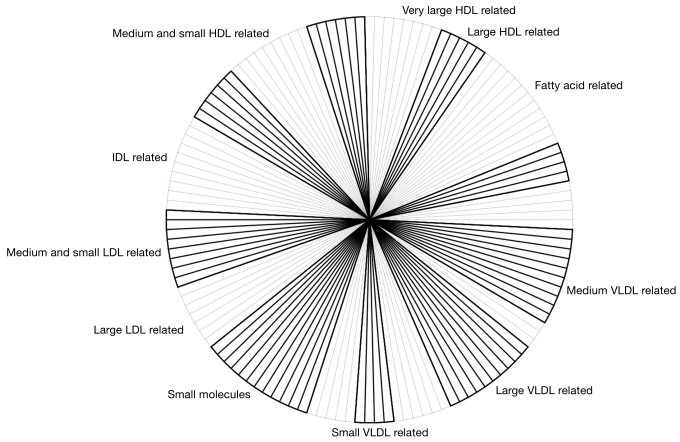
The topographic accuracy of the fitted map was estimated to be 0.94 with a 95% confidence interval of $(0.86, 1)$, whereas its embedding accuracy was approximately 0.7462. Hence, the estimated map convergence index was 0.84, implying a moderately good fit. Figure 3.4 shows the representative metabolomic profiles of each cell. It could be seen that people with higher levels of VLDL/IDL/LDL related metabolomic measurements tend to reside on the left side of the map, whereas the right part of the map (especially the bottom right corner) contains individuals with higher levels of HDL related measurements.

Two statistical colorings have been found to be significant through permutation testing. Weight (Figure 3.5) and BMI (Figure 3.6) both have an adjusted p-value of approximately 0.0448, indicating a significant association between them and the underlying metabolomic continuum. We could observe from the plots that weight and BMI seem to be positively

(negatively) correlated with VLDL/IDL/LDL (HDL) related metabolomic measurements. No significant associations have been found for the LDD related phenotypes.



(a) The characteristic metabolomic profiles of different cells are called the *codes* of the fitted SOM. The map is partitioned into 6 areas based on hierarchical clustering on its codes using complete linkage.



(b) Reference for the 135 star plots (subplots of Figure 3.4a shown in the hexagons). They visualize the codes of different map cells.

Fig. 3.4 Code plot of the fitted SOM.

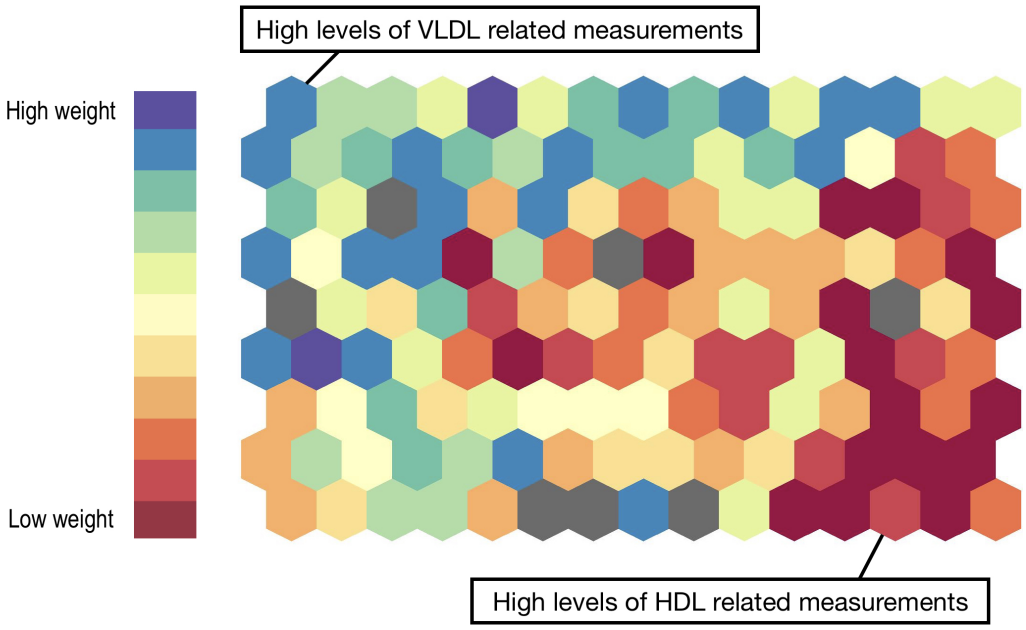


Fig. 3.5 Statistical coloring of the weight of samples on the SOM of 1H NMR spectra. Map units are colored according to the average weight of individuals allocated to it. Grey indicates NA due to missing phenotypic data in that cell. p-value ≈ 0.0134; adjusted p-value ≈ 0.0448.

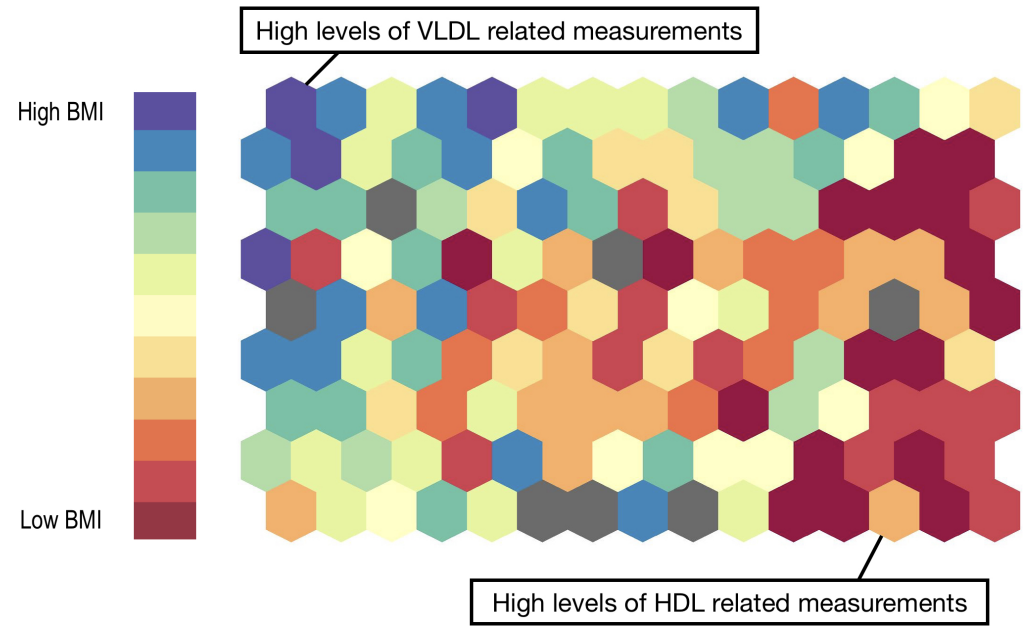


Fig. 3.6 Statistical coloring of the BMI of samples on the SOM of 1H NMR spectra. Map units are colored according to the average BMI of individuals allocated to it. Grey indicates NA due to missing phenotypic data in that cell. p-value ≈ 0.0224; adjusted p-value ≈ 0.0448.

3.4 Discussion

3.4.1 Correlation analysis

Through correlation analysis based on a population cohort, the first study presented in this chapter identified several significant associations between different metabolomic measurements and LDD related phenotypes. By conducting a population-based study, external validity[7] was ensured [Szklo, 1998; Wijmenga and Zhernakova, 2018]. Additionally, the LDD related phenotypes considered in this study were composite scores defined based on truncated normal converted MRI reads. The composite scores could better characterize the overall picture of an individual's disc degeneration status. The fact that most of the defined composite phenotypes were continuous (compared to the original MRI reads, which are binary or ordinal) also helped the study gain more statistical power [Altman and Royston, 2006].

The study found acetate, an epigenetic metabolite enhancing lipid synthesis [Gao et al., 2016], to be positively associated with the developmental component of LDD and upper LDD severity. Additionally, signal intensity loss (measured in terms of Schneiderman's score) and general LDD severity of the upper disc levels had a significantly positive correlation with small LDL related metabolites. These significant findings were all with respect to developmental LDD phenotypes – this is a possible indicator of shared genetic components between LDD and lipid metabolism.

Previous studies have also indicated that altered metabolism may contribute to LDD [Samartzis et al., 2013b; Ranjani et al., 2014]. Specifically, the association between serum lipid levels and disc herniation has been established in past research [Longo et al., 2011; Y. Zhang et al., 2016]. However, the precise underlying mechanism remains unclear. It has been hypothesized that dyslipidaemia[8] induces LDD through atherosclerosis or inflammatory pathways [Y. Zhang et al., 2016]. Further biological experiments need to be conducted to test the validity of this hypothesis and better understand the relationship between LDD and blood lipid levels.

My analysis also identified a variety of metabolites significantly associated with modic change (MC). The results regarding type 1 and type 2 MC were different; for instance, type

[7] The applicability of results of the study to a defined population (in our case, southern Han Chinese).
[8] Dyslipidaemia means an abnormal amount of lipids in the blood.

2 MC tended to be negatively associated with HDL related metabolites, whereas it was the opposite for type 1 MC.

It is well-established that type 2 MC is related to fatty degeneration [De Roos et al., 1987; Modic et al., 1988]. High amounts of oxidized LDLs (often accompanied by low HDL levels) activate TLR2/4 (toll-like receptor 2/4), and chronic stimulation of TLRs could facilitate fatty marrow conversion, inducing type 2 MC [Dudli et al., 2016]. Unfortunately, the pathology of type 1 MC is still not well understood. My results may indicate a distinct underlying metabolism of type 1 versus type 2 MC. Nevertheless, since in my data set, the class distribution of type 1 MC is highly imbalanced (only 2.16% positive), the results regarding type 1 MC may be biased. Future studies may benefit from a larger sample size (in order to diminish the drawback of imbalanced data) or a case-control study design (so that we have enough cases of type 1 MC). Besides, more biological research is needed to gain more insight into the pathology of type 1 MC.

3.4.2 Self-organizing map analysis

From the SOM analysis, we could observe a strong association between an individual's metabolomic profile (especially in terms of lipid-related measurements) and his or her weight/BMI. This is consistent with the current clinical knowledge [W. M. Miller et al., 2005].

In my study, no supervised feature selection was performed before the SOM analysis. This may lead to unsatisfactory predictive performance for clinical and LDD MRI phenotypes [Mäkinen et al., 2008]. Nevertheless, I have decided to adhere to the current unsupervised approach so that the fitted SOM is not methodologically dependent on any of the clinical phenotypes and hence easier to generalize to the southern Chinese population. Supervised models should be utilized in future research with disease risk prediction as the main purpose.

Unfortunately, no significant associations have been found for the LDD related phenotypes. This is probably due to the limited sample size of our metabolomic data. What's worse, the distributions of LDD phenotypes were generally highly skewed. For instance, a rather small proportion of people had modic changes, rendering a 0 average MC score for most of the SOM cells. Therefore, I may not have enough statistical power to identify true significant associations, if any. Upon collecting more metabolomic and LDD phenotypic data, it could be expected that representative metabolomic profiles from the new fitted SOM could be utilized

in personalized medicine as a high-throughput cost-effective alternative to a collection of specific metabolomic measurements [Ala-Korpela, 2008; Lindon et al., 2006].

4

Genome-wide association study for identification of single nucleotide polymorphisms associated with metabolomic measurements

4.1 Introduction

For a given biological sample, the metabolome refers to the complete set of chemicals within it at a given time. Rapid advances in the field of metabolomics now enable us to provide a "snapshot" of the human metabolome for cohorts with biological samples like serum, measuring hundreds of metabolomic traits at the same time. This "snapshot" could be read as a functional characterization of the metabolism and physiological state of the human body [Gieger et al., 2008], and it is intuitive to perform genome-wide scans of genetic biomarkers like single nucleotide polymorphisms (SNPs) for elements of this "snapshot" through genome-wide association studies (GWAS).

Many studies have integrated genomic and metabolomic data in human cohorts and identified a number of genetic loci associated with changes in metabolomic traits [Gieger et al., 2008; Illig et al., 2010; Rhee et al., 2013; Kettunen et al., 2016]. Since many of these identified genetic loci code enzymes or transport proteins directly affecting the disposition of a given metabolite [Gieger et al., 2008; Suhre and Gieger, 2012], the genetic variants typically

display much larger effect sizes compared to findings in GWAS for complex diseases [Gieger et al., 2008; Rhee et al., 2013]. The genetic biomarkers found could help us achieve a better understanding of the genetic roots of metabolomic measurements [Rhee et al., 2013], as well as the metabolomic context of different traits and conditions (e.g. lumbar disc degeneration) [Kettunen et al., 2016].

In this chapter, I scan the whole genome for SNPs significantly associated with different serum ^{1}H NMR metabolomic measurements, annotating all the significant SNPs identified. Following up the detected associations, meta-analysis is performed to increase power for polygenic scoring of metabolomic traits, which would be covered in the next chapter.

4.2 Materials and methods

4.2.1 Study sample

The serum samples of 814 individuals were obtained for the application of ^{1}H NMR spectroscopy, as described in Section 2.2.5. For each individual, we took 137 metabolomic measurements, which belong to one of the three molecular windows – lipoprotein lipids (LIPO), low-molecular-weight metabolites (LMWM) and lipid extracts (LIPID).

Next, the metabolomic data set underwent data filtering and normalization (c.f. Section 2.3.2) to reduce noise and increase the robustness of consequent analyses. After data pre-processing, the data set included 130 metabolomic measurements (c.f. Table 2.10) for 757 individuals. Among the 757 subjects, 571 also had GWAS data. Procedures for genotyping are described in Section 2.2.4.

4.2.2 Quality control

One thing about data analysis is that, your analysis results are only as good as (or, almost always, slightly worse than) your data. Garbage in, garbage out (GIGO) – when you feed the model with trash, all you receive in return is trash – the quality of output is partially determined by the quality of the input (c.f. Figure 4.1). That is why the success of GWAS, like that of any type of data analysis, depends on careful quality control (QC).

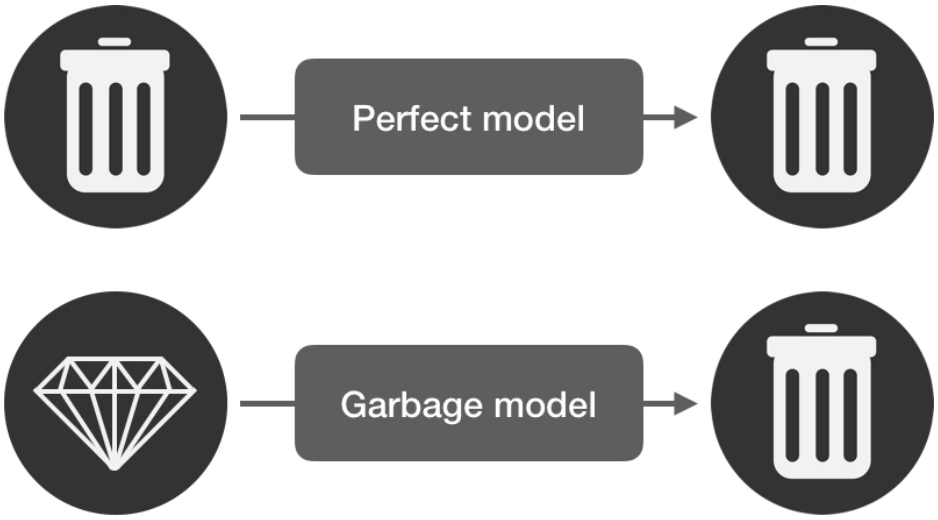


Fig. 4.1 Illustration of the GIGO principle. If the input (model) is garbage, even when the model (input) is perfect, which is unlikely in the real life, you would nevertheless end up with garbage results.

A lot of factors could go wrong in GWAS. Samples may be mislabeled, DNA might be contaminated, and genotyping is prone to error. Besides, the relatedness and underlying population structure of individuals in the cohort could lead to biased results. Since the numbers of SNPs and samples are generally quite big in GWAS, even if the error rate is suppressed to very low, there could be a very large number of false positive and false negative trait-variant associations. To ensure the accuracy of GWAS, both sample QC and variant QC need to be conducted.

4.2.2.1 Sample quality control

To guarantee the quality of samples used in GWAS, we filtered out bad quality samples with low SNP call rates, sample mislabeling, gender inconsistencies, sample contamination, relatedness and diverse ethnicity based on the genotype data for all $2,482$ individuals using PLINK [Purcell et al., 2007], following the pipeline used in [Y. Li, 2016].

SNP call rate checking

If a sample has a low SNP call rate (i.e. a high missing rate of SNP genotypes), it may be of poor quality or have undergone certain technical problems when genotyped. Therefore, such samples should be excluded from subsequent analysis.

We calculated the proportion of missing SNPs for each individual using PLINK [Purcell et al., 2007] and plotted the SNP call rates ($1-$ missingness) against their corresponding cumulative frequency of samples. As shown in Figure 4.2, the scatterplot elbows at around a 97% call rate. Hence, we dropped the 23 samples with over 3% missingness.

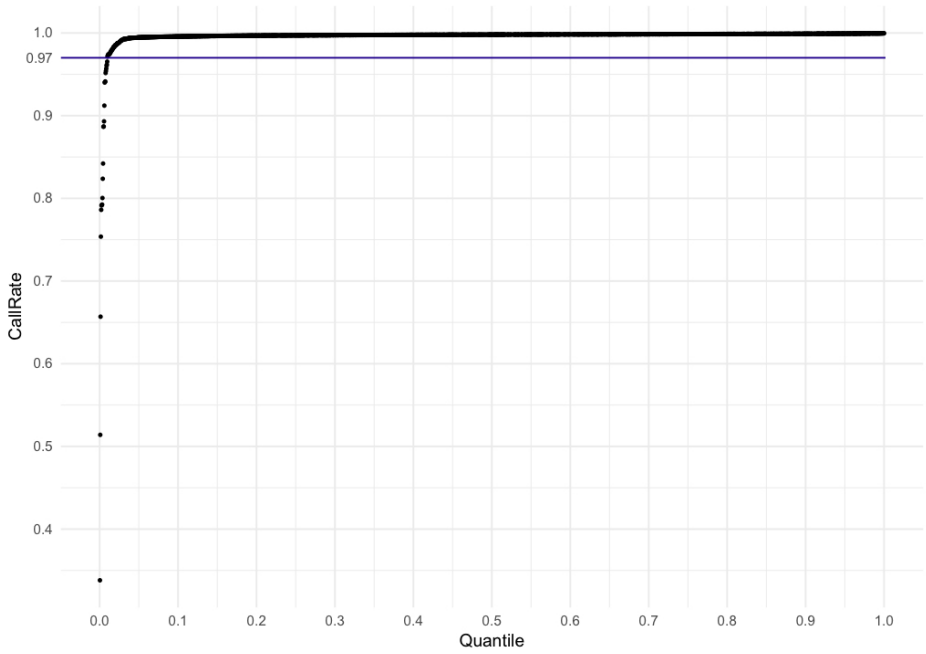


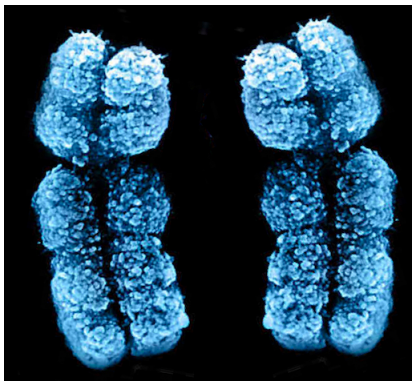
Fig. 4.2 Plot of SNP call rates of all the samples. The graph turns roughly at call rate $= 97\%$, which is chosen to be the QC threshold.

Sample mislabeling and gender checking

An individual's sex could be inferred based on his or her genotype. This inference could be made by estimating the X chromosome inbreeding (homozygosity) coefficient (F), which measures the difference between the observed and expected numbers of homozygous loci.

As shown in Figure 4.3, normal males have one X and one Y chromosome, while normal females have two X chromosomes. Since males have only one X chromosome, they are hemizygous for all sex-linked genes, making the observed homozygosity in males larger than expected by chance. On the contrary, females have two X chromosomes (just like all other autosomes), so the observed and expected homozygosity should be quite similar. Therefore, I classified samples with $F > 0.8$ (high observed homozygosity) as males and individuals with $F < 0.2$ (low observed homozygosity) as females using PLINK [Purcell et al., 2007]. If

$0.2 \leq F \leq 0.8$, PLINK would conclude that the sample has ambiguous gender [Purcell et al., 2007].



(a) Sex chromosomes of a normal female (XX). Modified based on the picture by [University of Toronto, 2014].

(b) Sex chromosomes of a normal male (XY). Modified based on the picture by [University of Toronto, 2014].

Fig. 4.3 Sex chromosomes of normal males and females.

Out of the 2,458 samples resulting from the previous step (SNP call rate QC), 43 individuals have genetically determined genders (inferred by PLINK [Purcell et al., 2007]) that are different from their "real" genders recorded in our database. Apart from those with gender discrepancy, 95 mislabeled samples (with unidentifiable ID) were pinpointed by referring to the phenotypic database. All the 138 $(43+95)$ individuals were dropped from our genotype data.

Heterozygosity and sample contamination checking

As mentioned in the previous section (gender checking), the inbreeding coefficient F is a measure of the extent of homozygosity in chromosomes. Hence intuitively, we could quantify the level of heterozygosity within samples by calculating F for only autosomal chromosomes (chromosomes 1 to 22 for humans). An elevated level of heterozygosity (i.e. very low F) may indicate cross-contamination of samples. On the other hand, if F is unusually high, there may be an excess of homozygous genotypes, which could be due to degraded DNA samples or inbred[1] subjects.

In my study, all the SNPs on autosomal chromosomes were used to estimate F with PLINK [Purcell et al., 2007]. 28 samples had F valued outside three standard deviations of the mean,

[1] Inbreeding means breeding between (close) relatives.

which is quite unlikely according to the three-sigma rule of thumb[2]. These samples may be contaminated or degraded, and were excluded from subsequent analysis.

Relatedness checking

One of the assumptions of GWAS is that all the samples for study are independent of each other. Biologically related individuals are likely to have a higher proportion of shared DNA sequences, and they may distort the overall distribution of allele frequencies, leading to biased GWAS conclusions. In light of this, we need to ensure that the data set for GWAS does not contain duplicated or genetically related people.

A common measure of relatedness (or duplication) between pairs of individuals is based on identity by descent (IBD). To understand this concept, first, think about the origin of life. Though how lives came into being is still a mystery, it is widely believed that all life today evolved by common descent from a single primitive life form. If we fast forward to today and only consider one of the current life forms – human beings, we could similarly conclude that in a finite population, all the individuals are related if traced back long enough[3].

If two people share certain nucleotide sequences in a DNA segment, this segment is identical by state (IBS) in them. If additionally, this segment is inherited from a common ancestor without recombination, the IBS segment is IBD in the individuals (c.f. Figure 4.4). Segments of IBD could be broken up by recombination during meiosis; therefore, the expected length of an IBD segment is related to the number of generations since the most recent common ancestor at the locus of the segment. Hence, the genetic relationship between two individuals could be tested based on the amount (both number and length) of IBD sharing. As long as a large number of SNPs is available, we could calculate genome-wide IBD given IBS information in a homogeneous sample [Purcell et al., 2007].

Prior to calculating IBD, we first applied LD pruning using a window size of $20,000$ SNPs, a step size of $2,000$ SNPs and a 0.5 pairwise SNP-SNP correlation (R^2) threshold[4]. LD pruning is a common way to keep only the markers not in LD with each other and reduce

[2] If X is an observation from a $N(\mu, \sigma^2)$ random variable, $Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$. Empirically, we could treat 99.7% probability as near certainty. Note also that $1 - \frac{28}{2320} \approx 0.9879$. If we perform a two-proportions z-test between 0.9879 and 0.9973 with sample sizes both equal to 2320, we could conclude that the two proportions are significantly different.

[3] Again, though I am an agnostic, I would like to say, it is a small world after all!

[4] This means we would (1) consider a window of $20,000$ SNPs, (2) calculate LD between each pair of SNPs in the window, (3) remove one of a pair of SNPs if LD > 0.5, (4) shift the window $2,000$ SNPs forward and (5) repeat steps 1 to 4.

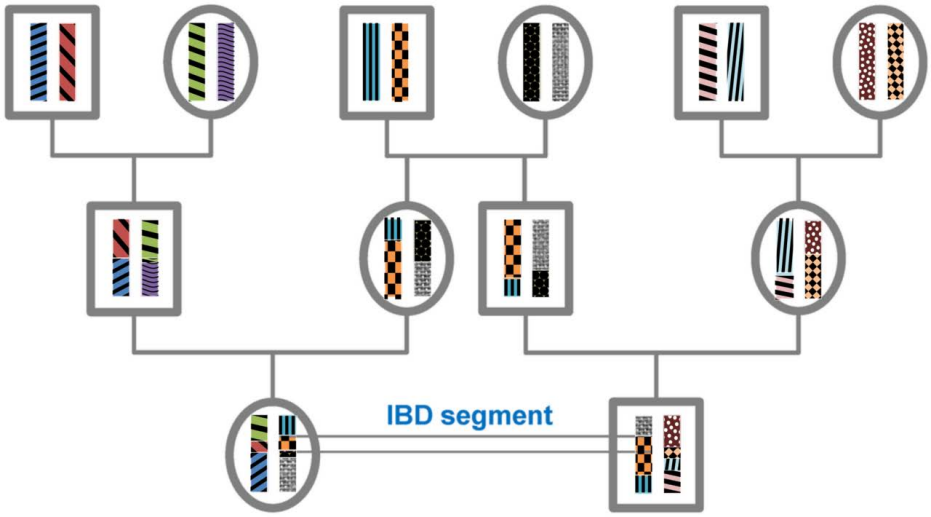


Fig. 4.4 IBD segments in a pedigree. We could see that the marked IBD segment in the two youngest individuals is from the maternal grandfather of the girl, who is one of the most recent common ancestors of the two individuals.

chromosomal artifacts' influences. It also renders the IBD calculation less computationally extensive.

Based on the pruned genotype data, IBD was calculated using PLINK [Purcell et al., 2007]. Let Z_0, Z_1 and Z_2 denote the probabilities of having IBD $= 0$, 1 or 2 over the loci. For a parent and his or her offspring, an ideal case would be $(Z_0, Z_1, Z_2) = (0, 1, 0)$, i.e. all loci have one IBD allele. As another example, for ideal full siblings, $(Z_0, Z_1, Z_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, i.e. 25% of loci have no IBD alleles, 50% have one IBD allele and the rest 25% have two IBD alleles.

The level of kinship could estimated by $\hat{\pi} = Z_2 + 0.5Z_1$. We could see that for both the parent-offspring and full siblings cases, an ideal $\hat{\pi}$ would be 0.5. Hence, a $\hat{\pi}$ close to 0.5 indicates first degree relatives. Some approximate values of $\hat{\pi}$ for other common relationships could be found in Table 4.1.

Table 4.1 Indications of the kinship coefficient $\hat{\pi}$.

Relationship	Approximate value of $\hat{\pi}$
Duplicated samples or monozygotic twins	1
First degree relatives	0.5
Second degree relatives	0.25
Third degree relatives	0.125
Completely unrelated	0

In our data set, we found 338 related or duplicated ($\hat{\pi} < 0.1$) samples falling into 159 groups. Out of each group, we randomly kept one of the individuals, and the rest (179) were removed from the data set.

Checking for diverse ethnicity

All the volunteers in our cohort identified themselves as Chinese. To check if any individuals are of a diverse ethnicity, samples were clustered using multidimensional scaling (MDS). Since there were no outliers in the MDS plot, no samples showed evidence of admixture and we may safely conclude that all the remaining $2,113$ individuals are indeed from a single population.

4.2.2.2 Variant quality control

After conducting QC on the individuals, variant QC was performed. SNPs of poor quality were removed, including those of low call rate, low minor allele frequency (MAF) and deviating from the Hardy-Weinberg equilibrium (HWE), based on the genotype data for all $900,015$ variants using PLINK [Purcell et al., 2007], following the pipeline used in [Y. Li, 2016].

SNP call rate and MAF checking

Minor allele frequency (MAF) refers to the frequency at which the minor allele (second most common) occurs in a given population. SNPs with a very low MAF (or even monomorphic) have little genetic variation and should be removed from the analysis.

In this study, all the rare variants (MAF < 0.005) were excluded. For common variants (MAF > 0.05), we dropped the SNPs with a $> 3\%$ missing rate. For the other variants ($0.005 \leq$ MAF ≤ 0.05), we removed the SNPs with a $> 1\%$ missing rate. As a result, $67,203$ SNPs were excluded from the following analysis.

Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) states that the allele and genotype frequencies in a population would remain constant in the absence of other evolutionary factors, including mate choice, mutation, natural selection, genetic drift and so on [Hardy et al., 1908; Weinberg, 1908]. Specifically, if we consider a single locus with two alleles denoted A and a, the genotype frequencies would have a stable ratio of $p^2 : 2pq : q^2$, where p and q are, respectively, the frequencies of A and a [Hardy et al., 1908; Weinberg, 1908].

Indications of significant deviation from the HWE include population stratification, selection, and genotyping errors. Therefore, we performed an exact test developed by [Wigginton et al., 2005] to exclude any SNPs deviating from the HWE using PLINK [Purcell et al., 2007]. $21,422$ SNPs did not pass the test (p-value $< 10^{-5}$) and were dropped from our data set.

Other exclusion criteria

In our data set, some SNPs were coded to be on chromosome 0 (control), 24 (chromosome Y), 25 (chromosome XY) or 26 (mitochondrial chromosome). These are generally not useful in GWAS so we dropped them. Besides, we also excluded the SNPs recorded to have a 0 morgan genetic distance. After excluding these variants, there were $805,525$ SNPs for further analysis.

4.2.2.3 Summary of GWAS quality control

All the steps of GWAS QC conducted in this thesis are summarized in Figure 4.5. After QC, the genotype data contained $2,113$ individuals and $805,525$ SNPs.

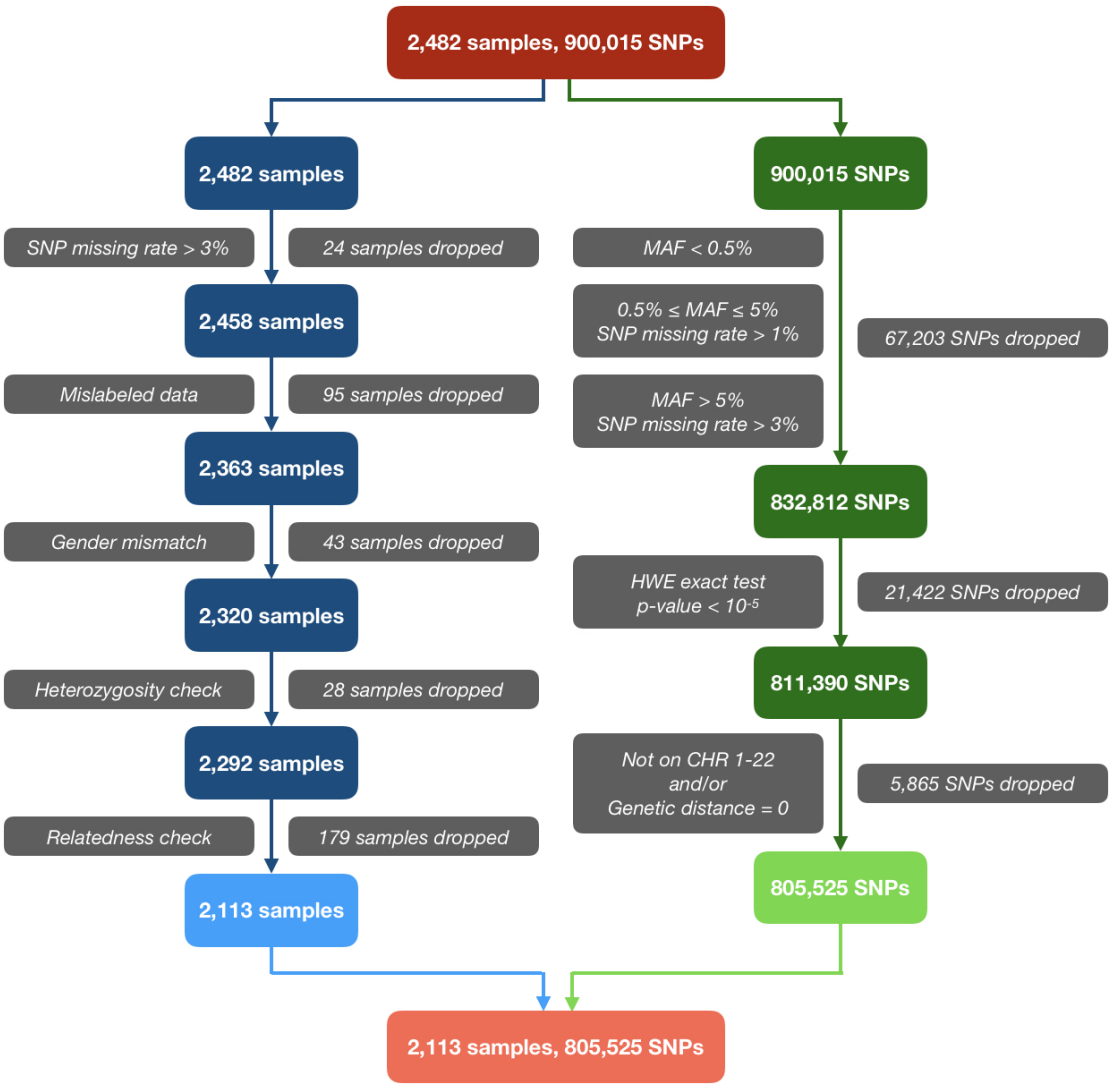


Fig. 4.5 Flow chart of GWAS QC. Variant QC is conducted after sample QC.

4.2.3 Correcting for population stratification

Population stratification refers to the existence of a systematic difference in allele frequencies between subpopulations within a given population. Hence, despite the fact that we have verified all the samples to be of Chinese ancestry during QC, we still need to adjust for population stratification in order to reduce false positives resulting from ancestral differences unrelated to metabolomic measurements (the traits we aim to analyze in GWAS).

After matching our metabolomic data and the GWAS data after QC, we had 571 individuals in total. EIGENSTRAT [Price et al., 2006] was used to model the ancestral differences based on the genomic data of these individuals via PCA. Using the top principal components

as covariates could correct for population stratification in GWAS [Price et al., 2006]. We selected the first 10 principal components as covariates since judging from the PCA's scree plot (c.f. Figure 4.6), the first 10 could account for most of the variation. Besides, we need to avoid using too many principal components and hence losing too much power since the sample size of our GWAS data set is not huge.

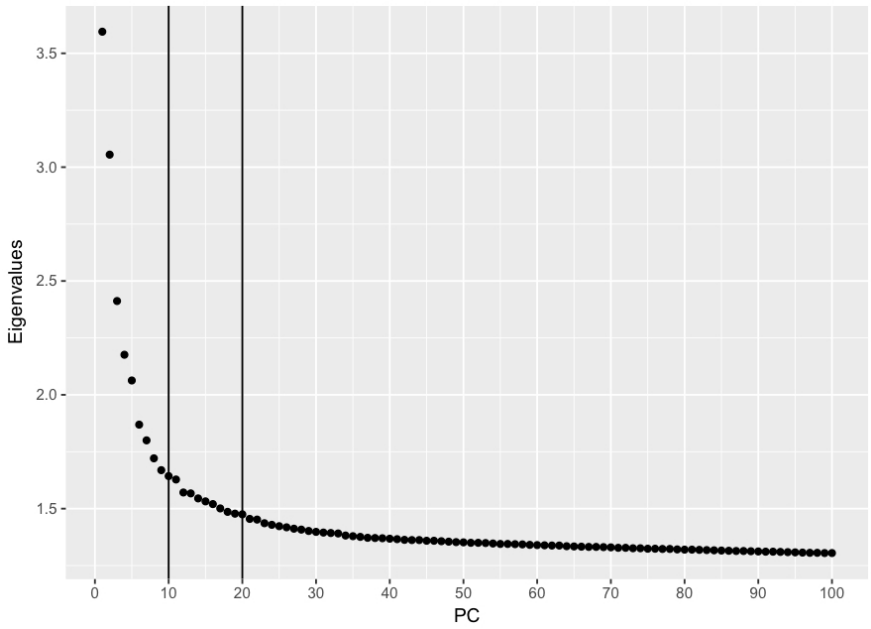


Fig. 4.6 Scree plot showing the eigenvalues of the first 100 principal components of samples.

4.2.4 Association testing

Denote the two alleles for a SNP A (major) and a (minor). In this thesis, I used an additive model for GWAS, assuming there is a uniform, linear increase in risk for each copy of the a allele [Bush and Moore, 2012]. For instance, if the risk is $2k$ for Aa, according to my assumption, there would be a $4k$ risk for aa. The association between a SNP X and a trait Y could then be examined through a generalized linear regression model for a phenotype.

$$g(Y) = b_0 + b_1 \cdot X + b_2 \cdot SEX + b_3 \cdot AGE + b_4 \cdot PC_1 + \cdots + b_{13} \cdot PC_{10} + \varepsilon \tag{4.1}$$

In Equation 4.1, the link function $g(\cdot)$ is an identity function for continuous traits and a *logit* function[5] for binary traits. b_0 is the intercept, which is normally ignored in GWAS. b_1

[5]The *logit* function $logit(p) = \log(\frac{p}{1-p})$. It is the inverse of the sigmoid function. When the link function is a *logit* function, the model becomes logistic regression.

is the regression coefficient of SNP X, and its direction represents the effect of each extra minor allele (i.e. a positive regression coefficient means that as there are more minor alleles, the phenotype or its *logit*, if it is dichotomous, would on average increase). Sex, age and the first ten principal components (c.f. Section 4.2.3) were used as covariates so that their confounding is adjusted for. Finally, the noise ε was assumed to follow a normal distribution $N(0, \sigma_\varepsilon^2)$.

Prior to fitting the model, $g(Y)$ was first standardized to mean 0 and unit variance so that the resulting regression coefficients would also be standardized and easier to explain. With the fitted model, we could calculate the t statistic of b_1, which is the fitted regression coefficient $\hat{b}_1$ divided by its standard error. The asymptotic p-value of the t statistic would determine whether trait Y is significantly associated with SNP X.

The model shown in Equation 4.1 was fitted for every metabolomic measurement against every SNP in the QC-ed genomic data using PLINK [Purcell et al., 2007]. To correct for multiple testing, we applied the widely accepted p-value threshold of 5×10^{-8} (c.f. Section 1.1.1.2).

4.2.5 Visualization of GWAS results

The association tests in Section 4.2.4 would return us a list of SNPs, their chromosomal positions and p-values (p) signifying the statistical significance of the associations.

In practice, Manhattan plots are commonly used to visualize these results. In the graph, each SNP is plotted as a point with its chromosomal position as x and its corresponding $-\log_{10}(p)$ as y. As a result, chromosomal regions with many highly significantly associated SNPs in LD stand tall like skyscrapers in the plot, contrasting with short "blocks" of relatively insignificant SNPs[6]. A horizontal line of $-\log_{10}(5 \times 10^{-8}) \approx 7.3$ is often drawn as a reference line in the graph – all the SNPs above it reach genome-wide significance.

Another widely adopted diagnostic plot is the quantile-quantile (QQ) plot. For all the SNPs, the observed p-values are plotted against their expected p-values following the uniform distribution[7]. The p-values are often first $-\log_{10}$-transformed for clearer visualization. If there exist some strongly associated SNPs, the scatterplot would deviate from the diagonal at the upper-right corner. On the other hand, if the plotted data points systematically deviate

[6] Indeed, the plot bears a resemblance to its namesake, the famous Manhattan skyline.

[7] Under the null hypothesis, there exists no association and all the p-values are uniformly distributed.

from the diagonal (e.g. inflation across the x-axis, even for very high p-values), the data may be problematic, e.g. suffering from population stratification, batch effect or cryptic relatedness [Turner, 2014].

In this study, the GWAS results from Section 4.2.4 were visualized via Manhattan plots and QQ plots using the R package "qqman" [Turner, 2014].

4.2.6 Variant annotation

Common within a population, SNPs are the primary biomarkers found in GWAS. Each SNP refers to a variation in a single nucleotide at a certain locus, and its location could have profound importance in predicting functional significance [T. H. Shen et al., 2009].

The process of predicting the function of a SNP is called variant annotation. In this study, I used ANNOVAR [K. Wang et al., 2010] to perform gene-based annotation for the SNPs identified to be significantly associated with metabolomic measurements. Furthermore, FUMA [Watanabe et al., 2017] was used to visualize the significant genomic risk loci through regional plots. All the significant SNPs were categorized into one of the variant classes listed in Table 4.2.

Table 4.2 Variant classes from gene-based annotation by ANNOVAR [K. Wang et al., 2010].

Variant class	Explanation
Exonic	Overlaps a coding
Splicing	Within 2 bp of a splicing junction
ncRNA	Overlaps a transcript without coding annotation in the gene definition
UTR5	Overlaps a 5' untranslated region
UTR3	Overlaps a 3' untranslated region
Intronic	Overlaps an intron
Upstream	Overlaps 1 kb region upstream of transcription start site
Downstream	Overlaps 1 kb region downstream of transcription end site
Intergenic	In intergenic region

For each variant category X, denote the percentage of significant SNPs falling into X as p_{sig}, and the percentage of SNPs classified as X in the whole genome as p_{all}. It would be interesting to see whether certain variant categories are enriched (or underrepresented) among the hits, i.e. we would like to test $H_A : p_{sig} = p_{all}$ against $H_B : p_{sig} \neq p_{all}$.

Note that the two populations are not independent[8]. Therefore, we could not directly perform a z-test. Fortunately, it is obvious that testing H_A against H_B is logically and statistically equivalent to testing $H_C : p_{sig} = p_{insig}$ against $H_D : p_{sig} \neq p_{insig}$, where p_{insig} is the percentage of insignificant SNPs falling into X. Hence, in this study, I tested H_C against H_D using a two-tailed two-proportion z-test for each variant category.

The test statistic is:

$$z = \frac{p_{sig} - p_{insig}}{\sqrt{p_{all}(1 - p_{all})(\frac{1}{n_{sig}} + \frac{1}{n_{insig}})}} \tag{4.2}$$

where n_{sig} (n_{insig}) is the total number of significant (insignificant) SNPs.

The p-value corresponding to the calculated z-statistic could then be derived according to the z-table. To circumvent multiple testing, the FDR was controlled at level $\alpha = 0.1$ through the B-H procedure [Hochberg and Benjamini, 1990].

Additional to gene-based annotation, I checked whether the hits in my study have been reported to be associated with certain diseases or traits in previous research using ANNOVAR [K. Wang et al., 2010] and the GWAS catalog [Welter et al., 2013].

4.2.7 Meta-analysis

In GWAS, we scan the whole genome trying to identify common variants significantly associated with traits of interest. Nevertheless, since the genetic effects of common alleles are typically small, large sample sizes are required to gain enough statistical power for signal detection [Evangelou and Ioannidis, 2013]. Unfortunately, the sample size of my GWAS-metabolomic data set was quite small (571 individuals), rendering the study underpowered. Therefore, meta-analysis[9] was performed to increase power and reduce false positives.

[Kettunen et al., 2016] performed GWAS on 123 metabolomic measurements (also ^{1}H NMR spectroscopy data) based on up to 24,925 individuals. The metabolomic data used in their study was extracted and quantified via the same high-throughput NMR metabolomics platform as ours; hence the metabolomic phenotypes used in the two studies (theirs and ours) could be matched properly.

[8]The set of significant SNPs is a subset of all the SNPs in the whole genome.

[9]Meta-analysis refers to statistically synthesizing information from multiple independent studies [Evangelou and Ioannidis, 2013].

116 metabolomic measurements were present in both studies. For each of these traits, the summary statistics from the two studies were matched by chromosome / base position; over 85% of the SNPs in our genetic data were also present in theirs. Meta-analysis was next performed in Plink [Purcell et al., 2007] using a random effect approach, assuming that the true effect sizes of SNPs may differ from study to study [Evangelou and Ioannidis, 2013]. The summary statistics from meta-analysis were used for polygenic scoring, which would be covered in the next chapter.

4.3 Results

The 130 genome-wide association studies (one for each metabolomic measurement) have identified 123 unique SNPs significantly associated with at least one of the metabolomic measurements. Summary statistics of all the significant results are shown in Table 4.4.

As could be seen in Figure 4.7, metabolomic measurements related to lipids and fatty acids tend to have more associated GWAS hits – the metabolomic measurement with the largest number of hits is mean diameter for VLDL particles (VLDL.D).

There were altogether 42 metabolomic measurements with one or more significantly associated SNP(s), and their association results are visualized in Appendix A. It is worth noting that metabolomic measurements tended to form "clusters" (e.g. Alb, S.HDL.L and S.HDL.P) – variants significantly associated with one metabolite in a cluster were quite likely to be significantly associated with the others in the cluster as well. This is probably due to the fact that certain metabolomic traits have high relatedness with each other and justifies Section 2.3.2.3, where I tried to perform dimensionality reduction on metabolites through hierarchical clustering.

The 123 significant variants found are of the types listed in Table 4.3. Among all the significant SNPs, exonic, intronic and UTR3 variants were enriched, whereas intergenic variants were underrepresented. A majority of the significant SNPs (47.15%) were intronic, and another 19.51% were exonic, ncRNA exonic/intronic, UTR3 or UTR5 – these variants hit 52 unique loci, which are listed in Table 4.5. It has been shown that in GWAS test statistics, (1) UTR5, exonic and UTR3 SNPs show the largest abundance of associations, (2) intronic SNPs are only moderately enriched, and (3) intergenic SNPs are relatively underrepresented [A. J. Schork et al., 2013]. This is more or less in line with my results, except that in my study, intronic variants are the most heavily enriched. Further research is needed to determine

whether the vast enrichment of intronic SNPs in my study is because of direct functional significance or it is simply due to the LD between the significant intronic SNPs and other unidentified functional SNPs nearby [McCauley et al., 2007; D. N. Cooper, 2010].

Table 4.3 Types of variants significantly associated with one or many metabolomic traits.

Type	p_{sig}	p_{insig}	p-value	q-value	Status
Exonic	6.50%	2.18%	0.0010	0.0042	Enriched
Intronic	47.15%	36.10%	0.0107	0.0214	Enriched
Non-coding RNA (exonic)	1.63%	0.50%	0.0785	0.1255	(Insignificant)
Non-coding RNA (intronic)	5.69%	5.88%	0.9280	0.9438	(Insignificant)
Intergenic	32.52%	51.67%	0.0000	0.0002	Underrepresented
Upstream	0.81%	0.76%	0.9438	0.9438	(Insignificant)
UTR3	4.88%	1.79%	0.0096	0.0214	Enriched
UTR5	0.81%	0.27%	0.2428	0.3237	(Insignificant)

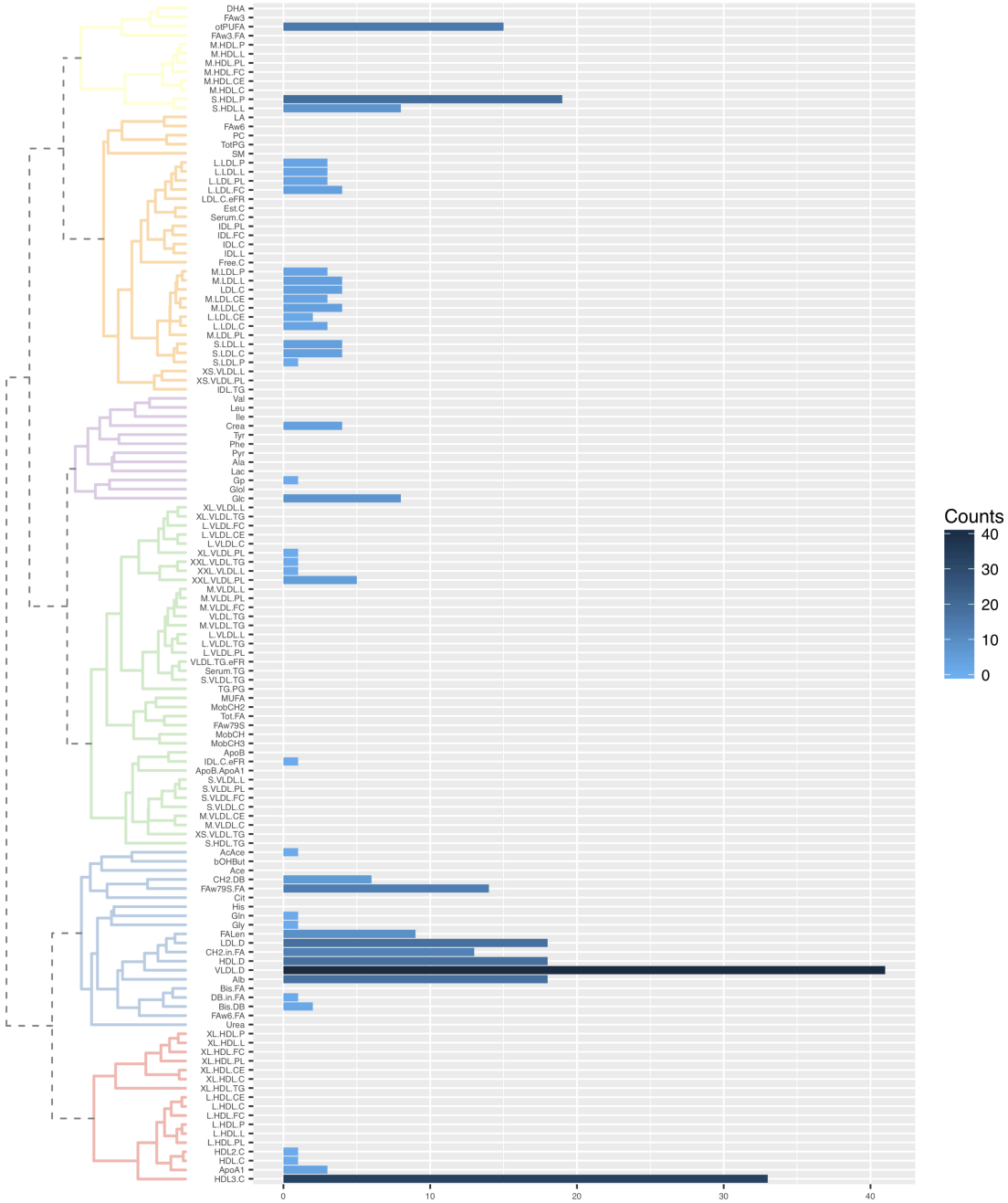


Fig. 4.7 Bar plot of the counts of significantly associated SNPs for 130 metabolomic measurements. The 130 metabolomic measurements are clustered using hierarchical clustering based on Kendall's correlation – the dendrogram is cut into six subtrees by height.

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs.

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
AcAce	INPP5D	2	234075691	rs7570320	C	A	intronic	0.24	(0.1595, 0.3206)	8.892E-09
Alb	TAF1A	1	222750544	kgp15137813	C	A	intronic	0.2399	(0.1615, 0.3182)	3.501E-09
Alb	NPHP1	2	110885620	rs13403558	C	A	intronic	0.2228	(0.1441, 0.3015)	4.476E-08
Alb	NPHP1	2	110937101	rs11898910	T	C	intronic	0.2444	(0.1661, 0.3228)	1.795E-09
Alb	LINC00116	2	110977357	rs10171646	G	A	ncRNA_intronic	0.2438	(0.1657, 0.3219)	1.771E-09
Alb	BHLHE40 (dist=13374), ARL8B (dist=123691)	3	5040239	rs6781674	G	A	intergenic	0.2231	(0.144, 0.3022)	4.95E-08
Alb	BHLHE40 (dist=23486), ARL8B (dist=113579)	3	5050351	rs4684458	C	A	intergenic	0.2231	(0.144, 0.3022)	4.95E-08
Alb	TRAPPC11 (dist=64391), NONE (dist=NONE)	4	184699138	kgp22750344	G	A	intergenic	0.2324	(0.1535, 0.3112)	1.259E-08
Alb	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2362	(0.1572, 0.3153)	8.11E-09
Alb	PDCD6	5	314935	rs7736	G	A	UTR3	0.236	(0.1571, 0.315)	7.999E-09
Alb	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.2324	(0.1541, 0.3106)	9.944E-09
Alb	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.2324	(0.1541, 0.3106)	9.944E-09
Alb	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.2324	(0.1541, 0.3106)	9.944E-09
Alb	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.2707	(0.1923, 0.3491)	3.328E-11
Alb	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.2409	(0.1624, 0.3194)	3.309E-09
Alb	CTBP2	10	126714256	rs2949367	G	A	intronic	0.2422	(0.1642, 0.3202)	2.138E-09
Alb	BEST3	12	70092000	rs2068191	G	A	intronic	0.2263	(0.1481, 0.3046)	2.318E-08
Alb	TBX3 (dist=969325), MED13L (dist=305087)	12	116091294	rs1427771	C	A	intergenic	0.2575	(0.1792, 0.3358)	2.47E-10
Alb	TBX3 (dist=978898), MED13L (dist=295514)	12	116100867	rs12425668	T	C	intergenic	0.2585	(0.1801, 0.3369)	2.26E-10
ApoA1	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.2664	(0.1898, 0.3431)	2.474E-11
ApoA1	REEP3 (dist=901145), ANXA2P3 (dist=299257)	10	66286028	rs7083780	G	A	intergenic	0.2231	(0.1459, 0.3002)	2.388E-08
ApoA1	LINC00908	18	74269894	rs7232061	G	A	ncRNA_intronic	0.2198	(0.1422, 0.2975)	4.438E-08
Bis.DB	ADRA2C (dist=94130), FAM86EP (dist=79104)	4	3864383	rs59079720	G	C	intergenic	-0.2241	(-0.303, -0.1451)	4.087E-08
Bis.DB	GUCY1A2 (dist=52382), CWF19L2 (dist=255519)	11	106941553	rs1840572	T	C	intergenic	-0.224	(-0.3033, -0.1446)	4.847E-08
CH2.DB	MSH3	5	79984714	rs863214	T	C	intronic	-0.2269	(-0.3071, -0.1467)	4.537E-08
CH2.DB	MSH3	5	80082865	rs6151838	G	A	intronic	-0.2328	(-0.3129, -0.1526)	2.046E-08
CH2.DB	MSH3	5	80132177	rs10075024	T	C	intronic	-0.2442	(-0.3236, -0.1649)	2.967E-09
CH2.DB	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	-0.2275	(-0.3069, -0.148)	3.181E-08
CH2.DB	PCK2	14	24564684	rs9783666	T	G	intronic	-0.2277	(-0.3076, -0.1478)	3.626E-08
CH2.DB	ABHD2	15	89689583	rs4932475	T	G	intronic	-0.2409	(-0.3203, -0.1614)	4.934E-09
CH2.in.FA	KBTBD8	3	67054649	rs13096789	T	C	exonic	-0.2283	(-0.308, -0.1486)	3.136E-08
CH2.in.FA	PISRT1 (dist=33379), MRPS22 (dist=77118)	3	138985743	rs80319952	T	C	intergenic	-0.2336	(-0.3145, -0.1526)	2.487E-08

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
CH2.in.FA	ADRA2C (dist=94130), FAM86EP (dist=79104)	4	3864383	rs59079720	G	C	intergenic	-0.2286	(-0.3079, -0.1493)	2.553E-08
CH2.in.FA	MSH3	5	80132177	rs10075024	T	C	intronic	-0.2422	(-0.3223, -0.1622)	5.204E-09
CH2.in.FA	LINC01861 (dist=55082), FAM114A2 (dist=36062)	5	153333629	rs75273818	T	C	intergenic	-0.2359	(-0.3159, -0.1559)	1.255E-08
CH2.in.FA	CTTNBP2	7	117389972	rs10247163	T	C	intronic	-0.2299	(-0.3095, -0.1503)	2.375E-08
CH2.in.FA	CTTNBP2	7	117401398	rs10258815	G	A	intronic	-0.2299	(-0.3095, -0.1503)	2.375E-08
CH2.in.FA	CTTNBP2	7	117406238	rs10254610	G	A	intronic	-0.2299	(-0.3095, -0.1503)	2.375E-08
CH2.in.FA	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	-0.2417	(-0.3215, -0.1619)	5.067E-09
CH2.in.FA	KIAA1217	10	24791582	rs2150650	T	C	intronic	-0.2292	(-0.3088, -0.1495)	2.769E-08
CH2.in.FA	KIAA1217	10	24798368	rs12252802	G	A	intronic	-0.2316	(-0.3113, -0.1519)	2.001E-08
CH2.in.FA	PCK2	14	24564684	rs9783666	T	G	intronic	-0.2364	(-0.3167, -0.156)	1.339E-08
CH2.in.FA	ABHD2	15	89689583	rs4932475	T	G	intronic	-0.2503	(-0.3302, -0.1705)	1.523E-09
Crea	GRIK2 (dist=1729378), HACE1 (dist=928632)	6	104247336	rs4546515	T	C	intergenic	0.1719	(0.1115, 0.2323)	3.845E-08
Crea	PRKG1	10	53356952	rs1917841	T	C	intronic	0.1892	(0.1287, 0.2496)	1.645E-09
Crea	GUCY1A2 (dist=52382), CWF19L2 (dist=255519)	11	106941553	rs1840572	T	C	intergenic	0.1861	(0.1259, 0.2464)	2.583E-09
Crea	LINC01065 (dist=651538), LINC00558 (dist=11981)	13	54377573	kgp16825644	A	G	intergenic	0.2026	(0.1424, 0.2627)	9.608E-11
DB.in.FA	ADRA2C (dist=94130), FAM86EP (dist=79104)	4	3864383	rs59079720	G	C	intergenic	-0.2282	(-0.3065, -0.15)	1.8E-08
FALen	ADRA2C (dist=94130), FAM86EP (dist=79104)	4	3864383	rs59079720	G	C	intergenic	-0.2422	(-0.3211, -0.1634)	3.107E-09
FALen	CTTNBP2	7	117389972	rs10247163	T	C	intronic	-0.2256	(-0.3051, -0.1462)	4.054E-08
FALen	CTTNBP2	7	117401398	rs10258815	G	A	intronic	-0.2256	(-0.3051, -0.1462)	4.054E-08
FALen	CTTNBP2	7	117406238	rs10254610	G	A	intronic	-0.2256	(-0.3051, -0.1462)	4.054E-08
FALen	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	-0.2358	(-0.3155, -0.1561)	1.118E-08
FALen	KIAA1217	10	24791582	rs2150650	T	C	intronic	-0.2414	(-0.3207, -0.1622)	4.207E-09
FALen	KIAA1217	10	24798368	rs12252802	G	A	intronic	-0.2437	(-0.323, -0.1645)	3.023E-09
FALen	GUCY1A2 (dist=52382), CWF19L2 (dist=255519)	11	106941553	rs1840572	T	C	intergenic	-0.2297	(-0.3092, -0.1502)	2.378E-08
FALen	ABHD2	15	89689583	rs4932475	T	G	intronic	-0.2291	(-0.3092, -0.149)	3.25E-08
FAw79S.FA	PISRT1 (dist=33379), MRPS22 (dist=77118)	3	138985743	rs80319952	T	C	intergenic	-0.2314	(-0.3124, -0.1504)	3.363E-08
FAw79S.FA	MSH3	5	80132177	rs10075024	T	C	intronic	-0.2451	(-0.3251, -0.1652)	3.386E-09
FAw79S.FA	CTTNBP2	7	117389972	rs10247163	T	C	intronic	-0.2256	(-0.3052, -0.1459)	4.39E-08
FAw79S.FA	CTTNBP2	7	117401398	rs10258815	G	A	intronic	-0.2256	(-0.3052, -0.1459)	4.39E-08
FAw79S.FA	CTTNBP2	7	117406238	rs10254610	G	A	intronic	-0.2256	(-0.3052, -0.1459)	4.39E-08
FAw79S.FA	LINC01605 (dist=16054), ZNF703 (dist=158311)	8	37394958	rs16886859	G	A	intergenic	-0.2312	(-0.3111, -0.1513)	2.288E-08
FAw79S.FA	SLC20A2	8	42296993	kgp20396476	A	G	exonic	-0.231	(-0.3113, -0.1508)	2.666E-08
FAw79S.FA	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	-0.2356	(-0.3155, -0.1557)	1.249E-08
FAw79S.FA	TBX3 (dist=978898), MED13L (dist=295514)	12	116100867	rs12425668	T	C	intergenic	-0.2288	(-0.3091, -0.1485)	3.681E-08

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
FAw79S.FA	PCK2	14	24564684	rs9783666	T	G	intronic	-0.2346	(-0.3149, -0.1542)	1.751E-08
FAw79S.FA	ABHD2	15	89689583	rs4932475	T	G	intronic	-0.2577	(-0.3374, -0.178)	4.779E-10
FAw79S.FA	NXN	17	718968	rs78895411	T	G	intronic	-0.2321	(-0.3126, -0.1517)	2.512E-08
FAw79S.FA	NXN	17	734321	rs74426014	G	A	intronic	-0.232	(-0.3131, -0.1509)	3.209E-08
FAw79S.FA	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	-0.2297	(-0.3091, -0.1502)	2.353E-08
Glc	ABHD5 (dist=24731), MIR138-1 (dist=366756)	3	43788948	kgp18146238	A	G	intergenic	0.2277	(0.1472, 0.3081)	4.495E-08
Glc	KIAA0895L	16	67213923	rs13339140	T	C	intronic	0.2278	(0.1473, 0.3083)	4.464E-08
Glc	E2F4	16	67229486	rs3730403	T	G	intronic	0.2287	(0.1478, 0.3096)	4.698E-08
Glc	LRRC29	16	67241282	rs12051247	A	G	UTR3	0.2305	(0.1498, 0.3112)	3.38E-08
Glc	LRRC29	16	67248831	rs13338688	G	A	intronic	0.267	(0.1872, 0.3468)	1.262E-10
Glc	CTCF	16	67627635	rs17686899	T	G	intronic	0.2332	(0.1523, 0.3142)	2.621E-08
Glc	CTCF	16	67655133	rs7191281	T	C	intronic	0.2329	(0.1521, 0.3137)	2.559E-08
Glc	CTCF	16	67671804	rs6499137	C	A	UTR3	0.2328	(0.1521, 0.3135)	2.511E-08
Gln	TRAPPC11 (dist=64391), NONE (dist=NONE)	4	184699138	kgp22750344	G	A	intergenic	0.2444	(0.1646, 0.3241)	3.458E-09
Gly	CPS1	2	211540507	rs1047891	C	A	exonic	0.2603	(0.1801, 0.3405)	4.17E-10
Gp	LINC01495	11	22472250	kgp12607087	A	G	ncRNA_intronic	0.2308	(0.1511, 0.3104)	2.219E-08
HDL.C	REEP3 (dist=901145), ANXA2P3 (dist=299257)	10	66286028	rs7083780	G	A	intergenic	0.2186	(0.1415, 0.2957)	4.264E-08
HDL.D	MPZ	1	161274905	rs16832786	T	C	UTR3	0.2268	(0.1467, 0.3068)	4.364E-08
HDL.D	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2268	(0.1466, 0.307)	4.557E-08
HDL.D	PDCD6	5	314935	rs7736	G	A	UTR3	0.2268	(0.1468, 0.3067)	4.221E-08
HDL.D	SLIT3	5	168484494	rs13154825	C	A	intronic	0.2245	(0.1449, 0.304)	4.993E-08
HDL.D	MAPK14 (dist=4984), MAPK13 (dist=14264)	6	36083997	kgp17318753	A	G	intergenic	0.2324	(0.1524, 0.3123)	1.977E-08
HDL.D	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.2613	(0.1827, 0.3398)	1.595E-10
HDL.D	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.2613	(0.1827, 0.3398)	1.595E-10
HDL.D	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.2613	(0.1827, 0.3398)	1.595E-10
HDL.D	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.2822	(0.2032, 0.3613)	7.503E-12
HDL.D	LOC392232	8	73150491	rs6989765	G	A	ncRNA_intronic	0.2785	(0.2, 0.3571)	1.034E-11
HDL.D	LOC392232	8	73153467	rs987242	G	A	ncRNA_intronic	0.2774	(0.1992, 0.3556)	1.032E-11
HDL.D	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.2579	(0.1788, 0.337)	3.501E-10
HDL.D	EDNRB-AS1	13	78451158	kgp16788601	A	G	ncRNA_intronic	0.2275	(0.1478, 0.3072)	3.501E-08
HDL.D	FARP1	13	99054527	rs9584835	T	C	intronic	0.2298	(0.1505, 0.3092)	2.231E-08
HDL.D	PCK2	14	24564684	rs9783666	T	G	intronic	0.2527	(0.173, 0.3323)	9.858E-10
HDL.D	PCK2	14	24569947	rs2759407	G	A	UTR3	0.2297	(0.1499, 0.3095)	2.663E-08
HDL.D	LINC01541	18	69215844	kgp16017220	G	A	ncRNA_exonic	0.258	(0.1785, 0.3375)	4.232E-10

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
HDL.D	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	0.2463	(0.1675, 0.3251)	1.683E-09
HDL2.C	REEP3 (dist=901145), ANXA2P3 (dist=299257)	10	66286028	rs7083780	G	A	intergenic	0.2262	(0.1492, 0.3033)	1.437E-08
HDL3.C	GRHL1	2	10095185	rs16867251	T	C	exonic	0.2347	(0.1547, 0.3147)	1.476E-08
HDL3.C	GRHL1	2	10101468	rs16867256	G	A	exonic	0.2355	(0.1555, 0.3155)	1.335E-08
HDL3.C	GRHL1	2	10103414	rs11902457	T	C	intronic	0.2347	(0.1547, 0.3147)	1.476E-08
HDL3.C	GRHL1	2	10116084	rs6735658	G	A	intronic	0.2347	(0.1547, 0.3147)	1.476E-08
HDL3.C	MIR4435-2HG	2	112123745	rs2292932	C	A	ncRNA_exonic	0.2391	(0.1594, 0.3189)	7.271E-09
HDL3.C	LOC101928551 (dist=1232), ADAM29 (dist=41162)	4	175798347	rs7667865	G	A	intergenic	0.2279	(0.1472, 0.3087)	4.903E-08
HDL3.C	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2436	(0.1635, 0.3237)	4.469E-09
HDL3.C	PDCD6	5	314935	rs7736	G	A	UTR3	0.2432	(0.1632, 0.3232)	4.53E-09
HDL3.C	LOC100133050 (dist=21809), FAM174A (dist=125242)	5	99745767	rs11738482	G	A	intergenic	0.2421	(0.1622, 0.322)	5.132E-09
HDL3.C	MAPK14 (dist=4984), MAPK13 (dist=14264)	6	36083997	kgp17318753	A	G	intergenic	0.2678	(0.1883, 0.3472)	9.14E-11
HDL3.C	THBS2 (dist=78638), WDR27 (dist=124456)	6	169732847	kgp17329668	A	G	intergenic	0.2613	(0.1827, 0.34)	1.678E-10
HDL3.C	THBS2 (dist=78912), WDR27 (dist=124182)	6	169733121	rs7381784	G	A	intergenic	0.2613	(0.1827, 0.34)	1.678E-10
HDL3.C	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.2908	(0.2127, 0.3689)	1.01E-12
HDL3.C	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.2908	(0.2127, 0.3689)	1.01E-12
HDL3.C	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.2908	(0.2127, 0.3689)	1.01E-12
HDL3.C	CTTNBP2	7	117445664	rs929668	G	A	intronic	0.2459	(0.1665, 0.3253)	2.373E-09
HDL3.C	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.3117	(0.2332, 0.3902)	3.553E-14
HDL3.C	LOC392232	8	73150491	rs6989765	G	A	ncRNA_intronic	0.2703	(0.1913, 0.3492)	4.992E-11
HDL3.C	LOC392232	8	73153467	rs987242	G	A	ncRNA_intronic	0.2695	(0.1908, 0.3482)	4.68E-11
HDL3.C	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.2915	(0.213, 0.3701)	1.177E-12
HDL3.C	ASTN2-AS1	9	119321341	kgp22748805	A	G	ncRNA_intronic	0.2355	(0.1558, 0.3152)	1.172E-08
HDL3.C	BEST3	12	70092000	rs2068191	G	A	intronic	0.2617	(0.1831, 0.3404)	1.571E-10
HDL3.C	MTMR6	13	25823451	rs17082035	G	A	exonic	0.2316	(0.1518, 0.3114)	2.103E-08
HDL3.C	MTMR6	13	25842806	rs17082070	T	C	intronic	0.2316	(0.1518, 0.3114)	2.103E-08
HDL3.C	PCK2	14	24564684	rs9783666	T	G	intronic	0.2632	(0.1835, 0.3429)	2.086E-10
HDL3.C	PCK2	14	24569947	rs2759407	G	A	UTR3	0.2342	(0.1543, 0.3142)	1.522E-08
HDL3.C	VRK1	14	97277922	rs76904997	C	A	intronic	0.2257	(0.1459, 0.3055)	4.584E-08
HDL3.C	INO80	15	41272913	kgp19746400	G	A	intronic	0.2293	(0.1491, 0.3094)	3.258E-08
HDL3.C	LINC01541	18	69215844	kgp16017220	G	A	ncRNA_exonic	0.274	(0.1946, 0.3534)	3.441E-11
HDL3.C	LINC01899 (dist=644331), CBLN2 (dist=110122)	18	70093793	kgp16131518	G	A	intergenic	0.2438	(0.164, 0.3235)	3.719E-09
HDL3.C	ZCCHC3 (dist=2912), NRSN2-AS1 (dist=16456)	20	283875	kgp22798012	A	G	intergenic	0.2378	(0.1582, 0.3175)	8.323E-09

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
HDL3.C	NONE (dist=NONE), NONE (dist=NONE)	23	138533773	rs4829986	G	A	intergenic	0.2332	(0.1532, 0.3133)	1.844E-08
HDL3.C	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	0.2686	(0.19, 0.3471)	5E-11
IDL.C.eFR	ETAA1 (dist=210062), LINC01812 (dist=175591)	2	67847595	rs7579880	G	A	intergenic	0.2336	(0.1538, 0.3134)	1.573E-08
L.LDL.C	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2527	(-0.3328, -0.1726)	1.208E-09
L.LDL.C	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2378	(-0.3183, -0.1573)	1.184E-08
L.LDL.C	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2536	(-0.3334, -0.1737)	9.629E-10
L.LDL.CE	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.24	(-0.3202, -0.1598)	7.788E-09
L.LDL.CE	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2402	(-0.3202, -0.1602)	6.904E-09
L.LDL.FC	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.2392	(-0.3202, -0.1583)	1.139E-08
L.LDL.FC	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.267	(-0.3472, -0.1869)	1.465E-10
L.LDL.FC	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.252	(-0.3326, -0.1715)	1.655E-09
L.LDL.FC	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2699	(-0.3498, -0.1901)	8.174E-11
L.LDL.L	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2493	(-0.3295, -0.1691)	2.106E-09
L.LDL.L	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2339	(-0.3145, -0.1532)	2.122E-08
L.LDL.L	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2505	(-0.3305, -0.1705)	1.585E-09
L.LDL.P	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2437	(-0.3241, -0.1633)	4.916E-09
L.LDL.P	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2279	(-0.3087, -0.1472)	4.908E-08
L.LDL.P	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2452	(-0.3254, -0.1651)	3.571E-09
L.LDL.PL	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2417	(-0.3223, -0.1612)	6.964E-09
L.LDL.PL	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2284	(-0.3093, -0.1475)	4.854E-08
L.LDL.PL	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2433	(-0.3236, -0.163)	5.089E-09
LDL.C	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.2331	(-0.3137, -0.1526)	2.252E-08
LDL.C	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2628	(-0.3426, -0.1831)	2.282E-10
LDL.C	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.247	(-0.3272, -0.1669)	2.835E-09
LDL.C	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2607	(-0.3403, -0.1811)	2.896E-10
LDL.D	BSDC1 (dist=44549), ZBTB8B (dist=26047)	1	32904611	rs11588809	T	C	intergenic	0.2323	(0.1528, 0.3117)	1.644E-08
LDL.D	BSDC1 (dist=65585), ZBTB8B (dist=5011)	1	32925647	rs16834988	T	C	intergenic	0.2323	(0.1528, 0.3117)	1.644E-08
LDL.D	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2271	(0.1467, 0.3076)	4.806E-08
LDL.D	PDCD6	5	314935	rs7736	G	A	UTR3	0.2271	(0.1468, 0.3073)	4.482E-08
LDL.D	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.2654	(0.1866, 0.3441)	9.205E-11
LDL.D	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.2654	(0.1866, 0.3441)	9.205E-11
LDL.D	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.2654	(0.1866, 0.3441)	9.205E-11
LDL.D	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.3042	(0.2254, 0.3829)	1.545E-13
LDL.D	LOC392232	8	73150491	rs6989765	G	A	ncRNA_intronic	0.2845	(0.2058, 0.3631)	4.223E-12

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
LDL.D	LOC392232	8	73153467	rs987242	G	A	ncRNA_intronic	0.2832	(0.2048, 0.3616)	4.255E-12
LDL.D	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.2424	(0.1627, 0.322)	4.45E-09
LDL.D	GFI1B	9	135855768	rs685783	G	A	intronic	0.2395	(0.1598, 0.3192)	6.709E-09
LDL.D	EDNRB-AS1	13	78451158	kgp16788601	A	G	ncRNA_intronic	0.2357	(0.1559, 0.3155)	1.195E-08
LDL.D	NRL (dist=3810), PCK2 (dist=5698)	14	24557642	rs11623285	T	G	intergenic	0.244	(0.1646, 0.3235)	3.17E-09
LDL.D	PCK2	14	24564684	rs9783666	T	G	intronic	0.2738	(0.1944, 0.3533)	3.537E-11
LDL.D	PCK2	14	24569947	rs2759407	G	A	UTR3	0.2423	(0.1625, 0.3221)	4.654E-09
LDL.D	LINC01541	18	69215844	kgp16017220	G	A	ncRNA_exonic	0.2628	(0.1832, 0.3425)	2.203E-10
LDL.D	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	0.2466	(0.1676, 0.3256)	1.805E-09
M.LDL.C	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.229	(-0.3094, -0.1487)	3.655E-08
M.LDL.C	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.26	(-0.3395, -0.1804)	3.205E-10
M.LDL.C	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2448	(-0.3248, -0.1648)	3.59E-09
M.LDL.C	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2555	(-0.3349, -0.176)	5.974E-10
M.LDL.CE	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2433	(-0.3231, -0.1634)	4.203E-09
M.LDL.CE	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2291	(-0.3093, -0.1488)	3.45E-08
M.LDL.CE	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2401	(-0.3198, -0.1604)	6.214E-09
M.LDL.L	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.2279	(-0.3084, -0.1475)	4.331E-08
M.LDL.L	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2573	(-0.337, -0.1777)	5.014E-10
M.LDL.L	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2407	(-0.3208, -0.1606)	6.732E-09
M.LDL.L	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2524	(-0.332, -0.1728)	9.87E-10
M.LDL.P	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2525	(-0.3323, -0.1727)	1.081E-09
M.LDL.P	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2351	(-0.3153, -0.1548)	1.552E-08
M.LDL.P	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2476	(-0.3273, -0.1679)	2.108E-09
otPUFA	MYRF	11	61547237	rs108499	T	C	intronic	0.2907	(0.2118, 0.3695)	1.666E-12
otPUFA	MYRF	11	61548559	rs509360	G	A	intronic	0.2888	(0.2101, 0.3674)	2.005E-12
otPUFA	MYRF	11	61551356	rs174535	C	A	exonic	0.3055	(0.2269, 0.384)	1.093E-13
otPUFA	TMEM258	11	61557803	rs102275	G	A	intronic	0.3008	(0.2223, 0.3793)	2.366E-13
otPUFA	TMEM258	11	61560081	rs174538	G	A	UTR5	0.2883	(0.2095, 0.3672)	2.473E-12
otPUFA	FADS1	11	61569830	rs174546	T	C	UTR3	0.3004	(0.222, 0.3789)	2.446E-13
otPUFA	FADS1	11	61570783	rs174547	T	C	intronic	0.3004	(0.222, 0.3789)	2.446E-13
otPUFA	FADS1	11	61571348	rs174548	G	C	intronic	0.2782	(0.1993, 0.3572)	1.353E-11
otPUFA	FADS1	11	61571478	rs174550	T	C	intronic	0.3004	(0.222, 0.3789)	2.446E-13
otPUFA	FADS2	11	61597212	rs174570	T	C	intronic	0.2966	(0.2173, 0.3759)	8.354E-13
otPUFA	FADS2	11	61597972	rs1535	G	A	intronic	0.3007	(0.2222, 0.3792)	2.425E-13

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
otPUFA	FADS2	11	61603510	rs174576	C	A	intronic	0.2985	(0.22, 0.377)	3.468E-13
otPUFA	FADS2	11	61604814	rs174577	C	A	intronic	0.302	(0.2237, 0.3804)	1.731E-13
otPUFA	FADS2	11	61609750	rs174583	T	C	intronic	0.3019	(0.2236, 0.3803)	1.8E-13
otPUFA	FADS3	11	61654092	rs77980989	C	A	intronic	0.2594	(0.1799, 0.339)	3.399E-10
S.HDL.L	TAF1A	1	222750544	kgp15137813	C	A	intronic	0.2524	(0.1725, 0.3324)	1.197E-09
S.HDL.L	NPHP1	2	110937101	rs11898910	T	C	intronic	0.2414	(0.1611, 0.3217)	6.65E-09
S.HDL.L	LINC00116	2	110977357	rs10171646	G	A	ncRNA_intronic	0.2387	(0.1586, 0.3188)	8.949E-09
S.HDL.L	TRAPPC11 (dist=64391), NONE (dist=NONE)	4	184699138	kgp22750344	G	A	intergenic	0.2544	(0.1741, 0.3347)	1.037E-09
S.HDL.L	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.3007	(0.2211, 0.3803)	5E-13
S.HDL.L	TBX3 (dist=969325), MED13L (dist=305087)	12	116091294	rs1427771	C	A	intergenic	0.2347	(0.154, 0.3154)	1.951E-08
S.HDL.L	TBX3 (dist=978898), MED13L (dist=295514)	12	116100867	rs12425668	T	C	intergenic	0.2413	(0.1606, 0.322)	7.998E-09
S.HDL.L	SNX29	16	12278595	kgp16526588	A	G	intronic	0.2288	(0.1478, 0.3098)	4.806E-08
S.HDL.P	RNF207	1	6276561	rs68032129	T	C	intronic	0.2318	(0.1509, 0.3128)	3.127E-08
S.HDL.P	TAF1A	1	222750544	kgp15137813	C	A	intronic	0.2637	(0.1837, 0.3437)	2.293E-10
S.HDL.P	NPHP1	2	110885620	rs13403558	C	A	intronic	0.243	(0.1625, 0.3236)	5.877E-09
S.HDL.P	NPHP1	2	110937101	rs11898910	T	C	intronic	0.2628	(0.1827, 0.3429)	2.734E-10
S.HDL.P	LINC00116	2	110977357	rs10171646	G	A	ncRNA_intronic	0.2604	(0.1805, 0.3403)	3.607E-10
S.HDL.P	TRAPPC11 (dist=64391), NONE (dist=NONE)	4	184699138	kgp22750344	G	A	intergenic	0.2691	(0.1889, 0.3494)	1.133E-10
S.HDL.P	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2366	(0.1554, 0.3178)	1.865E-08
S.HDL.P	PDCD6	5	314935	rs7736	G	A	UTR3	0.2362	(0.155, 0.3174)	1.929E-08
S.HDL.P	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.2386	(0.1582, 0.3189)	9.886E-09
S.HDL.P	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.2386	(0.1582, 0.3189)	9.886E-09
S.HDL.P	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.2386	(0.1582, 0.3189)	9.886E-09
S.HDL.P	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.3177	(0.2383, 0.3972)	2.313E-14
S.HDL.P	LOC392232	8	73150491	rs6989765	G	A	ncRNA_intronic	0.242	(0.1614, 0.3226)	7.023E-09
S.HDL.P	LOC392232	8	73153467	rs987242	G	A	ncRNA_intronic	0.2422	(0.1618, 0.3225)	5.982E-09
S.HDL.P	TBX3 (dist=969325), MED13L (dist=305087)	12	116091294	rs1427771	C	A	intergenic	0.2457	(0.1649, 0.3264)	4.383E-09
S.HDL.P	TBX3 (dist=978898), MED13L (dist=295514)	12	116100867	rs12425668	T	C	intergenic	0.2522	(0.1715, 0.333)	1.755E-09
S.HDL.P	PCK2	14	24564684	rs9783666	T	G	intronic	0.2336	(0.1523, 0.315)	2.834E-08
S.HDL.P	SNX29	16	12278595	kgp16526588	A	G	intronic	0.2366	(0.1554, 0.3177)	1.787E-08
S.HDL.P	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	0.2481	(0.1681, 0.3281)	2.277E-09
S.LDL.C	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.232	(-0.3123, -0.1516)	2.45E-08
S.LDL.C	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2618	(-0.3413, -0.1822)	2.45E-10
S.LDL.C	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.244	(-0.3241, -0.164)	4.081E-09

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
S.LDL.C	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2564	(-0.3358, -0.1769)	5.269E-10
S.LDL.L	NECTIN2	19	45389596	rs7254892	G	A	intronic	-0.2301	(-0.3104, -0.1498)	3.084E-08
S.LDL.L	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.2571	(-0.3366, -0.1775)	4.971E-10
S.LDL.L	APOE (dist=2988), APOC1 (dist=1864)	19	45415640	rs445925	T	C	intergenic	-0.2357	(-0.3158, -0.1556)	1.345E-08
S.LDL.L	APOC1 (dist=4186), APOC1P1 (dist=3268)	19	45426792	rs141622900	G	A	intergenic	-0.2501	(-0.3296, -0.1706)	1.362E-09
S.LDL.P	APOE (dist=1747), APOC1 (dist=3105)	19	45414399	rs72654473	C	A	intergenic	-0.229	(-0.3092, -0.1487)	3.487E-08
VLDL.D	BSDC1 (dist=44549), ZBTB8B (dist=26047)	1	32904611	rs11588809	T	C	intergenic	0.2448	(0.1659, 0.3238)	2.265E-09
VLDL.D	BSDC1 (dist=65585), ZBTB8B (dist=5011)	1	32925647	rs16834988	T	C	intergenic	0.2448	(0.1659, 0.3238)	2.265E-09
VLDL.D	LRRC7	1	70154480	rs17325299	T	C	intronic	0.2321	(0.1521, 0.312)	2.061E-08
VLDL.D	MPZ	1	161274905	rs16832786	T	C	UTR3	0.2267	(0.1466, 0.3068)	4.465E-08
VLDL.D	NPHP1	2	110885620	rs13403558	C	A	intronic	0.2244	(0.1448, 0.3039)	4.943E-08
VLDL.D	EPHB1	3	134826917	rs1554675	G	A	intronic	0.2413	(0.1613, 0.3213)	5.964E-09
VLDL.D	TRAPPC11 (dist=64391), NONE (dist=NONE)	4	184699138	kgp22750344	G	A	intergenic	0.2307	(0.1508, 0.3105)	2.366E-08
VLDL.D	AHRR, PDCD6	5	314518	rs1574220	G	A	intronic	0.2313	(0.1512, 0.3114)	2.448E-08
VLDL.D	PDCD6	5	314935	rs7736	G	A	UTR3	0.2312	(0.1513, 0.3111)	2.302E-08
VLDL.D	TIAM2	6	155411606	rs6915661	G	A	intronic	0.2293	(0.1499, 0.3088)	2.493E-08
VLDL.D	THBS2 (dist=78638), WDR27 (dist=124456)	6	169732847	kgp17329668	A	G	intergenic	0.2517	(0.173, 0.3304)	7.204E-10
VLDL.D	THBS2 (dist=78912), WDR27 (dist=124182)	6	169733121	rs7381784	G	A	intergenic	0.2517	(0.173, 0.3304)	7.204E-10
VLDL.D	BET1 (dist=269678), COL1A2-AS1 (dist=93152)	7	93903372	rs5014002	G	A	intergenic	0.2257	(0.1461, 0.3053)	4.281E-08
VLDL.D	CTTNBP2	7	117389972	rs10247163	T	C	intronic	0.303	(0.2255, 0.3806)	8.334E-14
VLDL.D	CTTNBP2	7	117401398	rs10258815	G	A	intronic	0.303	(0.2255, 0.3806)	8.334E-14
VLDL.D	CTTNBP2	7	117406238	rs10254610	G	A	intronic	0.303	(0.2255, 0.3806)	8.334E-14
VLDL.D	CTTNBP2	7	117445664	rs929668	G	A	intronic	0.2378	(0.1584, 0.3173)	7.482E-09
VLDL.D	XKR6 (dist=56426), MTMR9 (dist=26699)	8	11115301	kgp20243754	A	G	intergenic	0.338	(0.2605, 0.4156)	1.242E-16
VLDL.D	LOC392232	8	73150491	rs6989765	G	A	ncRNA_intronic	0.3087	(0.2309, 0.3865)	3.692E-14
VLDL.D	LOC392232	8	73153467	rs987242	G	A	ncRNA_intronic	0.3084	(0.2309, 0.3858)	3.02E-14
VLDL.D	HNF4G (dist=16278), LINC01111 (dist=823534)	8	76495355	kgp20331527	A	G	intergenic	0.248	(0.1687, 0.3273)	1.693E-09
VLDL.D	GFI1B	9	135855768	rs685783	G	A	intronic	0.2478	(0.1685, 0.3271)	1.726E-09
VLDL.D	FGF3 (dist=261194), LOC101928443 (dist=6950)	11	69895386	rs61886432	G	A	intergenic	0.2244	(0.1449, 0.304)	4.91E-08
VLDL.D	GRIK4	11	120547912	rs10502240	T	C	intronic	0.2303	(0.1511, 0.3096)	1.987E-08
VLDL.D	GRIK4	11	120548318	rs11607732	G	A	intronic	0.2308	(0.1514, 0.3102)	1.948E-08
VLDL.D	MTMR6	13	25823451	rs17082035	G	A	exonic	0.2286	(0.149, 0.3083)	2.934E-08
VLDL.D	MTMR6	13	25842806	rs17082070	T	C	intronic	0.2286	(0.149, 0.3083)	2.934E-08
VLDL.D	FARP1	13	99054527	rs9584835	T	C	intronic	0.234	(0.1546, 0.3133)	1.237E-08

(Continued on next page)

Table 4.4 Association results: metabolomic measurements and their significantly associated SNPs (cont'd).

Metab	Gene / Neighbouring gene(s)	Chr	Pos	SNP	A1	A2	Type	β	(95% CI)	P
VLDL.D	NRL (dist=3810), PCK2 (dist=5698)	14	24557642	rs11623285	T	G	intergenic	0.2875	(0.2093, 0.3657)	1.914E-12
VLDL.D	PCK2	14	24564684	rs9783666	T	G	intronic	0.3196	(0.2416, 0.3976)	5.741E-15
VLDL.D	PCK2	14	24569947	rs2759407	G	A	UTR3	0.267	(0.188, 0.346)	8.211E-11
VLDL.D	VRK1	14	97268547	rs10147248	G	A	intronic	0.2288	(0.1492, 0.3083)	2.773E-08
VLDL.D	VRK1	14	97277922	rs76904997	C	A	intronic	0.239	(0.1597, 0.3183)	6.086E-09
VLDL.D	INO80	15	41272913	kgp19746400	G	A	intronic	0.2275	(0.1475, 0.3074)	3.863E-08
VLDL.D	FA2H (dist=58713), WDR59 (dist=40029)	16	74867442	rs17672754	G	A	intergenic	0.2546	(0.1749, 0.3343)	7.658E-10
VLDL.D	LDLRAD4	18	13370572	kgp16195388	A	G	intronic	0.2283	(0.1486, 0.3081)	3.139E-08
VLDL.D	FAM210A	18	13702482	rs75948343	T	C	intronic	0.2268	(0.1475, 0.3061)	3.254E-08
VLDL.D	NONE (dist=NONE), MIR302F (dist=42322)	18	27836554	rs4514758	G	A	intergenic	0.2345	(0.1547, 0.3143)	1.402E-08
VLDL.D	LINC01541	18	69215844	kgp16017220	G	A	ncRNA_exonic	0.2658	(0.1864, 0.3451)	1.206E-10
VLDL.D	HPN	19	35531633	rs66878130	G	A	intronic	0.2503	(0.1704, 0.3301)	1.535E-09
VLDL.D	NONE (dist=NONE), NONE (dist=NONE)	23	142147716	rs5907384	G	A	intergenic	0.2786	(0.2005, 0.3567)	7.628E-12
XL.VLDL.PL	KLHL6 (dist=6819), KLHL24 (dist=73092)	3	183280319	rs17543764	G	A	intergenic	0.2316	(0.1538, 0.3093)	8.929E-09
XXL.VLDL.L	MPEG1 (dist=119725), OR5AN1 (dist=31713)	11	59100219	kgp12652619	G	A	intergenic	0.2201	(0.1431, 0.297)	3.261E-08
XXL.VLDL.PL	EPCAM (dist=163)	2	47596124	rs3814359	A	C	upstream	0.2356	(0.1578, 0.3134)	5.142E-09
XXL.VLDL.PL	SNORD137 (dist=52652), MPDZ (dist=80419)	9	13025284	rs10514822	T	G	intergenic	0.2227	(0.1454, 0.3001)	2.697E-08
XXL.VLDL.PL	MPEG1 (dist=119725), OR5AN1 (dist=31713)	11	59100219	kgp12652619	G	A	intergenic	0.2323	(0.1552, 0.3094)	6.141E-09
XXL.VLDL.PL	OR5A1 (dist=5457), OR4D6 (dist=7388)	11	59217046	rs17153737	T	C	intergenic	0.2232	(0.1457, 0.3007)	2.662E-08
XXL.VLDL.PL	OR4D6	11	59224738	rs17153770	T	C	exonic	0.2259	(0.1481, 0.3037)	2.018E-08
XXL.VLDL.TG	MPEG1 (dist=119725), OR5AN1 (dist=31713)	11	59100219	kgp12652619	G	A	intergenic	0.2174	(0.1405, 0.2943)	4.646E-08

Table 4.5 Association results: markers of genes significantly associated with ≥ 1 metabolomic trait(s).

Gene	Metab	Function	SNP	Chr	Pos	A1	A2
RNF207	S.HDL.P	intronic	rs68032129	1	6276561	T	C
LRRC7	VLDL.D	intronic	rs17325299	1	70154480	T	C
MPZ	HDL.D, VLDL.D	UTR3	rs16832786	1	161274905	T	C
TAF1A	Alb, S.HDL.L, S.HDL.P	intronic	kgp15137813	1	222750544	C	A
GRHL1	HDL3.C	exonic	rs16867251	2	10095185	T	C
	HDL3.C	exonic	rs16867256	2	10101468	G	A
	HDL3.C	intronic	rs11902457	2	10103414	T	C
	HDL3.C	intronic	rs6735658	2	10116084	G	A
NPHP1	Alb, S.HDL.P, VLDL.D	intronic	rs13403558	2	110885620	C	A
	Alb, S.HDL.L, S.HDL.P	intronic	rs11898910	2	110937101	T	C
LINC00116	Alb, S.HDL.L, S.HDL.P	ncRNA_intronic	rs10171646	2	110977357	G	A
MIR4435-2HG	HDL3.C	ncRNA_exonic	rs2292932	2	112123745	C	A
CPS1	Gly	exonic	rs1047891	2	211540507	C	A
INPP5D	AcAce	intronic	rs7570320	2	234075691	C	A
KBTBD8	CH2.in.FA	exonic	rs13096789	3	67054649	T	C
EPHB1	VLDL.D	intronic	rs1554675	3	134826917	G	A
PDCD6	Alb, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	intronic	rs1574220	5	314518	G	A
	Alb, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	UTR3	rs7736	5	314935	G	A

(Continued on next page)

Table 4.5 Association results: markers of genes significantly associated with ≥ 1 metabolomic trait(s) (cont'd).

Gene	Metab	Function	SNP	Chr	Pos	A1	A2
MSH3	CH2.DB	intronic	rs863214	5	79984714	T	C
	CH2.DB	intronic	rs6151838	5	80082865	G	A
	CH2.DB, CH2.in.FA, FAw79S.FA	intronic	rs10075024	5	80132177	T	C
SLIT3	HDL.D	intronic	rs13154825	5	168484494	C	A
TIAM2	VLDL.D	intronic	rs6915661	6	155411606	G	A
CTTNBP2	Alb, CH2.in.FA, FALen, FAw79S.FA, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	intronic	rs10247163	7	117389972	T	C
	Alb, CH2.in.FA, FALen, FAw79S.FA, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	intronic	rs10258815	7	117401398	G	A
	Alb, CH2.in.FA, FALen, FAw79S.FA, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	intronic	rs10254610	7	117406238	G	A
	HDL3.C, VLDL.D	intronic	rs929668	7	117445664	G	A
SLC20A2	FAw79S.FA	exonic	kgp20396476	8	42296993	A	G
LOC392232	HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	ncRNA_intronic	rs6989765	8	73150491	G	A
	HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	ncRNA_intronic	rs987242	8	73153467	G	A
ASTN2-AS1	HDL3.C	ncRNA_intronic	kgp22748805	9	119321341	A	G

(Continued on next page)

Table 4.5 Association results: markers of genes significantly associated with ≥ 1 metabolomic trait(s) (cont'd).

Gene	Metab	Function	SNP	Chr	Pos	A1	A2
GFI1B	LDL.D, VLDL.D	intronic	rs685783	9	135855768	G	A
KIAA1217	CH2.in.FA, FALen	intronic	rs2150650	10	24791582	T	C
	CH2.in.FA, FALen	intronic	rs12252802	10	24798368	G	A
PRKG1	Crea	intronic	rs1917841	10	53356952	T	C
CTBP2	Alb	intronic	rs2949367	10	126714256	G	A
LINC01495	Gp	ncRNA_intronic	kgp12607087	11	22472250	A	G
OR4D6	XXL.VLDL.PL	exonic	rs17153770	11	59224738	T	C
MYRF	otPUFA	intronic	rs108499	11	61547237	T	C
	otPUFA	intronic	rs509360	11	61548559	G	A
	otPUFA	exonic	rs174535	11	61551356	C	A
TMEM258	otPUFA	intronic	rs102275	11	61557803	G	A
	otPUFA	UTR5	rs174538	11	61560081	G	A
FADS1	otPUFA	UTR3	rs174546	11	61569830	T	C
	otPUFA	intronic	rs174547	11	61570783	T	C
	otPUFA	intronic	rs174548	11	61571348	G	C
	otPUFA	intronic	rs174550	11	61571478	T	C
FADS2	otPUFA	intronic	rs174570	11	61597212	T	C
	otPUFA	intronic	rs1535	11	61597972	G	A
	otPUFA	intronic	rs174576	11	61603510	C	A
	otPUFA	intronic	rs174577	11	61604814	C	A
	otPUFA	intronic	rs174583	11	61609750	T	C

(Continued on next page)

Table 4.5 Association results: markers of genes significantly associated with ≥ 1 metabolomic trait(s) (cont'd).

Gene	Metab	Function	SNP	Chr	Pos	A1	A2
FADS3	otPUFA	intronic	rs77980989	11	61654092	C	A
GRIK4	VLDL.D	intronic	rs10502240	11	120547912	T	C
	VLDL.D	intronic	rs11607732	11	120548318	G	A
BEST3	Alb, HDL3.C	intronic	rs2068191	12	70092000	G	A
MTMR6	HDL3.C, VLDL.D	exonic	rs17082035	13	25823451	G	A
	HDL3.C, VLDL.D	intronic	rs17082070	13	25842806	T	C
EDNRB-AS1	HDL.D, LDL.D	ncRNA_intronic	kgp16788601	13	78451158	A	G
FARP1	HDL.D, VLDL.D	intronic	rs9584835	13	99054527	T	C
PCK2	CH2.DB, CH2.in.FA, FAw79S.FA, HDL.D, HDL3.C, LDL.D, S.HDL.P, VLDL.D	intronic	rs9783666	14	24564684	T	G
	HDL.D, HDL3.C, LDL.D, VLDL.D	UTR3	rs2759407	14	24569947	G	A
VRK1	VLDL.D	intronic	rs10147248	14	97268547	G	A
	HDL3.C, VLDL.D	intronic	rs76904997	14	97277922	C	A
INO80	HDL3.C, VLDL.D	intronic	kgp19746400	15	41272913	G	A
ABHD2	CH2.DB, CH2.in.FA, FALen, FAw79S.FA	intronic	rs4932475	15	89689583	T	G
SNX29	S.HDL.L, S.HDL.P	intronic	kgp16526588	16	12278595	A	G
KIAA0895L	Glc	intronic	rs13339140	16	67213923	T	C
E2F4	Glc	intronic	rs3730403	16	67229486	T	G
LRRC29	Glc	UTR3	rs12051247	16	67241282	A	G

(Continued on next page)

Table 4.5 Association results: markers of genes significantly associated with ≥ 1 metabolomic trait(s) (cont'd).

Gene	Metab	Function	SNP	Chr	Pos	A1	A2
	Glc	intronic	rs13338688	16	67248831	G	A
CTCF	Glc	intronic	rs17686899	16	67627635	T	G
	Glc	intronic	rs7191281	16	67655133	T	C
	Glc	UTR3	rs6499137	16	67671804	C	A
NXN	FAw79S.FA	intronic	rs78895411	17	718968	T	G
	FAw79S.FA	intronic	rs74426014	17	734321	G	A
LDLRAD4	VLDL.D	intronic	kgp16195388	18	13370572	A	G
FAM210A	VLDL.D	intronic	rs75948343	18	13702482	T	C
LINC01541	HDL.D, HDL3.C, LDL.D, VLDL.D	ncRNA_exonic	kgp16017220	18	69215844	G	A
LINC00908	ApoA1	ncRNA_intronic	rs7232061	18	74269894	G	A
HPN	VLDL.D	intronic	rs66878130	19	35531633	G	A
NECTIN2	L.LDL.FC, LDL.C, M.LDL.C, M.LDL.L, S.LDL.C, S.LDL.L	intronic	rs7254892	19	45389596	G	A

4.3.1 Polyunsaturated fatty acids and the FADS1/FADS2 loci

As shown in Figure 4.8, for polyunsaturated fatty acids (PUFA) other than 18:2, the strongest association was found at the FADS2 locus on chromosome 11 (rs174577; p-value $= 1.731 \times 10^{-13}$). FADS2 is a protein-coding gene in the fatty acid desaturase (FADS) gene family. It is related to alpha-linolenic (omega3) and linoleic (omega6) acid metabolism pathways [Stelzer et al., 2016]. The association extended upstream, encompassing MYRF (a myelin regulatory factor gene), TMEM258 (transmembrane protein 258, a protein-coding gene) and FADS1 (fatty acid desaturase 1, another member of the FADS gene family).

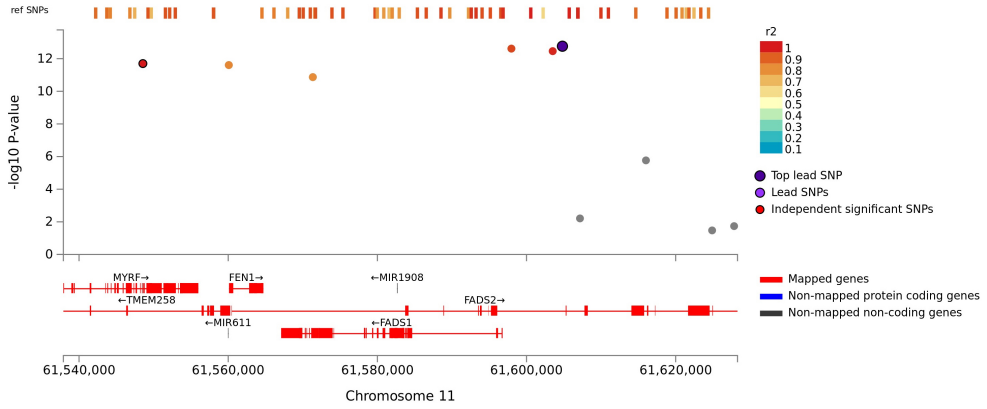


Fig. 4.8 Regional plot showing the genomic risk loci associated with otPUFA.

4.3.2 Lipid/FA related metabolites and the CTTNBP2 locus

The CTTNBP2 locus (the cortactin-binding protein 2 gene) on chromosome 7 was found to be strongly associated with a large number of metabolomic measurements related to fatty acids and lipids, including albumin, average number of methylene groups in a fatty acid chain, average fatty acid chain length, ratio of omega-9 and saturated fatty acids to total fatty acids, mean diameter for HDL particles, total cholesterol in HDL3, mean diameter for LDL/IDL particles, concentration of small HDL particles and mean diameter for VLDL particles (VLDL.D).

Take VLDL.D for example. It was significantly associated with rs10254610 (an intronic variant at the CTTNBP2 locus) with a p-value of 8.334×10^{-14}. The association weakened while extending upstream, covering CFTR, the cystic fibrosis (CF) transmembrane conductance regulator gene (Figure 4.9).

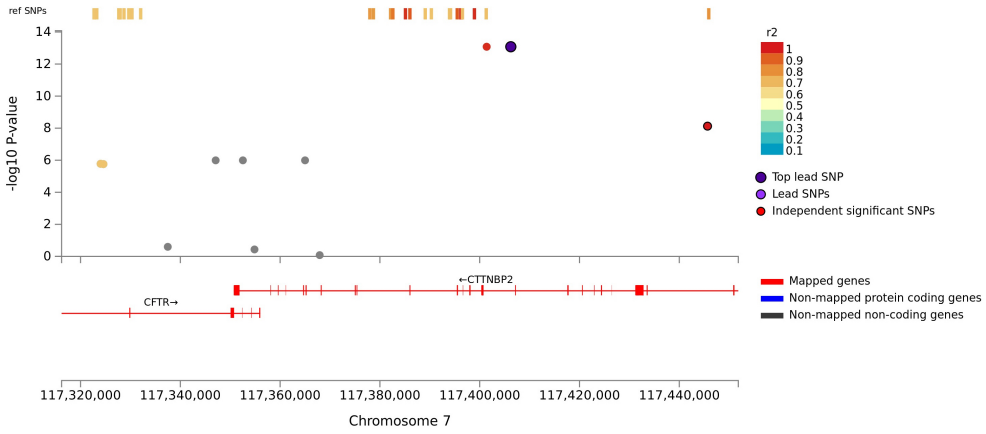


Fig. 4.9 Regional plot showing the association between VLDL.D and the CTTNBP2 locus.

4.3.3 Total cholesterol in HDL3 and the GRHL1 locus

As shown in Figure 4.10, total cholesterol in HDL3 (HDL3.C) was significantly associated with two independent SNPs at the GRHL1 locus (the grainyhead like transcription factor 1 gene) on chromosome 2 – rs16867256 (exonic; p-value $= 1.335 \times 10^{-8}$) and rs6735658 (intronic; p-value $= 1.476 \times 10^{-8}$).

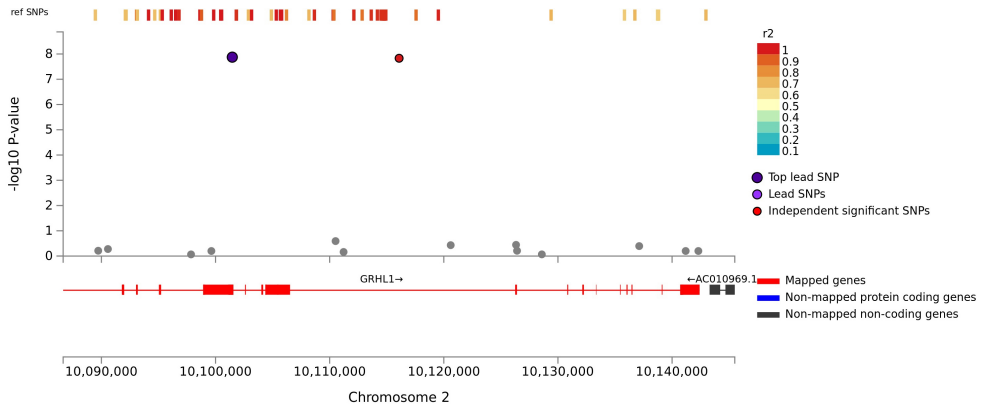


Fig. 4.10 Regional plot showing the association between HDL3.C and the GRHL1 locus.

4.3.4 Glucose and the LRRC29 locus

Glucose (Glc) was found to be significantly associated with rs13338688 (an intronic variant at the LRRC29 locus on chromosome 16) with a p-value of 1.262×10^{-10} (Figure 4.11).

LRRC29 (leucine rich repeat containing 29) encodes a member of the F-box protein family [Stelzer et al., 2016].

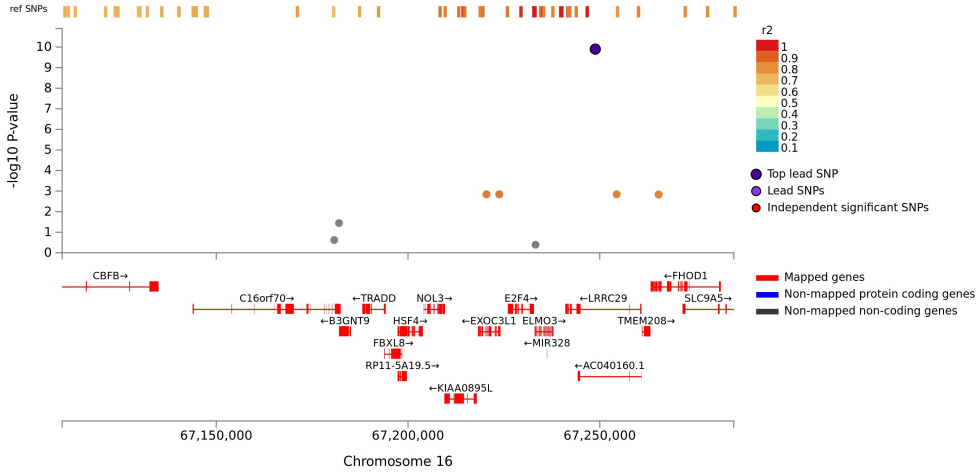


Fig. 4.11 Regional plot showing the association between Glc and the LRRC29 locus.

4.3.5 Overlap with previous studies

17 of the 123 SNPs significantly associated with one or more metabolomic measurements identified in my study have also been reported in previous GWAS studies. These are listed in Table 4.6. The two sets of findings are rather consistent with each other. Unfortunately, none of the previous reports is directly about LDD or other musculoskeletal disorders.

Table 4.6 Previously reported GWAS hits identified in this study.

Gene	Metab	SNP	Chr	Pos	Previous report(s)
CPS1	Gly	rs1047891	2	211540507	Metabolite levels (small molecules and protein measures), HDL cholesterol, plasma homocysteine levels (post-methionine load test), glomerular filtration rate (creatinine)
MYRF	otPUFA	rs108499	11	61547237	Trans fatty acid levels
	otPUFA	rs509360	11	61548559	Trans fatty acid levels, plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid)
	otPUFA	rs174535	11	61551356	Blood metabolite levels, red blood cell fatty acid levels, trans fatty acid levels, plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid)
TMEM258	otPUFA	rs102275	11	61557803	Crohn's disease, metabolic syndrome, metabolite levels, trans fatty acid levels, oleic acid (18:1n-9) levels, plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), glycerophospholipid levels, phospholipid levels (plasma), palmitoleic acid (16:1n-7) levels, plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), stearic acid (18:0) levels
	otPUFA	rs174538	11	61560081	Blood metabolite levels, trans fatty acid levels, plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid)

(Continued on next page)

Table 4.6 Previously reported GWAS hits identified in this study (cont'd).

Gene	Metab	SNP	Chr	Pos	Previous report(s)
FADS1	otPUFA	rs174546	11	61569830	C-reactive protein levels or triglyceride levels (pleiotropy), HDL cholesterol, metabolic syndrome, trans fatty acid levels, glycerophospholipid levels, triglycerides, total cholesterol, C-reactive protein levels or HDL-cholesterol levels (pleiotropy), LDL cholesterol, plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid)
	otPUFA	rs174547	11	61570783	Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), trans fatty acid levels, glycerophospholipid levels, metabolite levels (lipid measures), lipid metabolism phenotypes, plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid), comprehensive strength and appendicular lean mass, HDL cholesterol, metabolite levels, height, age-related disease endophenotypes, mortality and associated endophenotypes, metabolic traits, plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), triglycerides, plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), sphingolipid levels, resting heart rate

(Continued on next page)

Table 4.6 Previously reported GWAS hits identified in this study (cont'd).

Gene	Metab	SNP	Chr	Pos	Previous report(s)
	otPUFA	rs174548	11	61571348	Trans fatty acid levels, plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid), HDL cholesterol, metabolite levels, blood metabolite levels, delta-6 desaturase activity, hematology traits, triglycerides, blood metabolite ratios
	otPUFA	rs174550	11	61571478	Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid), trans fatty acid levels, plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid), fasting glucose-related traits (interaction with BMI), glycerophospholipid levels, red blood cell fatty acid levels, plasma omega-6 polyunsaturated fatty acid levels (adrenic acid), blood metabolite levels, triglyceride levels, HDL cholesterol levels, fasting glucose-related traits
FADS2	otPUFA	rs174570	11	61597212	Trans fatty acid levels, total cholesterol, HDL cholesterol, glycated hemoglobin levels, LDL cholesterol
	otPUFA	rs1535	11	61597972	Metabolic syndrome, trans fatty acid levels, glycerophospholipid levels, inflammatory bowel disease, total cholesterol levels, LDL cholesterol levels, plasma omega-3 polyunsaturated fatty acid levels (docosapentaenoic acid), plasma omega-3 polyunsaturated fatty acid levels (alphalinolenic acid), response to statin therapy
	otPUFA	rs174576	11	61603510	Trans fatty acid levels, bipolar disorder, glycerophospholipid levels

(Continued on next page)

Table 4.6 Previously reported GWAS hits identified in this study (cont'd).

Gene	Metab	SNP	Chr	Pos	Previous report(s)
	otPUFA	rs174577	11	61604814	Trans fatty acid levels, P wave duration, QRS duration, plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid), iron status biomarkers (transferrin levels)
	otPUFA	rs174583	11	61609750	Trans fatty acid levels, QT interval, response to statin therapy
NECTIN2	LDL set A[10]	rs7254892	19	45389596	Total cholesterol levels
Intergenic	LDL set B[11]	rs445925	19	45415640	Response to statins (LDL cholesterol change), apolipoprotein Levels, carotid intima media thickness, ideal cardiovascular health (clinical), LDL cholesterol levels, metabolite levels, blood metabolite ratios, lipoprotein-associated phospholipase A2 activity and mass

[10]LDL set A includes L.LDL.FC, LDL.C, M.LDL.C, M.LDL.L, S.LDL.C and S.LDL.L.

[11]LDL set B includes L.LDL.C, L.LDL.FC, L.LDL.L, L.LDL.P, L.LDL.PL, LDL.C, M.LDL.C, M.LDL.CE, M.LDL.L, M.LDL.P, S.LDL.C and S.LDL.L.

4.4 Discussion

In this study, 130 genome-wide association studies (GWAS) for serum ^{1}H NMR metabolomic measurements were performed based on a population cohort of 571 individuals, in order to gain knowledge on the underlying genetics of serum metabolism. 123 unique SNPs significantly associated with one or more metabolomic measurements were identified; among them, exonic, intronic and UTR3 variants were enriched, whereas intergenic variants were underrepresented. There were altogether 42 metabolomic measurements with one or more significantly associated SNP(s), most of them related to lipids and fatty acids.

4.4.1 Discussion of selected significant loci

4.4.1.1 Polyunsaturated fatty acids and the FADS1/FADS2 loci

A strong association between polyunsaturated fatty acids (PUFA) other than 18:2 and the FADS1/FADS2 loci has been identified. FADS2 is an important paralog of FADS1 [Stelzer et al., 2016], and FADS1 has been found to be associated with lipid metabolism disorders [Tian et al., 2016; Gromovsky et al., 2018].

Previous studies have also shown that the FADS gene cluster could influence lipid and PUFA levels in European [Glaser et al., 2010], Chinese [P. Li et al., 2018] and Korean [S. Lee et al., 2018] populations. Indeed, the FADS genes encode delta-5 desaturase and delta-6 desaturase enzymes, which are crucial in regulating the synthesis of long-chain PUFA [Lattka et al., 2010]. By playing an important role in synthesizing and regulating PUFAs, the FADS genes could largely affect the metabolism of essential fatty acids and potentially, the well-being of an individual [S. Lee et al., 2018].

4.4.1.2 Lipid/FA related metabolites and the CTTNBP2 locus

The CTTNBP2 locus was found be significantly associated with a group of metabolomic measurements related to fatty acids and lipids. The association extended upstream and weakly covers CFTR.

As a protein-coding gene, CTTNBP2 mainly controls dendritic spinogenesis and dendritic spine maintenance [Y.-K. Chen et al., 2012]. In mouse, it has been discovered to be related to

cytoskeletal protein binding [Hill et al., 2004]. Relatively little past literature exists regarding the role of CTTNBP2 in lipid/FA metabolomics, except for a recent study demonstrating the association between CTTNBP2 and triglycerides [Hebbar et al., 2018].

On the other hand, the CFTR gene is relatively more relevant to blood lipid levels since it is related to the CFTR metabolic syndrome. The gene encodes the CFTR protein, which builds a salt channel that allows chloride to move in and out of body cells [Borowitz et al., 2009]. If the salt channel malfunctions, chloride could not move freely as intended, and abnormal mucus would form in the pancreas and other organs, leading to the development of CF [Borowitz et al., 2009]. CF often harms digestion and hence proper absorption of fat and protein, resulting in nutrition problems and poor growth [Borowitz et al., 2009]. Nevertheless, this relationship between CFTR and lipid traits is not directly causal. Besides, in my study, no significant independent SNPs at the CFTR locus has been identified. Therefore, the strong association between lipid/FA related traits and CTTNBP2 might not be related to CFTR pathways. Further study is needed to replicate my findings and to understand the role of CTTNBP2 in FA/lipid metabolism better.

4.4.1.3 Total cholesterol in HDL3 and the GRHL1 locus

The GRHL1 gene encodes a transcription factor from the grainyhead family, and regulates lipid metabolism by peroxisome proliferator-activated receptor alpha [Stelzer et al., 2016]. It has also been found to be a risk locus for breast cancer [Michailidou et al., 2017] and prostate cancer [Eeles et al., 2013]. The link between GRHL1 and total cholesterol in HDL3 remains unclear.

4.4.1.4 Glucose and the LRRC29 locus

LRRC29 (the leucine-rich repeat containing 29 gene) is a protein-coding gene. Previous GWAS studies have identified its association with HDL cholesterol [Nagy et al., 2017; Nagy et al., 2017]. HDL has been found to be involved in glucose metabolism since it could control the homeostasis of glucose by mechanisms like insulin secretion [Drew et al., 2012]. It is possible that the signal found in my study is actually mediated by HDL.

4.4.2 Limitation of this study

Based on the common disease / common variant hypothesis, GWAS seeks to identify common variants significantly associated with traits of interest. Since common variants typically have relatively small effect sizes [Goldstein et al., 2009], GWAS requires a large sample size in order to be able to detect significant associations. One limitation of this study is that our sample size was relatively small (571 individuals in total). I only have a moderate confidence on the novel findings and future replication studies are much needed.

Unfortunately, measuring the metabolome is still quite expensive at the moment. As a result, most of the metabolomic GWAS suffer from small sample sizes. One possible way to boost the power of GWAS with limited data is employing multi-trait methods [Porter and O'Reilly, 2017] on high-dimensional metabolomic data, taking advantage of the fact that the metabolomic measurements are closely correlated to each other.

5

Associating different phenotypes with metabolomic measurements via polygenic scoring

5.1 Introduction

With the rapid development of nuclear magnetic resonance (NMR) spectroscopy techniques, hundreds of quantitative metabolomic measurements could now be taken for cohorts with biological samples [Suhre and Gieger, 2012]. These metabolomic measurements could be treated as "intermediate" phenotypes[1] linking genomic data to phenotypic data, which are valuable as potential biomarkers for different conditions and traits [Kettunen et al., 2016].

Since accurately and precisely measuring the human metabolome could be quite costly and require state-of-the-art equipment, cohorts with metabolomic data are often quite small compared with, say, cohorts with genomic data. Since GWAS loci for metabolomic traits typically have a large effect size [Gieger et al., 2008; Rhee et al., 2013], it is possible to "estimate" the human metabolome based on GWAS summary statistics for metabolomic measurements and genomic data through genetic risk prediction methods like polygenic scoring. By analyzing the relationship between different phenotypes and the estimated metabolomic traits, we could understand the metabolomic context of the phenotypes better.

[1] Recall that as shown in Figure 1.11, metabolomic data lies between genomic data and phenotypic data in the overall flow of omics information.

Additionally, this metabolome estimation model could be reused in cohorts (preferably of a similar population) with GWAS data and no metabolomic data, aiding future researchers.

In this chapter, the metabolome is estimated through polygenic scoring using genomic data and GWAS summary statistics from previous meta-analyses (c.f. Section 4.2.7). The association between estimated metabolomic traits and various phenotypes is next tested using regression methods, and potential metabolomic biomarkers for the phenotypes are identified.

5.2 Materials and methods

5.2.1 Study sample

Following the procedures stated in Section 2.2.5, the serum samples of 814 individuals were obtained for the application of ^{1}H NMR spectroscopy, and 137 metabolomic measurements were recorded for each individual.

After data filtering and normalization (c.f. Section 2.3.2) to reduce noise and increase the robustness of consequent analyses, the metabolomic data set included 130 metabolomic measurements (c.f. Table 2.10) for 757 individuals. Furthermore, to reduce the dimensionality of our data, 66 metabolomic features were defined through hierarchical clustering and dynamic tree cutting, which are listed in Table 2.11.

Among the 757 subjects, 571 also had GWAS data (i.e. "Data set I" in Figure 2.10). Procedures for genotyping are described in Section 2.2.4.

This study also utilized height, weight, body mass index (BMI), amount of cigarette smoking, five clinical phenotypes (c.f. Section 2.13) and the composite LDD phenotypes defined in Section 2.3.1.3 based on the MRI reads of two experienced physicians, Dr. Jaro Karppinen (JK) and Dr. Dino Samartzis (DS). The amount of subjects with both genotype and phenotype data, but no metabolomic data (i.e. "Data set II" in Figure 2.10) was phenotype-dependent, ranging from 632 to $1,214$.

5.2.2 Polygenic scoring

Lassosum [Mak et al., 2017], a genetic risk prediction (GRIP) software, was used to calculate polygenic risk scores (PRS) of the 116 metabolomic measurements[2] for all the people with genetic data in our cohort (2,113 individuals in total). Based on a penalized regression framework, Lassosum could account for linkage disequilibrium (LD) and constructs PRS utilizing both GWAS summary statistics and a reference panel [Mak et al., 2017]. It has been shown that Lassosum could achieve better predictive accuracy than other PRS-based GRIP methods [Mak et al., 2017].

In my study, the standard pipeline of Lassosum was used. The base data files were set to be the combined GWAS summary statistics from the meta-analyses (c.f. Section 4.2.7), whereas our genetic data was selected to be the reference panel. The LD regions defined for the east Asian population in [Berisa and Pickrell, 2016] were considered. Only chromosomes 1 to 22 were used in PRS construction, and pseudo-validation was performed to select the best set of parameters for Lassosum when calculating the PRS.

5.2.3 Regression analysis: one phenotype, one metabolomic PRS

In order to look into the relationship between metabolomic polygenic risk scores (PRS) and the 40 phenotypes listed in Table 2.13, a set of regression analyses was conducted. To avoid over-fitting, all the analyses were performed on a subset of people with both GWAS and phenotypic data, but no metabolomic data (i.e. I removed the people that are among the 571 I performed the original GWAS on – this resulted in "Data set II" in Figure 2.10).

For each metabolomic feature X,

- For each continuous phenotype, I regressed it on age, sex, the PRS for X and the first ten ancestry informative principal components $PC_1, \ldots, PC_{10}$.

$$\begin{aligned} Phen = b_0 + b_1 \cdot AGE + b_2 \cdot SEX + b_3 \cdot PRS_X \\ + b_4 \cdot PC_1 + \cdots + b_{13} \cdot PC_{10} + e \end{aligned} \tag{5.1}$$

[2] Recall that even though we have 130 metabolomic measurements, only 116 of them could be matched with the traits considered in the other group's study I use for meta-analysis. Please refer to Section 4.2.7 for more details.

- For each binary phenotype, I performed logistic regression regressing the phenotype status (TRUE or FALSE) on age, sex, the PRS for X and the first ten ancestry informative principal components $PC_1, \ldots, PC_{10}$.

$$\begin{aligned} logit(Phen) = b_0 + b_1 \cdot AGE + b_2 \cdot SEX + b_3 \cdot PRS_X \\ + b_4 \cdot PC_1 + \cdots + b_{13} \cdot PC_{10} + e \end{aligned} \tag{5.2}$$

Before running each regression, the PRS were standardized so that b_3, our focus of interest, is more interpretable. The p-value attached to b_3 was recorded, and a 95% confidence interval for b_3 was calculated.

5.2.3.1 Controlling for multiple testing

Since $4,640$ regression models (40 phenotypes $\times$ 116 metabolomic PRS[3]) were fitted in total, the issue of multiple testing arose. To circumvent this, two types of FDR-based approaches were adopted.

The aggregated FDR approach

In this approach, all $4,640$ p-values were lumped together, and the aggregated FDR was calculated following the B-H FDR procedure (c.f. Section 3.2.2.1). If for a *Phen-MetabPRS* (phenotype-metabolomic PRS) pair, the adjusted p-value of b_3 was less than 0.1, significance of the association between the phenotype and the metabolomic PRS was declared.

The adaptive group FDR approach

Since certain metabolites may have a group of associated phenotypes with relative low, but insignificant q-values and vice versa, it is more robust to also take into consideration group information when calculating the FDR. This could be achieved through the group B-H procedure [Hu et al., 2010].

Assume that the hypotheses could be partitioned into m groups, $g_1, \ldots, g_m$. Further assume that our scenario is an "oracle" case – for each group g_j ($j = 1, \ldots, m$), we already know π_j,

[3]In this study, I did not use the reduced metabolomic features (c.f. Section 2.3.2.3) mainly because (1) in this study, the sample size is large enough to provide a reasonably good statistical power for analysis; and (2) if I use the original measurements instead of the reduced features, my results would be more interpretable.

the proportion of true H_0, which is the number of true H_0 in group g_j divided by the total number of tests in g_j.

1. In group g_j, for each p-value $p_{i,j}$ ($i = 1, \ldots, n_j$, where n_j is the number of hypotheses in group g_j), calculate the corresponding weighted p-value $p^w_{i,j} = p_{i,j} \cdot \frac{\pi_j}{1-\pi_j}$.
   - If $\pi_j = 1$ (i.e. in group g_j all the H_0 are true), set $p^w_{i,j}$ to ∞.
   - If $\pi_j = 1$ for all g_j ($j = 1, \ldots, m$), accept all the H_0 and stop.
2. Pool all the weighted p-values and sort them in ascending order $p^w_{(1)} \leq \cdots \leq p^w_{(N)}$, where $N = \sum_{j=1}^{m} n_j$ is the total number of hypotheses.
3. For a given level of significance α, calculate the weighted α by $\alpha^w = \frac{\alpha}{1-\pi_0}$, where π_0 is the overall proportion of true H_0.
4. Find the largest k such that $p^w_{(k)} \leq \frac{k\alpha^w}{N}$.
   - If k exists, reject the k hypotheses associated with $p^w_{(1)}, \ldots, p^w_{(k)}$.
   - If k does not exist, reject none of the hypotheses.

In practice, $\pi_0, \pi_1, \ldots, \pi_m$ could be estimated by various techniques, e.g. the least slope method [Benjamini and Hochberg, 2000].

This study utilized the adaptive group B-H procedure to control the FDR at 0.1. First, for each group g_j, π_j was estimated by $\hat{\pi}_j$ using the least slope method. The above "oracle" group B-H algorithm was next applied with all π_j replaced by $\hat{\pi}_j$. Two grouping schemes were considered – (1) group by metabolomic measurement (116 groups, 40 p-values in each group); and (2) group by phenotype (40 groups, 116 p-values in each group).

5.2.4 Regression analysis: one phenotype, multiple metabolomic PRS

Section 5.2.3 fitted one regression model for each *Phen-MetabPRS* pair. Instead of doing this, we could also regress each phenotype on age, sex, the first ten ancestry informative principal components and all the PRS for metabolomic measurements. Only 40 models would be fitted in this way, and by incorporating more information, the resulting models would be able to account for more variation in the phenotypes.

One hindrance to this approach is that our sample size is quite limited (632 to $1,214$, depending on the phenotype), and there could be as many as $116+12=128$ explanatory variables in the model.

The model could be formulated as a hypothesis testing problem – H_0: none of the independent variables explain any of the variability in the phenotype vs. H_1: at least one of the regression coefficients is different from 0. This could be tested with an F test, and we could estimate the statistical power for this test.

Assume that our sample size for a certain phenotype is 800 and that the model could only explain 5% of the variance in the phenotype. Then the effect size is $0.05/(1-0.05)\approx 0.0526$.

For a full model with 128 explanatory variables, the numerator degrees of freedom for the F test is 128, and the denominator degrees of freedom is $800-128-1=671$. The power of the test with a significance level of 0.05 is approximately 0.6874, which is pretty low.

5.2.4.1 Dimensionality reduction on metabolomic PRS

To increase power, we could first perform dimensionality reduction on the metabolomic PRS according to the clusters defined in Table 2.11. Since the PRS was calculated based on the meta-analysis results and only 116 metabolomic measurements were considered, I removed the 14 missing traits from the defined clusters. Empty metabolomic groups were dropped, and the composite metabolomic PRS were re-calculated based on the new metabolomic groups. In this way, 59 new metabolomic PRS were defined, including 32 composite ones (calculated as the average of all the metabolomic PRS in that cluster) and 27 single ones (essentially the original metabolomic PRS).

Now, the full model only has $59+12=71$ explanatory variables. The numerator degrees of freedom for the F test is 71, and the denominator degrees of freedom is $800-71-1=728$. The power of the test with a significance level of 0.05 is approximately 0.8662, which is much better.

Another advantage of performing dimensionality reduction on metabolomic PRS first is the decrease in multicollinearity. Multiple regression models assume independence among the explanatory variables, and it is obvious that many of the metabolomic traits (and the corresponding PRS) are strongly associated with each other. For instance, it is safe to say that the PRS of S.HDL.L could be linearly predicted by the PRS of S.HDL.P with substantial accuracy. Since the sample size is too small, it is also unrealistic to add metabolite-metabolite

interaction terms to the model. By defining only one composite metabolomic PRS for a cluster of highly correlated metabolomic measurements, multicollinearity is controlled for, rendering the fitted model more precise.

5.2.4.2 Model fitting and selection

Again, to avoid overfitting, the regression analysis was performed on the people with both GWAS and phenotypic data, but no metabolomic data. Before running each regression, the "composite" metabolomic PRS were standardized.

The full regression models were quite straightforward. Each continuous phenotype was regressed on age, sex, the first ten ancestry informative principal components $PC_1, \ldots, PC_{10}$ and the 59 "composite" metabolomic PRS.

$$\begin{aligned} Phen = b_0 + b_1 \cdot AGE + b_2 \cdot SEX + b_3 \cdot PC_1 + \cdots + b_{12} \cdot PC_{10} \\ + b_{13} \cdot PRS_1 + \cdots + b_{71} \cdot PRS_{59} + e \end{aligned} \tag{5.3}$$

On the other hand, for each binary phenotype, we performed logistic regression regressing the phenotype status (TRUE or FALSE) on age, sex, the first ten ancestry informative principal components $PC_1, \ldots, PC_{10}$ and the 59 "composite" metabolomic PRS.

$$\begin{aligned} logit(Phen) = b_0 + b_1 \cdot AGE + b_2 \cdot SEX + b_3 \cdot PC_1 + \cdots + b_{12} \cdot PC_{10} \\ + b_{13} \cdot PRS_1 + \cdots + b_{71} \cdot PRS_{59} + e \end{aligned} \tag{5.4}$$

Using the R package "MASS", bidirectional stepwise model selection based on Akaike information criterion[4] (AIC) was performed.

Starting with a full model, at each step, the addition/deletion of each variable was tested with AIC as the criterion. When the inclusion of a variable gave the most significant improvement of the fitted model, it was added; when the exclusion of a variable improved the fit most significantly, it was dropped. This process was repeated until no further improvement of the model (via adding or dropping variables) can be made.

For each phenotype, the p-value of the final selected model was recorded. The aggregated FDR was controlled at the level of 0.1 through the B-H procedure.

[4]For a given data set, the AIC estimates the relative information loss of a statistical model [Akaike, 2011]. The smaller the AIC is, the less information the model loses, and the higher the quality of that model is.

5.2.5 Penalized regression analysis

Section 5.2.4 attempted to include all 116 metabolomic PRS as explanatory variables in one multiple regression model. Since the sample size is a bit limited for the giant saturated model, feature selection was conducted to increase power. The selection process was two-fold: (1) dimensionality reduction on metabolomic PRS prior to model fitting; and (2) bidirectional stepwise model selection after fitting the full model. This approach has some drawbacks. To begin with, it would not necessarily produce the best model if there are redundant predictors [Judd et al., 2011]. Furthermore, the estimated regression coefficients may be biased and require shrinkage [Tibshirani, 1996]. To circumvent these issues, a set of penalized regression analyses was performed.

5.2.5.1 A brief introduction to penalized regression methods

Given a predictor matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $Y \in \mathbb{R}^n$ (where n is the number of samples and p is the number of features), the ordinary least squares regression (e.g. the models shown in Equations 5.1 and 5.3) could be formulated as an optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 \tag{5.5}$$

Penalized regression methods achieve regularization simply by adding a penalty term. For example, ridge regression [Hoerl and Kennard, 1970] aims to solve the below convex optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \tag{5.6}$$

where $\lambda \geq 0$ is a tuning parameter and $\|\beta\|_2^2 = \sum_{i=1}^p |\beta_i|^2$ is the squared L_2 norm of β.

Similarly, Lasso [Tibshirani, 1996] aims to solve the below convex optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{5.7}$$

where $\lambda \geq 0$ is a tuning parameter and $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ is the L_1 norm of β.

The basic idea behind Lasso (ridge regression) is to force the L_1 (squared L_2) norm of β to be small so that the model is regularized. In the case of ridge regression, the size of the coefficients are shrunk – usually, none of them is set to zero. On the contrary, since Lasso

uses the L_1 norm instead, many of the regression coefficients would be forced to 0, effectively choosing a simpler model without those trivial coefficients. Therefore, Lasso could both avoid over-fitting and render the fitted model more sparse and interpretable (essentially feature selection).

In terms of prediction accuracy, Lasso has been shown to outperform stepwise regression when the signal-to-noise ratio (SNR) is low [Hastie et al., 2017].

5.2.5.2 Model fitting

Since Lasso is better suited for sparse scenarios (i.e. small number of non-zero β), it is the choice of penalized regression method in my study. To avoid over-fitting, the models were fitted using a subset of people with both GWAS and phenotypic data, but no metabolomic data (i.e. the people that were among the 571 I performed the original GWAS on were dropped).

Adapting the model shown in Equation 5.8 (note that all 116 metabolomic PRS were considered[5]), 40 Lasso models were fitted using the R package "glmnet" [Friedman et al., 2010]. The tuning parameter λ was selected via ten-fold cross validation. The loss used for cross validation was squared error for Gaussian models (i.e. continuous phenotypes) and deviance for logistic models (i.e. binary phenotypes).

$$\begin{aligned} Phen \text{ or } logit(Phen) = b_0 + b_1 \cdot AGE + b_2 \cdot SEX + b_3 \cdot PC_1 + \cdots + b_{12} \cdot PC_{10} \\ + b_{13} \cdot PRS_1 + \cdots + b_{128} \cdot PRS_{116} + e \end{aligned} \tag{5.8}$$

5.3 Results

5.3.1 Regression analysis: one phenotype, one metabolomic PRS

5.3.1.1 Based on aggregated FDR

Among the $4,640$ tests, 146 were significant at a FDR cut-off of 0.1. As could be seen in Table 5.1, it is fairly safe to say that almost all the p-values are truly significant if we set the

[5] Again, in this study, I did not use the reduced metabolomic features (c.f. Section 2.3.2.3) because (1) in this study, the sample size is large enough to provide a reasonably good statistical power for analysis; and (2) if I use the original measurements instead of the reduced features, my results would be more interpretable.

FDR cut-off to be 0.01. Nevertheless, we could still check all the *Phen* $\sim$ single *MetabPRS* pairs with q-value < 0.1, bearing in mind that some of them could be falsely significant. All the *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 at an aggregated FDR cut-off of 0.1 are listed in Table 5.2.

Table 5.1 # of significant *Phen* $\sim$ single *MetabPRS* pairs at different aggregated FDR cut-offs.

Interval	Count	# of expected falses
$(-\infty, 0.01)$	120	1.2
$(-\infty, 0.05)$	134	6.7
$(-\infty, 0.1)$	146	14.6

Height

There were 4 *Height* $\sim$ single *MetabPRS* regression models with a significant b_3. The related metabolomic measurements were the average number of methylene groups per a double bond (CH2.DB), other polyunsaturated fatty acids than 18:2 (otPUFA), omega-3 fatty acids (FAw3) and the average number of methylene groups in a fatty acid chain (CH2.in.FA). On average, if the polygenic scores for CH2.DB and CH2.in.FA increase, height would increase; on the contrary, if the PRS for otPUFA and FAw3 increase, height would decrease.

Weight and BMI

A majority of the significant results was related to weight (62 regression models) and BMI (75 regression models). This is quite intuitive since an individual's lipid profile is highly associated with his or her body fat distribution [Bertoli et al., 2003].

Figure 5.1 (Figure 5.2) demonstrates the magnitude and direction of the significant b_3's in *Weight* $\sim$ single *MetabPRS* (*BMI* $\sim$ single *MetabPRS*) regression models. The b_3's are sorted based on hierarchical clustering on their corresponding original metabolomic measurements. In both Figure 5.1 and Figure 5.2, we could observe two major clusters. The top left cluster mainly consists of high density lipoprotein (HDL) related metabolites, and the associated regression coefficients are negative, meaning on average, if the polygenic scores of HDL related metabolites increase, weight and BMI would decrease. On the contrary, the bottom right cluster mainly contains very low density lipoprotein (VLDL) related metabolites,

and the associated regression coefficients are positive, meaning on average, if the polygenic scores of VLDL related metabolites increase, weight and BMI would increase as well.

Sciatica and LDD MRI phenotypes

Among all the clinical phenotypes, only sciatica was found to be positively influenced by the PRS of total lipids in IDL (IDL.L) with a q-value of 0.0634.

There existed significant findings for four of the LDD MRI phenotypes. As the PRS of the mean diameter for VLDL particles (VLDL.D) decreases, overall MC (binary), lower MC (binary) and lower MC (continuous) would on average increase. The study also discovered that the PRS of sphingomyelin (SM) bears a significantly positive correlation with L3 LDD severity.

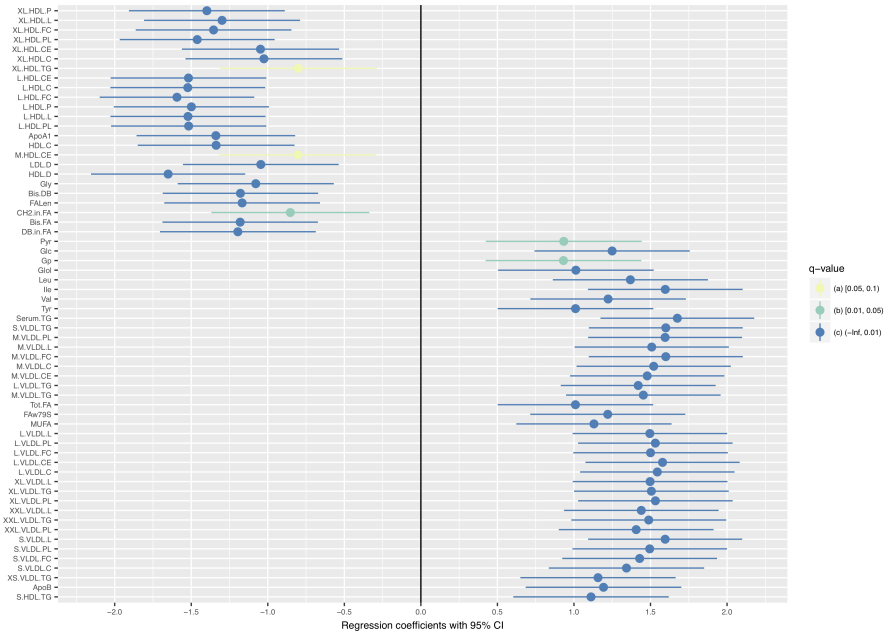


Fig. 5.1 Plot of the 95% CI of significant b_3's in *Weight* $\sim$ single *MetabPRS* regression models. The metabolomic PRS are sorted based on hierarchical clustering on their original metabolomic measurements with complete linkage.

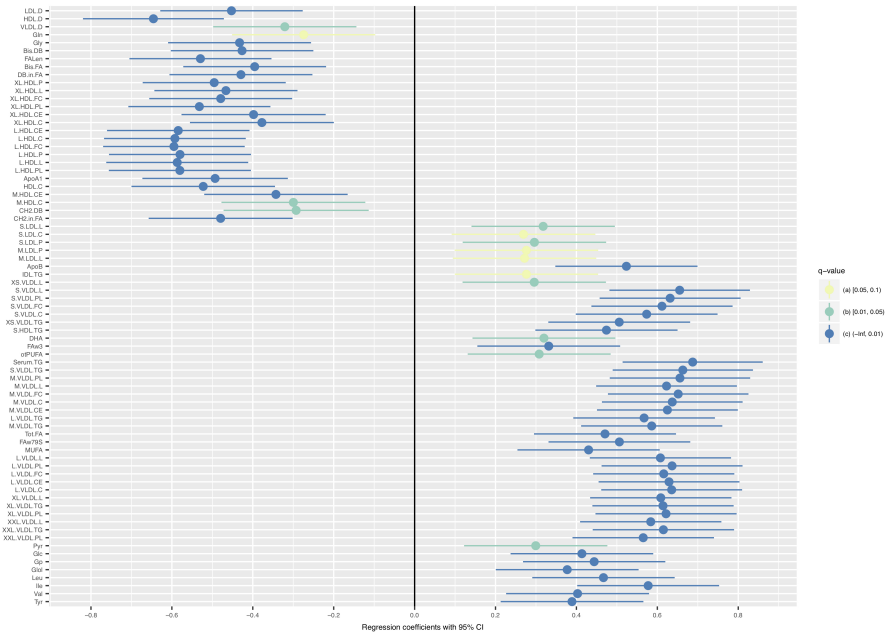


Fig. 5.2 Plot of the 95% CI of significant b_3's in *BMI* $\sim$ single *MetabPRS* regression models. The metabolomic PRS are sorted based on hierarchical clustering on their original metabolomic measurements with complete linkage.

Table 5.2 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (aggregated FDR).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
Height	CH2.DB	1214	0.0056	(0.0023, 0.0090)	9.3943E-04	3.4054E-02
Height	otPUFA	1214	-0.0055	(-0.0088, -0.0022)	1.0554E-03	3.7668E-02
Height	FAw3	1214	-0.0052	(-0.0085, -0.0020)	1.7770E-03	6.1078E-02
Height	CH2.in.FA	1214	0.0051	(0.0017, 0.0084)	2.9242E-03	9.3601E-02
Weight	Serum.TG	1214	1.6754	(1.1734, 2.1775)	8.6729E-11	1.5478E-08
Weight	HDL.D	1214	-1.6510	(-2.1546, -1.1475)	1.8073E-10	2.7051E-08
Weight	S.VLDL.TG	1214	1.5998	(1.0972, 2.1025)	5.9011E-10	7.8231E-08
Weight	M.VLDL.FC	1214	1.5998	(1.0968, 2.1028)	6.0728E-10	7.8272E-08
Weight	M.VLDL.PL	1214	1.5955	(1.0926, 2.0984)	6.6706E-10	8.2728E-08
Weight	S.VLDL.L	1214	1.5957	(1.0926, 2.0989)	6.7751E-10	8.2728E-08
Weight	Ile	1214	1.5965	(1.0910, 2.1021)	7.9563E-10	9.4660E-08
Weight	L.HDL.FC	1214	-1.5933	(-2.0984, -1.0883)	8.2474E-10	9.5670E-08
Weight	L.VLDL.CE	1214	1.5787	(1.0749, 2.0825)	1.0659E-09	1.2063E-07
Weight	L.VLDL.C	1214	1.5445	(1.0401, 2.0488)	2.4846E-09	2.7449E-07
Weight	L.VLDL.PL	1214	1.5316	(1.0269, 2.0363)	3.4389E-09	3.6264E-07
Weight	XL.VLDL.PL	1214	1.5320	(1.0271, 2.0368)	3.4362E-09	3.6264E-07
Weight	M.VLDL.C	1214	1.5208	(1.0168, 2.0248)	4.1904E-09	4.2268E-07
Weight	L.HDL.C	1214	-1.5226	(-2.0286, -1.0166)	4.6339E-09	4.4794E-07
Weight	L.HDL.L	1214	-1.5213	(-2.0281, -1.0145)	5.0143E-09	4.7483E-07
Weight	M.VLDL.L	1214	1.5081	(1.0042, 2.0120)	5.5764E-09	5.1550E-07
Weight	L.HDL.PL	1214	-1.5170	(-2.0241, -1.0099)	5.6661E-09	5.1550E-07
Weight	L.HDL.CE	1214	-1.5181	(-2.0266, -1.0096)	6.0676E-09	5.4142E-07

(Continued on next page)

Table 5.2 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
Weight	XL.VLDL.TG	1214	1.5057	(1.0004, 2.0111)	6.4810E-09	5.5858E-07
Weight	L.VLDL.FC	1214	1.5002	(0.9952, 2.0052)	7.1811E-09	6.0582E-07
Weight	XL.VLDL.L	1214	1.4973	(0.9918, 2.0028)	7.9219E-09	6.4586E-07
Weight	L.VLDL.L	1214	1.4958	(0.9908, 2.0008)	7.9341E-09	6.4586E-07
Weight	S.VLDL.PL	1214	1.4948	(0.9897, 1.9999)	8.1603E-09	6.5283E-07
Weight	L.HDL.P	1214	-1.4991	(-2.0061, -0.9920)	8.4641E-09	6.6565E-07
Weight	XXL.VLDL.TG	1214	1.4885	(0.9826, 1.9944)	9.9327E-09	7.5554E-07
Weight	M.VLDL.CE	1214	1.4786	(0.9741, 1.9830)	1.1266E-08	8.4311E-07
Weight	XL.HDL.PL	1214	-1.4610	(-1.9673, -0.9548)	1.8727E-08	1.3549E-06
Weight	M.VLDL.TG	1214	1.4532	(0.9485, 1.9579)	2.0175E-08	1.4184E-06
Weight	XXL.VLDL.L	1214	1.4398	(0.9347, 1.9449)	2.7741E-08	1.9212E-06
Weight	S.VLDL.FC	1214	1.4292	(0.9235, 1.9349)	3.6192E-08	2.4696E-06
Weight	L.VLDL.TG	1214	1.4196	(0.9135, 1.9256)	4.5468E-08	3.0576E-06
Weight	XXL.VLDL.PL	1214	1.4065	(0.9009, 1.9120)	5.8304E-08	3.8103E-06
Weight	XL.HDL.P	1214	-1.3980	(-1.9072, -0.8888)	8.6429E-08	5.5699E-06
Weight	Leu	1214	1.3684	(0.8620, 1.8749)	1.3692E-07	8.4706E-06
Weight	XL.HDL.FC	1214	-1.3542	(-1.8628, -0.8455)	2.0723E-07	1.2171E-05
Weight	S.VLDL.C	1214	1.3431	(0.8357, 1.8506)	2.4271E-07	1.3903E-05
Weight	HDL.C	1214	-1.3376	(-1.8496, -0.8255)	3.4732E-07	1.9416E-05
Weight	ApoA1	1214	-1.3392	(-1.8574, -0.8209)	4.6060E-07	2.5443E-05
Weight	XL.HDL.L	1214	-1.2988	(-1.8083, -0.7893)	6.5434E-07	3.5304E-05
Weight	Glc	1214	1.2493	(0.7422, 1.7564)	1.5134E-06	7.9799E-05
Weight	FAw79S	1214	1.2208	(0.7145, 1.7272)	2.5062E-06	1.2504E-04

(Continued on next page)

Table 5.2 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
Weight	Val	1214	1.2230	(0.7150, 1.7310)	2.5932E-06	1.2800E-04
Weight	DB.in.FA	1214	-1.1955	(-1.7044, -0.6866)	4.4880E-06	2.1920E-04
Weight	ApoB	1214	1.1931	(0.6847, 1.7016)	4.5926E-06	2.2196E-04
Weight	Bis.FA	1214	-1.1804	(-1.6882, -0.6725)	5.6352E-06	2.6681E-04
Weight	Bis.DB	1214	-1.1787	(-1.6863, -0.6710)	5.7660E-06	2.7025E-04
Weight	FALen	1214	-1.1676	(-1.6760, -0.6591)	7.2729E-06	3.3746E-04
Weight	XS.VLDL.TG	1214	1.1571	(0.6499, 1.6643)	8.3301E-06	3.8120E-04
Weight	MUFA	1214	1.1305	(0.6234, 1.6377)	1.3271E-05	5.8646E-04
Weight	S.HDL.TG	1214	1.1115	(0.6036, 1.6194)	1.8988E-05	8.2340E-04
Weight	Gly	1214	-1.0785	(-1.5884, -0.5687)	3.5567E-05	1.5003E-03
Weight	LDL.D	1214	-1.0457	(-1.5544, -0.5369)	5.8721E-05	2.4547E-03
Weight	XL.HDL.CE	1214	-1.0477	(-1.5613, -0.5342)	6.6475E-05	2.7540E-03
Weight	XL.HDL.C	1214	-1.0251	(-1.5375, -0.5127)	9.1616E-05	3.7619E-03
Weight	Glol	1214	1.0126	(0.5042, 1.5210)	9.8435E-05	4.0065E-03
Weight	Tot.FA	1214	1.0098	(0.5020, 1.5177)	1.0091E-04	4.0715E-03
Weight	Tyr	1214	1.0103	(0.5017, 1.5188)	1.0263E-04	4.1052E-03
Weight	Gp	1214	0.9314	(0.4228, 1.4400)	3.4044E-04	1.2948E-02
Weight	Pyr	1214	0.9332	(0.4239, 1.4426)	3.3789E-04	1.2948E-02
Weight	CH2.in.FA	1214	-0.8528	(-1.3686, -0.3370)	1.2122E-03	4.2290E-02
Weight	M.HDL.CE	1214	-0.8030	(-1.3144, -0.2916)	2.1113E-03	7.0990E-02
Weight	XL.HDL.TG	1214	-0.8013	(-1.3144, -0.2882)	2.2348E-03	7.4069E-02
BMI	Serum.TG	1214	0.6873	(0.5142, 0.8605)	1.4787E-14	6.8611E-11
BMI	S.VLDL.TG	1214	0.6630	(0.4896, 0.8364)	1.2315E-13	2.8570E-10

(Continued on next page)

Table 5.2 *Phen* ∼ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
BMI	M.VLDL.PL	1214	0.6559	(0.4824, 0.8295)	2.3056E-13	2.9333E-10
BMI	S.VLDL.L	1214	0.6552	(0.4815, 0.8289)	2.5287E-13	2.9333E-10
BMI	M.VLDL.FC	1214	0.6516	(0.4779, 0.8252)	3.4054E-13	3.1602E-10
BMI	HDL.D	1214	-0.6456	(-0.8197, -0.4715)	6.2726E-13	4.8508E-10
BMI	L.VLDL.C	1214	0.6355	(0.4613, 0.8096)	1.4091E-12	7.2647E-10
BMI	L.VLDL.PL	1214	0.6364	(0.4622, 0.8106)	1.3402E-12	7.2647E-10
BMI	M.VLDL.C	1214	0.6369	(0.4630, 0.8108)	1.1797E-12	7.2647E-10
BMI	S.VLDL.PL	1214	0.6316	(0.4574, 0.8059)	1.9784E-12	9.1799E-10
BMI	L.VLDL.CE	1214	0.6289	(0.4548, 0.8031)	2.3553E-12	9.9353E-10
BMI	M.VLDL.CE	1214	0.6249	(0.4508, 0.7990)	3.1671E-12	1.2246E-09
BMI	M.VLDL.L	1214	0.6227	(0.4487, 0.7967)	3.6729E-12	1.3110E-09
BMI	XL.VLDL.PL	1214	0.6215	(0.4471, 0.7959)	4.5342E-12	1.5028E-09
BMI	L.VLDL.FC	1214	0.6158	(0.4414, 0.7903)	7.0462E-12	2.1796E-09
BMI	XL.VLDL.TG	1214	0.6140	(0.4394, 0.7886)	8.4262E-12	2.2998E-09
BMI	XXL.VLDL.TG	1214	0.6150	(0.4403, 0.7897)	8.1173E-12	2.2998E-09
BMI	S.VLDL.FC	1214	0.6115	(0.4370, 0.7860)	9.9780E-12	2.5721E-09
BMI	XL.VLDL.L	1214	0.6085	(0.4339, 0.7832)	1.3055E-11	3.1111E-09
BMI	L.VLDL.L	1214	0.6076	(0.4331, 0.7821)	1.3410E-11	3.1111E-09
BMI	L.HDL.FC	1214	-0.5949	(-0.7699, -0.4199)	3.9146E-11	8.6493E-09
BMI	L.HDL.C	1214	-0.5925	(-0.7677, -0.4174)	4.8288E-11	1.0184E-08
BMI	M.VLDL.TG	1214	0.5861	(0.4116, 0.7606)	6.6334E-11	1.3382E-08
BMI	L.HDL.L	1214	-0.5867	(-0.7622, -0.4113)	7.9255E-11	1.5103E-08
BMI	XXL.VLDL.L	1214	0.5837	(0.4090, 0.7584)	8.1376E-11	1.5103E-08

(Continued on next page)

Table 5.2 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
BMI	L.HDL.CE	1214	-0.5842	(-0.7603, -0.4082)	1.1031E-10	1.8956E-08
BMI	L.HDL.P	1214	-0.5799	(-0.7554, -0.4043)	1.3356E-10	2.1369E-08
BMI	L.HDL.PL	1214	-0.5802	(-0.7559, -0.4046)	1.3200E-10	2.1369E-08
BMI	Ile	1214	0.5774	(0.4020, 0.7527)	1.5356E-10	2.3751E-08
BMI	S.VLDL.C	1214	0.5736	(0.3983, 0.7489)	1.9450E-10	2.8203E-08
BMI	L.VLDL.TG	1214	0.5673	(0.3922, 0.7424)	2.9245E-10	4.1120E-08
BMI	XXL.VLDL.PL	1214	0.5651	(0.3902, 0.7400)	3.2505E-10	4.4360E-08
BMI	XL.HDL.PL	1214	-0.5320	(-0.7077, -0.3564)	3.6727E-09	3.7870E-07
BMI	FALen	1214	-0.5294	(-0.7050, -0.3538)	4.3132E-09	4.2581E-07
BMI	ApoB	1214	0.5235	(0.3478, 0.6992)	6.5007E-09	5.5858E-07
BMI	HDL.C	1214	-0.5225	(-0.6999, -0.3451)	9.5799E-09	7.4085E-07
BMI	FAw79S	1214	0.5062	(0.3310, 0.6814)	1.8016E-08	1.3269E-06
BMI	XS.VLDL.TG	1214	0.5058	(0.3305, 0.6811)	1.8980E-08	1.3549E-06
BMI	XL.HDL.P	1214	-0.4954	(-0.6722, -0.3185)	4.7278E-08	3.1339E-06
BMI	ApoA1	1214	-0.4931	(-0.6729, -0.3133)	8.8960E-08	5.6545E-06
BMI	XL.HDL.FC	1214	-0.4793	(-0.6559, -0.3026)	1.2146E-07	7.6157E-06
BMI	S.HDL.TG	1214	0.4742	(0.2984, 0.6499)	1.4259E-07	8.7053E-06
BMI	CH2.in.FA	1214	-0.4795	(-0.6574, -0.3015)	1.4749E-07	8.8876E-06
BMI	Tot.FA	1214	0.4706	(0.2951, 0.6462)	1.6976E-07	1.0098E-05
BMI	Leu	1214	0.4668	(0.2908, 0.6429)	2.2970E-07	1.3322E-05
BMI	XL.HDL.L	1214	-0.4664	(-0.6433, -0.2895)	2.6919E-07	1.5232E-05
BMI	LDL.D	1214	-0.4526	(-0.6287, -0.2766)	5.2758E-07	2.8800E-05
BMI	Gp	1214	0.4439	(0.2680, 0.6197)	8.3913E-07	4.4753E-05

(Continued on next page)

Table 5.2 *Phen* ∼ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
BMI	Gly	1214	-0.4327	(-0.6094, -0.2560)	1.7490E-06	9.1183E-05
BMI	MUFA	1214	0.4301	(0.2542, 0.6060)	1.8109E-06	9.3362E-05
BMI	DB.in.FA	1214	-0.4293	(-0.6060, -0.2526)	2.1013E-06	1.0714E-04
BMI	Bis.DB	1214	-0.4265	(-0.6027, -0.2503)	2.2975E-06	1.1588E-04
BMI	Glc	1214	0.4135	(0.2372, 0.5898)	4.6400E-06	2.2196E-04
BMI	Val	1214	0.4029	(0.2262, 0.5795)	8.3799E-06	3.8120E-04
BMI	Bis.FA	1214	-0.3953	(-0.5719, -0.2188)	1.2136E-05	5.4672E-04
BMI	XL.HDL.CE	1214	-0.3980	(-0.5762, -0.2198)	1.2751E-05	5.6887E-04
BMI	Tyr	1214	0.3891	(0.2127, 0.5655)	1.6356E-05	7.1596E-04
BMI	Glol	1214	0.3773	(0.2008, 0.5537)	2.9310E-05	1.2592E-03
BMI	XL.HDL.C	1214	-0.3775	(-0.5553, -0.1996)	3.3557E-05	1.4285E-03
BMI	M.HDL.CE	1214	-0.3428	(-0.5200, -0.1655)	1.5630E-04	6.1985E-03
BMI	FAw3	1214	0.3316	(0.1551, 0.5081)	2.3810E-04	9.2837E-03
BMI	VLDL.D	1214	-0.3209	(-0.4979, -0.1439)	3.9008E-04	1.4715E-02
BMI	DHA	1214	0.3197	(0.1431, 0.4963)	3.9847E-04	1.4911E-02
BMI	S.LDL.L	1214	0.3178	(0.1406, 0.4950)	4.5026E-04	1.6714E-02
BMI	otPUFA	1214	0.3079	(0.1312, 0.4847)	6.4996E-04	2.3935E-02
BMI	Pyr	1214	0.2994	(0.1223, 0.4765)	9.3633E-04	3.4054E-02
BMI	M.HDL.C	1214	-0.2997	(-0.4775, -0.1219)	9.7214E-04	3.4967E-02
BMI	S.LDL.P	1214	0.2961	(0.1188, 0.4734)	1.0811E-03	3.8291E-02
BMI	XS.VLDL.L	1214	0.2958	(0.1185, 0.4730)	1.0901E-03	3.8319E-02
BMI	CH2.DB	1214	-0.2928	(-0.4722, -0.1135)	1.3957E-03	4.8329E-02
BMI	IDL.TG	1214	0.2765	(0.0994, 0.4535)	2.2320E-03	7.4069E-02

(Continued on next page)

Table 5.2 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (aggregated FDR) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	(95% CI)	p-value	q-value
BMI	M.LDL.P	1214	0.2764	(0.0990, 0.4538)	2.2817E-03	7.5084E-02
BMI	Gln	1214	-0.2742	(-0.4511, -0.0972)	2.4162E-03	7.8952E-02
BMI	M.LDL.L	1214	0.2717	(0.0945, 0.4490)	2.6886E-03	8.7239E-02
BMI	S.LDL.C	1214	0.2692	(0.0917, 0.4467)	2.9893E-03	9.5002E-02
Sciatica	IDL.L	795	0.2345	(0.0877, 0.3835)	1.8591E-03	6.3427E-02
L3 LDD severity	SM	751	0.1822	(0.0624, 0.3021)	2.9250E-03	9.3601E-02
Overall MC (binary)	VLDL.D	750	-0.3125	(-0.5125, -0.1159)	1.9815E-03	6.7112E-02
Lower MC (cont)	VLDL.D	750	-0.1148	(-0.1760, -0.0536)	2.4896E-04	9.6263E-03
Lower MC (binary)	VLDL.D	750	-0.3964	(-0.6075, -0.1901)	1.9127E-04	7.5210E-03

5.3.1.2 Based on adaptive group FDR

If we group the hypotheses by metabolomic measurement when running the adaptive group B-H procedure, at the FDR cut-off of 0.1, there were in total 22 (out of the $4,640$) regression models with a significant b_3 (Table 5.3).

The findings were quite similar to those in Section 5.3.1.1. The strongest signals were regarding weight and BMI – the PRS for HDL related metabolomic measurements tended to negatively influence weight and BMI, whereas the PRS for IDL, LDL, and VLDL related metabolites positively affected weight and BMI on average. The PRS for VLDL.D was found to have a negative relationship with BMI and positively influenced modic change, which was also consistent with previous findings.

It has also been found that on average, as the PRS for tyrosine increases, weight and BMI would increase, whereas the overall LDD would decrease.

If we group the hypotheses by phenotype when running the adaptive group B-H procedure, at the FDR cut-off of 0.1, there were in total 222 (out of the $4,640$) regression models with a significant b_3 (Table 5.4).

The findings were, again, quite similar to those in Section 5.3.1.1. Additionally, a group of LDL related metabolomic measurements (L.LDL.L, L.LDL.CE, L.LDL.C, L.LDL.P, L.LDL.PL) was found to be positively associated with modic change, which reinforced the potential of VLDL.D/LDL.D as a biomarker for MC.

Table 5.3 *Phen* $\sim$ single *MetabPRS* regressions with a significant b_3 (group FDR; by metab).

Phenotype	Metab PRS	Sample size	b3	p-value	q-value
Weight	L.HDL.CE	1214	-1.5181	6.0676E-09	2.9411E-06
Weight	L.HDL.FC	1214	-1.5933	8.2474E-10	5.5967E-07
Weight	L.HDL.L	1214	-1.5213	5.0143E-09	2.8356E-06
Weight	LDL.D	1214	-1.0457	5.8721E-05	1.2453E-02
Weight	Tyr	1214	1.0103	1.0263E-04	2.5629E-03
Weight	XXL.VLDL.PL	1214	1.4065	5.8304E-08	7.8200E-06
BMI	CH2.in.FA	1214	-0.4795	1.4749E-07	5.5603E-05
BMI	IDL.TG	1214	0.2765	2.2320E-03	1.7134E-02
BMI	L.HDL.CE	1214	-0.5842	1.1031E-10	9.3567E-08
BMI	L.HDL.FC	1214	-0.5949	3.9146E-11	9.3567E-08
BMI	L.HDL.L	1214	-0.5867	7.9255E-11	9.3567E-08
BMI	LDL.D	1214	-0.4526	5.2758E-07	1.7901E-04
BMI	M.LDL.L	1214	0.2717	2.6886E-03	9.5690E-02
BMI	Tyr	1214	0.3891	1.6356E-05	4.4558E-04
BMI	VLDL.D	1214	-0.3209	3.9008E-04	9.0499E-03
BMI	XXL.VLDL.PL	1214	0.5651	3.2505E-10	9.3567E-08
Overall LDD	Tyr	745	-0.5206	4.1475E-03	6.2143E-02
Lower SS	VLDL.D	763	-0.1455	5.5589E-03	9.2119E-02
Overall MC (cont)	VLDL.D	750	-0.1202	3.5565E-03	6.2143E-02
Overall MC (binary)	VLDL.D	750	-0.3125	1.9816E-03	3.8310E-02
Lower MC (cont)	VLDL.D	750	-0.1148	2.4896E-04	6.1884E-03
Lower MC (binary)	VLDL.D	750	-0.3964	1.9127E-04	5.1201E-03

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Height	CH2.DB	1214	0.0056	9.3943E-04	5.8367E-02
Height	otPUFA	1214	-0.0055	1.0554E-03	6.3970E-02
Height	FAw3	1214	-0.0052	1.7770E-03	9.9466E-02
Weight	Serum.TG	1214	1.6754	8.6729E-11	1.9687E-10
Weight	HDL.D	1214	-1.6510	1.8073E-10	3.3824E-10
Weight	S.VLDL.TG	1214	1.5998	5.9011E-10	1.0098E-09
Weight	M.VLDL.FC	1214	1.5998	6.0728E-10	1.0103E-09
Weight	M.VLDL.PL	1214	1.5955	6.6706E-10	1.0678E-09
Weight	S.VLDL.L	1214	1.5957	6.7751E-10	1.0678E-09
Weight	Ile	1214	1.5965	7.9563E-10	1.2218E-09
Weight	L.HDL.FC	1214	-1.5933	8.2474E-10	1.2348E-09
Weight	L.VLDL.CE	1214	1.5787	1.0659E-09	1.5570E-09
Weight	L.VLDL.C	1214	1.5445	2.4846E-09	3.5430E-09
Weight	XL.VLDL.PL	1214	1.5320	3.4362E-09	4.5185E-09
Weight	L.VLDL.PL	1214	1.5316	3.4389E-09	4.5185E-09
Weight	M.VLDL.C	1214	1.5208	4.1904E-09	5.3396E-09
Weight	L.HDL.C	1214	-1.5226	4.6339E-09	5.7817E-09
Weight	L.HDL.L	1214	-1.5213	5.0143E-09	6.1288E-09
Weight	M.VLDL.L	1214	1.5081	5.5764E-09	6.5258E-09
Weight	L.HDL.PL	1214	-1.5170	5.6661E-09	6.5258E-09

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Weight	L.HDL.CE	1214	-1.5181	6.0676E-09	6.8564E-09
Weight	XL.VLDL.TG	1214	1.5057	6.4810E-09	7.1879E-09
Weight	L.VLDL.FC	1214	1.5002	7.1811E-09	7.8196E-09
Weight	XL.VLDL.L	1214	1.4973	7.9219E-09	8.1926E-09
Weight	L.VLDL.L	1214	1.4958	7.9341E-09	8.1926E-09
Weight	S.VLDL.PL	1214	1.4948	8.1603E-09	8.2834E-09
Weight	L.HDL.P	1214	-1.4991	8.4641E-09	8.4486E-09
Weight	XXL.VLDL.TG	1214	1.4885	9.9327E-09	9.7520E-09
Weight	M.VLDL.CE	1214	1.4786	1.1266E-08	1.0882E-08
Weight	XL.HDL.PL	1214	-1.4610	1.8727E-08	1.7255E-08
Weight	M.VLDL.TG	1214	1.4532	2.0175E-08	1.8307E-08
Weight	XXL.VLDL.L	1214	1.4398	2.7741E-08	2.4797E-08
Weight	S.VLDL.FC	1214	1.4292	3.6192E-08	3.1876E-08
Weight	L.VLDL.TG	1214	1.4196	4.5468E-08	3.8901E-08
Weight	XXL.VLDL.PL	1214	1.4065	5.8304E-08	4.9180E-08
Weight	XL.HDL.P	1214	-1.3980	8.6429E-08	7.0908E-08
Weight	Leu	1214	1.3684	1.3692E-07	1.0513E-07
Weight	XL.HDL.FC	1214	-1.3542	2.0723E-07	1.5514E-07
Weight	S.VLDL.C	1214	1.3431	2.4271E-07	1.7727E-07
Weight	HDL.C	1214	-1.3376	3.4732E-07	2.5061E-07
Weight	ApoA1	1214	-1.3392	4.6060E-07	3.2453E-07

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Weight	XL.HDL.L	1214	-1.2988	6.5434E-07	4.5568E-07
Weight	Glc	1214	1.2493	1.5134E-06	1.0071E-06
Weight	FAw79S	1214	1.2208	2.5062E-06	1.6139E-06
Weight	Val	1214	1.2230	2.5932E-06	1.6522E-06
Weight	DB.in.FA	1214	-1.1955	4.4880E-06	2.7999E-06
Weight	ApoB	1214	1.1931	4.5926E-06	2.8356E-06
Weight	Bis.FA	1214	-1.1804	5.6352E-06	3.4438E-06
Weight	Bis.DB	1214	-1.1787	5.7660E-06	3.4881E-06
Weight	FALen	1214	-1.1676	7.2729E-06	4.3126E-06
Weight	XS.VLDL.TG	1214	1.1571	8.3301E-06	4.8911E-06
Weight	MUFA	1214	1.1305	1.3271E-05	7.4983E-06
Weight	S.HDL.TG	1214	1.1115	1.8988E-05	1.0628E-05
Weight	Gly	1214	-1.0785	3.5567E-05	1.9364E-05
Weight	LDL.D	1214	-1.0457	5.8721E-05	3.1683E-05
Weight	XL.HDL.CE	1214	-1.0477	6.6475E-05	3.5547E-05
Weight	XL.HDL.C	1214	-1.0251	9.1616E-05	4.8557E-05
Weight	Glol	1214	1.0126	9.8435E-05	5.1713E-05
Weight	Tot.FA	1214	1.0098	1.0091E-04	5.2552E-05
Weight	Tyr	1214	1.0103	1.0263E-04	5.2987E-05
Weight	Pyr	1214	0.9332	3.3789E-04	1.6712E-04
Weight	Gp	1214	0.9314	3.4044E-04	1.6712E-04

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Weight	CH2.in.FA	1214	-0.8528	1.2122E-03	5.5845E-04
Weight	M.HDL.CE	1214	-0.8030	2.1113E-03	9.4364E-04
Weight	XL.HDL.TG	1214	-0.8013	2.2348E-03	9.8415E-04
Weight	M.HDL.C	1214	-0.7311	5.2065E-03	2.1959E-03
Weight	S.LDL.L	1214	0.7240	5.5359E-03	2.3024E-03
Weight	VLDL.D	1214	-0.7077	6.6363E-03	2.7409E-03
Weight	PC	1214	-0.7036	6.9819E-03	2.8445E-03
Weight	S.LDL.P	1214	0.6290	1.5976E-02	6.0942E-03
Weight	XS.VLDL.L	1214	0.6141	1.8591E-02	7.0468E-03
Weight	M.HDL.FC	1214	-0.6059	2.0229E-02	7.5720E-03
Weight	M.HDL.L	1214	-0.5759	2.7508E-02	1.0195E-02
Weight	Ala	1214	0.5745	2.7713E-02	1.0195E-02
Weight	Gln	1214	-0.5731	2.7748E-02	1.0195E-02
Weight	M.LDL.P	1214	0.5701	2.8954E-02	1.0573E-02
Weight	SM	1214	-0.5674	3.1060E-02	1.1206E-02
Weight	M.HDL.PL	1214	-0.5604	3.1928E-02	1.1450E-02
Weight	M.LDL.PL	1214	0.5492	3.5326E-02	1.2593E-02
Weight	M.LDL.L	1214	0.5350	4.0283E-02	1.4026E-02
Weight	S.LDL.C	1214	0.5238	4.4980E-02	1.5571E-02
Weight	M.LDL.C	1214	0.5218	4.5631E-02	1.5706E-02
Weight	IDL.TG	1214	0.5068	5.1825E-02	1.7736E-02

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Weight	Phe	1214	0.4994	5.5539E-02	1.8792E-02
Weight	TotPG	1214	-0.4960	5.7254E-02	1.9188E-02
Weight	M.LDL.CE	1214	0.4944	5.8007E-02	1.9300E-02
Weight	M.HDL.P	1214	-0.4859	6.2931E-02	2.0823E-02
Weight	LDL.C	1214	0.4674	7.3699E-02	2.4120E-02
Weight	XS.VLDL.PL	1214	0.4645	7.5345E-02	2.4524E-02
Weight	DHA	1214	0.4593	7.7741E-02	2.5167E-02
Weight	Free.C	1214	-0.4523	8.3829E-02	2.6992E-02
Weight	Alb	1214	-0.4492	8.4860E-02	2.7178E-02
Weight	FAw3	1214	0.4352	9.4562E-02	2.9651E-02
Weight	Cit	1214	-0.4054	1.1988E-01	3.7199E-02
Weight	L.LDL.P	1214	0.4009	1.2467E-01	3.8487E-02
Weight	L.LDL.CE	1214	0.3798	1.4574E-01	4.4761E-02
Weight	L.LDL.L	1214	0.3770	1.4873E-01	4.5446E-02
Weight	otPUFA	1214	0.3607	1.6613E-01	5.0302E-02
Weight	L.LDL.C	1214	0.3617	1.6630E-01	5.0302E-02
Weight	bOHBut	1214	-0.3591	1.6922E-01	5.0927E-02
Weight	CH2.DB	1214	-0.3402	1.9794E-01	5.8979E-02
Weight	L.LDL.PL	1214	0.3292	2.0770E-01	6.1580E-02
Weight	Crea	1214	0.3252	2.1286E-01	6.2490E-02
Weight	Serum.C	1214	-0.2458	3.4779E-01	9.5109E-02

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
Weight	IDL.L	1214	0.2405	3.5820E-01	9.7071E-02
BMI	Serum.TG	1214	0.6873	1.4787E-14	7.1257E-13
BMI	S.VLDL.TG	1214	0.6630	1.2315E-13	2.9672E-12
BMI	M.VLDL.PL	1214	0.6559	2.3056E-13	3.0464E-12
BMI	S.VLDL.L	1214	0.6552	2.5287E-13	3.0464E-12
BMI	M.VLDL.FC	1214	0.6516	3.4054E-13	3.2821E-12
BMI	HDL.D	1214	-0.6456	6.2726E-13	5.0379E-12
BMI	M.VLDL.C	1214	0.6369	1.1797E-12	7.5448E-12
BMI	L.VLDL.PL	1214	0.6364	1.3402E-12	7.5448E-12
BMI	L.VLDL.C	1214	0.6355	1.4091E-12	7.5448E-12
BMI	S.VLDL.PL	1214	0.6316	1.9784E-12	9.5340E-12
BMI	L.VLDL.CE	1214	0.6289	2.3553E-12	1.0318E-11
BMI	M.VLDL.CE	1214	0.6249	3.1671E-12	1.2718E-11
BMI	M.VLDL.L	1214	0.6227	3.6729E-12	1.3615E-11
BMI	XL.VLDL.PL	1214	0.6215	4.5342E-12	1.5607E-11
BMI	L.VLDL.FC	1214	0.6158	7.0462E-12	2.2637E-11
BMI	XXL.VLDL.TG	1214	0.6150	8.1173E-12	2.3885E-11
BMI	XL.VLDL.TG	1214	0.6140	8.4262E-12	2.3885E-11
BMI	S.VLDL.FC	1214	0.6115	9.9780E-12	2.6713E-11
BMI	XL.VLDL.L	1214	0.6085	1.3055E-11	3.2311E-11
BMI	L.VLDL.L	1214	0.6076	1.3410E-11	3.2311E-11

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
BMI	L.HDL.FC	1214	-0.5949	3.9146E-11	8.9829E-11
BMI	L.HDL.C	1214	-0.5925	4.8288E-11	1.0577E-10
BMI	M.VLDL.TG	1214	0.5861	6.6334E-11	1.3898E-10
BMI	L.HDL.L	1214	-0.5867	7.9255E-11	1.5686E-10
BMI	XXL.VLDL.L	1214	0.5837	8.1376E-11	1.5686E-10
BMI	L.HDL.CE	1214	-0.5842	1.1031E-10	1.9687E-10
BMI	L.HDL.PL	1214	-0.5802	1.3200E-10	2.2193E-10
BMI	L.HDL.P	1214	-0.5799	1.3356E-10	2.2193E-10
BMI	Ile	1214	0.5774	1.5356E-10	2.4667E-10
BMI	S.VLDL.C	1214	0.5736	1.9450E-10	3.0235E-10
BMI	L.VLDL.TG	1214	0.5673	2.9245E-10	4.2706E-10
BMI	XXL.VLDL.PL	1214	0.5651	3.2505E-10	4.6071E-10
BMI	XL.HDL.PL	1214	-0.5320	3.6727E-09	4.1159E-09
BMI	FALen	1214	-0.5294	4.3132E-09	4.5185E-09
BMI	ApoB	1214	0.5235	6.5007E-09	6.2653E-09
BMI	HDL.C	1214	-0.5225	9.5799E-09	8.1926E-09
BMI	FAw79S	1214	0.5062	1.8016E-08	1.3781E-08
BMI	XS.VLDL.TG	1214	0.5058	1.8980E-08	1.4292E-08
BMI	XL.HDL.P	1214	-0.4954	4.7278E-08	3.3019E-08
BMI	ApoA1	1214	-0.4931	8.8960E-08	5.9541E-08
BMI	XL.HDL.FC	1214	-0.4793	1.2146E-07	7.9094E-08

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
BMI	S.HDL.TG	1214	0.4742	1.4259E-07	9.1616E-08
BMI	CH2.in.FA	1214	-0.4795	1.4749E-07	9.3518E-08
BMI	Tot.FA	1214	0.4706	1.6976E-07	1.0513E-07
BMI	Leu	1214	0.4668	2.2970E-07	1.4011E-07
BMI	XL.HDL.L	1214	-0.4664	2.6919E-07	1.6015E-07
BMI	LDL.D	1214	-0.4526	5.2758E-07	3.0267E-07
BMI	Gp	1214	0.4439	8.3913E-07	4.6479E-07
BMI	Gly	1214	-0.4327	1.7490E-06	9.5775E-07
BMI	MUFA	1214	0.4301	1.8109E-06	9.8052E-07
BMI	DB.in.FA	1214	-0.4293	2.1013E-06	1.1128E-06
BMI	Bis.DB	1214	-0.4265	2.2975E-06	1.2034E-06
BMI	Glc	1214	0.4135	4.6400E-06	2.3537E-06
BMI	Val	1214	0.4029	8.3799E-06	4.0382E-06
BMI	Bis.FA	1214	-0.3953	1.2136E-05	5.6781E-06
BMI	XL.HDL.CE	1214	-0.3980	1.2751E-05	5.9081E-06
BMI	Tyr	1214	0.3891	1.6356E-05	7.4983E-06
BMI	Glol	1214	0.3773	2.9310E-05	1.3078E-05
BMI	XL.HDL.C	1214	-0.3775	3.3557E-05	1.4836E-05
BMI	M.HDL.CE	1214	-0.3428	1.5630E-04	6.4376E-05
BMI	FAw3	1214	0.3316	2.3810E-04	9.7235E-05
BMI	VLDL.D	1214	-0.3209	3.9008E-04	1.5797E-04

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
BMI	DHA	1214	0.3197	3.9847E-04	1.6002E-04
BMI	S.LDL.L	1214	0.3178	4.5026E-04	1.7640E-04
BMI	otPUFA	1214	0.3079	6.4996E-04	2.5259E-04
BMI	Pyr	1214	0.2994	9.3633E-04	3.6097E-04
BMI	M.HDL.C	1214	-0.2997	9.7214E-04	3.7180E-04
BMI	S.LDL.P	1214	0.2961	1.0811E-03	4.1020E-04
BMI	XS.VLDL.L	1214	0.2958	1.0901E-03	4.1041E-04
BMI	CH2.DB	1214	-0.2928	1.3957E-03	5.2139E-04
BMI	IDL.TG	1214	0.2765	2.2320E-03	8.2107E-04
BMI	M.LDL.P	1214	0.2764	2.2817E-03	8.3297E-04
BMI	Gln	1214	-0.2742	2.4162E-03	8.7546E-04
BMI	M.LDL.L	1214	0.2717	2.6886E-03	9.5973E-04
BMI	S.LDL.C	1214	0.2692	2.9893E-03	1.0515E-03
BMI	M.LDL.PL	1214	0.2659	3.3316E-03	1.1634E-03
BMI	M.LDL.CE	1214	0.2642	3.5194E-03	1.2201E-03
BMI	M.LDL.C	1214	0.2615	3.9043E-03	1.3439E-03
BMI	M.HDL.L	1214	-0.2473	6.4128E-03	2.1917E-03
BMI	LDL.C	1214	0.2462	6.6380E-03	2.2369E-03
BMI	XS.VLDL.PL	1214	0.2393	8.3040E-03	2.7409E-03
BMI	M.HDL.FC	1214	-0.2295	1.1345E-02	3.6940E-03
BMI	L.LDL.CE	1214	0.2288	1.1592E-02	3.7410E-03

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
BMI	XL.HDL.TG	1214	-0.2298	1.1645E-02	3.7410E-03
BMI	L.LDL.P	1214	0.2262	1.2581E-02	3.9888E-03
BMI	M.HDL.PL	1214	-0.2248	1.3205E-02	4.1591E-03
BMI	L.LDL.L	1214	0.2199	1.5279E-02	4.7811E-03
BMI	Cit	1214	-0.2157	1.7132E-02	5.3262E-03
BMI	M.HDL.P	1214	-0.2124	1.9211E-02	5.9343E-03
BMI	L.LDL.C	1214	0.2054	2.3560E-02	7.1405E-03
BMI	L.LDL.PL	1214	0.1877	3.8511E-02	1.1206E-02
BMI	FAw6	1214	0.1817	4.5502E-02	1.2975E-02
BMI	Phe	1214	0.1805	4.6359E-02	1.3141E-02
BMI	Ala	1214	0.1782	4.9422E-02	1.3928E-02
BMI	Alb	1214	-0.1652	6.8083E-02	1.8641E-02
BMI	IDL.L	1214	0.1639	7.1273E-02	1.9188E-02
BMI	LA	1214	0.1542	8.9242E-02	2.3629E-02
BMI	bOHBut	1214	-0.1463	1.0676E-01	2.7365E-02
BMI	PC	1214	-0.1448	1.1026E-01	2.8114E-02
BMI	IDL.PL	1214	0.1435	1.1458E-01	2.9062E-02
BMI	SM	1214	-0.1366	1.3534E-01	3.3969E-02
BMI	IDL.C	1214	0.1024	2.6009E-01	6.1743E-02
BMI	Ace	1214	-0.0945	2.9848E-01	6.9152E-02
BMI	Free.C	1214	-0.0937	3.0274E-01	6.9804E-02

(Continued on next page)

Table 5.4 *Phen* $\sim$ single *MetabPRS* regression results with a significant b_3 (group FDR; by phen) (cont'd).

Phenotype	Metab PRS	Sample size	b_3	p-value	q-value
BMI	TotPG	1214	-0.0875	3.3433E-01	7.6355E-02
BMI	L.LDL.FC	1214	0.0857	3.4537E-01	7.8505E-02
Sciatica	IDL.L	795	0.2345	1.8591E-03	7.2692E-02
Overall MC (cont)	VLDL.D	750	-0.1202	3.5565E-03	8.0807E-02
Overall MC (cont)	L.LDL.L	750	0.1198	3.7906E-03	8.5520E-02
Overall MC (cont)	L.LDL.CE	750	0.1196	3.7992E-03	8.5520E-02
Lower MC (cont)	VLDL.D	750	-0.1148	2.4896E-04	3.8099E-03
Lower MC (cont)	L.LDL.L	750	-0.1148	5.8089E-03	6.5161E-02
Lower MC (cont)	L.LDL.CE	750	0.0868	5.9909E-03	6.6878E-02
Lower MC (cont)	L.LDL.C	750	0.0864	8.3270E-03	8.8673E-02
Lower MC (cont)	L.LDL.P	750	0.0832	9.2216E-03	9.6860E-02
MC exists (DS)	L.LDL.PL	632	0.2814	1.0142E-02	8.8673E-02
MC exists (DS)	LDL.D	632	-0.2730	1.0903E-02	9.4493E-02

5.3.2 Regression analysis: one phenotype, multiple metabolomic PRS

Due to the lack of positive data points, type 1 modic change was dropped from the analysis. All the other 39 models were found to be significant at the FDR threshold of 0.1, and the adjusted R^2 of the fitted models ranged from 0.0258 (upper HIZ) to 0.5547 (height).

Table 5.5 contains a summary of the regression results. For all 39 fitted models, the adjusted R^2 and model p-values are reported. Furthermore, within each model, all the metabolomic PRS having a regression coefficient b with p-value < 0.01 are listed. The components of each composite metabolomic feature could be found in Table 2.11.

Table 5.5 *Phen* $\sim$ multiple *MetabPRS* regression results.

Phenotype	Adjusted R^2	Model p-value	MetabPRS	b	p-value of b
Height	0.5547	<2.20E-16	VLDL.D	-0.0133	3.51E-03
			gr16	-0.0129	7.04E-03
			gr24	-0.0153	9.67E-03
Weight	0.3710	<2.20E-16	ApoB	4.0175	1.03E-04
			gr9	2.4662	1.56E-03
			gr24	-3.0273	3.62E-03
BMI	0.0989	4.91E-14	ApoB	1.1861	9.89E-05
			gr9	0.8774	9.58E-04
			gr24	-0.8622	1.25E-03
Smoking	0.1270	<2.20E-16	VLDL.D	-1.1970	6.07E-03
LBP	-	4.28E-07	gr25	-1.5415	3.01E-03
			gr6	0.7513	8.49E-03
			gr23	1.2914	9.01E-03
Sciatica	-	3.55E-06	gr26	0.8691	6.75E-04
			gr10	-0.5528	8.58E-04
Oswestry	0.0289	2.64E-04	gr12	-4.8398	6.17E-03
			gr3	4.0409	8.19E-03
VAS (test day)	0.0380	1.07E-04	CH2.DB	-3.8524	5.02E-03
VAS (severest)	0.0760	8.33E-12	gr14	-6.1462	8.49E-03
Overall LDD	0.1058	1.33E-13	Tyr	-0.9882	2.62E-03
			gr20	3.4267	5.83E-03
Deg score	0.0659	5.34E-09	Tyr	-0.6551	6.34E-03

(Continued on next page)

Table 5.5 *Phen* ∼ multiple *MetabPRS* regression results (cont'd).

Phenotype	Adjusted R^2	Model p-value	MetabPRS	*b*	p-value of *b*
Dev score	0.0725	1.50E-10	-	-	-
L1 LDD	0.0701	2.94E-09	ApoB	-0.2770	5.41E-04
			Free.C	0.2078	1.65E-03
			bOHBut	-0.2093	2.82E-03
L2 LDD	0.1347	<2.20E-16	gr26	-0.8115	5.07E-04
			gr17	0.5106	1.79E-03
			gr3	-0.5492	1.87E-03
			gr24	0.4302	6.45E-03
L3 LDD	0.1430	<2.20E-16	SM	0.3197	2.48E-03
			AcAce	0.3670	3.16E-03
L4 LDD	0.0674	4.07E-08	Tyr	-0.4093	3.43E-03
			Ile	0.3704	5.19E-03
L5 LDD	0.0376	3.99E-04	gr2	1.4026	7.73E-04
Upper LDD	0.1698	<2.20E-16	AcAce	0.6728	2.75E-03
Lower LDD	0.0416	4.93E-05	Ile	0.6620	1.47E-03
			gr20	2.1591	8.05E-03
Overall DB	0.0444	1.20E-05	-	-	-
Upper DB	0.0732	3.09E-11	Alb	0.2544	4.28E-03
Lower DB	0.0367	1.50E-04	Tyr	-0.2333	6.77E-03
			gr20	0.8312	7.79E-03
			Ile	0.2082	9.28E-03
Overall SS	0.1646	<2.20E-16	AcAce	0.5582	3.78E-03
Upper SS	0.2172	<2.20E-16	AcAce	0.4475	5.47E-04
			ApoB	-0.4506	6.49E-03
			bOHBut	-0.3360	9.57E-03
Lower SS	0.0542	5.92E-07	gr13	-0.6941	1.04E-03
			gr2	0.6649	1.36E-03
			VLDL.D	-0.2948	3.32E-03
Overall HIZ	0.0460	1.71E-05	gr11	0.6713	1.49E-03
			gr3	-0.4426	5.96E-03
			Ile	0.1996	7.94E-03
Upper HIZ	0.0258	1.71E-03	MUFA	-0.1814	3.71E-03

(Continued on next page)

Table 5.5 *Phen* $\sim$ multiple *MetabPRS* regression results (cont'd).

Phenotype	Adjusted R^2	Model p-value	MetabPRS	b	p-value of b
			gr13	0.1329	3.94E-03
Lower HIZ	0.0483	1.17E-05	Ile	0.1807	3.58E-03
			S.HDL.TG	0.2216	5.85E-03
			gr11	0.4696	7.70E-03
Overall MC (cont)	0.0813	4.19E-09	gr1	0.7859	9.26E-05
			gr28	-0.5024	2.52E-03
			gr13	-0.4588	6.88E-03
			gr14	-0.3760	7.21E-03
Overall MC (bool)	-	4.73E-07	gr4	1.6869	1.47E-03
			VLDL.D	-0.5756	4.78E-03
			gr7	1.5874	6.10E-03
Upper MC (cont)	0.0589	1.63E-06	gr26	-0.3216	4.04E-04
			gr20	0.4694	1.74E-03
			gr7	0.2175	3.29E-03
			gr27	0.1172	6.34E-03
Upper MC (bool)	-	1.16E-08	gr23	4.6925	7.21E-05
			gr25	-5.1404	1.29E-04
			gr28	-4.4234	3.84E-04
			gr20	9.8167	4.70E-04
			gr21	-7.3295	1.22E-03
			gr17	2.9191	1.26E-03
			gr24	3.6333	4.83E-03
			gr27	1.8733	5.06E-03
			Ile	-1.5137	8.19E-03
			His	-1.3497	9.01E-03
Lower MC (cont)	0.0810	4.51E-09	VLDL.D	-0.2104	6.93E-04
			gr14	-0.3345	1.70E-03
			gr1	0.1943	3.25E-03
			gr13	-0.5041	4.91E-03
Lower MC (bool)	-	3.38E-08	VLDL.D	-0.8461	2.39E-04
Any MC (DS)	-	1.44E-09	gr6	0.8353	3.43E-03
			His	-0.5638	4.33E-03

(Continued on next page)

Table 5.5 *Phen* $\sim$ multiple *MetabPRS* regression results (cont'd).

Phenotype	Adjusted R^2	Model p-value	MetabPRS	b	p-value of b
			gr14	-1.0231	5.05E-03
Type 2 MC (DS)	-	1.52E-07	gr1	1.2878	4.52E-03
Overall SN	0.0453	2.96E-06	gr4	1.5956	1.13E-03
			gr25	-0.6582	1.36E-03
Upper SN	0.0375	1.40E-04	gr27	-0.2858	6.61E-03
			gr4	1.2138	7.00E-03
Lower SN	0.0570	3.06E-07	gr26	-0.2193	3.19E-03
			gr25	-0.1809	3.73E-03
			gr23	0.1589	9.37E-03

5.3.3 Penalized regression analysis

In 12 fitted models[6], all the regression coefficients of metabolomic PRS terms were shrunk to zero. The R^2 of the other 28 models, where one or more metabolomic PRS had some predictive power for the phenotype, ranged from 0.0012 (lower DB) to 0.5531 (height).

Table 5.6 contains a summary of the Lasso results. The R^2 of all 40 fitted models are reported. Besides, within each model, all the metabolomic PRS with a non-trivial regression coefficient β are listed.

Table 5.6 *Phen* $\sim$ multiple *MetabPRS* Lasso results.

Phenotype	R^2	MetabPRS	β
Height	0.5531	VLDL.D	-0.0045
		Bis.FA	-0.0022
		otPUFA	-0.0019
		Free.C	-0.0018
		XL.VLDL.TG	-0.0018
		S.HDL.P	-0.0009
		M.HDL.C	-0.0005

(Continued on next page)

[6]The corresponding phenotypes are: lower back pain, Oswestry disability total score, LDD developmental score, L1 LDD, L5 LDD, lower LDD, upper DB, upper HIZ, type 1 MC, overall SN, upper SN and lower SN.

Table 5.6 *Phen* $\sim$ multiple *MetabPRS* Lasso results (cont'd).

Phenotype	R^2	MetabPRS	β
		S.HDL.TG	-0.0005
		SM	-0.0003
		XS.VLDL.L	-0.0001
		Gp	-0.0000
		Glc	0.0028
		Ace	0.0033
		CH2.in.FA	0.0052
Weight	0.3408	ApoB	0.0162
BMI	0.0727	S.HDL.L	0.0149
		Ala	0.0447
		Leu	0.0596
		HDL.C	0.0723
		ApoB	0.2134
Smoking	0.1133	VLDL.D	-0.2073
		Gp	0.2523
LBP	0.0231	-	-
Sciatica	0.0162	L.HDL.L	-0.1075
Oswestry	0.0014	-	-
VAS (test day)	0.0047	-	-
VAS (severest)	0.0525	XXL.VLDL.L	-0.2703
		L.VLDL.TG	-0.0558
		M.VLDL.TG	-0.0225
		bOHBut	0.0074
		Bis.FA	0.0330
		IDL.L	0.1779
Overall LDD	0.0785	Tyr	-0.3662
Deg score	0.0409	Tyr	-0.1449
Dev score	0.0501	-	-
L1 LDD	0.0232	-	-
L2 LDD	0.0992	Cit	0.0015
		Urea	0.0180
		Tyr	0.0701

(Continued on next page)

Table 5.6 *Phen* ∼ multiple *MetabPRS* Lasso results (cont'd).

Phenotype	R^2	MetabPRS	β
L3 LDD	0.1167	AcAce	0.0077
		SM	0.0449
L4 LDD	0.0246	Tyr	-0.0331
L5 LDD	0.0000	-	-
Upper LDD	0.1385	Tyr	-0.1042
Lower LDD	0.0000	-	-
Overall DB	0.0174	Tyr	-0.0500
Upper DB	0.0493	-	-
Lower DB	0.0012	Tyr	-0.0165
Overall SS	0.1344	Tyr	-0.0856
Upper SS	0.1954	Tyr	-0.1203
		bOHBut	-0.0082
		SM	0.0202
		Bis.DB	0.0472
		AcAce	0.0908
Lower SS	0.0182	VLDL.D	-0.0924
Overall HIZ	0.0296	M.HDL.PL	-0.1107
		Ace	-0.0456
		Tyr	-0.0319
		ApoA1	-0.0085
		M.VLDL.TG	0.0028
		Lac	0.0112
		otPUFA	0.0142
		S.HDL.TG	0.0302
		Ile	0.0788
Upper HIZ	0.0000	-	-
Lower HIZ	0.0247	M.HDL.PL	-0.0426
		ApoA1	-0.0257
		PC	-0.0222
		Ace	-0.0023
		Tyr	-0.0023
		bOHBut	0.0051

(Continued on next page)

Table 5.6 *Phen* ∼ multiple *MetabPRS* Lasso results (cont'd).

Phenotype	R^2	MetabPRS	β
		S.HDL.TG	0.0278
		Ile	0.0565
Overall MC (cont)	0.0047	VLDL.D	-0.0387
		L.LDL.CE	0.0022
Overall MC (bool)	0.0128	VLDL.D	-0.1780
Upper MC (cont)	0.0101	His	-0.0142
		IDL.TG	0.0128
		Est.C	0.0137
Upper MC (bool)	0.0306	His	-0.0381
		IDL.TG	0.0748
		Est.C	0.1048
Lower MC (cont)	0.0300	VLDL.D	-0.1118
		XXL.VLDL.PL	-0.0089
		L.LDL.CE	0.0218
Lower MC (bool)	0.0417	VLDL.D	-0.4138
		Urea	-0.0515
		XXL.VLDL.PL	-0.0216
		Bis.DB	0.0350
		L.LDL.L	0.0428
		L.LDL.FC	0.0500
Any MC (DS)	0.0573	His	-0.2465
		XXL.VLDL.PL	-0.1922
		LDL.D	-0.1261
		Alb	-0.0393
		L.LDL.CE	0.1239
		L.LDL.C	0.1464
		Bis.DB	0.1778
Type 1 MC (DS)	0.0000	-	-
Type 2 MC (DS)	0.0542	XXL.VLDL.PL	-0.3265
		LDL.D	-0.1981
		Pyr	-0.0950
		Alb	-0.0713

(Continued on next page)

Table 5.6 *Phen* ∼ multiple *MetabPRS* Lasso results (cont'd).

Phenotype	R^2	MetabPRS	β
		His	-0.0654
		VLDL.D	-0.0469
		Gly	-0.0313
		Bis.DB	0.0137
		Glol	0.0140
		S.HDL.TG	0.0279
		XL.HDL.P	0.0491
		L.LDL.PL	0.1099
		CH2.DB	0.1703
Overall SN	0.0058	-	-
Upper SN	0.0000	-	-
Lower SN	0.0228	-	-

5.4 Discussion

To my knowledge, this is the first study associating anthropometric, clinical and LDD MRI phenotypes with polygenic risk scores of metabolomic measurements. By integrating genomic, metabolomic and phenotypic data, potential biomarkers for LDD were identified with a purely data-driven approach, providing us with insights into the possible underlying metabolomic mechanism of LDD.

This study also serves as an illustration of the integrative framework proposed for the analysis of big omics data. Cohorts with metabolomic data are typically small compared to those with genomic data. Through the approach of this study, metabolomic data is no longer compulsory to future researchers if they would like to study the underlying metabolomic continuum of phenotypes of interest. The only data sets required are (1) their own genomic data and (2) GWAS summary statistics of metabolomic measurements based on a similar population from other studies.

Furthermore, the collection of metabolomic data for larger GWAS cohorts would enable us to build more accurate metabolome prediction models. Accompanied with transcriptome imputation methods like PrediXcan [Gamazon et al., 2015], this integrative framework could

potentially help us understand comprehensively[7] how information flows from the human genome to the transcriptome, to the metabolome and, finally, to the phenotype of interest. This knowledge is of great significance in deciphering complex diseases, which is pivotal in biomedical research as well as providing high-quality personalized health care.

5.4.1 Discussion of selected significant findings

5.4.1.1 The relationship between lipid levels and weight/BMI

As expected, the strongest signals found in this study concerned weight and BMI – the PRS for IDL, LDL, and VLDL related metabolites positively affected weight and BMI on average, whereas the PRS for HDL related metabolomic measurements tended to negatively influence weight and BMI.

To understand these results, consider the function of HDL, LDL, IDL and VLDL particles. As shown in Figure 5.3, lipids are transported in the circulatory system packed in different types of lipoproteins. Synthesized in the liver, VLDL delivers energy-rich triacylglycerol (TAG) to cells in the body [University of Washington, 2018]. As VLDL particles are derived of TAG, they become denser and are next remodeled at the liver, transforming into LDL, which delivers cholesterol to cells [University of Washington, 2018]. If there is any excess cholesterol from cells, HDL would bring it back to the liver [University of Washington, 2018]. The density of IDL is between that of VLDL and LDL particles. Like VLDL and LDL, it transports a variety of triglyceride fats and cholesterol to cells.

Therefore, if the HDL levels are too low, the process of reverse cholesterol transport would be hindered – the excess cholesterol from cells cannot be brought back to the liver in time, increasing the risk of obesity and artery diseases. On the other hand, a high level of VLDL/IDL/LDL particles indicates increased triglycerides. In patients with obesity, VLDL is often over-produced [Adiels et al., 2006].

5.4.1.2 Association between lipid levels and sciatica

In this study, sciatica was found to be positively influenced by the PRS for IDL related metabolites and negatively affected by the PRS for HDL related metabolites. This was in

[7] Partially comprehensive... Noise is always present.

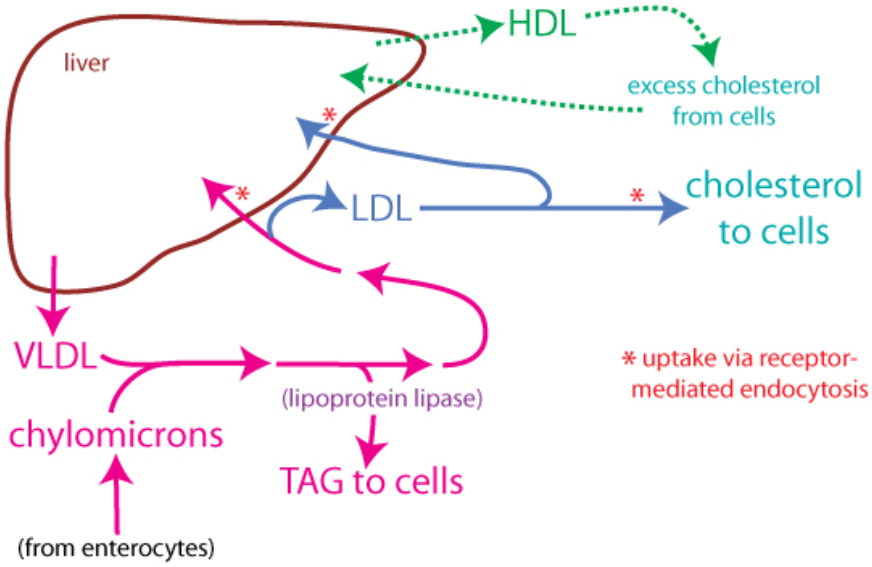


Fig. 5.3 The role of HDL, LDL, IDL, and VLDL particles in liver cholesterol transport [University of Washington, 2018].

line with [Leino-Arjas et al., 2008], which discovered the positive correlation between higher atherogenic[8] serum lipid levels and sciatica.

[Longo et al., 2011] has shown that patients with symptomatic herniated lumbar disc have significantly higher triglyceride and total cholesterol concentration. The herniated discs may compress the spinal nerve root, which is one of the causes of sciatica [Longo et al., 2011]. Therefore, blood lipid levels may be a potential biomarker for sciatica and LDD in general.

5.4.1.3 Potential biomarkers for LDD

Generally speaking, my metabolomic PRS related findings were more significant for LDD phenotypes defined on upper disc levels[9]. This is probably due to the fact that the upper disc levels are more developmental in nature; since the polygenic risk scores were calculated based on genomic data, they could explain the variance in developmental traits better (as opposed to degenerative conditions). This may also indicate shared genetic components between LDD and metabolomic traits. Further causal modeling studies are needed to determine whether the

[8] With the tendency to promote fatty deposits in the arteries.

[9] Except for the phenotypes with too highly imbalanced data – with too few positive cases, we lacked statistical power to detect such significant association, if any.

risk loci for certain metabolomic measurements are also significantly associated with LDD phenotypes.

Blood lipid levels and the mean diameter for VLDL particles

The study showed that as the PRS of the mean diameter for VLDL particles (VLDL.D) decreases, the degree of modic change would on average become more severe. A low VLDL.D is typically bad [Beard et al., 1996; Colhoun et al., 2002]. It has been found in samples with retinopathy, which is associated with higher VLDL particle levels in patients [Colhoun et al., 2002]. High amounts of oxidized LDLs activate TLR2/4 (toll-like receptor 2/4), and chronic stimulation of TLRs facilitates fatty marrow conversion as in type 2 modic change [Dudli et al., 2016]. Hence, both VLDL.D and VLDL/LDL levels may be potential biomarkers for LDD, more specifically type 2 MC.

A high level of VLDL/IDL/LDL particles also indicates increased triglycerides, one of the risk factors of atherosclerosis [Longo et al., 2011]. Studies have detected an association between atheromatous lesions in the aorta and LDD [Kauppila, 2009]. Moreover, LBP has been found to be associated with aortic calcification and stenosis of lumbar arteries [Kauppila, 2009]. Further clinical research is required to clarify the association of blood lipid levels, atherosclerosis, and LBP/LDD.

Sphingomyelins

From the regression results, we could see that the PRS of sphingomyelin (SM) bore a significantly positive correlation with L3 disc degeneration severity and signal intensity loss in the upper levels.

Early degeneration of the intervertebral disc is associated with a change in cellular differentiation from notochordal cells (NCs) to chondrocyte-like cells (CLCs) in the nucleus pulposus (NP) [Smolders et al., 2013]. Gene expression profiling studies have shown that the SM catabolic process is up-regulated in the transition from NC-rich NP to CLC-rich NP in chondrodystrophic dogs [Smolders et al., 2013]. Hence, SM may be one of the candidate biomarkers of LDD.

Tyrosine

The study demonstrated that on average, as the PRS for tyrosine increases, various LDD related phenotypes would decrease. Theoretically, as a precursor to neurotransmitters, tyrosine could elevate plasma neurotransmitter levels [Rasmussen et al., 1983] and hence boost metabolism. Therefore, it is widely believed to stimulate fat loss. Currently, there are no studies specifically showing that tyrosine could relieve LDD symptoms. Further research is needed to understand the role of tyrosine in LDD.

5.4.2 Limitations of this study

As discussed in Section 4.4.2, one limitation of this study was the small sample size (571 individuals in total) of our original GWAS. Therefore, the GWAS suffered from a relatively low statistical power in detecting significant genetic associations of metabolomic measurements, and the PRS generated in this study would be a less accurate estimate of the metabolomic traits. To improve the estimation performance, a larger GWAS cohort with metabolomic measurements is needed. As an alternative, we could also leverage pleiotropy through multi-trait studies to increase power.

Another possible limitation of this study was the subjectiveness of MRI reads by clinicians. Different clinicians may have distinct tendencies (biases) in reading the MRI scans, possibly making the resulting phenotypes not very reproducible. As the (fake) saying goes – to err is human, so why not use an AI[10]? Recent studies have proposed novel methods to automatically segment the lumbar vertebrae from computed tomography (CT) images [Janssens et al., 2018] and compute LDD gradings from MRIs via supervised learning [Jamaludin et al., 2017a; Jamaludin et al., 2017b]. Nevertheless, the latter methods for automatic LDD grading are still supervised and hence still quite dependent on the accuracy and consistency of clinician reads. Future LDD research may benefit from using semi-supervised or unsupervised computer vision algorithms[11] to define LDD phenotypes in a purely data-driven manner.

[10] Short for artificial intelligence. Nowadays, an abused buzzword.

[11] For instance, [Cho et al., 2015] and [Siva et al., 2013] are two methods for unsupervised object detection.

6

Conclusion

So we beat on, boats against the current, borne back ceaselessly into the past.

– F. Scott Fitzgerald, *The Great Gatsby*

6.1 Summary of main findings

As a prevalent global health problem[1], lower back pain (LBP) is one of the top conditions leading to disability [Vos et al., 2012]. A major reason for LBP is lumbar disc degeneration (LDD), which is measurable through magnetic resonance imaging (MRI) assessment. In Chapter 2, I demonstrated that (1) the severity of LDD is significantly associated with disc level; and (2) the five disc levels form two clusters. Accordingly, a scheme utilizing truncated normal distribution to quantify the degree of LDD from raw MRI reads was proposed. The composite MRI phenotypes calculated based on this scheme were analyzed in this thesis.

In recent years, researchers have gained interest in the role of altered metabolism in the development and progression of LDD [Samartzis et al., 2013a]. In Chapter 3, I carried out correlation analysis to study the relationship between LDD and metabolomic traits. One of the major findings was the positive correlation between acetate / small LDL and developmental LDD phenotypes. This might indicate shared genetic components between

[1] Occupational determinants of LBP include bending and carrying loads [Ozguler et al., 2000]. One particular population at risk of LBP consists of researchers and graduate students since they often have bad posture (Figure 6.1).

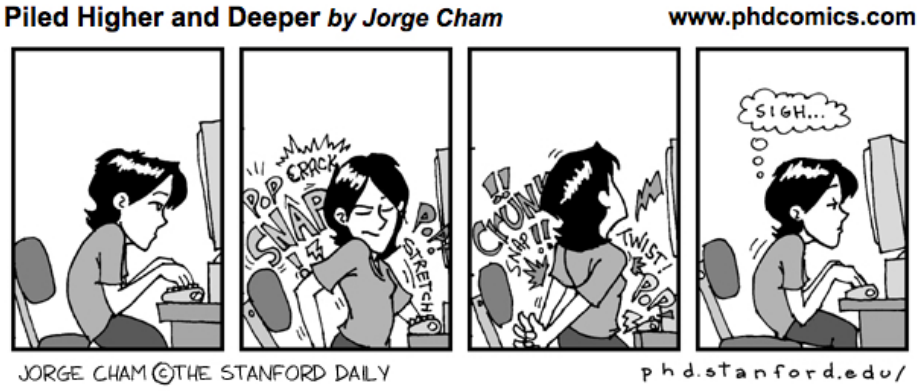


Fig. 6.1 Graduate student researching and suffering from LBP [Cham, 2000].

LDD and lipid metabolism since acetate is an epigenetic metabolite enhancing lipid synthesis [Gao et al., 2016]. Furthermore, I discovered that type 2 modic change (MC) tends to be negatively associated with HDL related metabolites. This is consistent with our current understanding of the pathology of type 2 MC – high amounts of oxidized LDLs, often accompanied by low HDL levels, could activate TLR2/4 (toll-like receptor 2/4), and chronic stimulation of TLRs could facilitate fatty marrow conversion, leading to the development of type 2 MC [Dudli et al., 2016].

Apart from the correlation analysis, self-organizing maps were also fitted to explore the underlying metabolomic continuum of LDD and several other phenotypes in Chapter 3. Unfortunately, there were no significant results related to LDD, which is probably due to the limited sample size and heavily imbalanced nature of LDD phenotypes. Nevertheless, we could observe a strong association between an individual's metabolomic profile (especially in terms of lipid-related measurements) and his or her weight/BMI. This reinforces the current clinical knowledge [W. M. Miller et al., 2005].

It is widely accepted that genetic variants associated with metabolomic traits typically have relatively large effect sizes [Gieger et al., 2008; Rhee et al., 2013]. In order to better understand the genetic roots of metabolomic measurements and in turn, the metabolomic context of different traits and conditions (e.g. LDD), I scanned the whole genome for SNPs significantly associated with different serum ^{1}H NMR metabolomic measurements in Chapter 4. 130 genome-wide association studies (GWAS) for the metabolomic traits were performed based on a population cohort of 571 individuals, identifying 123 unique SNPs significantly associated with one or more metabolomic measurements. Gene-based annotation showed that among all the hits, exonic, intronic and UTR3 variants were enriched, whereas intergenic

variants were underrepresented. There were altogether 42 metabolomic measurements with one or more significantly associated SNP(s), most of them related to lipids and fatty acids. For instance, polyunsaturated fatty acids were found to be significantly associated with the FADS1/FADS2 loci, and CTTNBP2 was identified as a potential risk locus for a cluster of lipid/FA related metabolites.

Following up the detected genetic associations, meta-analysis was performed and the human metabolome was estimated through polygenic scoring. Anthropometric, clinical and LDD MRI phenotypes were associated with polygenic risk scores of metabolomic traits in Chapter 5, identifying novel biomarkers for LDD including blood lipid levels, the mean diameter for VLDL particles, sphingomyelins, and tyrosine. These potential biomarkers for LDD could inspire us in understanding the possible underlying metabolomic mechanism of LDD and potentially aid personalized diagnosis and treatment of LBP.

The study in Chapter 5 also illustrates the framework for the integrative analysis of big omics data proposed in this thesis (Figure 1.16). In the framework, metabolome prediction models are first trained based on GWAS and metabolomic data. Future researchers could then utilize these fitted models to estimate metabolomic features for their GWAS cohort (preferably from a similar population to that the models were trained on) and study the underlying metabolomic continuum of phenotypes of interest. This would be quite useful in real-life scenarios since cohorts with metabolomic data are typically smaller and harder to obtain compared to those with genomic data. This process of the integration of big omics data could help us discover known and novel metabolomic biomarkers associated with complex traits and gain a better understanding of the biological mechanisms beneath these associations.

6.2 Future directions

6.2.1 GWAS-related future work

6.2.1.1 Increasing the power of GWAS for metabolomic traits

The population cohort studied in this thesis is one of the world's largest cohorts with LDD MRI data. However, since there were much fewer people with metabolomic data, my current genome-wide association studies suffered from a lack of statistical power in detecting truly positive associations due to a limited sample size (571 individuals).

The obvious way to circumvent this is to increase the sample size. Since measuring the human metabolome is still quite costly at the moment, another possible way to boost the power of GWAS is employing multi-trait methods [Porter and O'Reilly, 2017] on high-dimensional metabolomic data, making use of the fact that the metabolomic measurements are closely correlated to each other.

6.2.1.2 Further research

Following my current GWAS, we could also conduct various other studies.

First of all, to better understand the mechanisms underlying the metabolomic continuum, we could perform pathway-based and cell-type enrichment analyses based on my GWAS results.

Additionally, the findings of my study could be contrasted with those from other GWAS on metabolomic measurements in terms of, for instance, characteristics of the cohorts and the identified SNPs. In this way, we could assess our confidence in the current results as well as gain a better understanding of them.

Thirdly, causal variants pointing to metabolic mechanisms underlying the significant associations could be identified through fine-mapping algorithms like PAINTOR [Kichaev et al., 2014] and CAVIAR [Hormozdiari et al., 2014].

Finally, researchers have hypothesized that molecular adaptations in metabolic pathways have accompanied the dietary shift during evolutionary courses [Blekhman et al., 2014]. Hence, the genetic loci significantly associated with metabolomic features may be under selection pressure, which could be tested for statistically.

6.2.2 Automatic phenotyping for lumbar disc degeneration

In my study (and almost all the current LDD studies based on MRI), the MRI phenotypes were defined based on the reads by experienced clinicians.

Please imagine you are one of these clinicians. You load the black and white spine images on your computer. You carefully zoom in and out, checking if there is a tiny white dot[2] in an intervertebral disc. You do this, for every disc level shown in the image and for every individual with a scan. If we only consider the lumbar region, for our cohort, you need to

[2]This is an annular tear, seen as a high intensity zone.

perform $1,416 \times 5 = 7,080$ checks. Your eyes are now sore, back achy, arms are heavy, but wait up – that is only one of the LDD phenotypes. You still need to check for disc bulging, signal intensity loss and modic/endplate changes shown as zigzagged lines along the endplate. With five phenotypes, that is now a whopping total of $7,080 \times 5 = 35,400$ reads. Additionally, to attain better precision, not all of these readings are true or false ones. Certain phenotypes are ordinal, meaning, for example, you would need to distinguish different patterns/shades of gray[3] and grade them as 0, 1, 2, or 3.

This is a tedious[4] process. What's worse, sometimes clinicians need to proof-read the same cohort multiple times to reduce the error rate. To liberate the clinicians and increase the accuracy of MRI reads, machine learning methods could be adapted to directly extract LDD information from MRI scans [Jamaludin et al., 2017a; Jamaludin et al., 2017b].

Current methods for automatic LDD grading are still supervised and hence dependent on the accuracy and consistency of clinician reads. Unfortunately, since different clinicians are prone to distinct tendencies in reading the MRI scans (e.g. different distributions of 0, 1, 2, and 3 in the previous example), the current MRI reads may be a little subjective and not very reproducible. Future LDD research may benefit from using semi-supervised or unsupervised image recognition algorithms to define LDD phenotypes in a purely data-driven way.

6.2.3 Integrating transcriptomic data into the analysis framework

This thesis mainly focused on the integrative analysis of genomic, metabolomic and phenotypic data. This analysis framework may benefit from incorporating transcriptomic data as well (e.g. Figure 6.2) to make the overall model even more data-rich and interpretable.

Since this approach is completely driven by various types of big omics data, it benefits from being free of potential biases from hypothesizing based on current knowledge. However, as Aaron Levenstein once said, "Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital." Drawing conclusions solely relying on data without any further human thinking could be hazardous. Therefore, any findings from this framework (e.g. conclusions of this thesis) should be further examined by biologists and clinical experts through laboratory research to gather experimental evidence in the future.

[3] This example is for signal intensity loss. In my study, this is measured with Schneiderman's score.

[4] Again, a shout out to Dr. Jaro Karppinen and Dr. Dino Samartzis for making all this possible! I am really sorry if I have depicted the reading process imprecisely and/or offended you by being melodramatic.

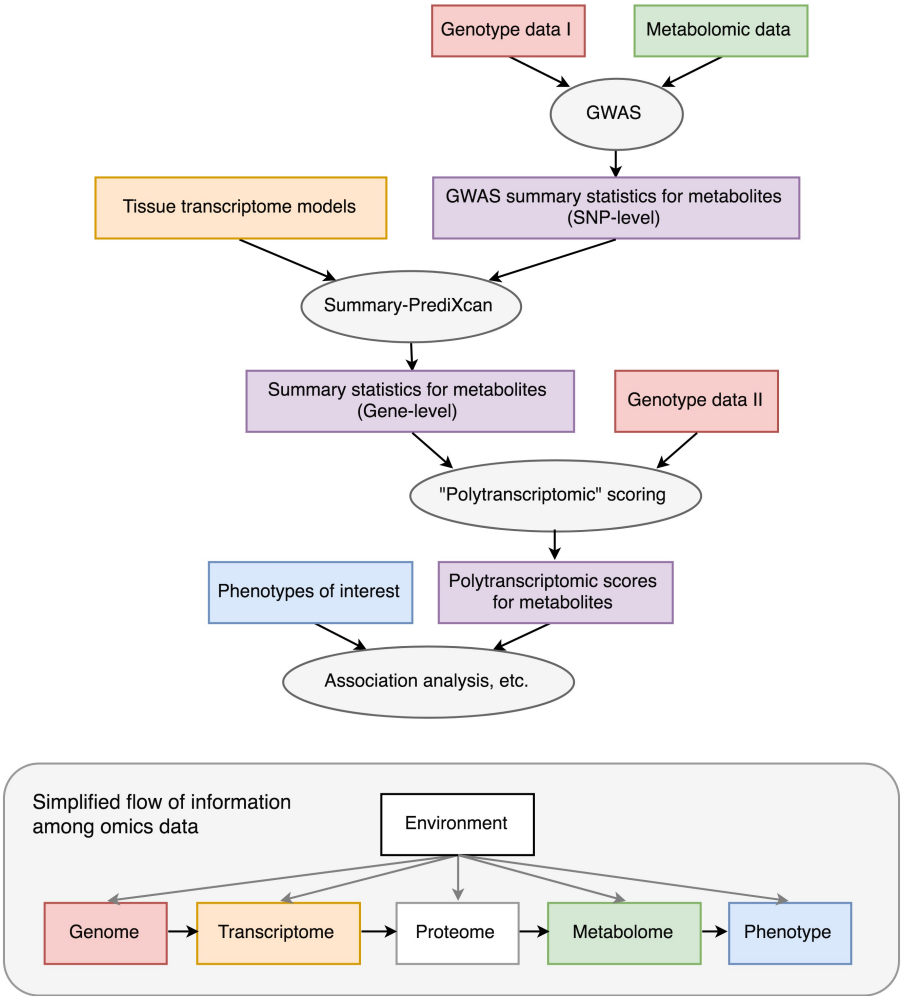


Fig. 6.2 One possible way to integrate transcriptomic data into my analysis framework. In this particular approach, summary-PrediXcan [Barbeira et al., 2016] is used for transcriptome prediction. As a side note, to improve the performance of our metabolome prediction model in the framework, more metabolomic data would be required.

6.3 Closing remarks: connecting the dots

A living organism is a strange, intricate and beautiful thing. Geneticists' efforts to dissect how its countless traits are developed and passed from generation to generation trace back to when Gregor Mendel planted his first pea plants. With the rapid development of science and technologies, computational researchers have joined forces with traditional biologists in genetic research, and nowadays, the availability of various types of big omics data is promoting a massive shift in the area.

Most of us have learned to solve the "connect the dots" puzzle as a kid. The mystery of inheritance is just like one of these puzzles waiting to be solved, with different types of omics data, phenotypes of interest and environmental factors as scattered dots. This puzzle, however, is much more complicated, since the solution to it, if any, is not a one-way path. The edges between two dots could be extraverted ($\leftrightarrow$), and the relationships are almost never one-to-one. Moreover, we need to understand the meaning behind each edge, which requires the study of several other downstream or upstream edges.

Hopefully, through connecting these dots, or rather, the process of seeking to do so, we could better understand how information flows from the human genome to the transcriptome, the proteome, the metabolome and, finally, to traits of interest. This knowledge would be valuable in aiding the personalized treatment of complex diseases and thus, providing affordable, high-quality health care.

The quote at the beginning of this chapter is the very last sentence of "The Great Gatsby", one of my favorite books back when I was a teenager. I have always resonated with the bittersweet sadness[5] – that inevitable "Omnia Vanitas" feeling – flowing in the poetic line. Yet still, its preceding sentence never fails to inspire me in the most harrowing days:

> *It eluded us then, but that's no matter – tomorrow we will run faster, stretch out our arms farther....And one fine morning –*
>
> – F. Scott Fitzgerald, *The Great Gatsby*

Therefore, at the end of this final chapter and my thesis, I would like to take the liberty of inelegantly twisting my favorite ending.

The solution to this omics "connect the dots" puzzle has always eluded us, and perhaps it forever will. Nevertheless, we actually never needed to attain a perfect solution. By gathering larger cohorts with high-quality omics data, performing integrative analysis and conducting lab experiments to examine the computational findings – by stretching out further, we would be approaching that best solution asymptotically –

Till the "one fine morning", that is truly ahead.

[5] Probably what Leopold Stotch would say as well.

* *https://goo.gl/nv3gJH*

Appendix A

Visualization of GWAS results of metabolomic measurements

Figures A.1 to A.42 visualize the GWAS results of the 42 metabolomic measurements with at least one significantly associated SNP.

In the Manhattan plots (left sub-figures), the blue line suggests moderate significance[1] and the red line indicates genome wide significance[2]. The SNPs reaching genome-wide significance are highlighted in green color; the top hit on each chromosome is also annotated.

All the QQ plots (right sub-figures) demonstrate significant deviations from the diagonal at the upper-right corner, indicating the existence of significantly associated SNPs. Additionally, there is no early deviation from the line of equality ($y = x$). Hence the significant findings are probably not due to an artifact.

[1] Blue line: $y = -\log_{10}(1 \times 10^{-5})$.
[2] Red line: $y = -\log_{10}(5 \times 10^{-8})$.

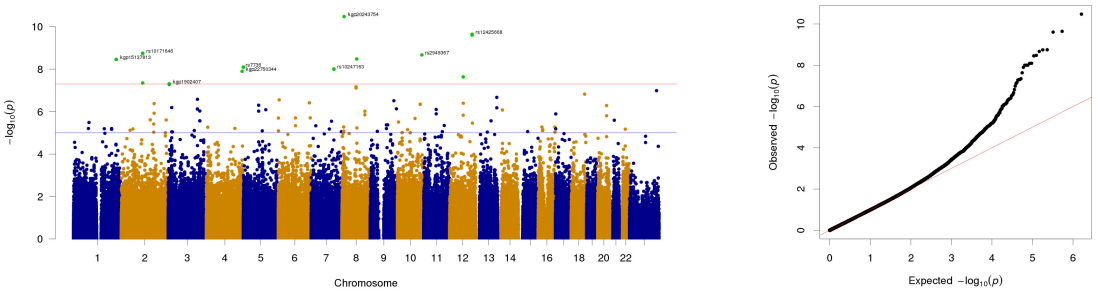


Fig. A.1 Visualization of the GWAS results of Alb.

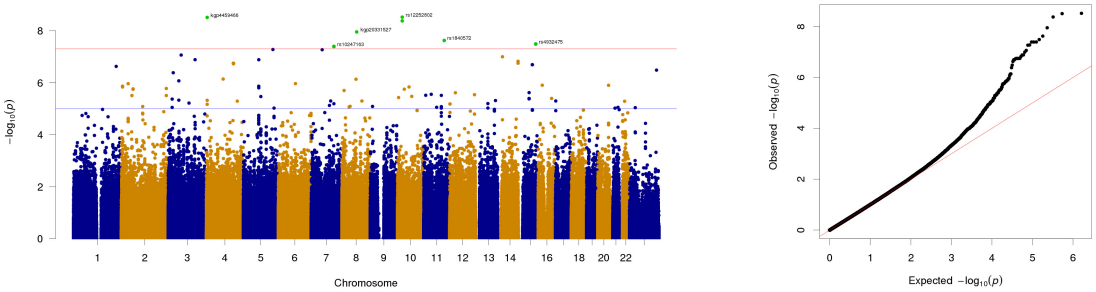


Fig. A.2 Visualization of the GWAS results of FALen.

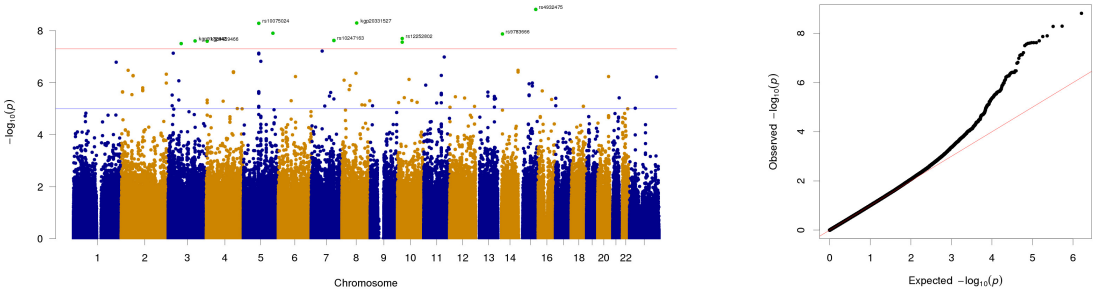


Fig. A.3 Visualization of the GWAS results of CH2.in.FA.

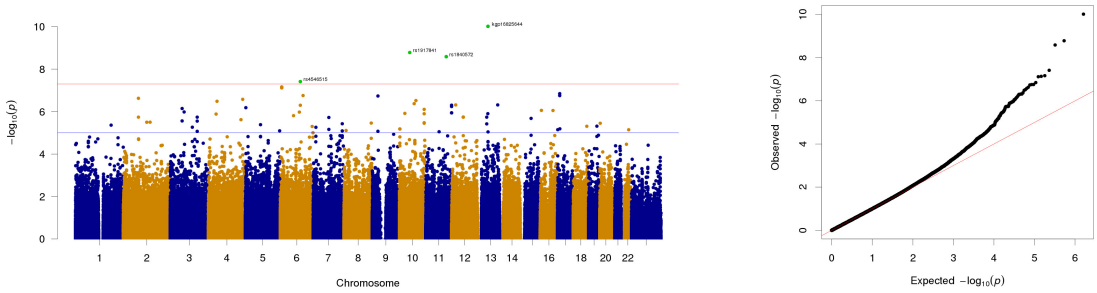


Fig. A.4 Visualization of the GWAS results of Crea.

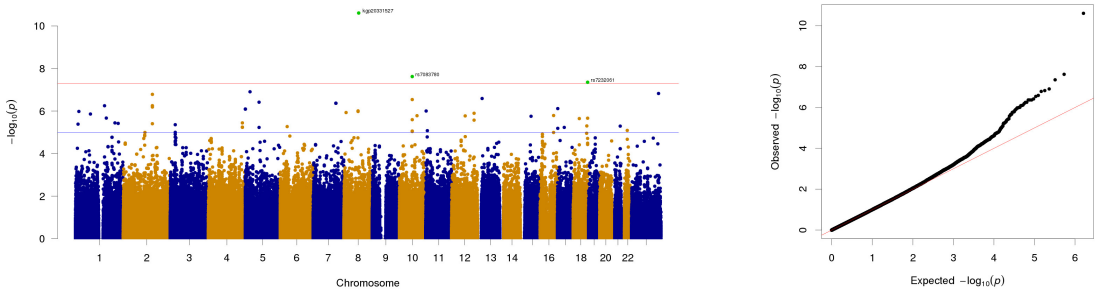


Fig. A.5 Visualization of the GWAS results of ApoA1.

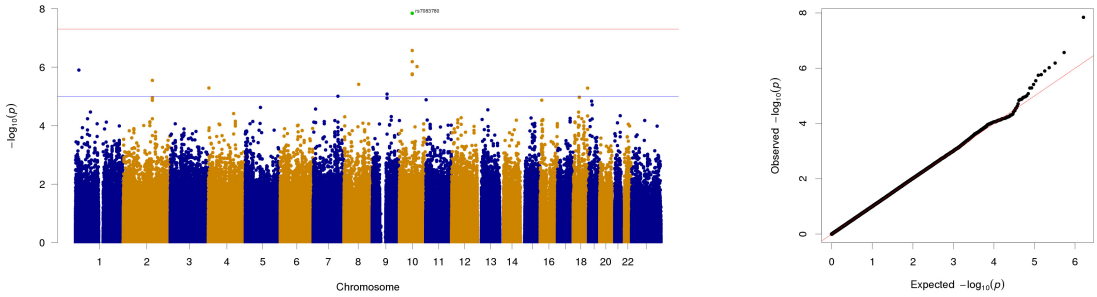


Fig. A.6 Visualization of the GWAS results of HDL2.C.

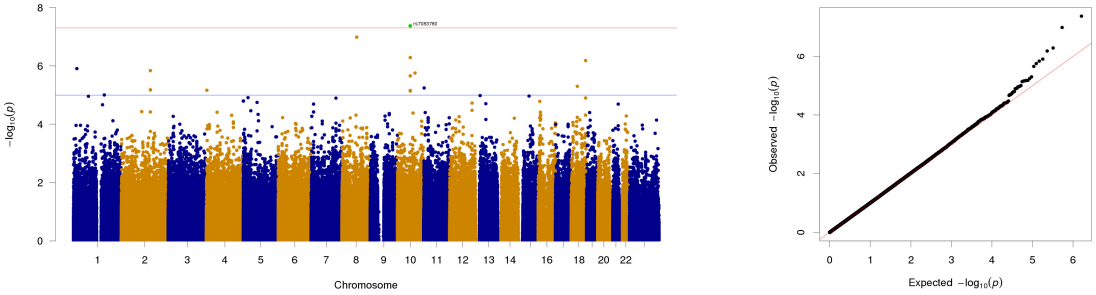


Fig. A.7 Visualization of the GWAS results of HDL.C.

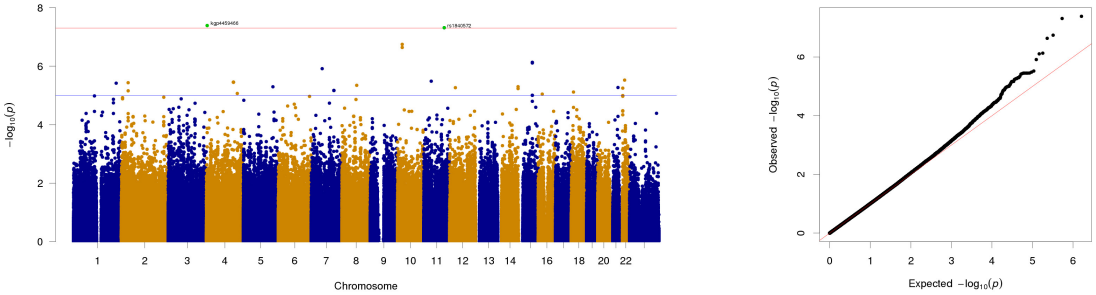


Fig. A.8 Visualization of the GWAS results of Bis.DB.

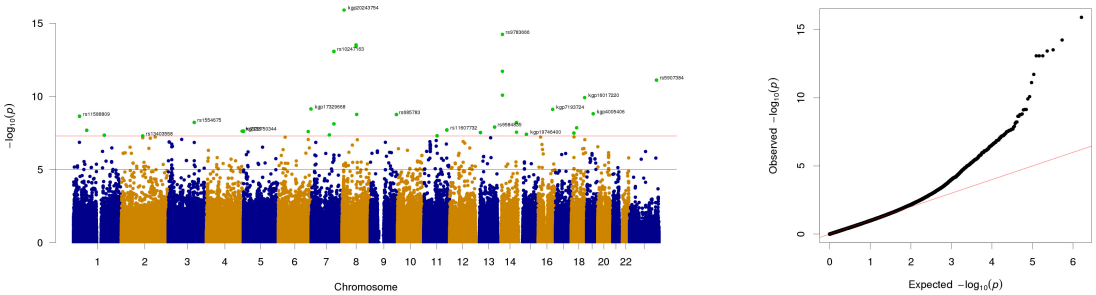


Fig. A.9 Visualization of the GWAS results of VLDL.D.

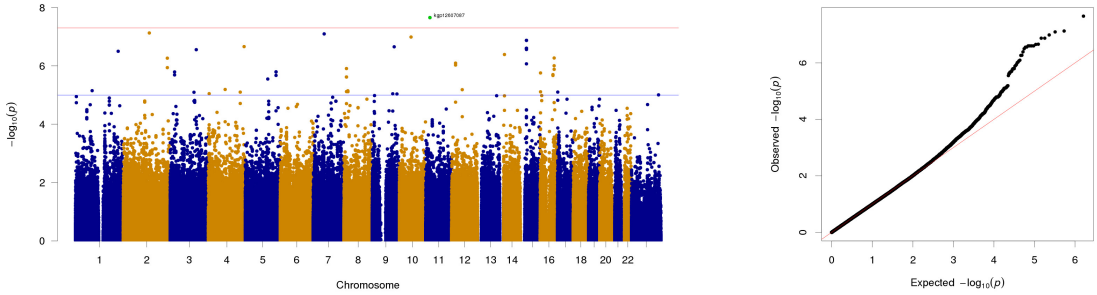


Fig. A.10 Visualization of the GWAS results of Gp.

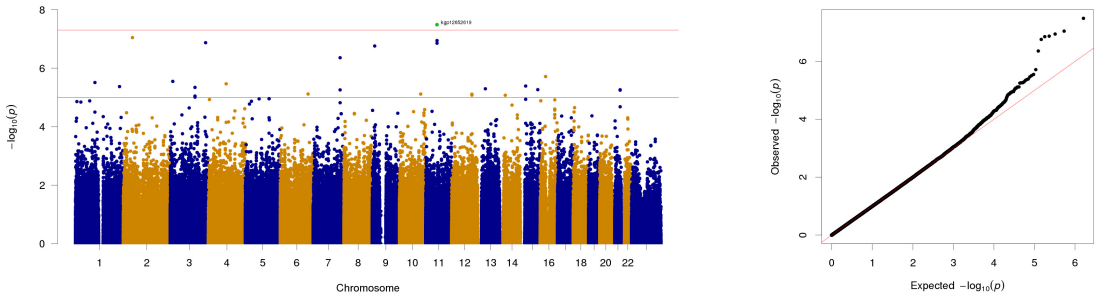


Fig. A.11 Visualization of the GWAS results of XXL.VLDL.L.

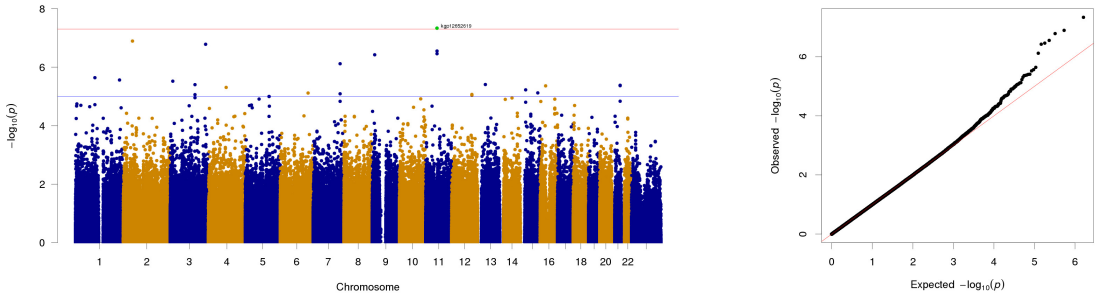


Fig. A.12 Visualization of the GWAS results of XXL.VLDL.TG.

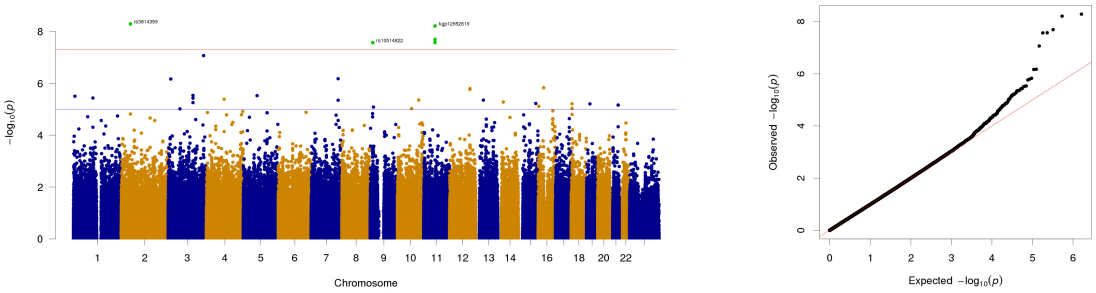


Fig. A.13 Visualization of the GWAS results of XXL.VLDL.PL.

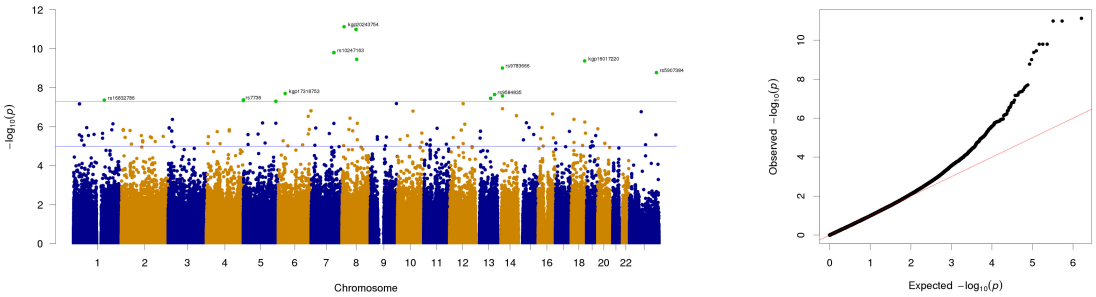


Fig. A.14 Visualization of the GWAS results of HDL.D.

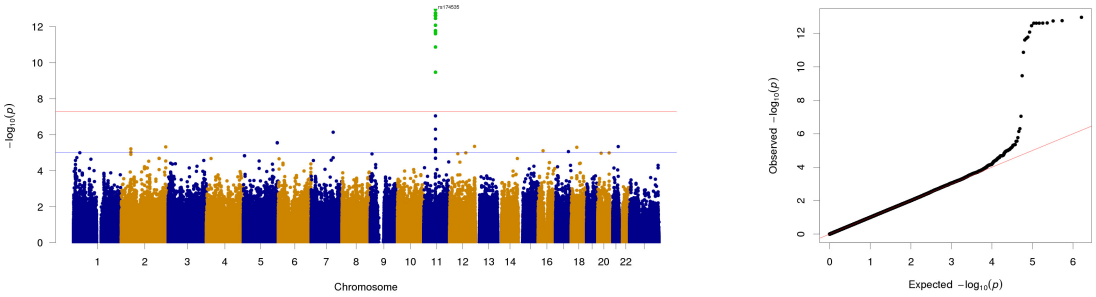


Fig. A.15 Visualization of the GWAS results of otPUFA.

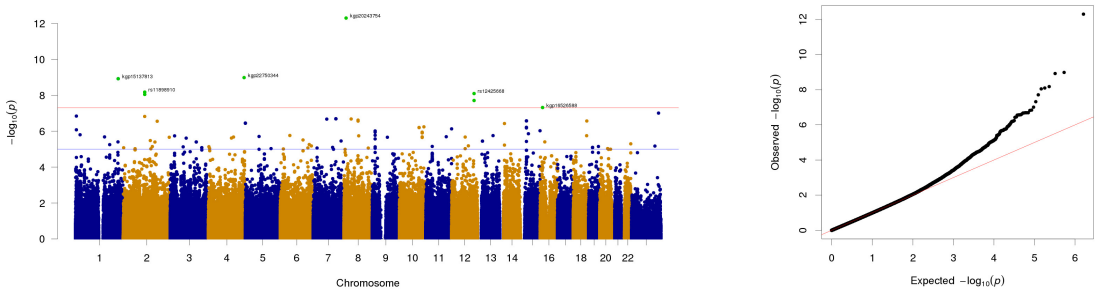


Fig. A.16 Visualization of the GWAS results of S.HDL.L.

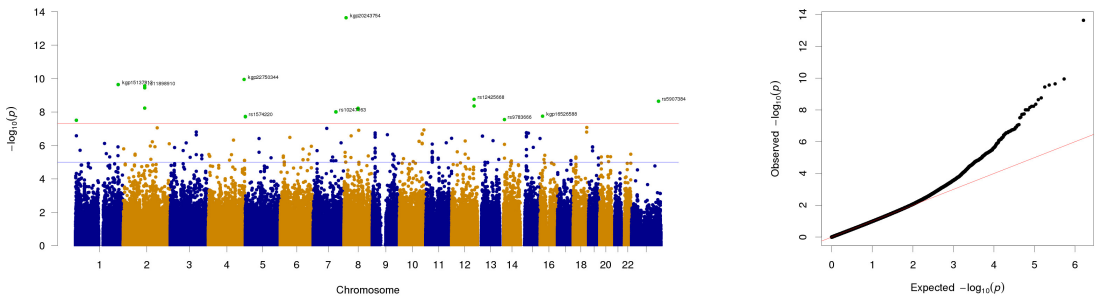


Fig. A.17 Visualization of the GWAS results of S.HDL.P.

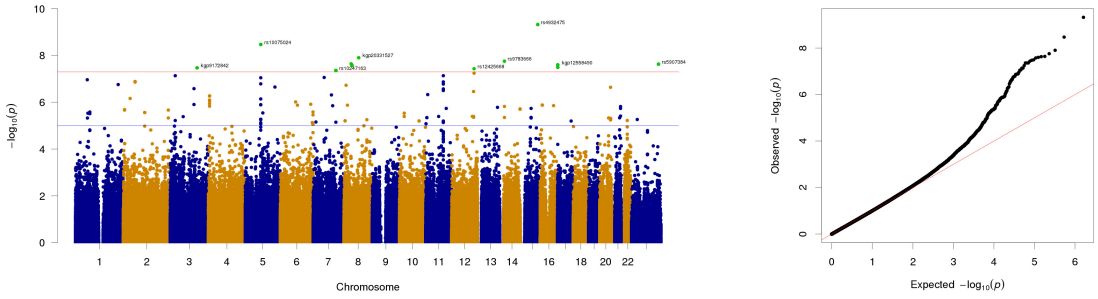


Fig. A.18 Visualization of the GWAS results of FAw79S.FA.

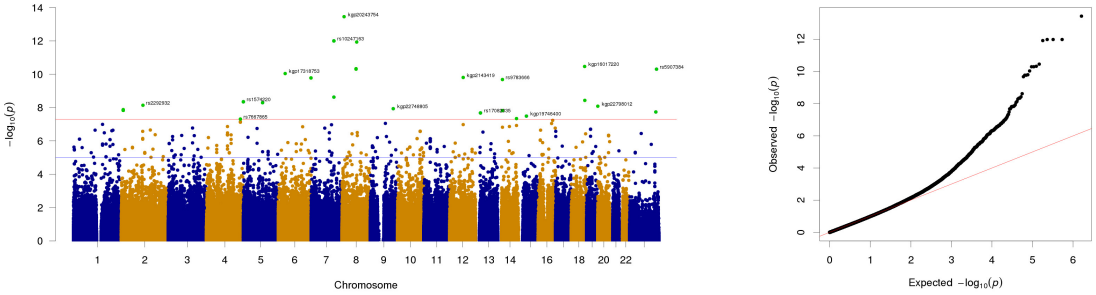


Fig. A.19 Visualization of the GWAS results of HDL3.C.

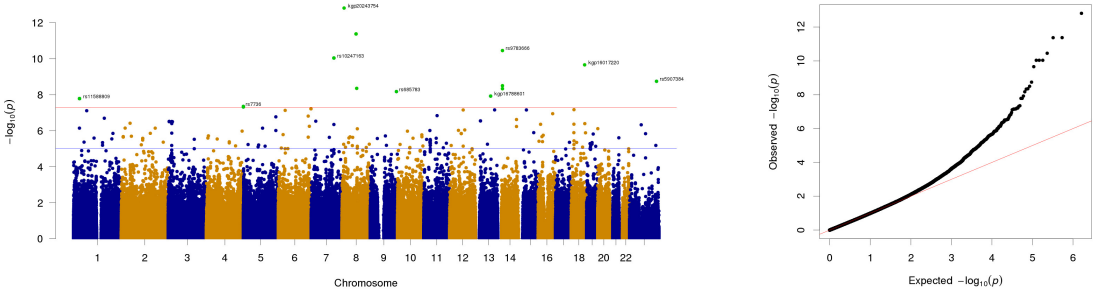


Fig. A.20 Visualization of the GWAS results of LDL.D.

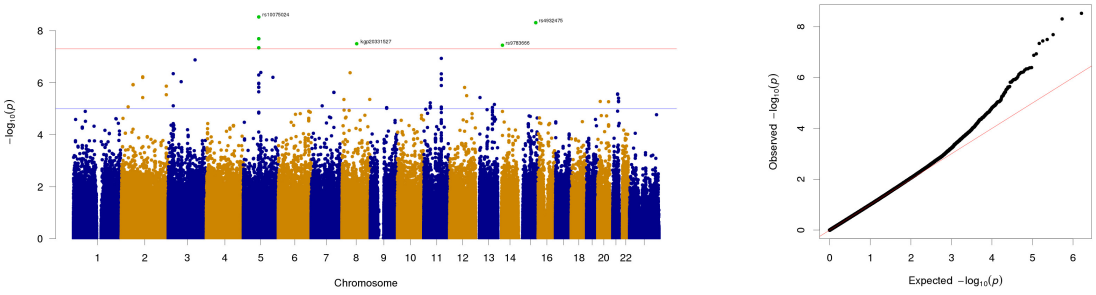


Fig. A.21 Visualization of the GWAS results of CH2.DB.

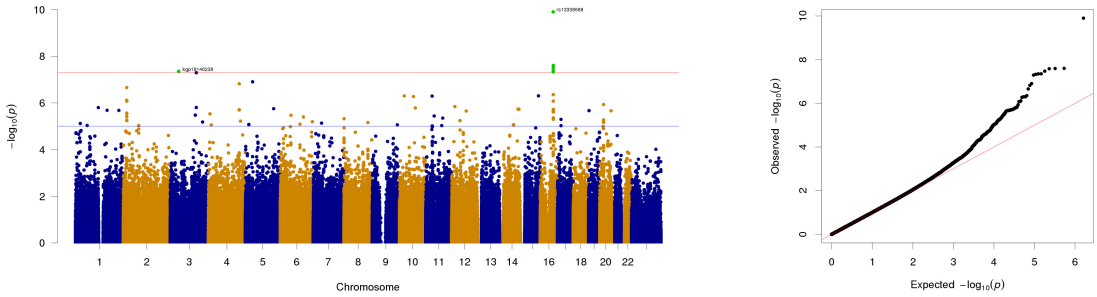


Fig. A.22 Visualization of the GWAS results of Glc.

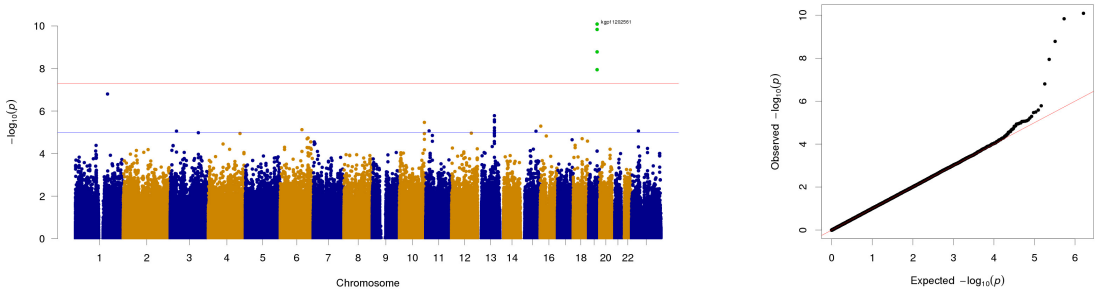


Fig. A.23 Visualization of the GWAS results of L.LDL.FC.

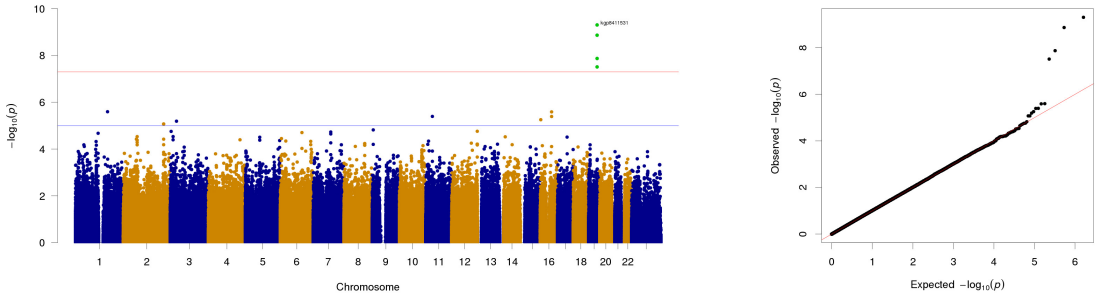


Fig. A.24 Visualization of the GWAS results of S.LDL.L.

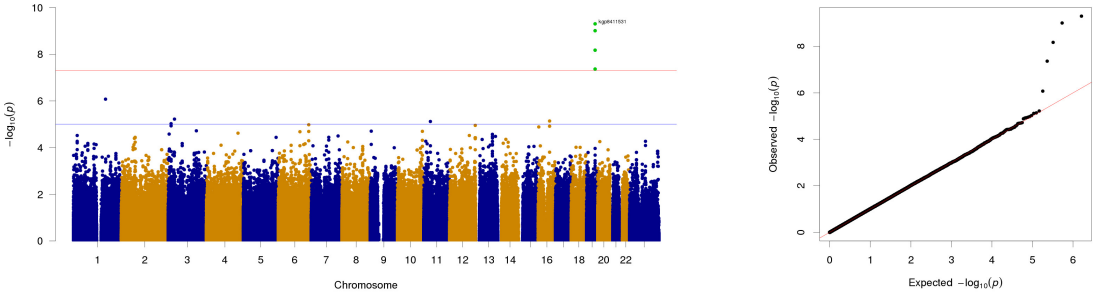


Fig. A.25 Visualization of the GWAS results of M.LDL.L.

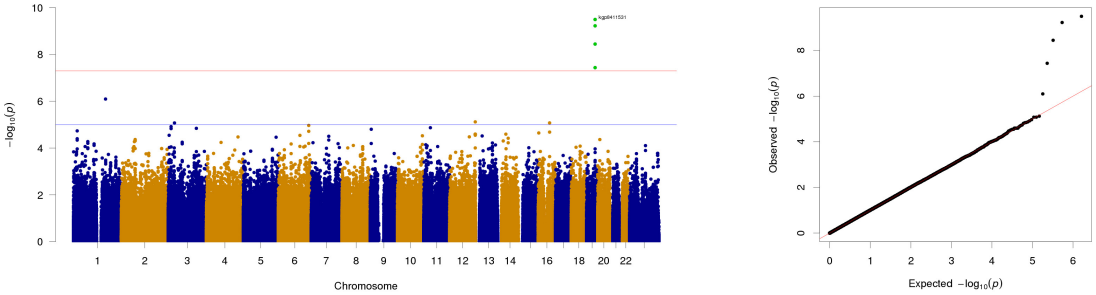


Fig. A.26 Visualization of the GWAS results of M.LDL.C.

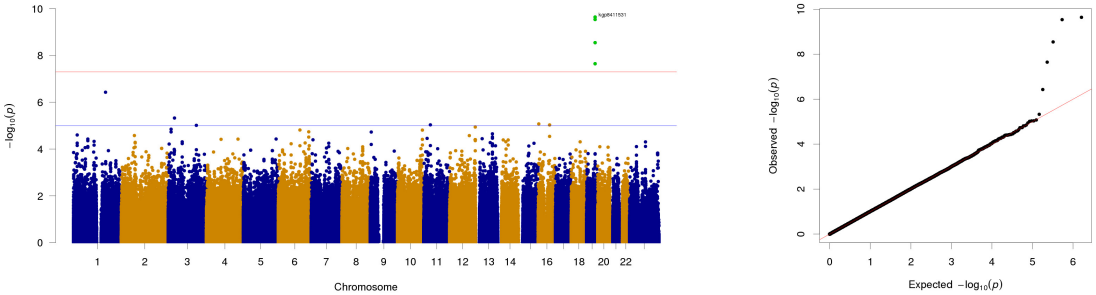


Fig. A.27 Visualization of the GWAS results of LDL.C.

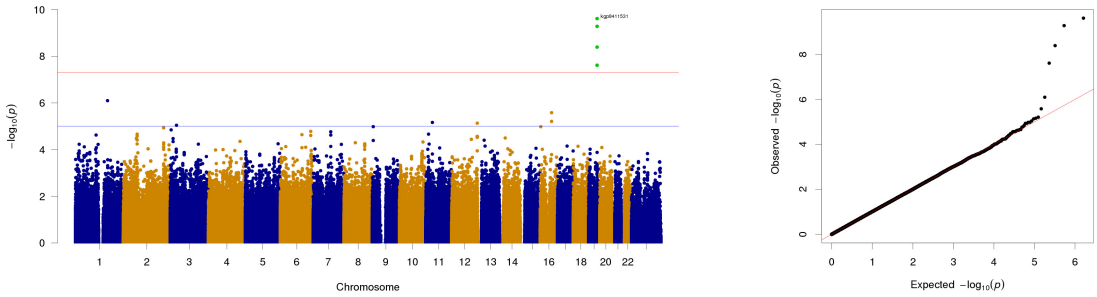


Fig. A.28 Visualization of the GWAS results of S.LDL.C.

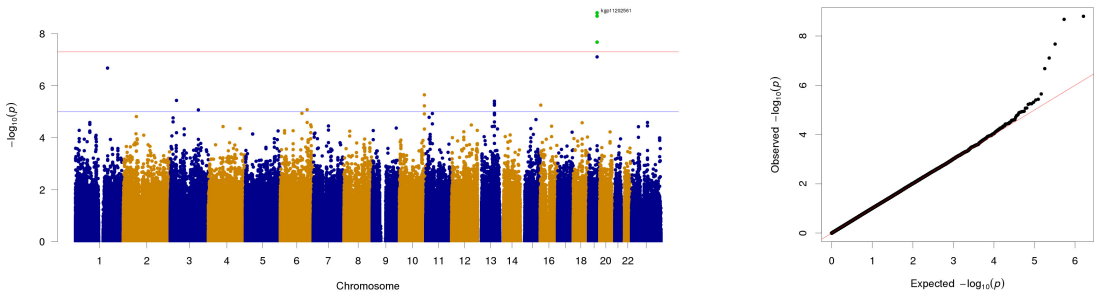


Fig. A.29 Visualization of the GWAS results of L.LDL.L.

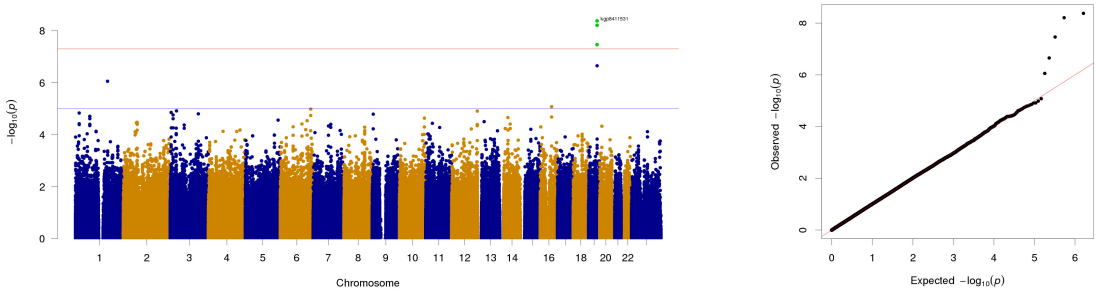


Fig. A.30 Visualization of the GWAS results of M.LDL.CE.

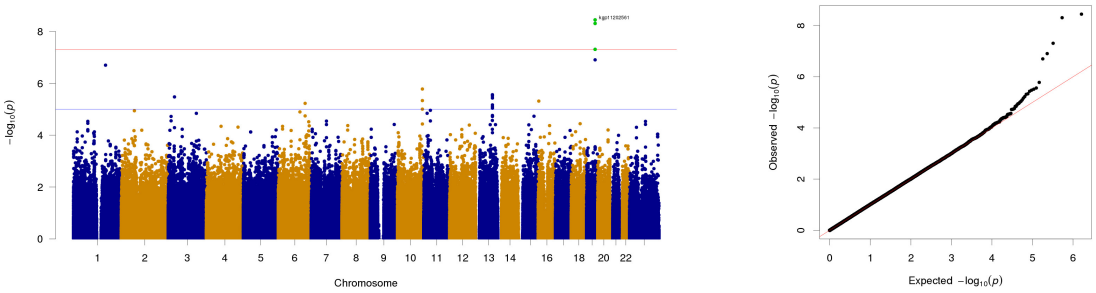


Fig. A.31 Visualization of the GWAS results of L.LDL.P.

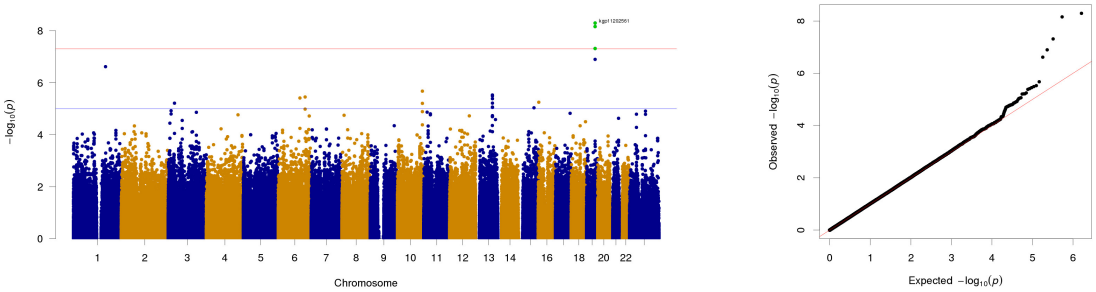


Fig. A.32 Visualization of the GWAS results of L.LDL.PL.

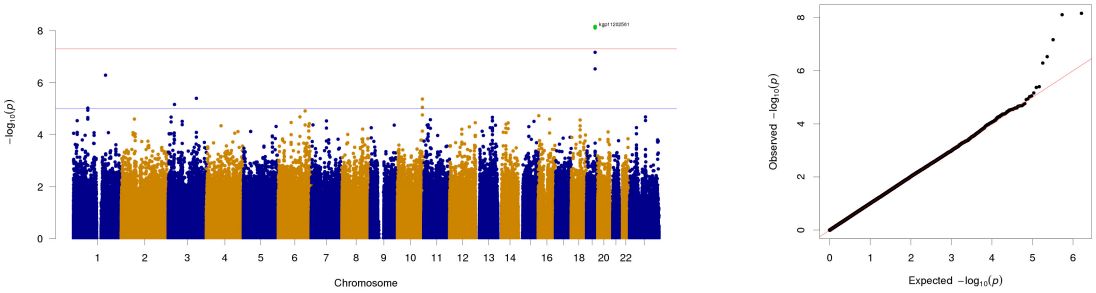


Fig. A.33 Visualization of the GWAS results of L.LDL.CE.

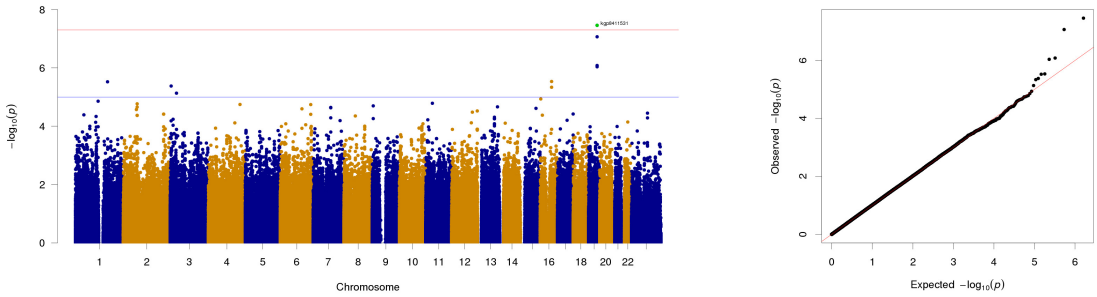


Fig. A.34 Visualization of the GWAS results of S.LDL.P.

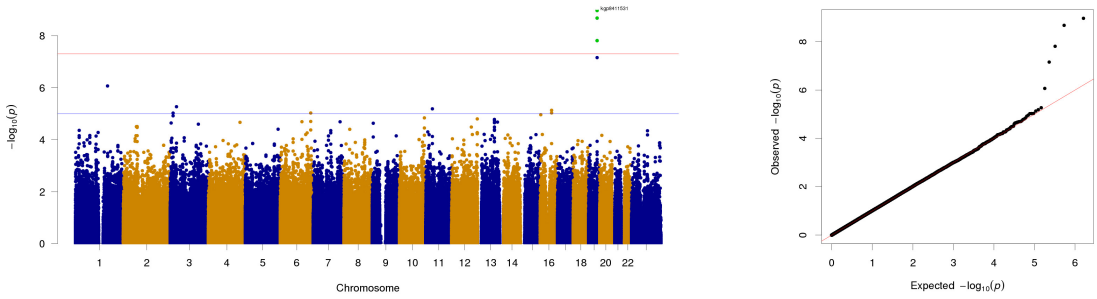


Fig. A.35 Visualization of the GWAS results of M.LDL.P.

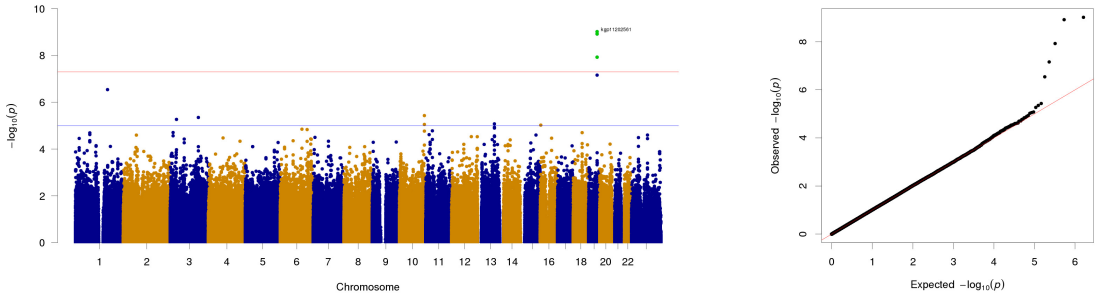


Fig. A.36 Visualization of the GWAS results of L.LDL.C.

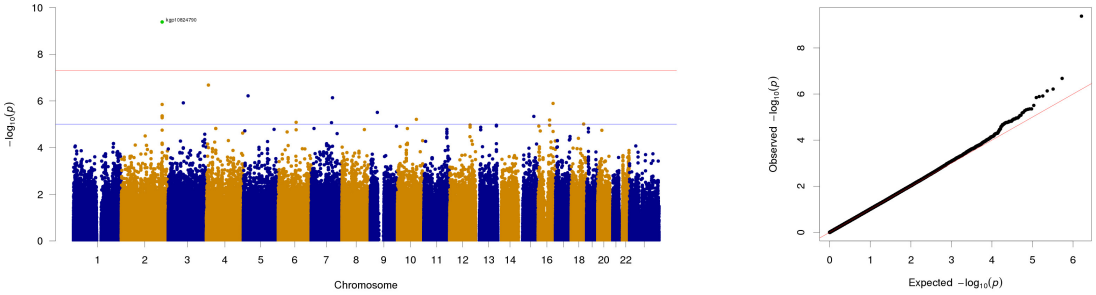


Fig. A.37 Visualization of the GWAS results of Gly.

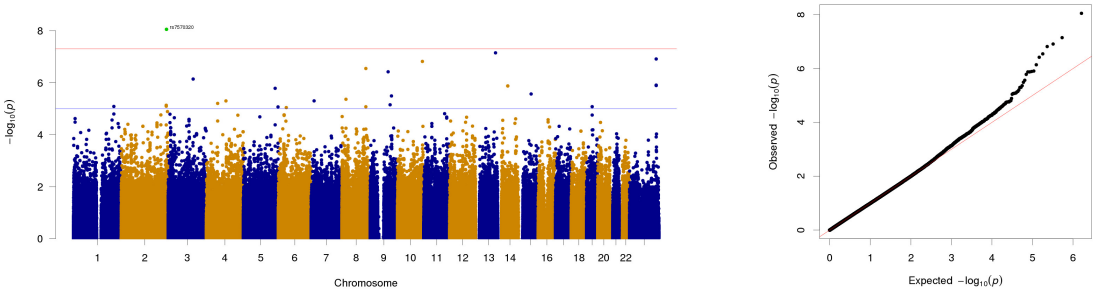


Fig. A.38 Visualization of the GWAS results of AcAce.

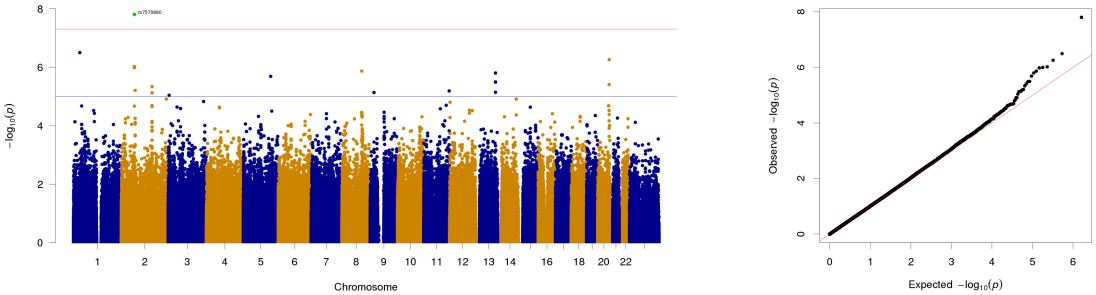


Fig. A.39 Visualization of the GWAS results of IDL.C.eFR.

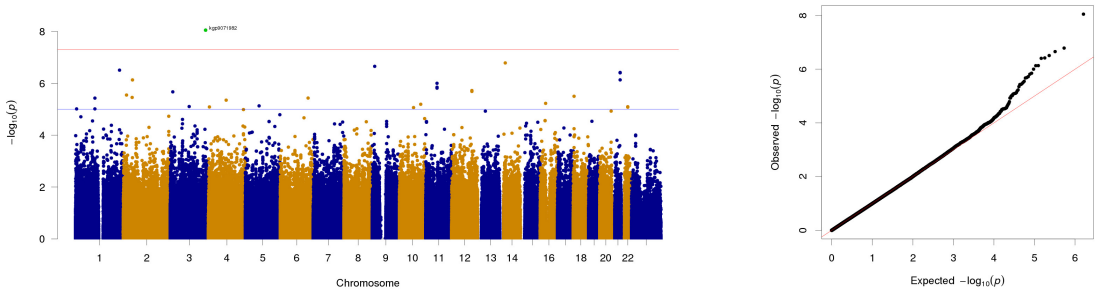


Fig. A.40 Visualization of the GWAS results of XL.VLDL.PL.

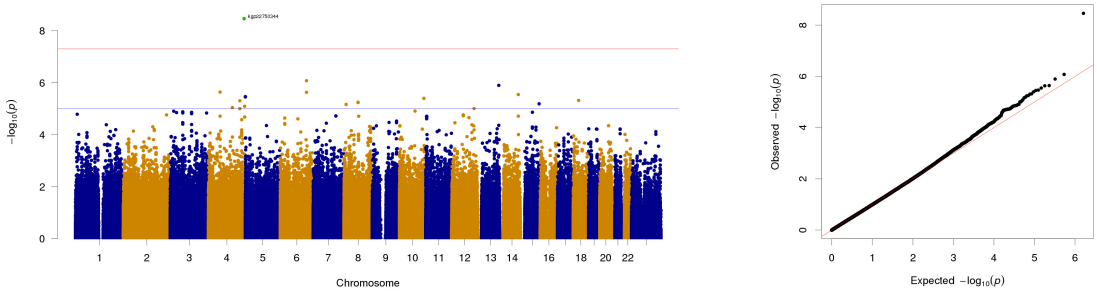


Fig. A.41 Visualization of the GWAS results of Gln.

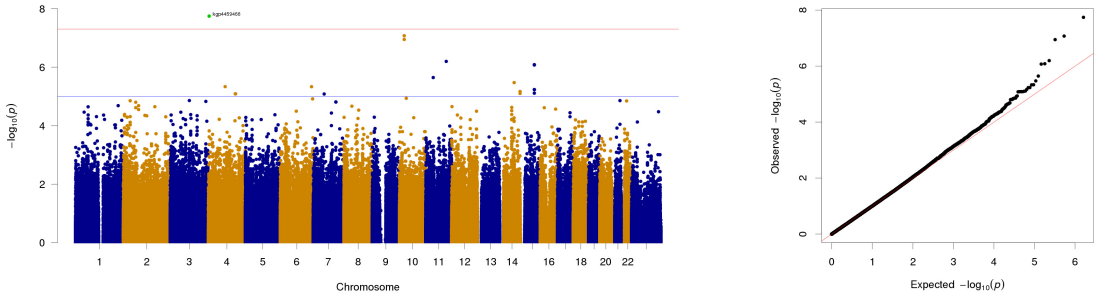


Fig. A.42 Visualization of the GWAS results of DB.in.FA.

References

Adams, M. A., Freeman, B. J., Morrison, H. P., Nelson, I. W., and Dolan, P. (2000). "Mechanical initiation of intervertebral disc degeneration". In: *Spine* 25.13, pp. 1625–1636.

Adams, M. A. and Roughley, P. J. (2006). "What is intervertebral disc degeneration, and what causes it?" In: *Spine* 31.18, pp. 2151–2161.

Adiels, M. et al. (2006). "Overproduction of large VLDL particles is driven by increased liver fat content in man". In: *Diabetologia* 49.4, pp. 755–765.

Adosraku, R., Choi, G., Constantinou-Kokotos, V., Anderson, M., and Gibbons, W. (1994). "NMR lipid profiles of cells, tissues, and body fluids: proton NMR analysis of human erythrocyte lipids." In: *Journal of lipid Research* 35.11, pp. 1925–1931.

Aittokallio, T. and Schwikowski, B. (2006). "Graph-based methods for analysing networks in cell biology". In: *Briefings in bioinformatics* 7.3, pp. 243–255.

Akaike, H. (2011). "Akaike's information criterion". In: *International encyclopedia of statistical science*. Springer, pp. 25–25.

Ala-Korpela, M. (2008). "Critical evaluation of 1H NMR metabonomics of serum as a methodology for disease risk assessment and diagnostics". In: *Clinical Chemistry and Laboratory Medicine* 46.1, pp. 27–42.

Altman, D. G. and Royston, P. (2006). "The cost of dichotomising continuous variables". In: *Bmj* 332.7549, p. 1080.

Arkin, A., Ross, J., and McAdams, H. H. (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells". In: *Genetics* 149.4, pp. 1633–1648.

Arkin, M. R. and Wells, J. A. (2004). "Small-molecule inhibitors of protein-protein interactions: progressing towards the dream". In: *Nature reviews Drug discovery* 3.4, p. 301.

Baenke, F., Peck, B., Miess, H., and Schulze, A. (2013). "Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development". In: *Disease models & mechanisms* 6.6, pp. 1353–1363.

Barabasi, A.-L. and Oltvai, Z. N. (2004). "Network biology: understanding the cell's functional organization". In: *Nature reviews genetics* 5.2, p. 101.

Barbeira, A. et al. (2016). "MetaXcan: summary statistics based gene-level association method infers accurate PrediXcan results". In: *bioRxiv*, p. 045260.

Barker, M. and Rayens, W. (2003). "Partial least squares for discrimination". In: *Journal of chemometrics* 17.3, pp. 166–173.

Barsh, G. S., Copenhaver, G. P., Gibson, G., and Williams, S. M. (2012). "Guidelines for genome-wide association studies". In: *PLoS genetics* 8.7, e1002812.

Battié, M. C., Videman, T., and Parent, E. (2004). "Lumbar disc degeneration: epidemiology and genetic influences". In: *Spine* 29.23, pp. 2679–2690.

Battié, M. C. et al. (1991). "1991 Volvo Award in clinical sciences. Smoking and lumbar intervertebral disc degeneration: an MRI study of identical twins." In: *Spine* 16.9, pp. 1015–1021.

Battié, M. C. et al. (2009). "The Twin Spine Study: contributions to a changing view of disc degeneration". In: *The Spine Journal* 9.1, pp. 47–59.

Beard, C. M., Barnard, R. J., Robbins, D. C., Ordovas, J. M., and Schaefer, E. J. (1996). "Effects of diet and exercise on qualitative and quantitative measures of LDL and its susceptibility to oxidation". In: *Arteriosclerosis, thrombosis, and vascular biology* 16.2, pp. 201–207.

Beckonert, O., Monnerjahn, J., Bonk, U., and Leibfritz, D. (2003). "Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps". In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 16.1, pp. 1–11.

Beckonert, O. et al. (2007). "Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts". In: *Nature protocols* 2.11, p. 2692.

Benjamini, Y. and Hochberg, Y. (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics". In: *Journal of educational and Behavioral Statistics* 25.1, pp. 60–83.

Berisa, T. and Pickrell, J. K. (2016). "Approximately independent linkage disequilibrium blocks in human populations". In: *Bioinformatics* 32.2, p. 283.

Berlemann, U., Gries, N., and Moore, R. (1998). "The relationship between height, shape and histological changes in early degeneration of the lower lumbar discs". In: *European Spine Journal* 7.3, pp. 212–217.

Bertoli, A. et al. (2003). “Lipid profile, BMI, body fat distribution, and aerobic fitness in men with metabolic syndrome”. In: *Acta diabetologica* 40.1, s130–s133.

Blekhman, R. et al. (2014). “Comparative metabolomics in primates reveals the effects of diet and gene regulatory variation on metabolic divergence”. In: *Scientific reports* 4, p. 5809.

Borowitz, D. et al. (2009). “Cystic Fibrosis Foundation practice guidelines for the management of infants with cystic fibrosis transmembrane conductance regulator-related metabolic syndrome during the first two years of life and beyond”. In: *The Journal of pediatrics* 155.6, S106–S116.

Boulton, A. A., Pollitt, R., and Majer, J. (1967). “Identity of a urinary “pink spot” in schizophrenia and Parkinson’s disease”. In: *Nature* 215.5097, p. 132.

Brazma, A. and Vilo, J. (2000). “Gene expression data analysis”. In: *FEBS letters* 480.1, pp. 17–24.

Broad Institute (2018). *What is mass spectrometry?* Retrieved 8-Mar-2018, from https://www.broadinstitute.org/proteomics/what-mass-spectrometry.

Broadhurst, D. I. and Kell, D. B. (2006). “Statistical strategies for avoiding false discoveries in metabolomics and related experiments”. In: *Metabolomics* 2.4, pp. 171–196.

Bush, W. S. and Moore, J. H. (2012). “Genome-wide association studies”. In: *PLoS computational biology* 8.12, e1002822.

Cambiaghi, A., Ferrario, M., and Masseroli, M. (2016). “Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration”. In: *Briefings in bioinformatics* 18.3, pp. 498–510.

Campos, G. de los, Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). “Prediction of complex human traits using the genomic best linear unbiased predictor”. In: *PLoS genetics* 9.7, e1003608.

Carey, N. (2015). *Junk DNA: A journey through the dark matter of the genome*. Columbia University Press.

Castle, A. L., Fiehn, O., Kaddurah-Daouk, R., and Lindon, J. C. (2006). “Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results”. In: *Briefings in Bioinformatics* 7.2, pp. 159–165.

Cavill, R., Jennen, D., Kleinjans, J., and Briedé, J. J. (2015). “Transcriptomic and metabolomic data integration”. In: *Briefings in bioinformatics* 17.5, pp. 891–901.

Cedars-Sinai (2018). *Lumbar spine*. Retrieved 1-May-2018, from https://www.cedars-sinai.org/health-library/diseases-and-conditions/l/lumbar-spine.html.

Cham, J. (2000). *Posture Back Cracking*. Retrieved 26-Aug-2018, from http://phdcomics.com/comics/archive.php?comicid=170.

Chen, Y.-K., Chen, C.-Y., Hu, H.-T., and Hsueh, Y.-P. (2012). "CTTNBP2, but not CTTNBP2NL, regulates dendritic spinogenesis and synaptic distribution of the striatin–PP2A complex". In: *Molecular biology of the cell* 23.22, pp. 4383–4392.

Chen, T. et al. (2013). "Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection". In: *Evidence-Based Complementary and Alternative Medicine* 2013.

Cheung, K. M. et al. (2009). "Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals". In: *Spine* 34.9, pp. 934–940.

Cho, M., Kwak, S., Schmid, C., and Ponce, J. (2015). "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1201–1210.

Chu, Y. and Corey, D. R. (2012). "RNA sequencing: platform selection, experimental design, and data interpretation". In: *Nucleic acid therapeutics* 22.4, pp. 271–274.

Colhoun, H. M. et al. (2002). "Lipoprotein subclasses and particle sizes and their relationship with coronary artery calcification in men and women with and without type 1 diabetes". In: *Diabetes* 51.6, pp. 1949–1956.

Collins, F. S. et al. (1987). "Construction of a general human chromosome jumping library, with application to cystic fibrosis". In: *Science* 235.4792, pp. 1046–1049.

Collins, R. L. et al. (2013). "Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis." In: *BioData mining* 6.4.

Conesa, A. et al. (2016). "A survey of best practices for RNA-seq data analysis". In: *Genome biology* 17.1, p. 13.

Consortium, I. H. et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, p. 851.

Cooper, D. N. (2010). *Functional intronic polymorphisms: buried treasure awaiting discovery within our genes*.

Cortes, C. and Vapnik, V. (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Crick, F. (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, p. 561.

De Roos, A., Kressel, H., Spritzer, C., and Dalinka, M. (1987). "MR imaging of marrow changes adjacent to end plates in degenerative lumbar disk disease". In: *American Journal of Roentgenology* 149.3, pp. 531–534.

Diagrams for all (2018). *Sciatic nerve location*. Retrieved 10-Apr-2018, from http://diagramiac.com/sciatic-nerve-location.

Dictionary, O. E. (2004). "Oxford English dictionary online". In: *Mount Royal College Lib., Calgary* 14.

Dietterich, T. G. (1997). "Machine-learning research". In: *AI magazine* 18.4, p. 97.

Donoho, D. L. et al. (2000). "High-dimensional data analysis: The curses and blessings of dimensionality". In: *AMS math challenges lecture* 1.2000, p. 32.

Drew, B. G., Rye, K.-A., Duffy, S. J., Barter, P., and Kingwell, B. A. (2012). "The emerging role of HDL in glucose metabolism". In: *Nature Reviews Endocrinology* 8.4, p. 237.

Dudbridge, F. (2013). "Power and predictive accuracy of polygenic risk scores". In: *PLoS Genet* 9.3, e1003348.

Dudli, S., Fields, A. J., Samartzis, D., Karppinen, J., and Lotz, J. C. (2016). "Pathobiology of Modic changes". In: *European Spine Journal* 25.11, pp. 3723–3734.

Dündar, F., Skrabanek, L., and Zumbo, P. (2015). "Introduction to differential gene expression analysis using RNA-seq". In: *Appl. Bioinformatics*, pp. 1–67.

Eeles, R. A. et al. (2013). "Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array". In: *Nature genetics* 45.4, p. 385.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). "Empirical Bayes analysis of a microarray experiment". In: *Journal of the American statistical association* 96.456, pp. 1151–1160.

Eskola, P. J. et al. (2012). "Genetic association studies in lumbar disc degeneration: a systematic review". In: *PLos one* 7.11, e49995.

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2014). "PRSice: Polygenic Risk Score software". In: *Bioinformatics*, btu848.

Evangelou, E. and Ioannidis, J. P. (2013). "Meta-analysis methods for genome-wide association studies and beyond". In: *Nature Reviews Genetics* 14.6, p. 379.

Fairbank, J. C. and Pynsent, P. B. (2000). "The Oswestry disability index". In: *Spine* 25.22, pp. 2940–2953.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). "Electrospray ionization for mass spectrometry of large biomolecules". In: *Science* 246.4926, pp. 64–71.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). "Structural variation in the human genome". In: *Nature Reviews Genetics* 7.2, p. 85.

Folch, J., Lees, M., Sloane Stanley, G., et al. (1957). "A simple method for the isolation and purification of total lipids from animal tissues". In: *J biol Chem* 226.1, pp. 497–509.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

Frobin, W., Brinckmann, P., Kramer, M., and Hartwig, E. (2001). "Height of lumbar discs measured from radiographs compared with degeneration and height classified from MR images". In: *European radiology* 11.2, pp. 263–269.

Gamazon, E. R. et al. (2015). "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature genetics* 47.9, p. 1091.

Gao, X. et al. (2016). "Acetate functions as an epigenetic metabolite to promote lipid synthesis under hypoxia". In: *Nature communications* 7, p. 11960.

Genome Research Limited (2016). *What is the 'Central Dogma'?* Retrieved 5-Apr-2018, from https://www.yourgenome.org/facts/what-is-the-central-dogma.

Gibson, G. (2012). "Rare and common variants: twenty arguments". In: *Nature Reviews Genetics* 13.2, p. 135.

Gieger, C. et al. (2008). "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum". In: *PLoS genetics* 4.11, e1000282.

Glaser, C., Heinrich, J., and Koletzko, B. (2010). "Role of FADS1 and FADS2 polymorphisms in polyunsaturated fatty acid metabolism". In: *Metabolism* 59.7, pp. 993–999.

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). "False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies". In: *Journal of clinical epidemiology* 67.8, pp. 850–857.

Goel, V., Monroe, B., Gilbertson, L., and Brinckmann, P. (1995). "Interlaminar Shear Stresses and Laminae Separation in a Disc: Finite Element Analysis of the L3-L4 Motion Segment Subjected to Axial Compressive Loads." In: *Spine* 20.6, pp. 689–698.

Goldstein, D. B. et al. (2009). "Common genetic variation and human traits". In: *New England Journal of Medicine* 360.17, p. 1696.

Gromovsky, A. D. et al. (2018). "Δ-5 fatty acid desaturase FADS1 impacts metabolic disease by balancing proinflammatory and proresolving lipid mediators". In: *Arteriosclerosis, thrombosis, and vascular biology* 38.1, pp. 218–231.

Hamanishi, C., Kawabata, T., Yosii, T., and Tanaka, S. (1994). "Schmorl's nodes on magnetic resonance imaging. Their incidence and clinical relevance." In: *Spine* 19.4, pp. 450–453.

Hamel, L. (2016). "Som quality measures: An efficient statistical approach". In: *Advances in Self-Organizing Maps and Learning Vector Quantization*. Springer, pp. 49–59.

Hamel, L. and Ott, B. (2012). "A population based convergence criterion for self-organizing maps". In: *Proceedings of the 2012 International Conference on Data Mining, Las Vegas, Nevada (July 2012) Google Scholar*.

Hardy, G. H. et al. (1908). "Mendelian proportions in a mixed population". In: *Science* 28.706, pp. 49–50.

Hartigan, J. A. (1975). "Clustering algorithms". In:

Hasin, Y., Seldin, M., and Lusis, A. (2017). "Multi-omics approaches to disease". In: *Genome biology* 18.1, p. 83.

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017). "Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso". In: *arXiv preprint arXiv:1707.08692*.

Hebbar, P. et al. (2018). "Genome-wide association study identifies novel recessive genetic variants for high TGs in an Arab population". In: *Journal of Lipid Research*, jlr–P080218.

Hill, D. P. et al. (2004). "The mouse Gene Expression Database (GXD): updates and enhancements". In: *Nucleic acids research* 32.suppl_1, pp. D568–D571.

Hochberg, Y. and Benjamini, Y. (1990). "More powerful procedures for multiple significance testing". In: *Statistics in medicine* 9.7, pp. 811–818.

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1, pp. 55–67.

Holm, S., Holm, A. K., Ekström, L., Karladani, A., and Hansson, T. (2004). "Experimental disc degeneration due to endplate injury". In: *Clinical Spine Surgery* 17.1, pp. 64–71.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). "Identifying causal variants at loci with multiple signals of association". In: *Genetics*, genetics–114.

Horner, H. A. and Urban, J. P. (2001). "2001 Volvo Award Winner in Basic Science Studies: effect of nutrient supply on the viability of cells from the nucleus pulposus of the intervertebral disc". In: *Spine* 26.23, pp. 2543–2549.

Hoy, D. et al. (2012). "A systematic review of the global prevalence of low back pain". In: *Arthritis & Rheumatology* 64.6, pp. 2028–2037.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). "False discovery rate control with groups". In: *Journal of the American Statistical Association* 105.491, pp. 1215–1227.

Hummel, J. (1996). "Linked bar charts: Analysing categorical data graphically". In: *Computational Statistics* 11.1, pp. 23–33.

Illig, T. et al. (2010). "A genome-wide perspective of genetic variation in human metabolism". In: *Nature genetics* 42.2, p. 137.

Illumina (2016). *Data sheet of Infinium OmniZhongHua-8 v1.3 BeadChip*. Retrieved 20-Apr-2018, from https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-omnizhonghua.pdf.

International HapMap Consortium (2003). "The international HapMap project". In: *Nature* 426.6968, p. 789.

Jamaludin, A., Kadir, T., and Zisserman, A. (2017a). "Self-supervised learning for spinal MRIs". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, pp. 294–302.

– (2017b). "SpineNet: Automated classification and evidence visualization in spinal MRIs". In: *Medical image analysis* 41, pp. 63–73.

James, A. T. and Martin, u. A. (1952). "Gas-liquid partition chromatography: the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid". In: *Biochemical Journal* 50.5, p. 679.

Janssens, R., Zeng, G., and Zheng, G. (2018). "Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks". In: *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, pp. 893–897.

Jensen, O. K., Nielsen, C. V., Sørensen, J. S., and Stengaard-Pedersen, K. (2014). "Type 1 modic changes was a significant risk factor for 1-year outcome in sick-listed low back pain patients: a nested cohort study using magnetic resonance imaging of the lumbar spine". In: *The Spine Journal* 14.11, pp. 2568–2581.

Jensen, T. S. et al. (2010). "Predictors of new vertebral endplate signal (Modic) changes in the general population". In: *European Spine Journal* 19.1, pp. 129–135.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

Jungersen, K. (2004). "The relation between text and colours in medieval urine wheels". In: *The relation between text and colours in medieval urine wheels*.

Kanai, M. et al. (2016). "Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set". In: *Journal of human genetics* 61.10, p. 861.

Kaplan, W. et al. (2013). *Priority medicines for Europe and the world – 2013 update*. World Health Organization.

Kauffman, S. A. (1969). "Metabolic stability and epigenesis in randomly constructed genetic nets". In: *Journal of theoretical biology* 22.3, pp. 437–467.

Kauppila, L. (2009). "Atherosclerosis and disc degeneration/low-back pain–a systematic review". In: *European journal of vascular and endovascular surgery* 37.6, pp. 661–670.

Kendall, M. G. (1938). “A new measure of rank correlation”. In: *Biometrika* 30.1/2, pp. 81–93.

Kenny, L. C. et al. (2010). “Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers”. In: *Hypertension* 56.4, pp. 741–749.

Keshari, K. R. et al. (2008). “Lactic acid and proteoglycans as metabolic markers for discogenic back pain”. In: *Spine* 33.3, pp. 312–317.

Kettunen, J. et al. (2016). “Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA”. In: *Nature communications* 7, p. 11122.

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). “Predicting disease risks from highly imbalanced data using random forest”. In: *BMC medical informatics and decision making* 11.1, p. 51.

Kichaev, G. et al. (2014). “Integrating functional data to prioritize causal variants in statistical fine-mapping studies”. In: *PLoS genetics* 10.10, e1004722.

Kiviluoto, K. (1996). “Topology preservation in self-organizing maps”. In: *Neural Networks, 1996., IEEE International Conference on*. Vol. 1. IEEE, pp. 294–299.

Kjaer, P., Leboeuf-Yde, C., Korsholm, L., Sorensen, J. S., and Bendix, T. (2005). “Magnetic resonance imaging and low back pain in adults: a diagnostic imaging study of 40-year-old men and women”. In: *Spine* 30.10, pp. 1173–1180.

Klug, W. S., Cummings, M. R., et al. (2003). *Concepts of genetics.* Ed. 7. Pearson Education, Inc.

Knox, J. H. et al. (1978). *High-performance liquid chromatography.* Edinburgh University Press.

Kohonen, T. (1998). “The self-organizing map”. In: *Neurocomputing* 21.1-3, pp. 1–6.

Kolmogorov, A. (1933). “Sulla determinazione empirica di una lgge di distribuzione”. In: *Inst. Ital. Attuari, Giorn.* 4, pp. 83–91.

Kruglyak, L. and Nickerson, D. A. (2001). “Variation is the spice of life”. In: *Nature genetics* 27.3, p. 234.

Kukurba, K. R. and Montgomery, S. B. (2015). “RNA sequencing and analysis”. In: *Cold Spring Harbor protocols* 2015.11, pdb–top084970.

Kuska, B. (1998). *Beer, Bethesda, and biology: how “genomics” came into being*.

Langfelder, P., Zhang, B., and Horvath, S. (2007). “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R”. In: *Bioinformatics* 24.5, pp. 719–720.

Larmo, P. S. et al. (2013). “Effects of sea buckthorn and bilberry on serum metabolites differ according to baseline metabolic profiles in overweight women: a randomized crossover trial–”. In: *The American journal of clinical nutrition* 98.4, pp. 941–951.

Last, J. M., Abramson, J. H., and Freidman, G. D. (2001). *A dictionary of epidemiology*. Vol. 4. Oxford University Press New York.

Lattka, E., Illig, T., Koletzko, B., and Heinrich, J. (2010). "Genetic variants of the FADS1 FADS2 gene cluster as related to essential fatty acid metabolism". In: *Current opinion in lipidology* 21.1, pp. 64–69.

Lee, S. et al. (2018). "Dietary n-3 and n-6 polyunsaturated fatty acids, the FADS gene, and the risk of gastric cancer in a Korean population". In: *Scientific reports* 8.1, p. 3823.

Leino-Arjas, P. et al. (2008). "Serum lipids in relation to sciatica among Finns". In: *Atherosclerosis* 197.1, pp. 43–49.

Leotraining (2016). *Low back pain in rowers: part 2*. Retrieved 9-Apr-2018, from https://www.rowperfect.co.uk/low-back-pain-rowers-part-2/.

Li, P. et al. (2018). "A regulatory insertion-deletion polymorphism in the FADS gene cluster influences PUFA and lipid profiles among Chinese adults: a population-based study". In: *The American journal of clinical nutrition* 107.6, pp. 867–875.

Li, Y. (2016). "Identification of Common and Rare Genetic Risk Factors for Lumbar Disc Degeneration". PhD thesis. The University of Hong Kong.

Li, Y. et al. (2016). "Two subtypes of intervertebral disc degeneration distinguished by large-scale population-based study". In: *The Spine Journal* 16.9, pp. 1079–1089.

Lindon, J. C., Holmes, E., and Nicholson, J. K. (2006). "Metabonomics techniques and applications to pharmaceutical research & development". In: *Pharmaceutical research* 23.6, pp. 1075–1088.

Liu, Y. et al. (2014). "Microarray-based identification of nerve growth-promoting genes in neurofibromatosis type I". In: *Molecular medicine reports* 9.1, pp. 192–196.

Longo, U. G. et al. (2011). "Symptomatic disc herniation and serum lipid levels". In: *European Spine Journal* 20.10, pp. 1658–1662.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). "Transcriptomics technologies". In: *PLoS computational biology* 13.5, e1005457.

Lumen Learning (2007). *Biology for majors II: Human axial skeleton*. Retrieved 9-Apr-2018, from https://courses.lumenlearning.com/wm-biology2/chapter/human-axial-skeleton/.

Luoma, K., Vehmas, T., Riihimäki, H., and Raininko, R. (2001). "Disc height and signal intensity of the nucleus pulposus on magnetic resonance imaging as indicators of lumbar disc degeneration". In: *Spine* 26.6, pp. 680–686.

Luoma, K. et al. (2000). "Low back pain in relation to lumbar disc degeneration". In: *Spine* 25.4, pp. 487–492.

Ma, S., Kemmeren, P., Gresham, D., and Statnikov, A. (2014). "De-novo learning of genome-scale regulatory networks in S. cerevisiae". In: *Plos one* 9.9, e106479.

Määttä, J. H. et al. (2016). "Refined phenotyping of modic changes: imaging biomarkers of prolonged severe low back pain and disability". In: *Medicine* 95.22.

Mahadevan, S., Shah, S. L., Marrie, T. J., and Slupsky, C. M. (2008). "Analysis of metabolomic data using support vector machines". In: *Analytical Chemistry* 80.19, pp. 7562–7570.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). "Polygenic scores via penalized regression on summary statistics". In: *Genetic epidemiology* 41.6, pp. 469–480.

Mäkinen, V.-P. et al. (2008). "1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death". In: *Molecular systems biology* 4.1, p. 167.

Malovini, A., Barbarini, N., Bellazzi, R., and De Michelis, F. (2012). "Hierarchical Naïve Bayes for genetic association studies". In: *BMC bioinformatics* 13.Suppl 14, S6.

Manolio, T. A. et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, p. 747.

Manz, A. et al. (1992). "Planar chips technology for miniaturization and integration of separation techniques into monitoring systems: capillary electrophoresis on a chip". In: *Journal of Chromatography A* 593.1-2, pp. 253–258.

Manzoni, C. et al. (2016). "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences". In: *Briefings in bioinformatics*, bbw114.

McCarthy, M. I. and Hirschhorn, J. N. (2008). "Genome-wide association studies: potential next steps on a genetic journey". In: *Human molecular genetics* 17.R2, R156–R165.

McCauley, J. L. et al. (2007). "SNPs in Multi-species Conserved Sequences (MCS) as useful markers in association studies: a practical approach". In: *BMC genomics* 8.1, p. 266.

McNaught, A. D. (1997). *Compendium of chemical terminology*. Vol. 1669. Blackwell Science Oxford.

Meyer, D., Zeileis, A., and Hornik, K. (2017). *vcd: Visualizing Categorical Data*. R package version 1.4-4.

Michailidou, K. et al. (2017). "Association analysis identifies 65 new breast cancer risk loci". In: *Nature* 551.7678, p. 92.

Miller, J. A., Schmatz, C., and Schultz, A. (1988). "Lumbar disc degeneration: correlation with age, sex, and spine level in 600 autopsy specimens." In: *Spine* 13.2, pp. 173–178.

Miller, W. M., Nori-Janosz, K. E., Lillystone, M., Yanez, J., and McCullough, P. A. (2005). "Obesity and lipids". In: *Current cardiology reports* 7.6, pp. 465–470.

Mills, R. E. et al. (2006). “An initial map of insertion and deletion (INDEL) variation in the human genome”. In: *Genome research* 16.9, pp. 1182–1190.

Modic, M., Steinberg, P., Ross, J., Masaryk, T., and Carter, J. (1988). “Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging.” In: *Radiology* 166.1, pp. 193–199.

Mok, F. P. et al. (2016). “Modic changes of the lumbar spine: prevalence, risk factors, and association with disc degeneration and low back pain in a large-scale population-based cohort”. In: *The Spine Journal* 16.1, pp. 32–41.

Mondul, A. M., Shui, I. M., Yu, K., et al. (2013). “Genetic variation in the vitamin D pathway in relation to risk of prostate cancer—results from the breast and prostate cancer cohort consortium”. In: *Cancer Epidemiology Biomarkers & Prevention* 22.4, pp. 688–696.

Nagy, R. et al. (2017). “Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants”. In: *Genome medicine* 9.1, p. 23.

National Human Genome Research Institute (2007). *A guide to your genome*. Retrieved 5-Mar-2018, from https://www.genome.gov/pages/education/allaboutthehumangenomeproject/guidetoyourgenome07_vs2.pdf.

National Institute of Neurological Disorders and Stroke (2017). *Low back pain fact sheet*. Retrieved 10-Apr-2018, from https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Low-Back-Pain-Fact-Sheet.

Newson, R. (2002). “Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences”. In:

Newton, F. (1994). “Constantine the African and Monte Cassino: New Elements and the Text of the Isagoge”. In: *Burnett and Jacquart*, pp. 16–47.

Nicholson, J. K. and Wilson, I. D. (2003). “Understanding’global’systems biology: metabonomics and the continuum of metabolism”. In: *Nature Reviews Drug Discovery* 2.8, p. 668.

Okser, S. et al. (2014). “Regularized machine learning in the genetic prediction of complex traits”. In: *PLoS genetics* 10.11, e1004754.

Olsson, U., Drasgow, F., and Dorans, N. J. (1982). “The polyserial correlation coefficient”. In: *Psychometrika* 47.3, pp. 337–347.

OpenStax (2013). *Anatomy and Physiology*. Retrieved 10-Apr-2018, from http://cnx.org/content/col11496/1.6/.

Osti, O., Vernon-Roberts, B., Moore, R., and Fraser, R. (1992). “Annular tears and disc degeneration in the lumbar spine. A post-mortem study of 135 discs”. In: *Bone & Joint Journal* 74.5, pp. 678–682.

Ozguler, A., Leclerc, A., Landre, M.-F., Pietri-Taleb, F., and Niedhammer, I. (2000). "Individual and occupational determinants of low back pain according to various definitions of low back pain". In: *Journal of Epidemiology & Community Health* 54.3, pp. 215–220.

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). "Innovation: Metabolomics: the apogee of the omics trilogy". In: *Nature reviews Molecular cell biology* 13.4, p. 263.

Pearson, T. A. and Manolio, T. A. (2008). "How to interpret a genome-wide association study". In: *Jama* 299.11, pp. 1335–1344.

Peng, B., Hou, S., Wu, W., Zhang, C., and Yang, Y. (2006). "The pathogenesis and clinical significance of a high-intensity zone (HIZ) of lumbar intervertebral disc on MR imaging in the patient with discogenic low back pain". In: *European Spine Journal* 15.5, pp. 583–587.

Pianka, E. (2007). *Symptoms of overpopulation: land*. Retrieved 3-Apr-2018, from http://www.zo.utexas.edu/courses/thoc/land.html.

Porter, H. F. and O'Reilly, P. F. (2017). "Multivariate simulation framework reveals performance of multi-trait GWAS methods". In: *Scientific reports* 7, p. 38837.

Price, A. L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8, p. 904.

Purcell, S. et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575.

Rabi, I. I., Millman, S., Kusch, P., and Zacharias, J. R. (1939). "The molecular beam resonance method for measuring nuclear magnetic moments". In: *Physical review* 55.6, p. 526.

Rajasekaran, S. et al. (2004). "ISSLS prize winner: a study of diffusion in human lumbar discs: a serial magnetic resonance imaging study documenting the influence of the endplate on diffusion in normal and degenerate discs". In: *Spine* 29.23, pp. 2654–2667.

Ranjani, R. V. et al. (2014). "Profiling of metabolites from human intervertebral disc through gas chromatography-Mass spectrometry". In: *Indian Journal of Science and Technology* 7.8, pp. 1228–1235.

Rasmussen, D., Ishizuka, B., Quigley, M., and Yen, S. (1983). "Effects of tyrosine and tryptophan ingestion on plasma catecholamine and 3, 4-dihydroxyphenylacetic acid concentrations". In: *The Journal of Clinical Endocrinology & Metabolism* 57.4, pp. 760–763.

Reich, D. E. and Lander, E. S. (2001). "On the allelic spectrum of human disease". In: *TRENDS in Genetics* 17.9, pp. 502–510.

Rhee, E. P. et al. (2013). "A genome-wide association study of the human metabolome in a community-based cohort". In: *Cell metabolism* 18.1, pp. 130–143.

Risch, N. and Merikangas, K. (1996). "The future of genetic studies of complex human diseases". In: *Science* 273.5281, pp. 1516–1517.

Rivas, M. A. et al. (2015). "Effect of predicted protein-truncating genetic variants on the human transcriptome". In: *Science* 348.6235, pp. 666–669.

Rozen, S. et al. (2005). "Metabolomic analysis and signatures in motor neuron disease". In: *Metabolomics* 1.2, pp. 101–108.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., and Hendriks, M. M. (2014). "Reflections on univariate and multivariate analysis of metabolomics data". In: *Metabolomics* 10.3, pp. 361–374.

Samartzis, D., Karppinen, J., Chan, D., Luk, K. D., and Cheung, K. (2012). "The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: A population-based study". In: *Arthritis & Rheumatology* 64.5, pp. 1488–1496.

Samartzis, D., Karppinen, J., Cheung, J. P. Y., and Lotz, J. (2013a). "Disk degeneration and low back pain: are they fat-related conditions?" In: *Global spine journal* 3.3, pp. 133–143.

Samartzis, D. et al. (2011). "A population-based study of juvenile disc degeneration and its association with overweight and obesity, low back pain, and diminished functional status". In: *JBJS* 93.7, pp. 662–670.

Samartzis, D. et al. (2013b). "SERUM METABOLOMIC BIOMARKERS AND LUMBAR DISC DEGENERATION: O16." In: *Spine Journal Meeting Abstracts*. LWW, pp. 11–12.

Sambrook, P., MacGregor, A., and Spector, T. (1999). "Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins". In: *Arthritis & Rheumatology* 42.2, pp. 366–372.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235, pp. 467–470.

Schneiderman, G. et al. (1987). "Magnetic resonance imaging in the diagnosis of disc degeneration: correlation with discography." In: *Spine* 12.3, pp. 276–281.

Schork, A. J. et al. (2013). "All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs". In: *PLoS genetics* 9.4, e1003449.

Schork, N. J. (1997). "Genetics of complex disease: approaches, problems, and solutions". In: *American journal of respiratory and critical care medicine* 156.4, S103–S109.

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). "Common vs. rare allele hypotheses for complex diseases". In: *Current opinion in genetics & development* 19.3, pp. 212–219.

Schrodi, S. J., Mukherjee, S., Shan, Y., et al. (2014). "Genetic-based prediction of disease traits: prediction is very difficult, especially about the future". In: *Frontiers in Genetics* 5.162. ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00162. URL: http://www.frontiersin.org/applied_genetic_epidemiology/10.3389/fgene.2014.00162/abstract.

Schwender, H., Krause, A., and Ickstadt, K. (2003). *Comparison of the empirical bayes and the significance analysis of microarrays*. Tech. rep. Technical Report, SFB 475: Komplexitätsreduktion in multivariaten, Datenstrukturen, TU Dortmund.

Science Learn Hub (2011). *DNA, chromosomes and gene expression*. Retrieved 5-Apr-2018, from https://www.sciencelearn.org.nz/resources/206-dna-chromosomes-and-gene-expression.

Shen, T. H., Carlson, C. S., and Tarczy-Hornoch, P. (2009). "SNPit: a federated data integration system for the purpose of functional SNP annotation". In: *Computer methods and programs in biomedicine* 95.2, pp. 181–189.

Shulaev, V. (2006). "Metabolomics technology and bioinformatics". In: *Briefings in bioinformatics* 7.2, pp. 128–139.

Siva, P., Russell, C., Xiang, T., and Agapito, L. (2013). "Looking beyond the image: Unsupervised learning for object saliency and detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3238–3245.

Smirnov, N. (1948). "Table for estimating the goodness of fit of empirical distributions". In: *The annals of mathematical statistics* 19.2, pp. 279–281.

Smolders, L. A. et al. (2013). "Gene expression profiling of early intervertebral disc degeneration reveals a down-regulation of canonical Wnt signaling and caveolin-1 expression: implications for development of regenerative strategies". In: *Arthritis research & therapy* 15.1, R23.

Soderberg, T. (2016). *Organic chemistry with a biological emphasis*. LibreTexts.

Speliotes, E. K., Willer, C. J., et al. (2010). "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index". In: *Nature genetics* 42.11, pp. 937–948.

Stelzer, G. et al. (2016). "The GeneCards suite: from gene data mining to disease genome sequence analyses". In: *Current protocols in bioinformatics* 54.1, pp. 1–30.

Storey, J. D. (2002). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3, pp. 479–498.

Sugimoto, M., Kawakami, M., Robert, M., Soga, T., and Tomita, M. (2012). "Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis". In: *Current bioinformatics* 7.1, pp. 96–108.

Suhre, K. and Gieger, C. (2012). "Genetic variation in metabolic phenotypes: study designs and applications". In: *Nature reviews genetics* 13.11, p. 759.

Sun, L., Craiu, R. V., Paterson, A. D., and Bull, S. B. (2006). "Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies". In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 30.6, pp. 519–530.

Szklo, M. (1998). "Population-based cohort studies". In: *Epidemiologic reviews* 20.1, pp. 81–90.

Szymańska, E., Saccenti, E., Smilde, A. K., and Westerhuis, J. A. (2012). "Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies". In: *Metabolomics* 8.1, pp. 3–16.

Tang, H., Wang, Y., Nicholson, J. K., and Lindon, J. C. (2004). "Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma". In: *Analytical biochemistry* 325.2, pp. 260–272.

Teraguchi, M. et al. (2014). "Prevalence and distribution of intervertebral disc degeneration over the entire spine in a population-based cohort: the Wakayama Spine Study". In: *Osteoarthritis and cartilage* 22.1, pp. 104–110.

Teraguchi, M. et al. (2016). "Classification of high intensity zones of the lumbar spine and their association with other spinal MRI phenotypes: the Wakayama spine study". In: *PloS one* 11.9, e0160111.

Tian, Y. et al. (2016). "FADS1-FADS2 gene cluster confers risk to polycystic ovary syndrome". In: *Scientific reports* 6, p. 21195.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Toyone, T. et al. (1994). "Vertebral bone-marrow changes in degenerative lumbar disc disease. An MRI study of 74 patients with low back pain". In: *The Journal of bone and joint surgery. British volume* 76.5, pp. 757–764.

Tsai, P.-C. et al. (2013). "Serum microRNA-21 and microRNA-221 as potential biomarkers for cerebrovascular disease". In: *Journal of vascular research* 50.4, pp. 346–354.

Tukiainen, T. et al. (2008). “A multi-metabolite analysis of serum by 1 H NMR spectroscopy: Early systemic signs of Alzheimer’s disease”. In: *Biochemical and biophysical research communications* 375.3, pp. 356–361.

Turner, S. D. (2014). “qqman: an R package for visualizing GWAS results using QQ and manhattan plots”. In: *BioRxiv*, p. 005165.

Tusher, V. G., Tibshirani, R., and Chu, G. (2008). *Significance analysis of microarrays*. US Patent 7,363,165.

U.S. National Library of Medicine (2018). *Urine – abnormal color*. Retrieved 2-Apr-2018, from https://medlineplus.gov/ency/article/003139.htm.

University of Toronto (2014). *Evolutionary biologists glimpse early stages of Y-chromosome degeneration*. Retrieved 5-May-2018, from https://phys.org/news/2014-05-evolutionary-biologists-glimpse-early-stages.html.

University of Washington (2018). *Cholesterol, Lipoproteins and the Liver*. Retrieved 10-Aug-2018, from https://courses.washington.edu/conj/bess/cholesterol/liver.html.

Urban, J. P., Smith, S., and Fairbank, J. C. (2004). “Nutrition of the intervertebral disc”. In: *Spine* 29.23, pp. 2700–2709.

Venter, J. C., Adams, M. D., et al. (2001). “The sequence of the human genome”. In: *science* 291.5507, pp. 1304–1351.

Vesanto, J. and Alhoniemi, E. (2000). “Clustering of the self-organizing map”. In: *IEEE Transactions on neural networks* 11.3, pp. 586–600.

Videman, T. et al. (2003). “Associations between back pain history and lumbar MRI findings”. In: *Spine* 28.6, pp. 582–588.

Vohradsky, J. (2001). “Neural model of the genetic network”. In: *Journal of Biological Chemistry* 276.39, pp. 36168–36173.

Vos, T. et al. (2012). “Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010”. In: *The lancet* 380.9859, pp. 2163–2196.

Wang, K., Li, M., and Hakonarson, H. (2010). “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. In: *Nucleic acids research* 38.16, e164–e164.

Wang, Y., Videman, T., and Battié, M. C. (2012). “Modic changes: prevalence, distribution patterns, and association with age in white men”. In: *The Spine Journal* 12.5, pp. 411–416.

Warren, S. T., Zhang, F., Licameli, G. R., and Peters, J. F. (1987). “The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites”. In: *Science* 237.4813, pp. 420–423.

Watanabe, K., Taskesen, E., Bochoven, A., and Posthuma, D. (2017). "Functional mapping and annotation of genetic associations with FUMA". In: *Nature communications* 8.1, p. 1826.

Wehrens, R. and Buydens, L. (2007). "Self- and Super-organising Maps in R: the kohonen package". In: *J. Stat. Softw.* 21.5. URL: http://www.jstatsoft.org/v21/i05.

Wei, T. and Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*. (Version 0.84). URL: https://github.com/taiyun/corrplot.

Wei, Z., Wang, K., Qu, H.-Q., et al. (2009). "From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes". In: *PLoS genetics* 5.10, e1000678.

Weinberg, W. (1908). "ber den Nachweis der Vererbung beim Menschen". In: *Jahres. Wiertt. Ver. Vaterl. Natkd.* 64, pp. 369–382.

Welter, D. et al. (2013). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic acids research* 42.D1, pp. D1001–D1006.

Westerhuis, J. A., Velzen, E. J. van, Hoefsloot, H. C., and Smilde, A. K. (2010). "Multivariate paired data analysis: multilevel PLSDA versus OPLSDA". In: *Metabolomics* 6.1, pp. 119–128.

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005). "A note on exact tests of Hardy-Weinberg equilibrium". In: *The American Journal of Human Genetics* 76.5, pp. 887–893.

Wijmenga, C. and Zhernakova, A. (2018). "The importance of cohort studies in the post-GWAS era". In: *Nature genetics*, p. 1.

Wilks, D. S. (2011). "Cluster analysis". In: *International geophysics*. Vol. 100. Elsevier, pp. 603–616.

Williams, F., Manek, N., Sambrook, P., Spector, T., and Macgregor, A. (2007). "Schmorl's nodes: common, highly heritable, and related to lumbar disc disease". In: *Arthritis Care & Research* 57.5, pp. 855–860.

Williams, R. J. (1956). "Biochemical individuality; the basis for the genetotrophic concept." In:

Wimmer, V., Lehermeier, C., Albrecht, T., et al. (2013). "Genome-wide prediction of traits with different genetic architecture through efficient variable selection". In: *Genetics* 195.2, pp. 573–587.

Wittek, P., Gao, S. C., Lim, I. S., and Zhao, L. (2013). "Somoclu: An efficient parallel library for self-organizing maps". In: *arXiv preprint arXiv:1305.1422*.

Worldometers (2008). *Current world population*. Retrieved 3-Apr-2018, from http://www.worldometers.info/world-population/.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., and Wishart, D. S. (2012). "MetaboAnalyst 2.0 —- a comprehensive server for metabolomic data analysis". In: *Nucleic acids research* 40.W1, W127–W133.

Xia, J. and Wishart, D. S. (2010a). "MetPA: a web-based metabolomics tool for pathway analysis and visualization". In: *Bioinformatics* 26.18, pp. 2342–2344.

– (2010b). "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data". In: *Nucleic acids research* 38.suppl_2, W71–W77.

Yale University (2018). *Visual analogue scale*. Retrieved 20-Apr-2018, from https://assessment-module.yale.edu/im-palliative/visual-analogue-scale.

Yang, Q., Khoury, M. J., Botto, L., et al. (2003). "Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes". In: *The American Journal of Human Genetics* 72.3, pp. 636–649.

Younes, M. et al. (2006). "Prevalence and risk factors of disk-related sciatica in an urban population in Tunisia". In: *Joint Bone Spine* 73.5, pp. 538–542.

Zenker, S., Rubin, J., and Clermont, G. (2007). "From inverse problems in mathematical physiology to quantitative differential diagnoses". In: *PLoS computational biology* 3.11, e204.

Zhang, A., Sun, H., Wang, P., Han, Y., and Wang, X. (2012). "Modern analytical techniques in metabolomics analysis". In: *Analyst* 137.2, pp. 293–300.

Zhang, Y. et al. (2016). "Serum lipid levels are positively correlated with lumbar disc herniation—a retrospective study of 790 Chinese patients". In: *Lipids in health and disease* 15.1, p. 80.