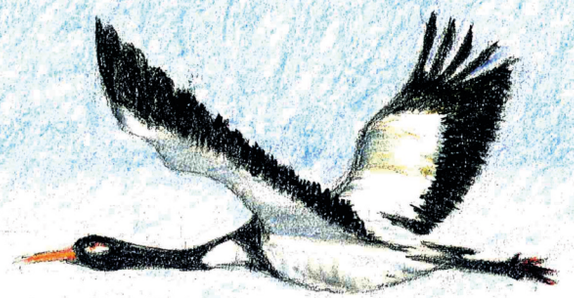
Studies with high dimensional omics data

Pathways, networks and statistical analysis

Zhang, X.



Studies with High Dimensional Omics Data Xiang Zhang

Studies with High Dimensional Omics Data

Pathways, Networks and Statistical Analysis

Xiang Zhang

Propositions

1. Mathematics is the logic of certainty; probability is the logic of uncertainty.
2. All models are wrong some are useful.
3. The choice of statistical models matters, no matter how big your data set is. (This thesis)
4. Statistical analysis has to be bespoke. (This thesis)
5. Prior knowledge can be translated into English, as well as databases and probability distributions. (This thesis)
6. We can do much better about missing data than just dropping them. (This thesis)
7. Want to understand a model? Run simulations! (This thesis)

Studies with High Dimensional Omics Data:

Pathways, Networks and Statistical Analysis

Xiang Zhang

Colophon

ISBN: 978-94-6375-411-8

Cove design by Jinyu Zhang

Layout design by Xiang Zhang

Print by Ridderprint BV, www.ridderprint.nl

This document was typeset in LaTeX

Studies with High Dimensional Omics Data:

Pathways, Networks and Statistical Analysis

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K. I. J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Aula der Universiteit

op woensdag 19 juni 2019, te 11.00 uur

door Xiang Zhang

geboren te Chengdu

Promotiecommissie

Promotores:	prof. dr. A.K. Groen	AMC-UvA
	prof. dr. A.H. Zwinderman	AMC-UvA
Overige leden:	prof. dr. A. Abu-Hanna	AMC-UvA
	prof. dr. I.C.W. Arts	Universiteit Maastricht
	prof. dr. N.A.W. van Riel	AMC-UvA
	dr. B.J.H. van den Born	AMC-UvA
	prof. dr. M. Nieuwdorp	AMC-UvA
	dr. ir. G.M. Dallinga-Thie	AMC-UvA

Faculteit der Geneeskunde

Contents

Acknowledgements

I would like to acknowledge all people who were involved with the work presented in this thesis. First and foremost, I would like to thank my promoters prof. A.K. Groen and prof. A.H. Zwinderman. I am truly grateful to both of them for their massive support and inspirational guidance over the years. Thank you Bert for giving me the chance to start my PhD in Groningen and later another opportunity to join AMC. Thank you for everything. Without your support I cannot have such luxury position and fully concentrate on my research. Thank you Koos for accepting me as one of your PhD student when I just moved out Groningen. I am truly grateful and enjoy to having so many fruitful discussions with you on statistics and science.

I want to thank the committee members, prof. A. Abu-Hanna, prof. I.C.W. Arts, prof. N.A.W. van Riel, dr. B.J.H. van den Born, prof. M. Nieuwdorp, and dr. G.M. Dallinga-Thie for their time and efforts assessing this Ph.D. dissertation.

I would also like to thank my colleagues in Groningen. In particular I want to thank prof. J.A. Kuivenhoven for his guidance during my days in Groningen. I also want to thank dr. Antoine Rimbert and Willem Balder for their dedication and help in the Lifelines women study.

I would like to thank dr. Adil Mardinoglu for introducing me everything about genome-scale metabolic modeling in early days of my Ph.D. program.

I want to thank prof. G.K. Hovingh for giving me the opportunity to participate in research projects in his group.

I would like to thank my AMC colleagues, Hilde Herrema, Mark Davids, Wil Kopatz,

Miranda de Boer Versloot, Hans Jansen, Hans van Beek, Stefan Havik, Jorge Peter, Han Levels, Alinda Schimmel, Maaike Winkelmeijer, Renate Hoogeveen, Lubna Ali, Jeffery Kroon, Mahnouch Bahjat, Kim Dzobo, Aldo Grefhorst, Torsten Scheithauer, Jan Schnitzier and Joao Belo Pereira for creating a cheerful working atmosphere. Special thanks to my paranymphs, Rens Reeskamp and Koen Wortelboer.

I would like to thank the Netherlands CardioVascular Research Initiative (CVON2011-2016; Acronym Genius), and the European Union (FP7 305707; Acronym RESOLVE) for funding this work.

I want to thank Federico Oldoni for his true friendship. When I look back my days in Groningen, I always feel grateful to have such a wonderful person as my colleague, and more importantly as my best friend.

My earnest and heartfelt gratitude goes to my families, especially my parents. Their support, encouragement and love are always there no matter where I live.

Last but not least, I want to thank my wife Sicong for her warm and ongoing support. Thank you for the joy you bring into my life. I am excited to start a new life with you and our upcoming daughter.

1 General introduction and outline of the thesis

1.1 High dimensional data

Biomedical research relies on data produced by either experimental or observational studies. In an experimental study, biological units such as subjects, animals or cells are sampled from a population of interest and assigned to receive treatment and control. Because in an experiment scientists can make the groups receiving different treatments comparable, they can evaluate causal effects of treatments on outcomes. However, some experiments are impractical or unethical. In such scenarios, scientists often use observational data to make causal inference of treatment effects on outcomes. In contrast to an experimental study, scientists cannot control differences between groups that receive different treatments in an observational study. As a result, scientists need to rely on more data than just treatments and outcomes and more complicated statistical models.

This thesis contains one experimental and three observational studies. Although these studies focus on different phenotypes such as inflammation, hypercholesterolemia and type 2 diabetes, they all rely on high-throughput technologies to measure many variables simultaneously from a single biological sample. Since every biological sample is now characterized by a large number of variables, this type of data is called high dimensional data. In particular, this dissertation will focus on three types of high dimensional data, namely transcriptomics, metabolomics and gut microbiome. I will first introduce each of them.

1.1.1 Transcriptomics

Transcriptomics uses high-throughput technologies, such as microarray and next generation sequencing, to identify and determine abundance of the entire collection of RNA molecules in biological samples. The shape of transcriptomics data is either a matrix of fluorescence signal values (microarray platforms) or a matrix of frequencies (RNA sequencing platforms). In both cases, every row of the matrix represents a transcript or gene, and every column represents a biological sample. The most common data analysis in transcriptomics studies is to identify genes (or transcripts) that are differentially expressed in response to different treatments. In order to identify biological pathways that respond to the treatments, gene expression data are often integrated with biological databases

such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2012). Recent developments of genome scale metabolic models (GEMs) enable us to predict metabolic *functions* based on gene expression data (Hoppe 2012). GEMs represent our current knowledge of all established metabolic reactions involved in human energy metabolism and macromolecule biosynthesis (Thiele et al. 2013; Mardinoglu et al. 2014). Both KEGG database and GEMs help us grouping genes that participate in the same metabolic pathway. GEMs can further help us grouping genes that regulate the same metabolite. Later in chapter 2, I will show an example in which gene expression data were integrated with both the KEGG database and with GEMs to study metabolic reprogramming in immune cells.

1.1.2 Metabolomics

Metabolomics uses mass spectrometry and NMR (nuclear magnetic resonance) spectroscopy to identify and quantify small molecules in cells, tissues or biological fluids (e.g. plasma and urine). The resulting metabolic profile is regarded as a snapshot of the metabolic state, and end product of both genetic and environmental factors. Similar to transcriptomics, the shape of metabolomics data is a matrix, in which every row is a metabolic feature and every column is a biological sample. Metabolomics has been successfully used to identify novel biomarkers for disease diagnostics and has improved our understanding of pathophysiologic mechanisms (Newgard 2017; Würtz et al. 2017). This dissertation will focus on metabolomics data generated by the Nightingale metabolomics platform (Nightingale Health, Finland). This platform quantifies 225 metabolic features including lipids, lipoproteins, fatty acids, amino acids, and glycolysis precursor molecules. A major challenge of metabolomics data analysis is how to functionally interpret metabolite measurements (Bartel, Krumsiek, and Theis 2013). One strategy is to integrate metabolomics with other omics data such as transcriptomics (Bartel et al. 2015a). In this thesis, we aimed to use plasma metabolomics to improve our understanding of hypercholesterolemia (chapter 3) and type 2 diabetes (chapter 4), by integrating metabolomics with genetic mutation and clinical data.

1.1.3 Gut microbiome

The human gut microbiome plays an important role in health and disease (Lynch and Pedersen 2016). Currently two major approaches, 16S ribosomal RNA (rRNA) gene amplicons and shotgun metagenomics, are used to profile the gut microbiome (Jovel et al. 2016). The difference between these two approaches is that 16S rRNA sequencing is restricted to bacteria and archaea whereas shotgun metagenomics measures all microorganisms present in a sample. In this dissertation, we will focus on microbiome data generated by 16S rRNA gene amplicons. 16S microbiome data is represented as a frequency matrix giving the number of times each microbe (every row) is observed in each sample (every column). In general microbiome data have the following features: 1) library sizes can vary by orders of magnitude across samples. 2) microbiome data often have excess zero counts. These zero counts can be due to either biological absence of a microbe, or insufficient sequencing. 3) microbiome data are compositional data, meaning that the obtained counts do not reflect the absolute number of microbes that are present. 4) microbiome data are often over-dispersed, characterized as some taxa (e.g., *Bacteroides* and *Lactobacillus* species) are common among samples, many other taxa are present at much lower abundances.

A particular interest of human gut microbiome research is to link nutrition intake and microbiome, because diet is modifiable and shapes the composition of human microbiota (Turnbaugh et al. 2009). For this purpose, associations of dietary intake and microbial abundance were evaluated in various cohort studies (Wu et al. 2011; Deschasaux et al. 2018). These association studies can improve our understanding of the relationships between the human microbiome and nutrient intake, as well as help development of new therapeutic interventions. In chapter 5, I will show a study in which we evaluated diet-microbe associations based on a large cohort microbiome data.

After introducing the high dimensional data, I would like to highlight some statistical strategies used in this thesis in the next section.

1.2 Statistical analysis strategies

In this dissertation, I applied different statistical analysis strategies. In chapter 2, I applied sample permutations for statistical inference. In chapter 3, I chose hierarchical clustering to group hypercholesterolemic females based on their plasma metabolic profiles. In chapter 4, I performed Bayesian data analysis in order to deal missing observations. In chapter 5, I carried out statistical simulations to compare different analysis methods. In following paragraphs, I will give a short description of permutation, quantification of similarity, Bayesian imputation and statistical simulation.

1.2.1 Permutation

I often ask my colleagues "what is your motivation of producing this data set?". Very likely the answer is "I want to know if there is any difference between groups.". For many people, hypothesis testing is the most common data analysis performed in biomedical research. What we care about is whether our observed difference is by chance or reflecting a true biological difference. My preferred strategy is to do permutation analysis. In permutation analysis, we randomly shuffle the labels of "control" and "treatment", and then calculate the difference between these randomly nested control and treatment groups. Any difference we observe after this shuffling is due to chance. If we repeat this process many times, we can count how many times we observe a difference as big as or even bigger than the one based on the original labeling. The corresponding proportion is called the (one sided) permutation P value. In my opinion, this strategy alleviates some anxiety on distribution assumptions built in many statistical tests. Sometimes permutation analysis can be computationally expensive and it is possible that we cannot enumerate all the permutations due to the limit of time and computational resources. In such case we take a sample of all permutations to perform the calculation. This is called a randomization approach.

1.2.2 Quantification of similarity

One advantage of high dimensional data is that it provides us a large number of variables to describe biological samples. Mathematically each biological sample is represented by

a point inside a high dimensional space. In that high dimensional space, we can quantify the distance between any two points by calculating the Euclidean or any other type of distance. This distance can then be used to quantify similarity between pairs of samples. Popular clustering algorithms, such as k-nearest neighbors and hierarchical clustering, are built upon this idea. In chapter 3, I will show an example in which I used this approach to study hypercholesterolemia with and without known genetic mutations.

1.2.3 Bayesian imputation

High dimensional data often contain missing values. For example, metabolomics can have missing values because the concentration of a metabolite is below the limit of detection, or because values were rejected by the automatic sample and measurement quality control procedure. Dropping observations with missing data is the default setting used by many data analysis programs. However, this is almost never appropriate because the dropped cases can bias the results. One of the ways to deal with missing data is by using a Bayesian approach. In the Bayesian framework, what we observed is called data, and what we did not observe is called parameters. Because missing data are not observed, they are treated as parameters in that analysis. All parameters in Bayesian models require a prior distribution to incorporate our prior information before seeing the data. As a result, Bayesian imputation does not give us one or a few imputed values but a whole posterior distribution for each missing observation.
In chapter 4, I will show an example in which I used Bayesian imputation to deal with missing values in both response variables and predictor variables.

1.2.4 Statistical simulation

Statistical simulation plays a unique role in data analysis. The uniqueness is that in a simulation study we know the truth. This is why when a new statistical method is developed, simulation studies are used to show that the method is working properly. One has two options to perform a statistical simulation. The first option is to construct a probabilistic model which mimics the data generating process. The data generating process describes how the data came to be, and it can be inspired by a biological theory or model. The next challenge is to translate the data generating process in the language of

probability. When we know little about the data generating process, we can alternatively generate the simulated data by repeated sampling with replacement from the data we have. This approach is sometimes called resampling. Like permutation, resampling can be sometimes computationally expensive. In chapter 5, I will show an example in which I applied both approaches to simulate human gut microbiome data.

1.3 Outline of this thesis

This dissertation contains four biomedical research projects in which we analyzed various type of high dimensional data, including transcriptomics of immune cells, human plasma metabolomics and human gut microbiome.

Our research examples started with a case-control study, which is a common study design in biomedical research. In chapter 2, we studied differential metabolic regulation in peripheral blood mononuclear cells (PBMCs) of healthy volunteers challenged by *Candida albicans*, *Borrelia burgdorferi*, lipopolysaccharide, and *Mycobacterium tuberculosis* in vitro. The gene expression data were generated by microarray. The goal of this study was to identify discriminating metabolic pathways and metabolites in human PBMCs stimulated by various pathogenic agents. To this end, we performed gene set enrichment analysis in the context of KEGG pathways and a human genome scale metabolic model. A genome-scale metabolic model represents a curated knowledgebase of all established metabolic reactions involved in human energy metabolism and macromolecule biosynthesis (O'Brien, Monk, and Palsson 2015). Our analysis generated a list of pathways and metabolites that can be used to discriminate PBMCs stimulated by *Candida albicans*, *Borrelia burgdorferi*, lipopolysaccharide and *Mycobacterium tuberculosis*.

If case-control studies are the most classical biomedical research topic in the wet lab, the cohort study might be the most common study in the dry lab. In chapter 3, we studied 119 females with high circulating cholesterol (also called hypercholesterolemia) in Lifelines, a large cohort study and biobank that includes a total of 167,729 individuals from the north of the Netherlands (Scholtens et al. 2015). The 119 females were selected if they were apparently healthy and had high circulating cholesterol levels. The data generated in this study were plasma metabolomics. Prior knowledge on hypercholesterolemia tells us that

the high plasma LDL cholesterol is often caused by genetic mutations in the LDL receptor (*LDLR*), apolipoprotein B (*APOB*), or proprotein convertase subtilisin/kexin type 9 (*PCSK9*). However, a substantial proportion of hypercholesterolemic subjects do not have any mutations in these canonical genes. Based on prior knowledge, we assumed that the cohort of 119 hypercholesterolemic females contained at least two subgroups. One of the subgroups should contain subjects carrying mutations in the canonical genes. To uncover the subgroups, we performed hierarchical clustering based on plasma metabolomics data, and identified four subtypes of hypercholesterolemia. Overlapping the clustering outcomes with true genetic information, we identified that subjects with mutations in *LDLR* or *APOB* preferentially clustered together, suggesting that patients with defects in the LDLR pathway show a distinctive metabolomics profile. We also identified that subjects without mutations in *LDLR* or *APOB* were characterized by two clusters, with or without elevated triglyceride concentration. In conclusion, we show the potential of using metabolomics to segregate hypercholesterolemic subjects into different clusters, which helps in targeting genetic analysis.

Chapters 1 and 2 showed examples that are completely driven by experimental data. Prior knowledge was introduced either by using the biological databases (chapter 2) or motivating the choice of clustering analysis (chapter 3). In chapter 4, we showed an example in which we introduced prior knowledge by specifying model structure. In this chapter, we were asking "why are subjects with African ethnic background more vulnerable to develop type 2 diabetes than subjects with an European ethnic background?". We profiled metabolomics of 773 subjects with European, Ghanaian or African Surinamese background. We then performed Bayesian lognormal regression analyses to assess associations between hemoglobin A1C (HbA1c) and plasma metabolites. In order to evaluate the effect of ethnicity, an interaction term between ethnicity and HbA1c was introduced into the lognormal regression model. Compared to European subjects, we found that subjects with Ghanaian and African Surinamese background had reversed associations between HbA1c and circulating acetoacetate and small HDLs. We hypothesized that these metabolic abnormalities may link to impaired cholesterol efflux capacity of HDL that may explain the excess type 2 diabetes in the subjects with African background.

Chapters 2-4 describe how data can be used to answer research questions as well as to generate novel hypothesis. Running statistical models is an essential step in this process.

Sometimes more than one method can be chosen for the purpose. So which methods are to be preferred? Surprisingly, this is not a trivial question. In fact application of particular methods seems to be based on the tradition of a particular research group, availability of experience with particular software, or dependent on the outcomes of the analysis. Statistical simulation is the preferred way to scientifically benchmark different methods. In chapter 5, based on the microbiome data of HELiUS (HEalthy Life in an Urban Setting), a large-scale prospective cohort study which included 25,000 inhabitants (18-70 years) from the city of Amsterdam, the Netherlands (K. Stronks et al. 2013a), we showed that depending on the choice of statistical methods, significant associations between microbe and nutrition intake varied dramatically. Our subsequent statistical simulations showed that no single analysis method was optimal. To achieve better control of false discovery rate, the best we can do is to run multiple analyses and focus on the significant findings identified by all methods.

A single type of omics profile can provide a useful glimpse, but cannot capture the entire biological complexity of most human diseases (Karczewski and Snyder 2018). Multi-omics integration is considered as the next step to achieve a holistic picture of human phenotypes and disease. In chapter 6, we discussed three systems biology platforms for multi-omics integration. These platforms include gene regulatory networks, protein-protein interaction networks and genome-scale metabolic modeling.

In chapter 7, we discuss the results of the various studies included in this dissertation and end with some future perspectives.

2

Identification of discriminating metabolic pathways and metabolites in human PBMCs stimulated by various pathogenic agents

This chapter has been published as Zhang X, Mardinoglu A, Joosten LAB, Kuivenhoven JA, Li Y, Netea MG and Groen AK (2018) Identification of Discriminating Metabolic Pathways and Metabolites in Human PBMCs Stimulated by Various Pathogenic Agents. Front. Physiol. 9:139. doi: 10.3389/fphys.2018.00139

2.1 Abstract

Immunity and cellular metabolism are tightly interconnected but it is not clear whether different pathogens elicit specific metabolic responses. To address this issue, we studied differential metabolic regulation in peripheral blood mononuclear cells (PBMCs) of healthy volunteers challenged by *Candida albicans*, *Borrelia burgdorferi*, lipopolysaccharide and *Mycobacterium tuberculosis* in vitro. By integrating gene expression data of stimulated PBMCs of healthy individuals with the KEGG pathways, we identified both common and pathogen-specific regulated pathways depending on the time of incubation. At 4 hour of incubation, pathogenic agents inhibited expression of genes involved in both the glycolysis and oxidative phosphorylation pathways. In contrast, at 24 hour of incubation, particularly glycolysis was enhanced while genes involved in oxidative phosphorylation remained unaltered in the PBMCs. In general, differential gene expression was less pronounced at 4 hour compared to 24 hour of incubation. KEGG pathway analysis allowed differentiation between effects induced by *Candida* and bacterial stimuli. Application of genome-scale metabolic model further generated a *Candida*-specific set of 103 reporter metabolites (e.g. desmosterol) that might serve as biomarkers discriminating *Candida*-stimulated PBMCs from bacteria-stimulated PBMCs. Our analysis also identified a set of 49 metabolites that allowed discrimination between the effects of *Borrelia burgdorferi*, lipopolysaccharide and *Mycobacterium tuberculosis*. We conclude that analysis of pathogen-induced effects on PBMCs by a combination of KEGG pathways and genome-scale metabolic model provides deep insight in the metabolic changes coupled to host defense.

Keywords: innate immunity, metabolism, peripheral blood mononuclear cell, *Candida albicans*, lipopolysaccharides, *Mycobacterium tuberculosis*, *Borrelia burgdorferi*, genome scale metabolic model

2.2 Introduction

As the first line of host defense, the innate immune system can immediately sense and combat foreign pathogens (McGettrick and O'Neill 2013; Mills and O'Neill 2014). Cells of the innate immune system, such as monocytes and neutrophils recognize pathogens via pat-

tern recognition receptors (PRRs) (McGettrick and O'Neill 2013; Mills and O'Neill 2014; Cheng, Joosten, and Netea 2014). These PRRs, such as Toll-like receptors, NOD-like receptors, C-type lectin receptors, and RigI-helicases, are found on the plasma membrane of innate immune cells (McGettrick and O'Neill 2013; Mills and O'Neill 2014; Cheng, Joosten, and Netea 2014). Activation of these PRRs leads to profound changes in gene expression and subsequent production of inflammatory mediators such as cytokines and chemokines (McGettrick and O'Neill 2013; Cheng, Joosten, and Netea 2014; Pearce and Pearce 2013). Once innate immune cells are activated, they can trigger responses of the adaptive immune system (e.g. activate T lymphocytes) (Mills and O'Neill 2014; Pearce et al. 2013).

Although often not realized, the responses of immune cells against pathogens are tightly linked to endogenous changes of metabolism (Mills and O'Neill 2014). It is known that upon activation, immune cells (e.g. monocytes and T lymphocytes) dramatically shift from oxidative phosphorylation to aerobic glycolysis, in order to meet the rapidly increasing energy demand by processes such as cytokine production and cell proliferation (McGettrick and O'Neill 2013; Cheng, Joosten, and Netea 2014; Pearce and Pearce 2013; Pearce et al. 2013). In addition, immune cells also increase the activity of the pentose phosphate pathway to provide sufficient nucleotide precursors for accelerated cell proliferation (e.g. T lymphocytes) (Mills and O'Neill 2014; Pearce et al. 2013). Also, in lipopolysaccharide (LPS) challenged macrophages, succinate and citrate accumulate to regulate production of IL-1 β (Tannahill et al. 2013). Thus far, however, metabolism of activated immune cells has been mainly investigated after challenges with LPS which only activates Toll-like receptor 4 (McGettrick and O'Neill 2013; Tannahill et al. 2013; Bordbar et al. 2012). A recent study on the modulation of glycolysis and oxidative phosphorylation in immune cells stimulated with LPS and other TLR stimuli supported the concept that different stimuli may induce various metabolic programs in immune cells (Lachmandas et al. 2016).

To our knowledge, a comprehensive understanding of the metabolism of immune cells after stimulation of various PRRs (e.g. TLRs, NOD-like receptors - NLRs, C-type lectin receptors - CLRs, and RigI-helicases) has not yet been reported. In the current study, we interrogate which metabolic pathways and metabolites are altered upon activation by various pathogens. To this end, we systematically measured gene expression profiles

in human PBMCs (peripheral blood mononuclear cells) stimulated by heat inactivated *Candida albicans* (*Candida*), *Borrelia burgdorferi* (*Borrelia*), *Escherichia coli*-derived LPS and *Mycobacterium tuberculosis* (MTB). These four are typical stimuli of innate immune pathways. LPS is the prototypical stimulus recognized by TLR 4 (Ngkelo et al. 2012). *Candida* is recognized by TLRs and CLRs, and causes mucosal and systematic infection in immunocompromised individuals (Mayer, Wilson, and Hube 2013). *Borrelia* is recognized by TLRs, NLRs, CLRs and RigI-helicases and causes Lyme disease (Oosting et al. 2016). MTB is recognized by TLRs, NLRs, and CLRs and causes tuberculosis (Kleinnijenhuis et al. 2011).

To identify gene expression changes involved in metabolism, we ran Kyoto Encyclopedia of Genes and Genomes (KEGG) based metabolic pathway analysis and genome-scale metabolic model (GEM) based reporter metabolite analysis, respectively. KEGG pathway analyses are widely and successfully used in biomedical research over the last decade as a routine step of interpreting gene expression data (Kanehisa et al. 2012). As an alternative, genome scale metabolic models (GEMs) are increasingly used to interpret large-scale gene expression data sets. GEMs are represented by networks in which the nodes are metabolites and the connecting edges are metabolic reactions (Mardinoglu, Gatto, and Nielsen 2013; Bordbar et al. 2014). Generic human GEMs, such as Recon2 (Thiele et al. 2013) and HMR2 (Mardinoglu et al. 2014) represent our current knowledge of all established metabolic reactions involved in human energy metabolism and macromolecule biosynthesis. GEMs have mostly been used to identify key enzymes and metabolites that may serve as potential biomarkers and drug targets for non-alcoholic fatty liver disease, obesity, Alzheimer's disease, and cancer (Mardinoglu et al. 2014; Mardinoglu et al. 2013; Lewis et al. 2010; Yizhak et al. 2014; Agren et al. 2014). Our analysis showed that KEGG pathway analysis allowed differentiation between effects induced by *Candida* and bacterial stimuli, and application of genome-scale metabolic model further generated a *Candida*-specific set of 103 reporter metabolites that might serve as biomarkers discriminating *Candida*-stimulated PBMCs from bacteria-stimulated PBMCs.

2.3 Material and Methods

2.3.1 Study populations

As described in the previous study (Smeekens et al. 2013), blood was collected after written informed consent from healthy volunteers. The study was approved by the Institutional Review Boards at Radboud University Nijmegen Medical Centre (RUNMC, Nijmegen, The Netherlands). The study was performed in accordance with the declaration of Helsinki. After informed consent was given, blood was collected by venipuncture into 10 ml EDTA syringes (Monoject, s-Hertogenbosch, The Netherlands).

2.3.2 Gene expression microarray data of stimulated PBMCs

As reported in in the previous study (Smeekens et al. 2013), we isolated PBMCs from healthy subjects by density centrifugation and stimulated them with heat-killed *C. albicans* (UC 820) (1×10^6 per ml), heat-killed *B. burgdorferi*, *E. coli*-derived LPS (10 ng per ml) or heat-killed MTB (1 μg per ml) respectively for 4 or 24 hours. PBMCs that were cultured in only RPMI medium were used as controls. Illumina Human HT-12 Expression BeadChips were used to measure gene expression levels at 4 and 24 hour. Details about the experiment and processed data are available in GSE42606 archived by Gene Expression Omnibus.

2.3.3 Identification of differentially expressed genes

The raw gene expression data were preprocessed by using the lumi R package with default settings, which includes background correction, variance stabilizing transformation and quantile normalization (Lin et al. 2008). Principal component analysis was performed with the full gene expression data set by using the function prcomp in R. Valid paired samples were selected to perform differential expression analysis at 4 and 24 hour separately. At 4 hour, the size of paired samples for each stimulation were 19 (*Candida*), 25 (*Borrelia*), 19 (LPS), and 18 (MTB). At 24 hour, the size of paired samples were 29 (*Candida*), 29 (*Borrelia*), 20 (LPS), and 30 (MTB). Illumina probe IDs were mapped to Ensembl gene IDs (Ensembl version 73) or Entrez gene IDs by using the lumiHumanIDMapping

and biomaRt R packages (Du et al. 2016; Durinck et al. 2009). To exclude the influence of ambiguous probes (a probe ID corresponding to two or more gene IDs), only the probes that have unique gene IDs were used for differential gene expression analysis. Moreover, the hidden batch effect originated from microarray analysis were adjusted by applying surrogate variable analysis which is built in the sva R package (Leek and Storey 2007; Leek and Storey 2008; Leek et al. 2012). Gene expression levels of stimulated PBMCs were then compared to controls by using linear models and empirical Bayes statistics (Smyth 2004). Both methods were implemented in the limma R package (Ritchie et al. 2015). Significance inference of differential expression was done with moderated t test (Ritchie et al. 2015) and the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was performed to calculate False Discovery Rate (FDR). In cases when a gene has multiple probes on the chip, the probe-level statistical test results were aggregated into a single gene-level statistic based on the smallest FDR.

2.3.4 Gene set enrichment analysis

In this study, the KEGG pathways and the generic human genome-scale metabolic model, HMR2 were used to analyze the gene expression data of human PBMCs stimulated by different pathogenic agents for 4 or 24 hours. The KEGG pathway information was downloaded from Molecular Signature Database v5.1 (Subramanian et al. 2005). There are in total 186 pathways and the related gene identifiers are Entrez gene IDs. Here we focused on 68 metabolic pathways since this study aims to identify metabolic signatures of stimulated human PBMCs. The HMR2 (SBML format) was downloaded from Human Metabolic Atlas (Pornputtapong, Nookaew, and Nielsen 2015). HMR2 contains 3,765 genes, 6,007 metabolites and 8,181 reactions (Mardinoglu et al. 2014). Essentially, KEGG pathway analysis and reporter metabolite analysis are two gene set enrichment analysis methods. The difference between them is that KEGG pathway analysis uses protein constituted pathways to group genes, whereas reporter metabolite analysis uses metabolites to define gene sets. Since every metabolite serves as a gene set in reporter metabolite analysis, the information of which genes belonged to which metabolite was attained through using the piano R package (Väremo, Nielsen, and Nookaew 2013a). The gene identifiers in HMR2 were annotated by Ensembl gene IDs (version 73). When

KEGG pathways were used as gene sets, we computed average t statistics of pathways as the summary statistics:

$$Z_{pathway} = \frac{\sum_{i=1}^{N_{pathway}} t_i}{\sqrt{N_{pathway}}} \tag{2.1}$$

This simple approach was first introduced by (Irizarry et al. 2009). $Z_{pathway}$ is the summary statistic of a pathway. $N_{pathway}$ is the number of genes in the pathway and t_i is the modified t statistics of gene i in the pathway. When metabolites of HMR2 were translated to gene sets, the original reporter metabolite algorithm (Patil and Nielsen 2005) was adapted to calculate summary statistics for metabolites. (Patil and Nielsen 2005) defined reporter metabolites of which the expression levels were significantly changed. In the original reporter metabolite algorithm (Patil and Nielsen 2005), the gene-level P values were first converted to Z scores by using the inverse normal cumulative distribution. Then an aggregated Z score (gene set summary statistic) was calculated for each metabolite from the gene-level Z scores of its associated genes. Here we calculated summary statistics for metabolites directly with the gene-level modified t statistics:

$$Z_{metabolite} = \frac{\sum_{i=1}^{N_{metabolite}} t_i}{\sqrt{N_{metabolite}}} \tag{2.2}$$

$Z_{metabolite}$ is the summary statistics of a metabolite, and t_i is the t statistics of gene i associated with the metabolite. $N_{metabolite}$ is the number of genes associated with the metabolite.

Regarding statistical inference, we calculated a P value for each gene set based on its background distribution of summary statistics. However, unlike the original reporter metabolite algorithm (Patil and Nielsen 2005), which derived background distributions by randomly sampling genes from the GEM, we applied sample permutations to derive such background distributions. Comparing gene/sample permutations is out of the scope of this manuscript. (Goeman and Bühlmann 2007) extensively discussed this topic previously. The sample labels (stimulated or control) were randomly shuffled within each pair of samples (PBMCs derived from the same donor). As the next step, we repeated the same procedures as described previously to recalculate the gene-level as well as the summary statistics. In total, we performed such permutations 10,000 times for each stimulation

case. The resulted permutation Z scores were used to represent the enrichment:

$$\text{Enrichment score} = \frac{Z - mean(Z_{null})}{sd(Z_{null})} \tag{2.3}$$

Z is the summary statistic of a gene set (either $Z_{pathway}$ or $Z_{metabolite}$). Z_{null} refer to the summary statistics of that gene set based on the sample permutations.

Permutation P values were then calculated by using the function permp in the statmod R package. The algorithm underlying the permp function was developed by Phipson and Smyth (Phipson and Smyth 2010). Since we tested a number of pathways or metabolites simultaneously, we performed the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) to derive the FDR. When a metabolite had a FDR value below 0.05, we defined that particular metabolite as a reporter metabolite.

2.3.5 Identification of discriminating metabolic pathways and reporter metabolites

We were interested in metabolic pathways and metabolites that can discriminate *Candida*-stimulated PBMCs from *Borrelia*, LPS and MTB-stimulated PBMCs. We were also interested in metabolic pathways and metabolites that can discriminate *Borrelia*, LPS and MTB-stimulated PBMCs. To this end, we first compared gene set enrichment results across PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB after treatment at 4 and 24 hour. We compared the 4-hour gene expression profile of PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB to the paired RPMI-treated PBMCs. We did the same regarding the 24-hour gene expression profile. When a pathway or a metabolite had a FDR value below 0.05 and a positive enrichment score, we labeled its transcriptional regulation as "Up". When a pathway or a metabolite had a FDR value below 0.05 and a negative enrichment score, we marked its transcriptional regulation as "Down". The remaining pathways and metabolites were then denoted as "N.S.", meaning no significant transcriptional changes. In the following analysis, comparisons of pathways or metabolites in PBMCs stimulated by various pathogens were done based on their "Up", "Down", and "N.S." patterns. The euclidean distance was calculated to quantify similarity between two metabolic pathway gene expression patterns. The ggdendro R package was used to

produce the dendrogram and the cmdscale function of the stat R package was used to produce the multidimensional scaling plot. To identify metabolic pathways and metabolites that were differentially regulated in a specific bacterial stimulation at both 4 and 24 hour, we also compared gene set enrichment results across PBMCs stimulated by *Borrelia*, LPS and MTB. Considering difficulty of interpretation, HMR2 subsystems (equivalent to pathways), including "Isolated", "Artificial reactions", "Exchange reactions", "Pool reactions", "Miscellaneous", "Other amino acid", and "Blood group biosynthesis" were not included in the analysis. To simplify data visualization, all the transport subsystems were not included as well. If a metabolite could be mapped to multiple subsystems, all the subsystems were included in the final results.

To evaluate whether pathogen-specific metabolism corresponded to a specific immune response, we focused on innate immunity genes provided by the database innateDB (Breuer et al. 2013). According to the innateDB, there are 1,057 innate immune genes in human. Our microarray platform measured 850 of these innate immune genes. Similar to the procedures in pathway analysis, when an innate immune gene had a FDR value below 0.05 and a positive t statistic, we labeled its transcriptional regulation as "Up". When an innate immune gene had a FDR value below 0.05 and a negative t statistic, we marked its transcriptional regulation as "Down". The remaining innate immune genes were then denoted as "N.S.", meaning no significant transcriptional changes. Again we performed the multidimensional scaling analysis.

2.4 Results

2.4.1 Transcriptional regulation in metabolic pathways of human PBMCs stimulated by various pathogenic challenges

Depending on the duration and type of pathogenic stimulant, gene expression patterns of human PBMCs varied considerably. Along the axis of the first principal component, a clear separation of 4 and 24 hour gene expression patterns was observed (Figure 2.1). To identify differentially regulated metabolic pathways in human PBMCs stimulated by heat-killed *Candida*, heat-killed *Borrelia*, LPS and heat-killed MTB, we ran gene set enrichment analysis with KEGG metabolic pathways. In general, we observed more down

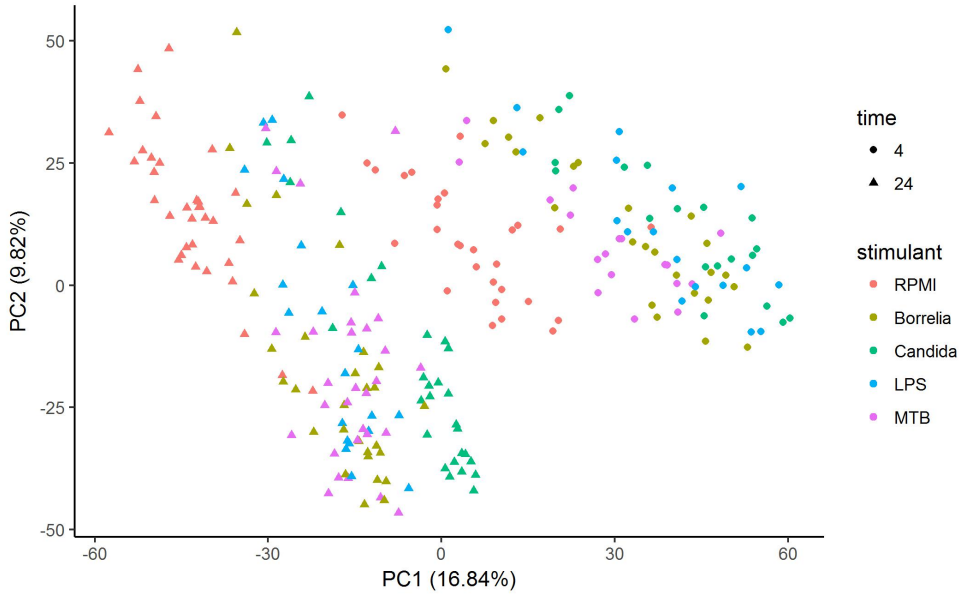


Figure 2.1: Principal component analysis of gene expression of human PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB for 4 and 24 hour.

than up-regulated metabolic pathways in stimulated PBMCs at 4 hour. However, this was reversed at 24 hour (Figure 2.2). Hierarchical clustering analysis revealed that metabolic pathway metabolic pathway regulations were very different between 4 and 24 h irrespective of the stimuli used (Figure 2.3). Multidimensional scaling analysis confirmed the result of hierarchical clustering analysis. Furthermore, we observed that the clustering result based on metabolic pathways was consistent with the clustering outcome based on innate immunity genes at 24 hour after stimulation (Figure 2.4).

2.4.2 Transcriptional regulation of energy metabolism in human PBMCs stimulated by various pathogenic challenges

At 4 hour after stimulation, glycolysis pathway was down-regulated in *Candida* (Enrichmment score = -5.88, FDR = 2.41×10^{-4}), *Borrelia* (Enrichment score = -5.96, FDR = 3.09×10^{-4}), LPS (Enrichment score = -5.83, FDR = 3.21×10^{-4}) and MTB-stimulated (Enrichment score = -4.17. FDR = 0.0013) PBMCs. Oxidative phosphorylation pathway was also down-regulated in *Candida* (Enrichment score = -4.90, FDR = 2.41×10^{-4}),

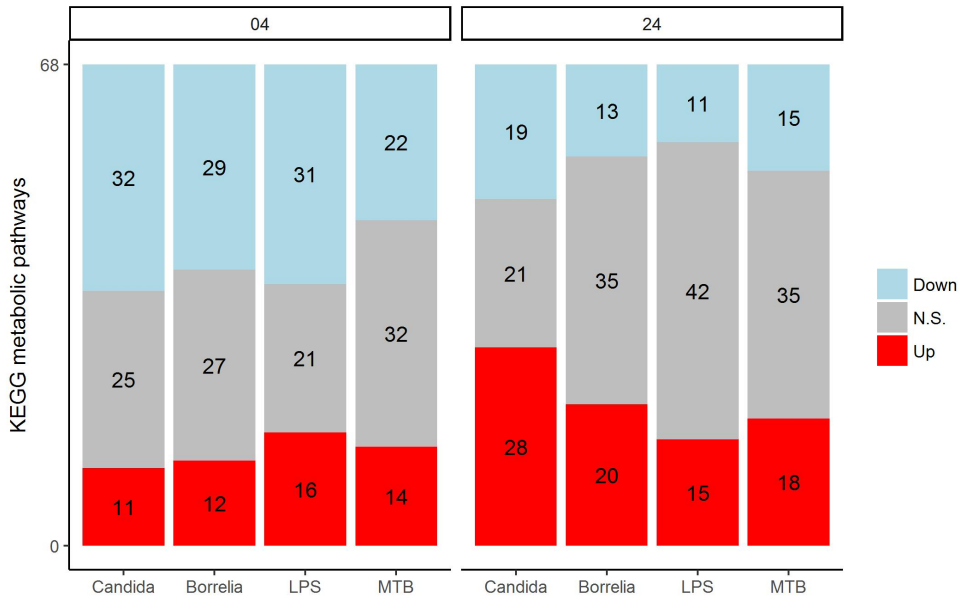


Figure 2.2: Distribution of significantly up-regulated (red), down-regulated (blue), and not significantly changed (grey) pathways in 68 KEGG metabolic pathways for *Candida*, *Borrelia*, LPS and MTB-stimulated human PBMCs at 4 and 24 hour. Any metabolic pathway is significantly changed if its FDR < 0.05.

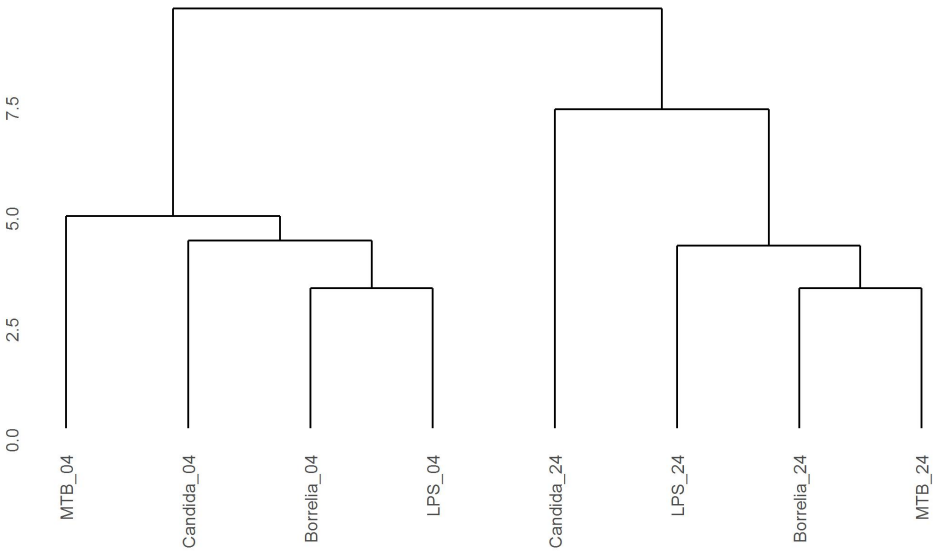


Figure 2.3: Hierarchical clustering gene expression pattern in KEGG metabolic pathways derived from human PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB at 4 and 24 hour. Euclidean distance is calculated to quantify similarity between two metabolic pathway gene expression pattern.

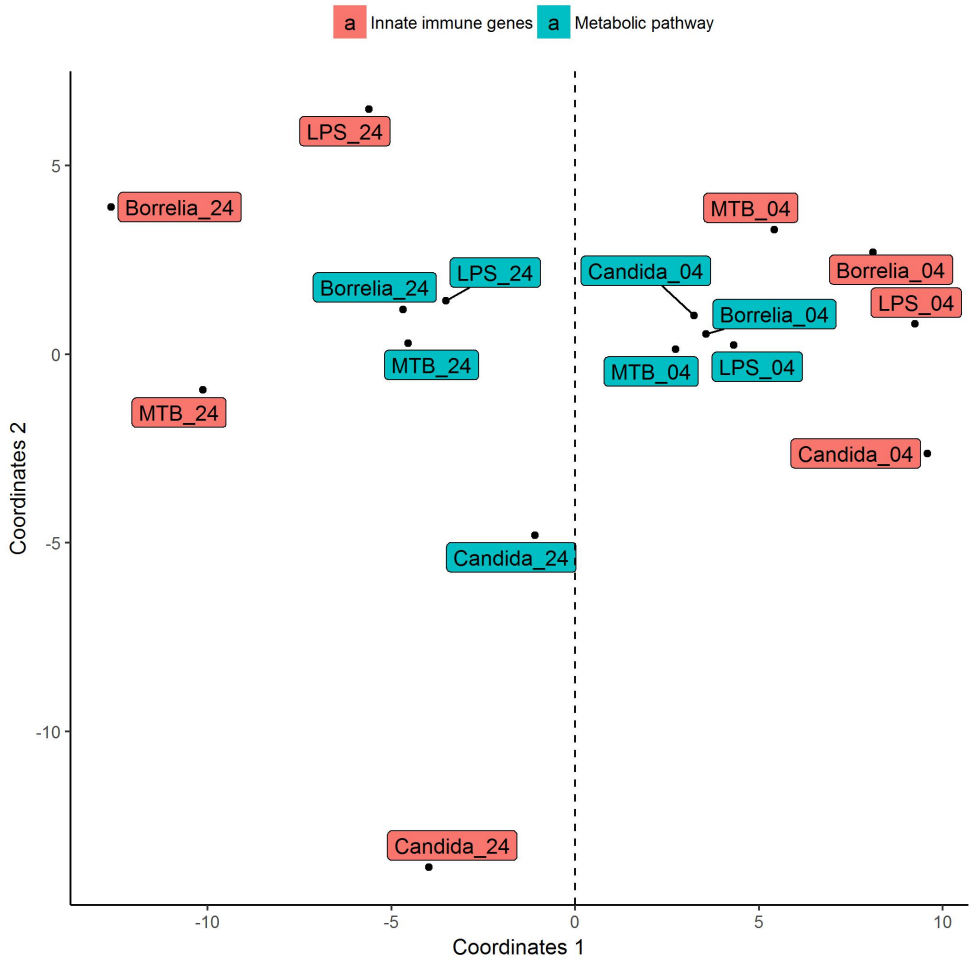


Figure 2.4: Multidimensional scaling of differential expression patterns of human PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB at 4 and 24 hour. Differential expression patterns were derived from genes involved in KEGG metabolic pathways and innate immunity. Euclidean distance is calculated to quantify similarity between two metabolic pathway gene expression pattern.

Borrelia (Enrichment score = -4.60, FDR = 3.09×10^{-4}), LPS (Enrichment score = -5.21, FDR = 3.21×10^{-4}) and MTB-stimulated (Enrichment score = -3.82. FDR = 0.0013) PBMCs. At 24 hour after stimulation, glycolysis pathway was up-regulated in *Candida* (Enrichment score = 4.33, FDR = 2.12×10^{-4}), *Borrelia* (Enrichment score = 7.52, FDR = 3.09×10^{-4}), LPS (Enrichment score = 2.99, FDR = 0.0019) and MTB-stimulated (Enrichment score = 7.51, FDR = 4.25×10^{-4}) PBMCs. However, oxidative phosphorylation was not significantly changed in PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB.

2.4.3 Discriminating metabolic pathways in human PBMCs stimulated by various pathogenic challenges

We focused on metabolic pathways that had the same transcriptional patterns in PBMCs stimulated by *Borrelia*, LPS and MTB, but differed from *Candida*-stimulated PBMCs at both 4 and 24 hour. The detail statistics for pathways were provided in the Supplementary Table 1. The pentose phosphate pathway was down-regulated in *Borrelia*, LPS and MTB-stimulated PBMCs, but not in *Candida*-stimulated PBMCs at 4 hour (Figure 2.5). However, at 24 hour, the pentose phosphate pathway was up-regulated in *Candida*-stimulated PBMCs, but had no significant change in *Borrelia*, LPS and MTB-stimulated PBMCs (Figure 2.5). Riboflavin, beta alanine and histidine metabolism were differentially regulated in *Candida*-stimulated PBMCs, but not significantly changed in *Borrelia*, LPS and MTB-stimulated PBMCs at both 4 and 24 hour (Figure 2.5). Aminoacyl tRNA biosynthesis was up-regulated in *Borrelia*, LPS and MTB-stimulated PBMCs but not significantly changed in *Candida*-stimulated PBMCs at 4 hour. However, this pathway was up-regulated in *Candida*-stimulated PBMCs but down-regulated in *Borrelia*, LPS and MTB-stimulated PBMCs at 24 hour (Figure 2.5).

Regarding the metabolic pathways that discriminated *Borrelia*, LPS and MTB-stimulated PBMCs, we observed that glycosylphosphatidylinositol GPI anchor biosynthesis was up-regulated in LPS-stimulated PBMCs but did not change in *Borrelia* and MTB-stimulated PBMCs at 4 hour. However, at 24 hour, this pathway was down-regulated in *Borrelia* and MTB-stimulated PBMCs whereas it remained unchanged in LPS-stimulated PBMCs (Figure 2.6). Similarly, fatty acid metabolism and glycerolipid metabolism were down-

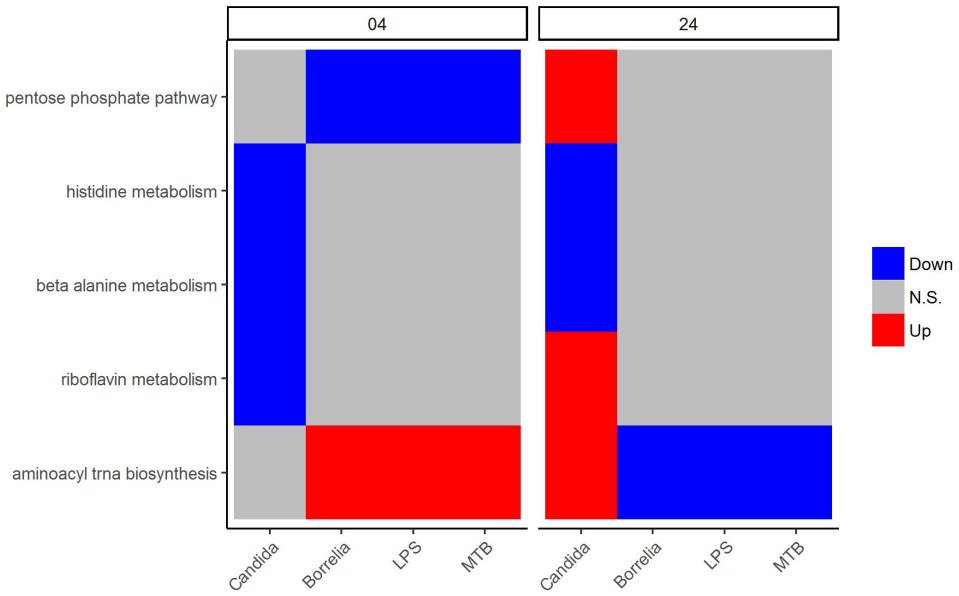


Figure 2.5: KEGG metabolic pathways that discriminated *Candida*-stimulated PBMCs from *Borrelia*, LPS and MTB-stimulated human PBMCs. Blue refers to significantly down regulation. Red refers to significantly up regulation. Grey means not significantly changed. A pathway is significantly changed if its FDR < 0.05.

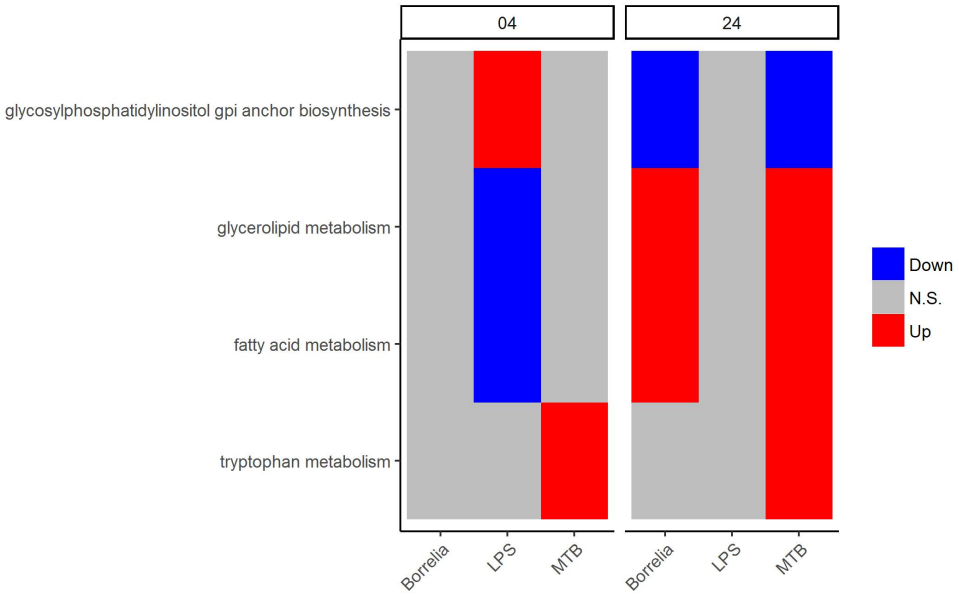


Figure 2.6: KEGG metabolic pathways that discriminated between *Borrelia*, LPS and MTB-stimulated human PBMCs. Blue refers to significantly down regulation. Red refers to significantly up regulation. Grey means not significantly changed. A pathway is significantly changed if its FDR < 0.05.

regulated in LPS-stimulated PBMCs but not in *Borrelia* and MTB-stimulated PBMCs at 4 hour. This pathway was up-regulated in *Borrelia* and MTB-stimulated PBMCs but did not change in LPS-stimulated PBMCs at 24 hour (Figure 2.6). Tryptophan metabolism was differentially regulated in MTB-stimulated PBMCs, but not significantly changed in *Borrelia* and LPS-stimulated PBMCs at both 4 and 24 hour (Figure 2.6). We did not identify a metabolic pathway that can discriminate *Borrelia*-stimulated PBMCs from LPS and MTB-stimulated PBMCs.

2.4.4 Discriminating metabolites in human PBMCs stimulated by various pathogenic challenges

In an attempt to identify metabolites that discriminated PBMCs with various stimuli, we ran reporter metabolite analysis with the human genome-scale metabolic model, HMR2. A total number of 4,548 metabolites were involved in the reporter metabolite analysis. We

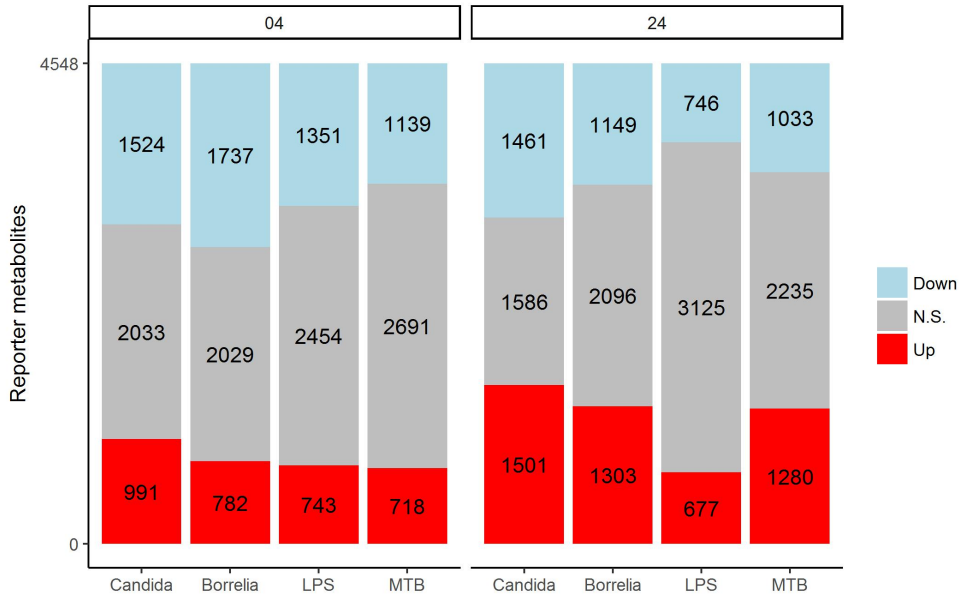


Figure 2.7: Distribution of significantly up-regulated (red), down-regulated (blue), and not significantly changed (grey) reporter metabolites for *Candida*, *Borrelia*, LPS and MTB-stimulated human PBMCs at 4 and 24. When a reporter metabolite has FDR $<$ 0.05, it is significant.

observed more down-regulated than up-regulated reporter metabolites in the stimulated PBMCs at 4 hour. However, this pattern was reversed at 24 hour (Figure 2.7). In a next step, we focused on reporter metabolites that were differentially regulated in *Candida*-stimulated PBMCs but not in PBMCs with bacterial stimuli at both 4 and 24 hour. Among the identified reporter metabolites at 4 and 24 hour, 103 of them were found specific for *Candida*-stimulated PBMCs. These 103 *Candida*-specific reporter metabolites participated in 45 pathways including nucleotide metabolism (15 reporter metabolites), and fatty acid biosynthesis (10 reporter metabolites) (Figure 2.8). We also focused on reporter metabolites that can discriminate between *Borrelia*, LPS and MTB-stimulated PBMCs at both 4 and 24 hour. We identified 32, 7 and 10 reporter metabolites that were specific for *Borrelia*, LPS and MTB-stimulated PBMCs, respectively (Figure 2.9). Statistics of all the pathogen-specific reporter metabolites were provided in Supplementary Table 2.

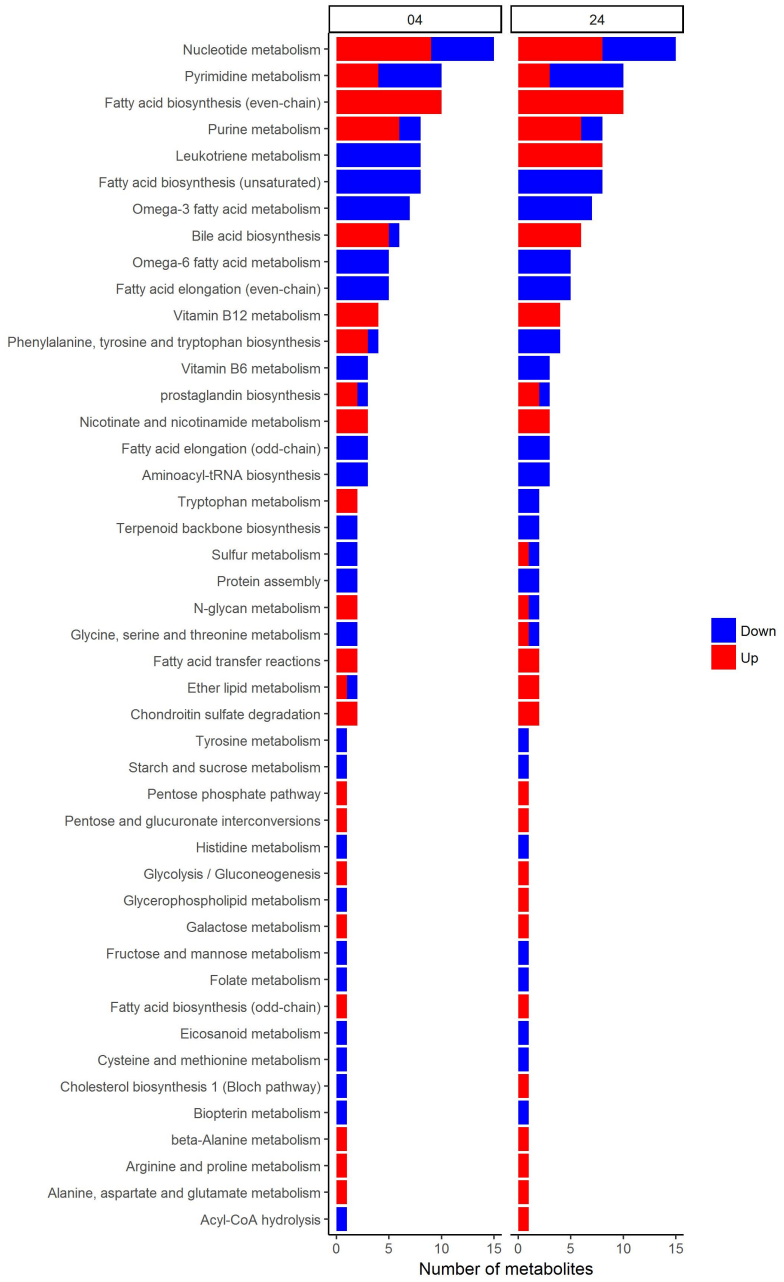


Figure 2.8: Reporter metabolites that discriminate Candida-stimulated PBMCs from Borrelia, LPS and MTB-stimulated PBMCs at 4 and 24 hour. These reporter metabolites were grouped based on their associated subsystems in HMR2. Blue denotes significant down-regulation. Red denotes significant up-regulation.

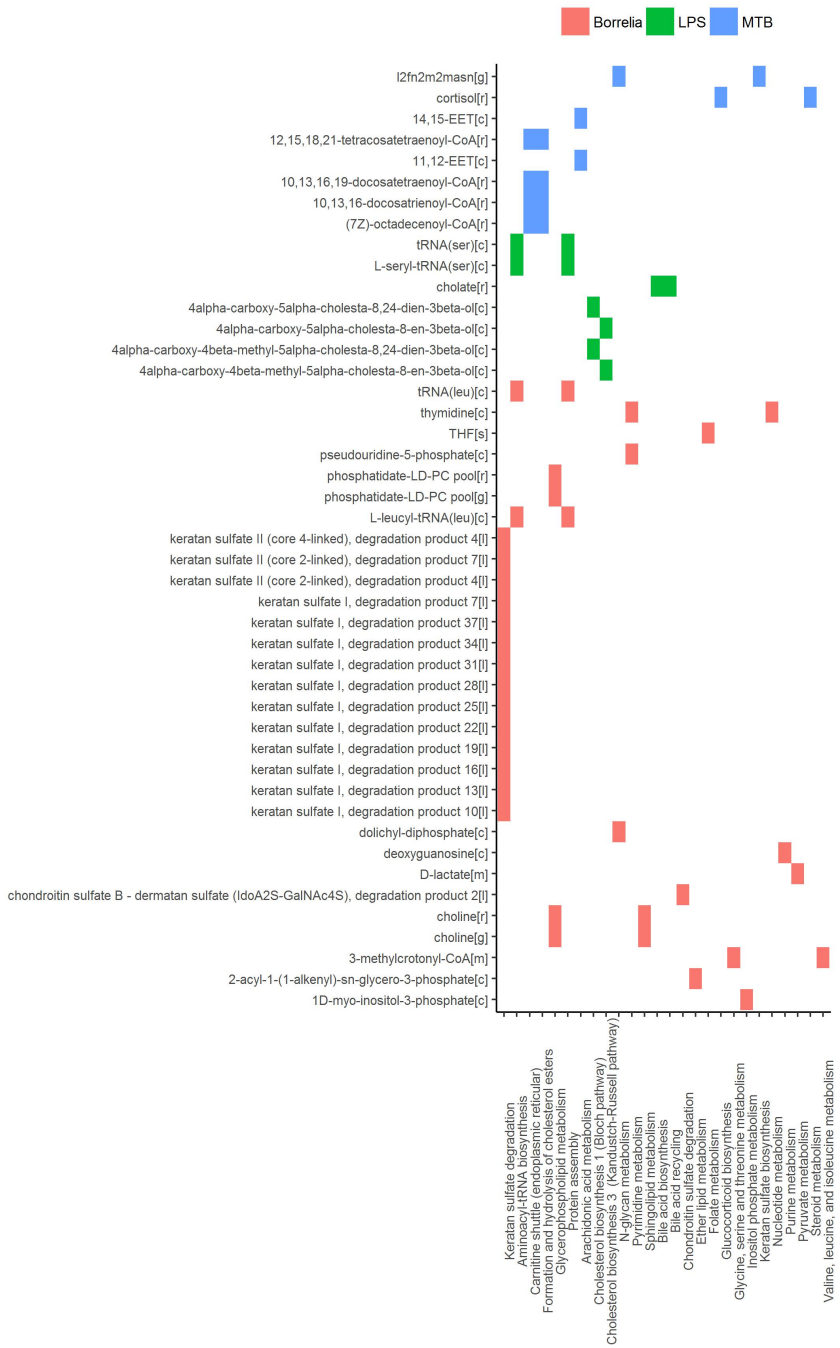


Figure 2.9: Reporter metabolites that discriminated Borrelia, LPS and MTB-stimulated PBMCs at 4 and 24 hour. Associated subsystems of these reporter metabolites are identified in HMR2.

2.5 Discussion

The main finding of our study is that characterization of pathogen-dependent metabolic reprogramming in immune cells treated by various stimuli of innate immune pathway. For this purpose, we performed gene set enrichment analysis on gene expression data of human PBMCs treated with heat-killed *Candida*, heat-killed *Borrelia*, *E. coli*-derived LPS and heat-killed MTB. Either KEGG metabolic pathways or metabolites in human genome-scale metabolic models were used as gene sets. Our particular experimental setup with one fungal pathogen (*Candida*) and three bacterial inflammatory stimuli (*Borrelia*, LPS and MTB) allowed us to identify metabolic signatures of *Candida*-induced host response, but also host response differences between bacterial challenges.

A very strong temporal effect on the expression of metabolic genes was observed. This observation is in line with the concept that stimulation period is a critical factor in immune response (Nagy and Haschemi 2015; Hotamisligil 2017). At 4 hour after stimulation, both oxidative phosphorylation and glycolysis were down-regulated. At 24 hour, however, gene expression of glycolysis showed up-regulation, whereas gene expression of oxidative phosphorylation remained unaltered in PBMCs. The observation of down-regulation of glycolysis genes after 4 hour of stimulation is novel, and its impact for cell function warrants future studies. However, the observation at 24 hour is consistent with literature data showing that activated immune cells shift towards glycolysis and away from oxidative phosphorylation (McGettrick and O'Neill 2013; Cheng, Joosten, and Netea 2014; Pearce and Pearce 2013; Pearce et al. 2013).

For the purpose of identifying pathogen-dependent metabolic reprogramming in immune cells, we focused on metabolic pathways and metabolites that allow discrimination between various stimuli at both 4 and 24 hour.

2.5.1 Five metabolic pathways can discriminate *Candida*-stimulated PBMCs from *Borrelia*, LPS and MTB-stimulated PBMCs

Five pathways, i.e. the pentose phosphate pathway, histidine metabolism, beta alanine metabolism, riboflavin metabolism and aminoacyl tRNA biosynthesis, were identified

to discriminate *Candida*-stimulated PBMCs from *Borrelia*, LPS and MTB-stimulated PBMCs. Interestingly, we observed that the pentose phosphate pathway was differentially regulated in PBMCs stimulated by *Borrelia*, LPS and MTB but not in *Candida*-stimulated PBMCs at 4 hour. In contrast, at 24 hour, this pathway was differentially regulated only in *Candida*-stimulated PBMCs but not significantly changed in *Borrelia*, LPS and MTB-stimulated PBMCs. The pentose phosphate pathway was reported to support cytokine secretion in dendritic cells (Everts et al. 2014). Since cytokine production of human PBMCs depends on the type of stimulus (Henderson and Rippin 1995), our observation of differential regulation in the pentose phosphate pathway likely indicates a specific function for *Candida* stimulated cytokine production. Indeed, our findings corroborate those of a recent study in which *Candida*-stimulated PBMCs were identified to have different cytokine profiles from bacteria-stimulated PBMCs (Li et al. 2016). On the other hand, little is known about the specific roles of the other four *Candida*-specific metabolic pathways in regulation of the immune response, and further investigation is warranted to validate these novel findings.

2.5.2 Four metabolic pathways can differentiate between *Borrelia*, LPS and MTB-stimulated PBMCs

We further noted that three pathways (glycosylphosphatidylinositol GPI anchor biosynthesis, glycerolipid metabolism, fatty acid metabolism) discriminated LPS-stimulated PBMCs from *Borrelia* and MTB-stimulated PBMCs. Meanwhile, tryptophan metabolism differentiates MTB-stimulated PBMCs from *Borrelia* and LPS-stimulated PBMCs. We failed to identify pathways that allow discrimination *Borrelia*-stimulated PBMCs from LPS and MTB-stimulated PBMCs. Activation of tryptophan metabolism was previously reported in human marcophages *in vitro* upon MTB stimulation (Blumenthal et al. 2012), and a recent study (Laarhoven et al. 2018) has identified a crucial role of tryptophan metabolism for the pathophysiology of tuberculous meningitis. In addition, enhancement of tryptophan catabolism is an IFN (interferon) γ-induced immune response in many different host cell types, and has been postulated to reduce the supply of tryptophan to bacterial pathogens (Moffett and Namboodiri 2003; O'Neill, Kishton, and Rathmell 2016). A reduced supply of tryptophan is linked to suppress T cell proliferation (Munn

et al. 1999). Our observation of differential regulation of tryptophan in MTB-stimulated PBMCs might be related to different T cell proliferation after stimulation of MTB, compared to *Borrelia* and LPS.

2.5.3 Genome-scale metabolic model provides metabolic pathways with details

The KEGG pathway based analysis failed to identify metabolic pathways that discriminate *Borrelia*-stimulated PBMCs from LPS- and MTB-stimulated PBMCs. To explore potential differences in more depth, we ran the reporter metabolite analysis, which is a gene set enrichment analysis with a genome-scale metabolic model. A genome-scale metabolic model is comprised of metabolites and reactions between them. Compared to KEGG metabolic pathway information, the genome-scale metabolic model makes use of detailed information on biochemical reactions of pathways. For instance, for any enzyme catalyzing reaction, we can retrieve the genes encoding that enzyme in the genome-scale metabolic model. Moreover, metabolites can be products of some reactions and meanwhile act as substrates in other reactions. Consequently, reporter metabolite analysis based on genome-scale metabolic model does not repeat but complement results from KEGG pathway analysis. We used HMR2 in our analysis since we did not perform flux balance analysis.

2.5.4 103 reporter metabolite can discriminate *Candida*-stimulated PBMCs from *Borrelia*, LPS and MTB-stimulated PBMCs

In this study, we identified 103 reporter metabolites that were differentially regulated in *Candida*-stimulated PBMCs, but not in PBMCs stimulated with bacterial stimuli at both 4 and 24 hour. A considerable number of these *Candida*-specific reporter metabolites were found to be related to lipid metabolism. The previous study (Smeekens et al. 2013) reported that *Candida* induced a type I IFN response that was distinct from *Borrelia*, LPS and MTB stimulation. Interestingly, type I IFN was identified to influence *de novo* cholesterol biosynthesis and fatty acids biosynthesis in murine marcophages (York et al.

2015). Desmosterol, one of the *Candida*-specific reporter metabolites, is the last intermediary metabolite in the Bloch pathway of cholesterol biosynthesis. This metabolite was previously reported to coordinate cholesterol and fatty acid homeostasis, and affect anti-inflammatory function in macrophage (Spann et al. 2012). Taken together, we proposed that desmosterol might serve as a metabolic read out of the type I IFN response in *Candida*-stimulated PBMCs.

2.5.5 49 reporter metabolites can discriminate between *Borrelia*, LPS and MTB-stimulated PBMCs

In PBMCs stimulated by *Borrelia*, LPS and MTB, 49 metabolites were identified to discriminate different kinds of pathogenic challenges. Within LPS-specific reporter metabolites, we observed intermediate metabolites present in the Bloch pathway and Kandutsch-Russell pathway (e.g. 4α-carboxy-5α-cholesta-8,24-dien-3β-ol). With mass spectrometry and isotope labeling techniques, (Mitsche et al. 2015) previously showed that different tissues or cell types were characterized by different flux distributions in the Bloch and Kandutsch-Russell pathway. Our observation indicates that there also might be condition-specific flux distribution in these two parallel cholesterol biosynthesis pathways. Within MTB-specific reporter metabolites, we observed two kind of epoxyeicosatrienoic acids, synthesized from arachidonic acid. Epoxyeicosatrienoic acids were reported to inhibit inflammatory gene expression in immune cells and animal models (Thomson, Askari, and Bishop-Bailey 2012).

2.6 Conclusions

In summary, by integrating gene expression data with KEGG metabolic pathways in combination with the human genome-scale metabolic model, a very sensitive method to characterize metabolic reprogramming in immune cells is obtained. Applying this methodology, we were able to discriminate metabolic pathways and metabolites in human PBMCs stimulated by *Candida*, *Borrelia*, LPS and MTB. For instance, in the case of *Candida* we identified five differentially regulated pathways spanning metabolic regions from the pentose phosphate pathway to aminoacyl tRNA biosynthesis. Our analysis

here, for the first time, provides insight into pathogen-specific metabolism which affects stimulus-dependent signal transduction and cytokine production in stimulated human PBMCs.

2.7 Funding

This work was supported by grants CVON-Genius (CVON2011-19) and RESOLVE (FP7 305707).

2.8 Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2018.00139/full#supplementary-material

3

Use of plasma metabolomics to analyze phenotype-genotype relationships in young hypercholesterolemic females

This chapter has been published as Zhang, X., A. Rimbert, W. Balder, A. H. Zwinderman, J. A. Kuivenhoven, G. M. Dallinga-Thie, and A. K. Groen. Use of plasma metabolomics to analyze phenotype-genotype relationships in young hypercholesterolemic females. J. Lipid Res. 2018. 59: 2174–2180

3.1 Abstract

Hypercholesterolemia is characterized by high plasma low density lipoprotein (LDL) cholesterol and often caused by genetic mutations in *LDLR*, *APOB* or proprotein convertase subtilisin/kexin type 9 (*PCSK9*). However, a substantial proportion of hypercholesterolemic subjects do not have any mutations in these canonical genes, leaving the underlying pathobiology to be determined. In this study, we investigated whether combining plasma metabolomics with genetic information increases insight in the biology of hypercholesterolemia. For this proof of concept study, we combined plasma metabolites from 119 hypercholesterolemic females with genetic information on the LDL canonical genes. Using hierarchical clustering we identified four subtypes of hypercholesterolemia, which could be distinguished along two axes represented by triglyceride and large LDL particle concentration. Subjects with mutations in *LDLR* or *APOB* preferentially clustered together suggesting that patients with defects in the LDL receptor pathway show a distinctive metabolomics profile. In conclusion, we show the potential of using metabolomics to segregate hypercholesterolemic subjects in different clusters which may help in targeting genetic analysis.

Keywords: hypercholesterolemia; triglyceride; low density lipoprotein; genetics; metabolomics;

3.2 Introduction

Hypercholesterolemia due to a high concentration of plasma low density lipoprotein (LDL) cholesterol has been shown to be a causal factor in accelerating atherosclerosis in a plethora of studies (Ference et al. 2017; Goldstein and Brown 2015). The liver plays a pivotal role in the regulation of cholesterol metabolism. It secretes cholesterol packaged in VLDL particles that are subsequently converted into IDL and LDL particles largely by the action of different lipases in the periphery (Packard and Shepherd 1997). A key step in the uptake of cholesterol is the internalization of LDL via the LDL receptor (LDLR) (Brown and Goldstein 1986). Mutations in the *LDLR* as well as mutations in genes encoding apolipoprotein B (*APOB*) or proprotein convertase subtilisin/kexin type 9 (*PCSK9*), are causally related with hypercholesterolemia (Soutar and Naoumova 2007). These genetic

mutations, however, do not explain all hypercholesterolemic cases. For instance, in the UK pilot cascade project, 403 of 635 (63.5%) hypercholesterolemic subjects did not have mutations in *LDLR*, *APOB*, or *PCSK9* (Taylor et al. 2010). In a recent large scale study designed to evaluate the prevalence of a familial hypercholesterolemia (FH) mutation among individuals with severe hypercholesterolemia (Khera, Won, et al. 2016), only 24 of 1,386 subjects with LDL cholesterol above 5 mmol/L were identified to have mutations in these three canonical genes. Although the prevalence of genetically defined hypercholesterolemia varies across studies (Wang et al. 2016), a substantial proportion of hypercholesterolemic subjects do not have mutations in *LDLR*, *APOB* or *PCSK9*. A major reason for this finding could be the presence of disease-causing mutations in other genes involved in cholesterol homeostasis either affecting the LDL receptor pathway or other yet to be defined mechanisms. Interestingly, whole exome sequencing of a cohort with FH subjects without mutations in *LDLR*, *APOB* and *PCSK9* did not identify novel causal mutations (Futema et al. 2014).

Recently, we analyzed a cohort of 119 young females with plasma LDL cholesterol above the 99th percentile for their age. In 20 hypercholesterolemic females, we identified 12 causal heterozygous mutations in *LDLR* and one causal heterozygous mutation in *APOB* (Balder et al. 2018). In the 99 remaining females we found eight subjects carrying a variant in *LDLR* or *APOB* with unknown clinical significance (Balder et al. 2018). This left us with 91 females that suffered from hypercholesterolemia caused by either a polygenic (Talmud et al. 2013) or epigenetic (Dekkers et al. 2016) mechanism, or presence of pathogenic variant in yet unknown genes. To get further insight in the underlying pathobiology of hypercholesterolemia of unknown origin, we performed plasma metabolite analysis on all the 119 hypercholesterolemic females. We hypothesized that mutations in genes belonging to the same metabolic pathway (e.g. the LDL receptor pathway) should render a similar plasma metabolome. This analysis differentiated four subgroups, which could be distinguished along two axes represented by plasma triglyceride and large LDL particle concentration.

3.3 Materials and Methods

3.3.1 Participants

The selection of the participants (N = 119) in this study is described in detail elsewhere (Balder et al. 2018). In brief, these women were apparently healthy, aged 25 to 40 year and had plasma LDL cholesterol level above 4.7 mmol/l. Exclusion criteria were diagnosis of cardiovascular disease (e.g. myocardial infarction, stroke or coronary surgery), diabetes mellitus, use of lipid-lowering drug, or having aberrant thyroid, liver or kidney function. The study protocol was approved by the Medical Ethical Committee of the University Medical Center Groningen in The Netherlands and all participants provided written informed consent.

3.3.2 Next generation sequencing

With a custom target sequencing array developed based on the SureSelect capture system, we sequenced the coding regions of 11 genes, including *LDLR*, *APOB*, *PCSK9*, *LDLRAP1*, *APOE*, *ABCG5*, *LIPA*, *STAP1*, *MTTP*, *ANGPTL3*, and *SAR1B* to assess a monogenic cause of hypercholesterolemia. If a mutation had minor allele frequency below 0.1% in the Genome of Netherlands (Netherlands Consortium 2014), it was considered a rare mutation. Mutations that are verified to cause hypercholesterolemia were listed in our previous publication (Balder et al. 2018).

Detection of copy number variations (CNV) was performed using the CoNVaDING (Copy Number Variation Detection in Next-generation sequencing Gene panels) (Johansson et al. 2016). Detected CNVs were validated using either multiplex ligation-dependent probe amplification, or by long-range PCR or real-time PCR (Balder et al. 2018).

3.3.3 Genetic risk score calculation

To study a possible polygenic cause of hypercholesterolemia, we calculated the weighted genetic risk score (wGRS). The Global Lipid Genetic Consortium (GLGC) metaanalysis of genome-wide association studies identified 95 loci affecting LDL cholesterol concentration (Teslovich et al. 2010). Among these loci, 12 SNPs had the highest power to

discriminate between FH mutation-negative individuals and the general population (Talmud et al. 2009; Talmud et al. 2013). For each individual, we calculated the wGRS using the weighted sum of the risk allele (the LDL cholesterol-raising allele) (Balder et al. 2018). The weights used were the corresponding per-allele effect in plasma LDL cholesterol changes reported by the Global Lipid Genetic Consortium (Teslovich et al. 2010).

3.3.4 Lifestyle score calculation

To investigate the association between lifestyle and plasma metabolome in hypercholesterolemic females, we used a recently described healthy lifestyle score (Khera, Emdin, et al. 2016). Points were given for the major lifestyle parameters including smoking status and eating habits. The details were described in our previous publication (Balder et al. 2018). In short, a maximum of four points reflects a very healthy lifestyle: the smaller the score, the less healthy the lifestyle. The minimum point is zero.

3.3.5 Metabolite measurements

Fasting plasma samples were routinely collected by Lifelines (www.lifelines.nl) and stored at $-80°C$ until analysis on the Nightingale metabolomics platform (Nightingale Health, Finland). This platform includes 225 metabolic features including lipids, lipoproteins, fatty acids, amino acids, and glycolysis precursor molecules listed on https://nightingalehealth.com/biomarkers, using a NMR spectroscopy platform (Fischer et al. 2014; Soininen et al. 2015).

3.3.6 Statistical analysis

To explore subtypes of hypercholesterolemia, we performed hierarchical clustering based on the plasma metabolomics data. Since the metabolomics data contains measurements of different units, we first scaled the data so that every variable had mean 0 and standard deviation 1. Next, we ran the hierarchical clustering with the function hclust from R. We used Euclidean distance as the dissimilarity measure and complete linkage as the similarity measure between the clusters. The dendrogram was made by using the ggdendro and

ggplot2 (Wickham 2016) R package. Finally, we cut the dendrogram into four clusters by using cutree function in R.

To identify the cluster corresponding to hypercholesterolemia due to defects in the LDL receptor pathway, we performed principal component analysis (PCA) on the metabolomics data. Since the data contains measurements of different units, we converted the metabolomics data into ranks, so that every metabolite had value ranging between 1 and 119. We then calculated the covariance matrix and performed eigenvector decomposition. Entries of every eigenvector is also called loadings. Based on the loadings, we identified metabolites that most correlated to the first and second principal components by calculating the Spearman correlation coefficients.

To evaluate associations between genetic risk/lifestyle scores and metabolite concentrations, we applied a nonparametric method, namely the Kendall's tau correlation test. We reported the Kendall's tau correlation coefficient and P value. A P value below 0.05 is considered significant.

3.4 Result

A group of 119 young women with hypercholesterolemia, defined as plasma LDL cholesterol levels above the 99th percentile for their age, was selected from the Lifelines cohort. The baseline characteristics are presented in Table 3.1. To analyze the underlying pathobiology of the hypercholesterolemic phenotype, plasma metabolomics was performed using the Nightingale platform. Although the absolute values measured in the Nightingale platform are lower than the conventional measured plasma lipids, we showed that the correlation between both measurements are high (Table 3.1). A summary of all the results of metabolite analysis is presented in supplemental Table 1. Hierarchical clustering analysis of the metabolomics data set revealed three main clusters and one cluster containing only one sample (Figure 3.1). The size of the cluster 1, 2, 3, and 4 was 43, 15, 60 and 1, respectively.

To analyze the divergence of the different clusters, we ran principal component analysis. The first and second principal component explained 38% and 21% of the total variance of the metabolic variables across the 119 individuals, respectively (Figure 3.2). To under-

Table 3.1: Characteristics of 119 hypercholesterolemic females

	Clinical Chemistry	Nightingale Metabolomics	Spearman Correlation Coefficient
LDL cholesterol (mmol/l)	5.25 ± 0.50	2.27 ± 0.26	0.66
Total cholesterol (mmol/l)	7.17 ± 0.64	5.57 ± 0.43	0.68
Triglyceride (mmol/l)	1.50 ± 0.68	1.45 ± 0.47	0.96
HDL cholesterol (mmol/l)	1.39 ± 0.28	1.47 ± 0.22	0.84
ApoB (g/l)	1.25 ± 0.14	1.10 ± 0.11	0.78

Note:
Data are expressed as mean ± SD; N = 119; Age (year), 32.90 ± 4.37; BMI, 27.9 ± 5.10

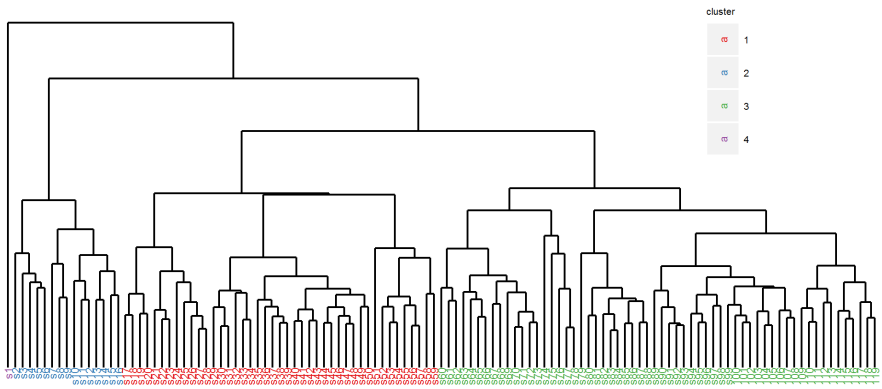


Figure 3.1: Hierarchical clustering of plasma metabolomics data derived from 119 hypercholesterolemic females. Euclidean distance was used as the dissimilarity measure and complete linkage was used as the dissimilarity measure between the clusters.

stand which metabolites correspond to the first and second principal component the most, we calculated the Spearman correlation coefficients between original variables and principal components (Supplemental Table 2). We observed that plasma triglyceride and large LDL particle concentration were the most correlated variables with the PC1 (Spearman correlation coefficient -0.988) and PC2 (Spearman correlation coefficient -0.978), respectively. Therefore, we used these two variables to represent the axes of PC1 and PC2 (Figure 3.3). Our next question was whether the 4 clusters derived from the hierarchical clustering analysis (Figure 3.1) were indeed separated by PC1 and PC2. To answer that, we added the hierarchical clustering results to the scatterplot (Figure 3.3). Inspection re-

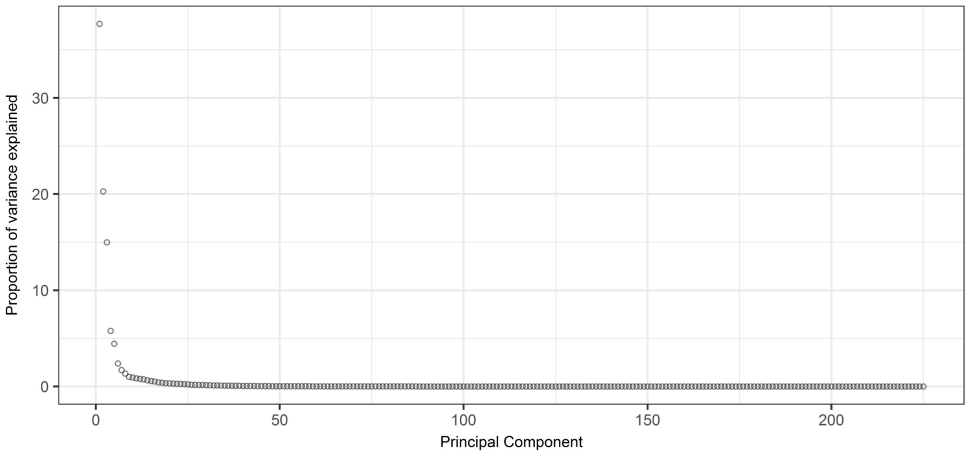


Figure 3.2: Proportion of variance explained by PCs derived from plasma metabolomics data of 119 hypercholesterolemic females.

veals that the females in cluster 3 are separated from the other groups by showing a high plasma large LDL particle concentration coupled with relatively low plasma triglyceride, suggesting a defect in hepatic LDL uptake.

Because we sequenced *LDLR*, *APOB* and *PCSK9* in all subjects, we could verify whether the females with known heterozygous mutations in the LDL receptor pathway plot in the region of cluster 3. Indeed, from 20 subjects with heterozygous mutations in *LDLR* or *APOB* 15 subjects were located in cluster 3 (Figure 3.4). The other 5 carriers were found in cluster 1 (n = 3) and cluster 2 (n = 2). In addition, we identified 8 women who were heterozygous carrier of a novel variant in *LDLR* or *APOB* from which the pathogenicity has not yet been determined. Five of these 8 subjects were positioned in cluster 3 and three in cluster 1 (Figure 3.5).

To improve our understanding of the underlying pathobiology of the elevated plasma LDL cholesterol in the remaining 91 women, we calculated the weighted genetic risk score (wGRS) and lifestyle score, and assessed the associations between both scores and plasma concentrations of large LDL particle and triglyceride. As shown in supplemental Figure 3.6 and 3.7, no relation could be demonstrated between both scores and plasma large LDL particle concentration (wGRS: Kendall tau correlation coefficient -0.017, P value = 0.80. Lifestyle score: Kendall tau correlation coefficient -0.04, P value = 0.57). Both scores

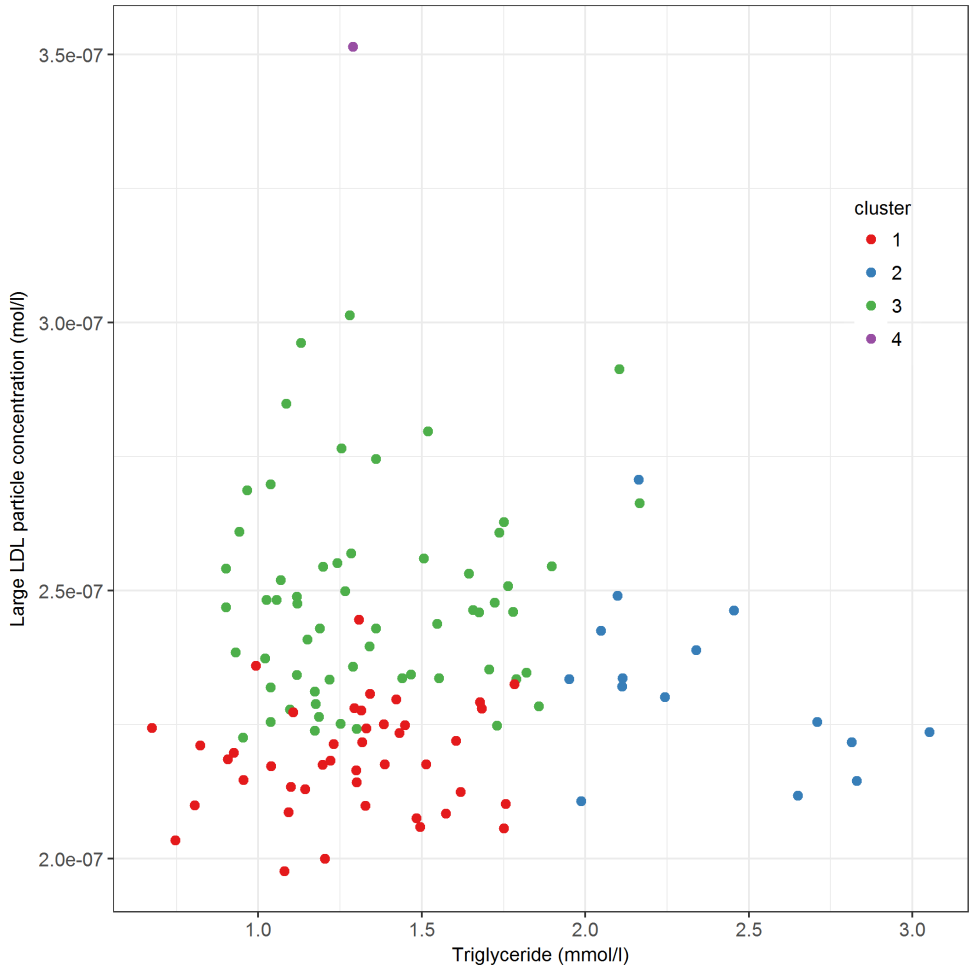


Figure 3.3: Plasma triglyceride against large LDL particle concentration in 119 hypercholesterolemic females. Different colors refer to the hierarchical clustering outcomes (red, cluster 1; blue, cluster 2; green, cluster 3; purple, cluster 4).

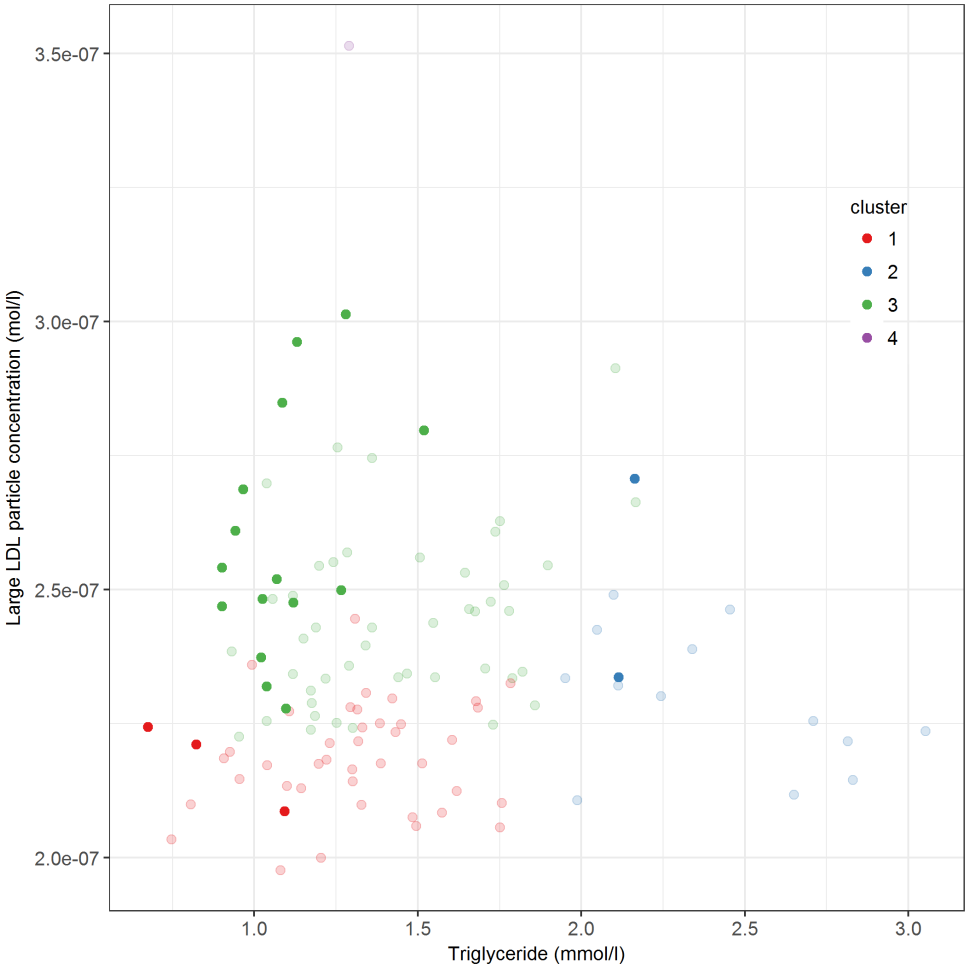


Figure 3.4: Plasma triglyceride against large LDL particle concentration in 119 hypercholesterolemic females. Different colors refer to the hierarchical clustering outcomes (red, cluster 1; blue, cluster 2; green cluster 3; purple, cluster 4). The hypercholesterolemic females with mutations that were known to affect the LDLR pathway were highlighted.

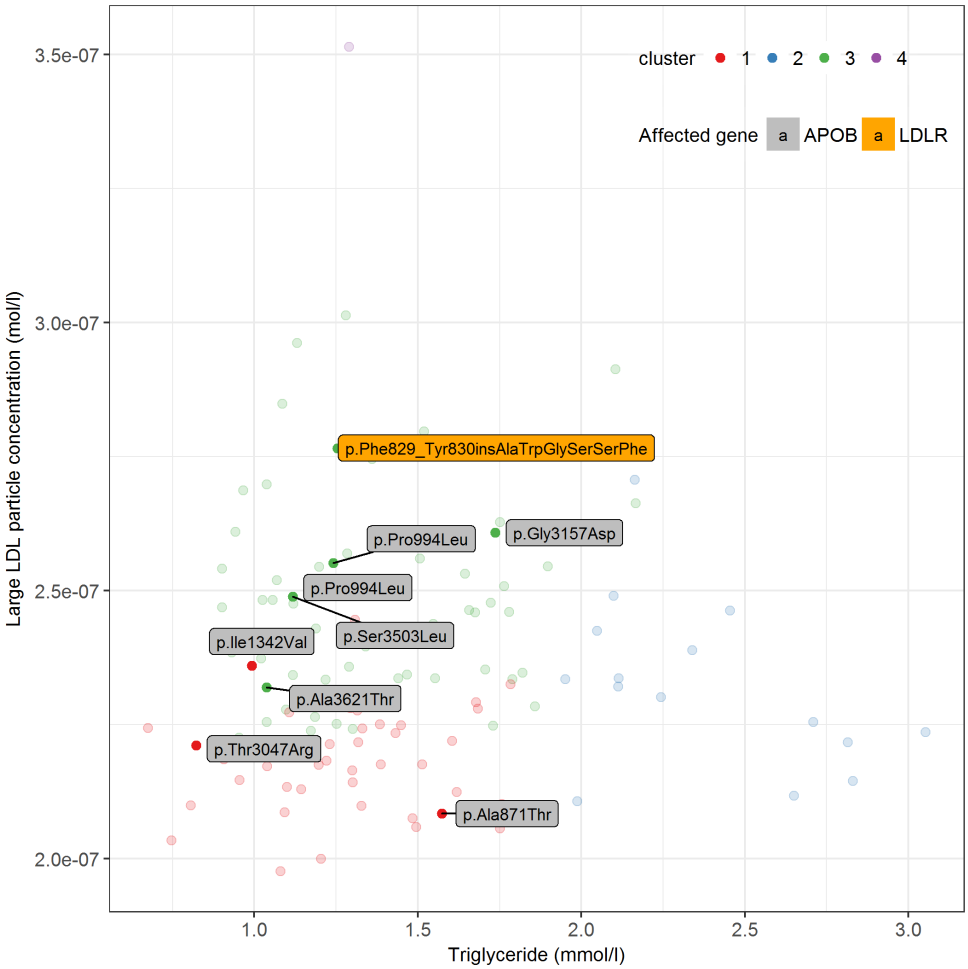


Figure 3.5: Plasma triglyceride against large LDL particle concentration in 119 hypercholesterolemic females. Different colors refer to the hierarchical clustering outcomes (red, cluster 1; blue cluster 2; green cluster 3; purple, cluster 4). The highlighted dots represent eight individuals who carry a heterozygous variant in *LDLR* or *APOB* of unknown clinical significance. The specific variant in *LDLR* or *APOB* is shown.

showed moderate association with plasma triglyceride concentration (wGRS: Kendall tau correlation coefficient -0.156, P value = 0.02. Lifestyle score: Kendall tau correlation coefficient -0.198, P value = 0.0099).

3.5 Discussion

In the current study, we showed that combining plasma metabolomics data with genetic information can improve our understanding of the origin of severe hypercholesterolemia in young healthy women. These analyses may help the diagnosis and personalized treatment of patients with hypercholesterolemia in which no causal mutations in the canonical LDL genes can be identified.

Metabolic profiling has been used in a large number of cohort studies to assess the value of circulating metabolites in prediction of risk for cardiovascular events (Würtz et al. 2015; Holmes et al. 2018). More specifically, metabolomics has been used to study associations between circulating metabolites and statin usage (Würtz et al. 2016), CETP inhibition (Kettunen et al. 2018) and PCSK9 inhibition (Sliz et al. 2018), generating insight in the broad metabolic effects of these interventions. Nightingale metabolomics data contain not only concentrations in different units, but also other quantities such as ratios, percentages, degrees of saturation and lipoprotein particle size. Therefore, in the current study, we scaled all the metabolic variables to make them have equal importance in the hierarchical clustering.

The hierarchical clustering analysis revealed four clusters in the 119 hypercholesterolemic females with plasma LDL cholesterol above 99th percentile for their age. We hypothesized that mutations in genes belonging to the same metabolic pathway (e.g. the LDL receptor pathway) should render a similar plasma metabolome (one cluster). The principal component analysis revealed that plasma triglyceride and large LDL particle concentrations are the major discriminators for the four clusters. Since cluster 3 is characterized by a high concentration of large LDL particle and relatively low triglyceride in plasma, we hypothesized that this cluster represented the hypercholesterolemia due to defective LDL clearance. Incorporation of the genetic information provided us the verdict, because we expected the 20 subjects carrying a known functional heterozygous mutation in *LDLR* or

APOB to position in cluster 3. Indeed, 15 subjects fit this hypothesis and were located in cluster 3.

Then we came up with the question "Can we get insight if a novel variant in *LDLR* or *APOB* is the underlying cause for the severe hypercholesterolemia based on the metabolome profile?". Indeed, 6 out of 8 carriers of a novel mutation fit in cluster 3, suggesting potential effects of these variants on LDL receptor mediated uptake. This observation suggests that metabolic profiling is useful to delineate the subjects with a pathogenic mutation from those that do not carry any variant in either *LDLR* or *APOB*. However, not all subjects in cluster 3 do carry a variant in *LDLR* or *APOB*. We realize that the pathway of LDL receptor mediated endocytosis and intracellular cholesterol trafficking contains many more genes (Marques-Pinheiro et al. 2010; Bartuzi et al. 2016; Paththinige, Sirisena, and Dissanayake 2017) than we have sequenced in our cohort. So expansion of the number of genes on the chip or choosing whole genome sequencing will ultimately improve the information on all genes involved in the LDL receptor pathway and may thus help to identify additional genetic variants underlying the pathobiology in the remaining 40 females in cluster 3. Meanwhile, we cannot exclude other processes underlying the hypercholesterolemia such as epigenetic changes (Dekkers et al. 2016), lincRNA (Hu et al. 2014), microRNA (Irani et al. 2018) or combinations thereof.

Cluster 4 contained only one subject, and the individual had the highest large LDL particle concentration among the 119 hypercholesterolemic females. Interestingly, we did not identify any mutations in the sequenced genes including *LDLR*, *APOB* and *PCSK9*. This female subject was 28 years with BMI 21.7 kg/m^2. Her waist circumference was 69 centimeters. When we compared her plasma metabolomics data to the other 118 hypercholesterolemic females, we identified 77 outlier variables [either below the 1st quantile (1.5×interquartile range) or above the 3rd quantile (1.5×interquartile range). supplemental Table 3]. We noticed that this female had a high proportion of esterified cholesterol in VLDL and HDL particles compared to the remaining 118 subjects. Interestingly, the CETPtg/apoCI-/- mouse model showed a very similar phenotype (Gautier et al. 2002). Apolipoprotein C1 is an important regulator for CETP activity, which may partly underlie the observed phenotype (Pillois et al. 2012). So far no mutations in *APOC1* have been described.

A recent study (Lorenzo et al. 2018) showed that hypercholesterolemic subjects without

any known genetic defect had lower levels of LDL cholesterol than those with a mutation. Therefore, we hypothesized that the origin of the hypercholesterolemia in cluster 1 may be either polygenic or due to lifestyle factors. After additional analysis of relationships between the wGRS or lifestyle score and triglyceride or large LDL particle concentration, we observed that only genetic risk scores were negatively associated with triglyceride concentration (Kendall tau correlation coefficient -0.23, P value = 0.04). This observation suggests that this cluster of hypercholesterolemic subjects may be caused by less damaging mutations in genes involved in the LDL receptor pathway. The major observation in the subjects located in cluster 2 is that they had elevated plasma triglyceride. The genetic array used in the current study does not contain the genes involved in triglyceride metabolism. Our data suggest that generation of a triglyceride specific gene array may generate interesting results in the subjects in this cluster.

In summary, this study shows that bioinformatic analysis of metabolomics data derived from hypercholesterolemic subjects generates interesting clusters of patients that may help to guide targeted genomics approaches hypercholesterolemia.

3.6 Acknowledgments

We would like to thank all participants of the Lifelines study. This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. This work was supported by the Netherlands CardioVascular Research Initiative: “the Dutch Heart Foundation, Dutch Federation of University Medical Centers, the Netherlands Organization for Health Research and Development and the Royal Netherlands Academy of Sciences” (CVON2011-2016; Acronym Genius1 and CVON2017-2020; Acronym Genius2 to Dr Kuivenhoven), the European Union (FP7-603091; Acronym TransCard to Dr Kuivenhoven). Dr Kuivenhoven is Established Investigator of the Netherlands Heart Foundation (2015T068).

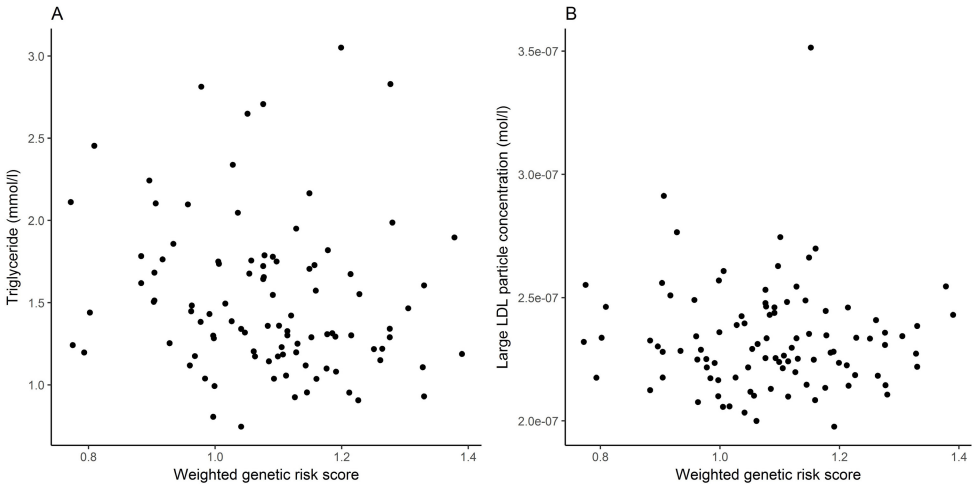


Figure 3.6: Scatterplots of weighted genetic risk score against triglyceride or large LDL particle concentration in 91 hypercholesterolemic females without canonical mutations.

3.7 Supplementary Material

The Supplementary tables for this article can be found online at: http://www.jlr.org/content/59/11/2174/suppl/DC1

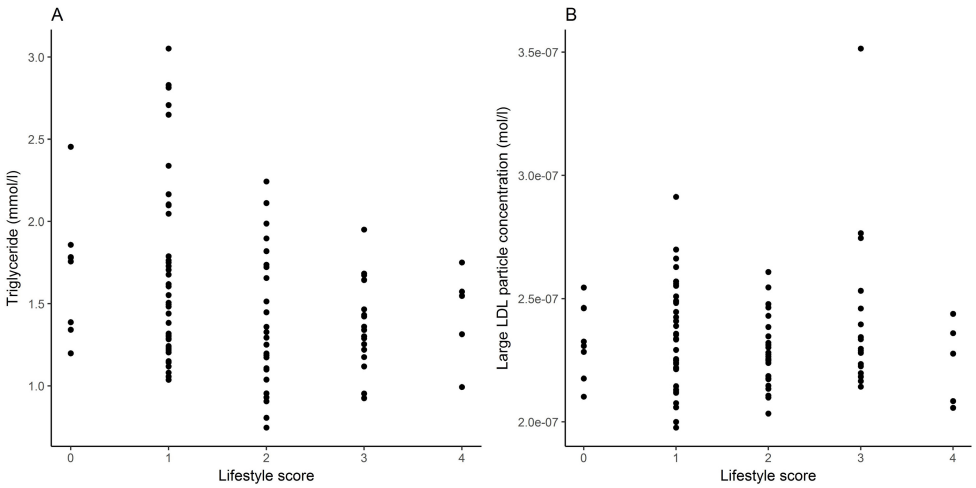


Figure 3.7: Scatterplots of weighted genetic risk score against triglyceride or large LDL particle concentration in 91 hypercholesterolemic females without canonical mutations.

4

Association of hemoglobin A1C with circulating metabolites in Dutch with European, African Surinamese and Ghanaian background

This chapter has been accepted by Nutrition & Diabetes

4.1 Abstract

Background: The prevalence of type 2 diabetes mellitus (T2DM) varies significantly across ethnic groups. A better understanding of the mechanisms underlying the variation in different ethnic groups may help to elucidate the pathophysiology of T2DM. The present work aims to generate a hypothesis regarding "why do subjects with African background have excess burden of T2DM?".

Methods: In the current study, we performed metabolite profiling of plasma samples derived from 773 subjects of three ethnic groups (Dutch with European, Ghanaian and African Surinamese background). We performed Bayesian lognormal regression analyses between HbA1c and circulating metabolites.

Results: We showed that subjects with Ghanaian and African Surinamese background had similar associations of HbA1c with circulating amino acids as subjects with European background. But subjects with Ghanaian and African Surinamese background had reversed association between HbA1c and acetoacetate, compared to the subjects with European background. Moreover, we also observed that HbA1c was associated with small HDL particles in subjects with African Surinamese background.

Conclusions: Based on the observations, we hypothesize that subjects with African background may have impaired cholesterol efflux capacity of HDL, linking to their excess burden of T2DM.

Keywords: metabolomics, diabetes mellitus, ethnicity, Bayesian, association

4.2 Introduction

The prevalence of type 2 diabetes mellitus (T2DM) increased rapidly worldwide during the past decades, and is strongly associated with the developing obesity pandemic (NCD Risk Factor Collaboration (NCD-RisC) 2016). Apart from the common risk factors that prevail in all populations, ethnic background is a risk factor as well (Maskarinec et al. 2009). In the Netherlands, subjects with a migration background showed a considerable higher incidence of T2DM compared to subjects with European background (Bindraban et al. 2008; Ujcic-Voortman et al. 2009). Among men, individuals with European,

Ghanaian and African Surinamese background had T2DM prevalence of 5.0%, 14.9% and 10.4%, respectively (Meeks et al. 2015). Among women, the prevalence is 2.3% (European), 11.1% (Ghanaian) and 11.5% (African Surinamese) (Meeks et al. 2015). The differences in T2DM prevalence across ethnic groups could not be explained by genetic variants alone (Waters et al. 2010). In contrast, other studies showed that the ethnic differences in T2DM prevalence were accompanied by differences in plasma amino acids and lipids. In line with this, compared to Europeans, the serum concentrations of isoleucine, phenylalanine, tyrosine and alanine were significantly higher in South Asians (Tillin et al. 2015). In another study (Valkengoed et al. 2017), individuals with Surinamese background were identified to have lower sphingolipids, but higher unsaturated acylcarnitines and higher amino acid levels, than Europeans.

Metabolite profiling, or metabolomics, has been widely applied to identify new biomarkers for T2DM (Roberts, Koulman, and Griffin 2014; Guasch-Ferré et al. 2016), predict T2DM risk (Wang et al. 2011; Rebholz et al. 2018) and improve our understanding of pathophysiologic mechanisms (Newgard et al. 2009; Newgard 2017). The most frequently observed metabolic abnormalities in insulin resistant and T2DM subjects include elevated circulating branched amino acids (BCAAs) and aromatic amino acids (AAAs) (Würtz, Mäkinen, et al. 2012; Würtz, Tiainen, et al. 2012; Würtz et al. 2013). A hypothesized mechanism linking elevated BCAAs to T2DM is that disturbed BCAA metabolism leads to accumulation of BCAA metabolites (e.g. 2-Aminoadipic acid), resulting in pancreatic β cell dysfunction (Lynch and Adams 2014; Wang et al. 2013). High triglyceride and low HDL cholesterol is another frequently observed metabolic abnormality in insulin resistance and T2DM subjects (Fizelova et al. 2015; Festa et al. 2005; Mackey et al. 2015; Wang et al. 2012; Garvey et al. 2003). Often dyslipidemias are viewed as consequences rather than cause of T2DM (Fizelova et al. 2015; Festa et al. 2005; Mackey et al. 2015). However, cholesterol homeostasis plays an important role in regulating pancreatic β cell function (Eckardstein and Sibler 2011; Fryirs et al. 2010; Kruit et al. 2011). Cholesterol is taken up by pancreatic β cells via the LDL receptor and exported back to plasma via the ATP-binding cassette transporter A1 (ABCA1) (Kruit et al. 2010). Accumulation of cholesterol in pancreatic β cells leads to impairment of glucose tolerance and defective insulin secretion (Kruit et al. 2010; Eckardstein and Sibler 2011; Kruit et al. 2011). The present work aimed to generate a hypothesis regarding "why do subjects with African

background have excess burden of T2DM?". As an initial step, we investigated whether the relationship between circulating metabolites and glucose tolerance varies depending on ethnicity.

The current study focused on Dutch with European, Ghanaian and African Surinamese background. Here we used the hemoglobin A1C (HbA1c) level as the surrogate of glucose tolerance, since HbA1c is an index of chronic glycemia and a predictor of T2DM (Nathan et al. 1984; Diabetes Prevention Program Research Group 2015). By running Bayesian lognormal regression analyses, we show that compared to the European origin, Dutch with Ghanaian and African Surinamese background have similar associations of HbA1c with circulating amino acids, but reversed association between HbA1c and concentration of acetoacetate. We also observed that HbA1c was associated with small HDL particles in subjects with African Surinamese background. We hypothesized that subjects with African background may have impaired cholesterol efflux capacity of HDL, linking to their excess burden of T2DM.

4.3 Materials and Methods

4.3.1 Study population

The study was composed of three ethnic groups in the Dutch population. In particular, 217 African Surinamese and 255 Ghanaian were from the HELIUS study (K. Stronks et al. 2013a; Snijder et al. 2017), and 301 European Dutch were from the 300-obesity cohort from the Human Functional Genomics Project (Netea et al. 2016). The HELIUS study was complied with all relevant ethical regulations and in accordance with the Declaration of Helsinki (6th, 7th revisions); it was approved by the Academic Medical Center (AMC) Medical Ethics Committee and all participants provided written informed consent.

4.3.2 Metabolite profiling

Fasting plasma samples were collected in the clinic and stored at $-80°$C. Quantification of 8 amino acids, 2 ketone bodies and 14 lipoproteins was performed by using a high-throughput NMR metabolomics platform (Nightingale, Austria) (Inouye et al. 2010).

The following 14 lipoprotein subclasses were quantified: extremely large (average particle diameter >75 nm), very large (average particle diameter 64.0 nm), large (53.6 nm), medium (44.5 nm), small (36.8 nm) and very small VLDL (31.3 nm). intermediate-density lipoprotein (IDL; 28.6 nm); three LDL subclasses, i.e. large (25.5 nm), medium (23.0 nm) and small LDL (18.7 nm); and four HDL subclasses, i.e. very large (14.3 nm), large (12.1 nm), medium (10.9 nm) and small HDL (8.7 nm). The following components of the lipoprotein particles were quantified: phospholipids (PL), triglycerides (TG), cholesterol, free cholesterol (FC) and cholesterol esters (CE).

4.3.3 Statistical analysis

Because this study contains three ethnic groups from two different cohort studies with different time of sampling and measurement, we cannot directly compare metabolite variables between ethnic groups. As an alternative, we performed association analyses between HbA1c and circulating metabolites within each ethnic group. The outcome variable (y) was concentration of a metabolite, such as amino acids, ketone bodies, or lipoprotein particles. The predictor variable (x) was the HbA1c level. To assess the strength of associations between HbA1c and metabolites, we ran lognormal regression because the outcome variables are positive continuous data with skewed distributions. To study the dependency of ethnicity on the relationship between HbA1c and metabolic variables, we introduced ethnicity-specific intercepts and slopes into the model. We also adjusted for covariates including gender and age. We centered and scaled HbA1c and age by subtracting their mean values and dividing by their standard deviations. After scaling, one unit HbA1c means 10 mmol/mol and one unit age means 10 (years). Due to missing observations in both outcome and predictor variables, we applied a Bayesian lognormal regression to handle the missing data. There are two types of missing values: 1) when the concentration of a metabolite is below the limit of detection, or 2) when values were rejected by the automatic sample and measurement quality control procedure in the Nightingale pipeline. All the missing observations were assumed missing at random and treated as parameters. Values were randomly drawn from a normal distribution with ethnicity-dependent mean and standard deviation. Based on a previous study (Dekker et al. 2015) the mean value and standard deviation of HbA1c is about 40 mmol/mol and 5 mmol/mol, therefore, we

used $exponential(0.025)$ and $exponential(0.2)$ as prior distributions. Regarding the missing values that were below the limit of detection, the imputed values were constrained between zero and the minimal observed value. For the rest parameters, we used default prior distributions and fitted the model by running Hamiltonian Markov Chain Monte Carlo in the program Stan (version 2.18.0) (Carpenter et al. 2017). The detailed model is given below:

$$y[i] \sim Lognormal(\mu[i], \sigma) \tag{4.1}$$

$$x[i] \sim Normal(\mu_{HbA1c,ethnicity[i]}, \sigma_{HbA1c,ethnicity[i]}) \tag{4.2}$$

$$zx[i] = \frac{x[i] - 40}{10} \tag{4.3}$$

$$zAge[i] = \frac{Age[i] - 59}{10} \tag{4.4}$$

$$\mu[i] = A[i] + B_H[i] \times zx[i] + \beta_A \times zAge[i] \tag{4.5}$$

$$A[i] = \alpha + \alpha_{ethnicity[i]} + \alpha_{gender[i]} \tag{4.6}$$

$$B_H[i] = \beta_H + \beta_{H,ethnicity[i]} + \beta_{H,gender[i]} \tag{4.7}$$

$$\sigma \sim Exponential(1) \tag{4.8}$$

$$\mu_{HbA1c,ethnicity[i]} \sim Exponential(0.025) \tag{4.9}$$

$$\sigma_{HbA1c} \sim Exponential(0.2) \tag{4.10}$$

$$\beta_A \sim Normal(0, 10) \tag{4.11}$$

$$\alpha \sim Normal(0, 10) \tag{4.12}$$

$$\alpha_{ethnicity[i]} \sim Normal(0, 0.1) \tag{4.13}$$

$$\alpha_{gender[i]} \sim Normal(0, 0.1) \tag{4.14}$$

$$\beta_H \sim Normal(0, 10) \tag{4.15}$$

$$\beta_{H,ethnicity[i]} \sim Normal(0, 0.1) \tag{4.16}$$

$$\beta_{H,gender[i]} \sim Normal(0, 0.1) \tag{4.17}$$

We ran four Markov chains with 4000 iterations in each chain. Results were presented with the posterior mean of lognormal regression coefficient with 95% credible interval (CI). The regression coefficient represents the expected difference in log(metabolite concentration)

due to a difference of one unit HbA1c.

4.3.4 Availability of Data

The metabolomics and clinical data of subjects with Ghanaian and African Surinamese background are available by submitting a proposal to the HELIUS Executive Board as outlined at http://www.heliusstudy.nl/en/researchers/ collaboration. Requests for further information and proposals can be submitted to Marieke Snijder. The metabolomics and clinical data of subjects with European background are available by contacting human functional genomics project www.humanfunctionalgenomics.org.

4.3.5 Code availability

The model file and analysis code are available at https://github.com/XiangZhangSC/nutrition_and_diabetes_paper.

4.4 Result

4.4.1 Participant characteristics

This study included in total 773 subjects from three ethnic groups in the Dutch population. Specifically, the study population consisted of 301 European Dutch, 255 Dutch with Ghanaian background, and 217 Dutch with African Surinamese background (Table 4.1). Dutch with European background were older than the other two ethnic groups. There were relatively more male participants in the Dutch with European background, and in the Dutch with African Surinamese background, there were relatively more female participants. To control for these possible confounding factors, all the results shown below were after adjusting for gender and age.

4.4.2 Association of HbA1c with circulating amino acids

Since circulating amino acids are robust markers of T2DM, we first evaluated their associations with HbA1c in Dutch with European, Ghanaian and African Surinamese back-

Table 4.1: Characteristics of 773 participants with different ethnic backgrounds

	European	Ghanaian	African Surinamese
N	301	255	217
Female (%)	44.5	52.5	65.9
Age (years)	67.1 ± 5.4	51.2 ± 8.3	55.2 ± 7.2
BMI	30.7 ± 3.5	29.6 ± 4.4	31.2 ± 5.8
Waist circumference (cm)	107.0 ± 9.8	98.5 ± 10.6	103 ± 13.4
HbA1c (mmol/mol)	41.7 ± 7.9	56.1 ± 19.7	46.6 ± 15.4

ground. We observed that in Dutch with European background increasing HbA1c concentrations were associated with increasing concentrations of circulating isoleucine (regression coefficient in males 0.14 with 95% credible interval [0.10 0.18], in females 0.15 [0.10 0.19]), leucine (males 0.07 [0.04 0.10], females 0.07 [0.04 0.10]), valine (males 0.06 [0.03 0.09], females 0.07 [0.04 0.10]) and alanine (males 0.07 [0.04 0.11], females 0.08 [0.05 0.12]), as well as decreasing levels of glutamine (males -0.07 [-0.10 -0.03], females -0.08 [-0.12 -0.04]) (Figure 4.1).

In Dutch with Ghanaian background, we observed that increasing HbA1c concentrations were associated with increasing concentrations of circulating isoleucine (females 0.03 [0.01 0.06]), and valine (females 0.03 [0.01 0.04]), as well as decreasing levels of glutamine (females -0.04 [-0.06 -0.01]) (Figure 4.1).

In Dutch with African Surinamese background, we observed that increasing HbA1c concentrations were associated with increasing concentrations of circulating isoleucine (males 0.03 [0.003 0.06], females 0.04 [0.02 0.06]), leucine (males 0.02 [0.01 0.04], females 0.03 [0.01 0.04]), and valine (males 0.03 [0.02 0.05]), female 0.08 [0.05 0.12]), as well as decreasing levels of glutamine (males -0.02 [-0.05 -0.003], females -0.04 [-0.06 -0.02]) (Figure 4.1).

4.4.3 Association of HbA1c with circulating ketone bodies

In the next step, we assessed the association between HbA1c and ketone bodies. We observed that increasing levels of HbA1c were associated with decreasing levels of acetoacetate in subjects with European background (males -0.11 [-0.21 -0.01], females -0.15 [-0.25 -0.05]). However, increasing levels of HbA1c were associated with increasing levels of acetoacetate in subjects with Ghanaian (males 0.08 [0.02 0.15]) and African Surinamese

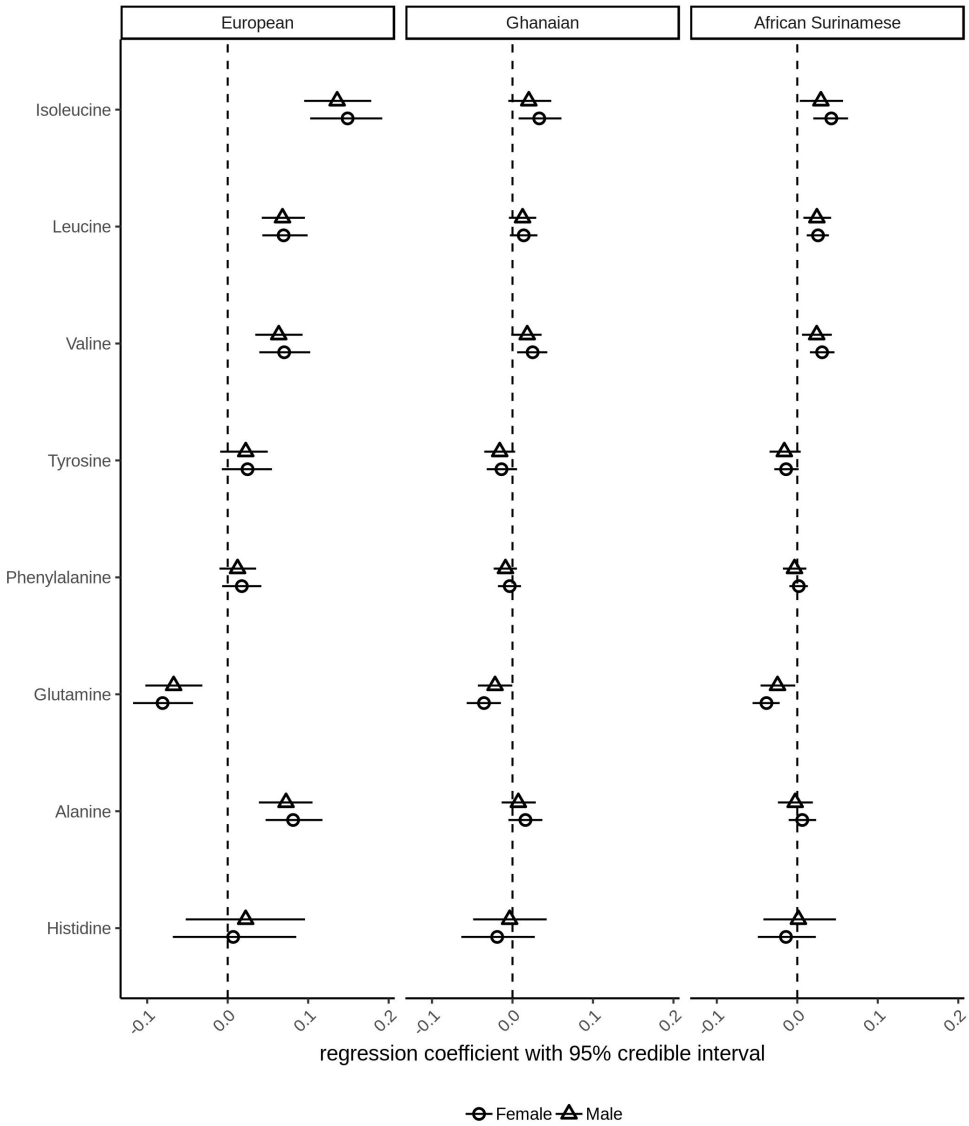


Figure 4.1: Regression parameter estimates between plasma amino acids and HbA1c in subjects with European, Ghanaian and African Surinamese background. Circles (female) or triangles (male) and horizontal lines represent the posterior means of the regression coefficient between plasma amino acids and HbA1c and 95% credible intervals.

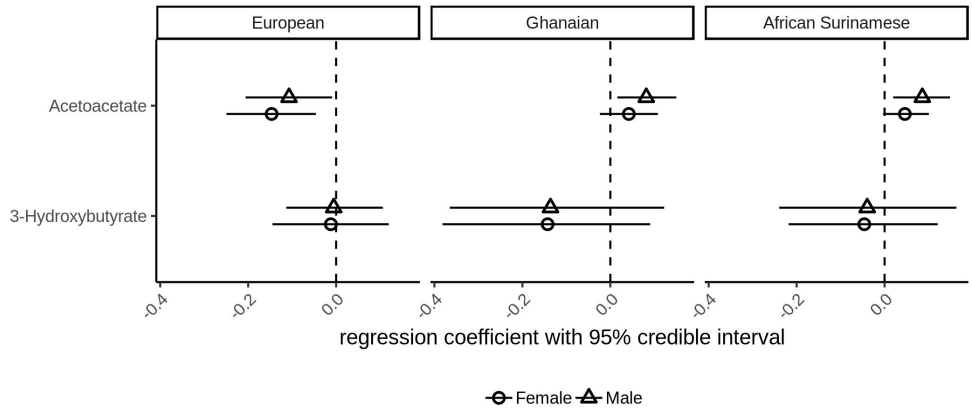


Figure 4.2: Regression parameter estimates between plasma ketone bodies and HbA1c in subjects with European, Ghanaian and African Surinamese background. Circles (female) or triangles (male) and horizontal lines represent the posterior means of the regression coefficient between plasma ketone bodies and HbA1c and 95% credible intervals.

(males 0.09 [0.02 0.15]) background (Figure 4.2).

4.4.4 Association of HbA1c with circulating HDL particles

In subjects with European background, we observed that increasing levels of HbA1c were associated with decreasing concentrations of large (males -0.18 [-0.25 -0.12], females -0.16 [-0.23 -0.09]) and medium (males -0.04 [-0.07 -0.01], females -0.04 [-0.07 -0.01]) HDL particles (Figure 4.3).

In subjects with African Surinamese background, however, we observed that increasing levels of HbA1c were associated with increasing concentration of small HDL particle (males 0.01 [0.0004 0.02], females 0.01 [0.004 0.02]) (Figure 4.3).

4.5 Discussion

"Why do subjects with African background have excess burden of T2DM?" To answer this question, we performed metabolite profiling in plasma of 773 subjects from three ethnic groups in the Dutch population (Dutch with European, Ghanaian and African Surinamese background). Consistent with a recent meta-analysis (Guasch-Ferré et al.

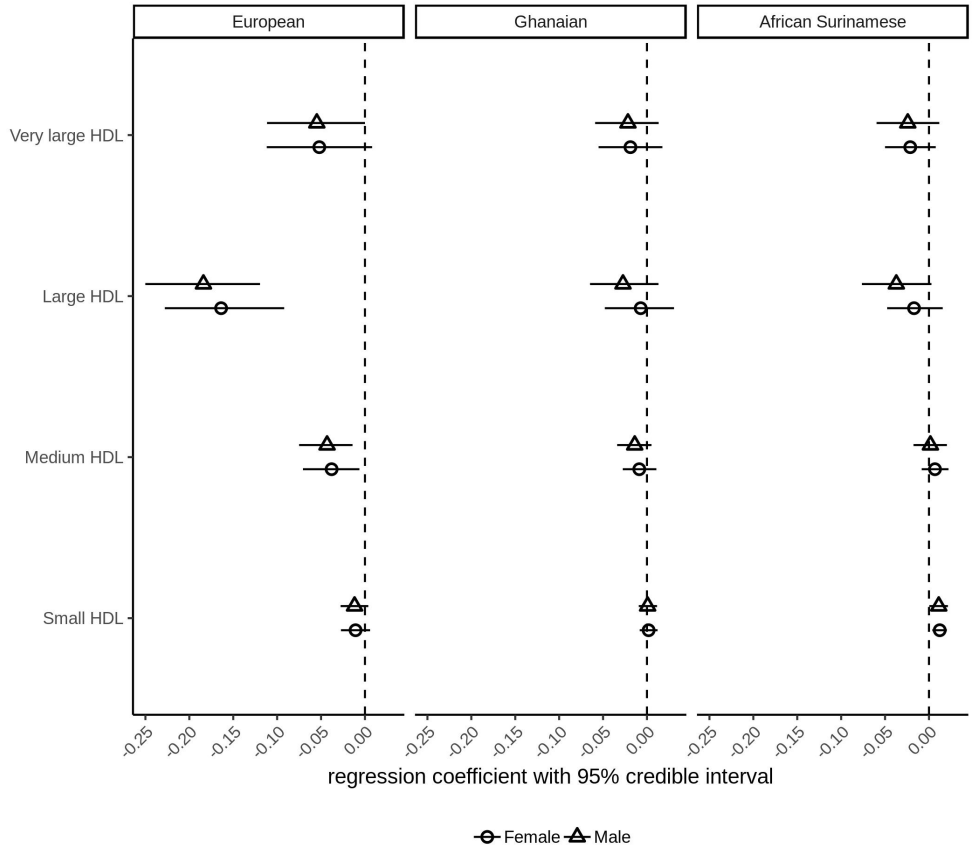


Figure 4.3: Regression parameter estimates between plasma HDL particle concentration and HbA1c in subjects with European, Ghanaian and African Surinamese background. Circles (female) or triangles (male) and horizontal lines represent the posterior means of the regression coefficient between plasma HDL particle concentration and HbA1c and 95% credible intervals.

2016), we observed that increasing levels of BCAAs, as well as decreasing levels of glutamine were associated with worsening of glycemia in Dutch with European background. We observed similar associations of HbA1c with circulating amino acids in subjects with Ghanaian and African Surinamese background. These observations suggested that amino acid metabolism was similar in subjects with European and African background. However, we observed that increasing levels of HbA1c were associated with increasing levels of acetoacetate in Dutch with Ghanaian and African Surinamese background, but associated with decreasing levels of acetoacetate in Dutch with European background. In liver, acetoacetate can be produced from tyrosine (KNOX and LeMAY-KNOX 1951). Tyrosine was repeatedly identified to be associated with glycemia and insulin resistance in the general population (Würtz, Tiainen, et al. 2012; Würtz, Mäkinen, et al. 2012; Würtz et al. 2013; Hellmuth et al. 2016). In South Asian men, tyrosine was identified as a strong predictor of T2DM incidence (Tillin et al. 2015). Interestingly, we observed a high probability (93% in males and 94% in females) of positive association between HbA1c and tyrosine in Dutch with European background. In contrast, we observed a high probability (95% in males and 92% in females) of negative association between HbA1c and tyrosine in Dutch with Ghanaian background. In Dutch with African Surinamese background, we also observed a high probability (95% in males and 96% in females) of negative association between HbA1c and tyrosine. Acetoacetate was shown to strongly inhibit expression of ABCA1, a key player mediating cellular cholesterol efflux (Uehara et al. 2002; Phillips 2014). Inactivation of ABCA1 leads to cholesterol accumulation in the pancreatic β cells and impaired glucose tolerance (Brunham et al. 2007; Vergeer et al. 2010; Kruit et al. 2011). The main acceptors of ABCA1-mediated cholesterol efflux are small HDL particles (Du et al. 2015). Interestingly, we observed that increasing concentrations of small HDL particles tended to associate with worsening of glycemia in subjects with African Surinamese background. Taking all these considerations together, we speculate that the excess burden of T2DM in subjects with African Surinamese and Ghanaian background might be due to impaired cholesterol efflux capacity of HDL caused by acetoacetate-induced inhibition of ABCA1 gene expression. Future research is needed to test our hypothesis.

One limitation of this study is that the three ethnic groups were from two different cohort studies, with different time of sampling and measurement. As a consequence, we cannot directly compare metabolite variables between ethnic groups but did run association

analyses within each ethnic group.

In conclusion, subjects with Ghanaian and African Surinamese background showed reversed associations of HbA1c with circulating acetoacetate, compared to Dutch with European background. We hypothesized that acetoacetate-induced inhibition of ABCA1 gene expression may link to impaired cholesterol efflux capacity to HDL in the ethnic groups of African background.

4.6 Acknowledgments

This work was supported by grants CVON-Genius (CVON2011-19) and RESOLVE (FP7 305707).

4.7 Conflict of Interest

There are no competing financial interests in relation to the work described.

5 Statistical evaluation of diet-microbe associations

This chapter has been accepted by BMC Microbiology

5.1 Abstract

Background: Statistical evaluation of the association between microbial abundance and dietary variables can be done in various ways. Currently, there is no consensus on which methods are to be preferred in which circumstances. Application of particular methods seems to be based on the tradition of a particular research group, availability of experience with particular software, or dependent on the outcomes of the analysis.

Results: We applied four popular methods including edgeR, limma, metagenomeSeq and shotgunFunctionalizeR, to evaluate the association between dietary variables and abundance of microbes. We found large difference in results between the methods. Our simulation studies revealed that no single method was optimal.

Conclusions: We advise researchers to run multiple analyses and focus on the significant findings identified by multiple methods in order to achieve a better control of false discovery rate.

Keywords: microbiome; diet; association; simulation; sequencing

5.2 Background

With the help of high-throughput sequencing technologies, human microbiota have been profiled and studied extensively (Duvallet et al. 2017). Since diet shapes the composition of human microbiota and influences human health, linking abundance of microbes to dietary variables is a common practice in human microbiome studies (Wu et al. 2011; Deschasaux et al. 2018). These association studies not only can improve our understanding of the relationships between the human microbiome and nutrient intake, but also may help development of new therapeutic interventions.

Microbiome data are often generated by targeted sequencing of the 16S ribosomal RNA (rRNA) gene, and represented as a frequency matrix giving the number of times each microbe is observed in each sample. In general microbiome data have following features: 1) library sizes can vary by orders of magnitude across samples. 2) microbiome data often have excess zero counts. These zero counts can be due to either biological absence of a microbe, or insufficient sequencing. 3) microbiome data are compositional data, meaning

that the obtained counts do not reflect the absolute number of microbes that are present. 4) microbiome data are over-dispersed, characterized as some taxa (e.g., *Bacteroides* and *Lactobacillus* species) are common among samples, many other taxa are present at much lower abundances.

various statistical methods have been developed for microbiome data analysis, but there are no standard procedures to perform association analyses (Xia and Sun 2017). Previous benchmark works (Thorsen et al. 2016; Jonsson et al. 2016) focused on case-control studies, and revealed that the choice of statistical methods considerably affected outcomes of differential relative abundance tests. Unlike case-control studies, association studies work also on continuous variables. To our best knowledge, the influence of choosing different methods on outcomes of association studies has not been evaluated. To assess the influence, we analyzed the associations between dietary variables and gut microbiota in 1090 individuals from the HELIUS-cohort study (Amsterdam, the Netherlands) (K. Stronks et al. 2013b; Vermeulen et al. 2017). Since the focus of the current work is on robustness of the statistical results rather than biological or epidemiological associations, biological interpretation of diet-microbe associations is out of the scope of this work. We used four methods including those based on Poisson (shotgunFunctionalizeR), negative binomial (edgeR), zero-inflated Gaussian (metagenomeSeq) distributions, as well as a weighted linear regression model (voom + limma). We compared the results derived from the four methods and observed large differences. To find out which method we should choose in which circumstances, we ran simulation studies and found that no single method was optimal for all microbiome data sets. Therefore, we advise researchers to run multiple statistical analyses and focus on the significant findings identified by multiple methods in order to achieve better control of false discovery rate.

5.3 Methods

5.3.1 Subjects and HELIUS cohort

Subjects were participants in the HEalthy Life in an Urban Setting (HELIUS) cohort study. This study used a stratified-random sampling approach to include between 2011 and 2015 25,000 inhabitants (18-70 years) from the city of Amsterdam, the Netherlands

(K. Stronks et al. 2013b). Stratification was done on six subgroups with different ethnic origins (African Surinamese, South Asian Surinamese, Ghanaian, Turkish, Moroccan, and Dutch). Subgroups were about equally large.

5.3.2 Dietary intakes assessment

As described previously (Dekker et al. 2011; Beukers et al. 2015), a subsample of voluntary participants of Dutch, Moroccan, Turkish, South-Asian Surinamese and African Surinamese origin were asked to participate in the HELIUS-Dietary Patterns study, with objective to collect detailed information on their diet. Usual dietary intakes were assessed through the completion of ethnic-specific semi-quantitative food frequency questionnaires (FFQs) with a reference period of 4 weeks. These FFQs were adapted from an existing Dutch FFQ and comprised about 200 items. Food items were collapsed into 73 food groups based on similarity in nutrient profile and culinary use. In this study ethnic-specific food groups were not included in this analysis and 67 food items were used for the analyses.

5.3.3 16S processing

We used the 16S ribosomal RNA (rRNA) sequencing data generated in a previous study based on the HELIUS cohort (Deschasaux et al. 2018). In short, the composition of fecal microbiota was profiled by sequencing the V4 region of the 16S rRNA gene on a MiSeq system. The 16S rRNA gene reads were processed on a mothur pipeline (version 1.39.5) (Schloss et al. 2009). The OTU clustering was done by using the vsearch (version 2.6) (Rognes et al. 2016) and FastTree 2.1 (Price, Dehal, and Arkin 2010). The details of the sequencing and bioinformatic pipelines were described in (Deschasaux et al. 2018).

5.3.4 Statistical analyses

Our analysis is based on 1090 subjects who had both fecal microbiome and FFQ data. Following (Duvallet et al. 2017), here we removed OTUs with fewer than 10 reads in total, as well as OTUs which were present in fewer than 1% of samples. The final OTU table contains 1090 samples and 2073 OTUs. We used four widely used methods for sequencing data analysis to quantify the strength of the associations between dietary variables (x)

and OTU counts (y). Because the large number of associations (67×2073), we used multidplyr R package (https://github.com/hadley/multidplyr) for parallel computation. The selected methods were as follows:

shotgunFunctionalizeR (Kristiansson, Hugenholtz, and Dalevi 2009) is a popular R package used in microbiome research community, and based on the Poisson generalized linear model (implemented in glm function in R). We used the glm function with log(total counts) as offset to quantify associations between dietary variables and OTU counts.

Negative binomial model, also called gamma-Poisson model, is popular for statistical modeling of OTU count data (McMurdie and Holmes 2014; Zhang et al. 2017). Phyloseq (McMurdie and Holmes 2013) is a popular R package used by the microbiome research community. The core of Phyloseq is based on another popular R package DESeq2 (Love, Huber, and Anders 2014), which is based on negative binomial model. However, when the sample size is big (above 100), the computation becomes slow in DESeq2. Therefore, in this work we used another negative binomial based R package, edgeR (Robinson, McCarthy, and Smyth 2010). The observed OTU count was modeled by a negative binomial distribution with two parameters, the mean and the dispersion. OTU specific dispersion was estimated by running estimateDisp function (Chen, Lun, and Smyth 2014) implemented in the edgeR package (Robinson, McCarthy, and Smyth 2010). The associations between dietary variables and OTU counts were quantified by running glmFit function of the edgeR package (Robinson, McCarthy, and Smyth 2010). The log(total counts) was used as offset.

In contrast to above methods modeling the counts with exact probabilistic distributions, others have advocated weighted linear regression analysis with precision weights derived from the mean variance relationship (Law et al. 2014). This approach has been implemented in the voom function of the popular R package limma (Ritchie et al. 2015). The weighted linear regression was done to estimate the linear association between dietary variables and OTU counts with precision weights estimated by the voom (Law et al. 2014) and lmFit functions in the limma package.

The last method, metagenomeSeq (Paulson et al. 2013) is also a popular R package used by microbiome research community. It is based on the zero-inflated Gaussian model. This approach has been implemented in the fitZig function of the popular R package

metagenomeSeq (Paulson et al. 2013). The cumulative-sum scaling method was used to take care library size difference.

In a typical association study, the primary goal is to identify some candidate associations for future research. Therefore, regarding multiple testing we calculated false discovery rate (FDR). If an association had FDR value below 0.05, we considered it as a significant association. Since the research question is focused only on robustness of the statistical results and not on biological or epidemiological associations, we did not adjust for possible confounding or selection factors.

5.3.5 Simulation framework

We use y to represent the simulated microbiome data with n rows and J columns. Every column of y represents a microbe and every row of y represents a subject. Here, we simulated associations of a dietary variable, denoted as x, with gut microbiota. x is a vector of length n, and was randomly sampled from real FFQ data with replacement. The FFQ data was published in (Wu et al. 2011) and contained 214 dietary variables that were scaled to having mean 0 and standard deviation 1. For each simulated microbiome data, we used one dietary variable and in total generated 214 simulated data sets. Our simulation framework included the steps below:

$$\eta[j] \sim Bernoulli(0.5) \tag{5.1}$$

$$\gamma[j] \sim T_7(0, 2.5) \tag{5.2}$$

$$\beta[j] = (1 - \eta[j]) \times 0 + \eta[j] \times \gamma[j] \tag{5.3}$$

$$\theta[i, 1:J] \sim Dirichlet(\pi[1:J]) \tag{5.4}$$

$$\alpha[i, 1:J] = logit(\theta[i, 1:J]) \tag{5.5}$$

$$logit(\mu[i, j]) = \alpha[i, j] + \beta[j] \times x[i] \tag{5.6}$$

$$N[i] \sim Lognormal(\mu_L, \sigma_L) \tag{5.7}$$

$$y[i, 1:J] \sim Multinomial(N[i], \mu[i, 1:J]) \tag{5.8}$$

1. The indicator variable, $\eta[j]$, indicates if a dietary variable influences the abundance of the microbe j. For microbe j, we randomly drew $\eta[j]$ from a Bernoulli distribution

with parameter 0.5 (Equation (5.1)).

2. $\gamma[j]$ represents the effect of the dietary variable on the abundance for OTU j, and was sampled from a t distribution with 7 degrees of freedom, location 0 and scale 2.5 (Gelman et al. 2008) (Equation (5.2)).

3. Then the true association between the diet and microbe j was captured by $\beta[j]$ defined in Equation (5.3).

4. The matrix θ has n rows and J columns. $\theta[i, j]$ corresponds to the baseline abundance level for the microbe j in subject i. For subject i, we randomly drew a vector of length J from a Dirichlet distribution (Equation (5.4)).

5. The parameter of the Dirichlet distribution π is a vector of length J. We used R package DirichletMultinomial (Holmes, Harris, and Quince 2012) and the Human Microbiome Project 16S rRNA stool data (Schiffer et al. 2018) to estimate the π.

6. The true microbe j proportion in subject i, $\mu[i, j]$ was modeled as a logistic regression of $x[i]$ (Equation (5.6)).

7. Similar to (Paulson et al. 2013), library size of subject i, $N[i]$, was randomly drawn from a lognormal distribution with mean μ_L and standard deviation σ_L. μ_L is the logarithm of target sequencing depth (Equation (5.7)). We estimated σ_L based on the HMP stool 16S rRNA data set by using the fitdistr function implemented in the MASS package.

8. Finally, for subject i, the observed microbe counts were randomly generated from a multinomial distribution (Equation (5.8)).

Our HELIUS microbiome data set had 1090 subjects and the median sequencing depth was about 50,000. To mimic HELIUS data, we simulated the 16S microbiome data sets, with each data set having 1000 subjects and mean of sequencing depth 50,000. Performance metrics included true positive rate, false positive rate and error probability for identifying a significant association between microbe and dietary variable. They are calculated per simulation and defined as below:

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.9}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{5.10}$$

$$\text{Error probability} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{5.11}$$

TP, FP, TN and FN refer to true positive, false positive, true negative and false negative, respectively. A true positive finding is defined as having a significant estimated $\beta \neq 0$ (Equation (5.6)) in case the true $\beta \neq 0$. A false positive finding is defined as having a significant estimated $\beta \neq 0$ (Equation (5.6)) in case the true $\beta = 0$. The error probability quantified the probability that a significant association is false. Here we did not use "false discovery rate" but used the term "error probability" in order to avoid confusion, because we also calculated the false discovery rate during analyses of associations between OTUs and dietary variables.

5.4 Results

5.4.1 Large difference in results between statistical analyses

To evaluate effect of choosing different methods on outcomes in association studies, we performed association analyses between 67 dietary variables and 2073 OTUs derived from 1090 HELIUS participants with four methods. Out of 138,891 association tests, we identified 3,535, 20,081, 62,581 and 71,371 associations with FDR below 0.05 in edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR, respectively. There were 1,296 associations identified to be significant by all the four methods. In addition, there were 14, 3,703, 23,666, and 29,327 associations that were identified as significant only by edgeR, voom + limma, metagenomeSeq or shotgunFunctionalizeR (Figure 5.1).

5.4.2 16S rRNA microbiome data simulation

After realizing such considerably different results between the methods, we attempted to find out which method we should choose. To this end, we simulated 16S rRNA micro-

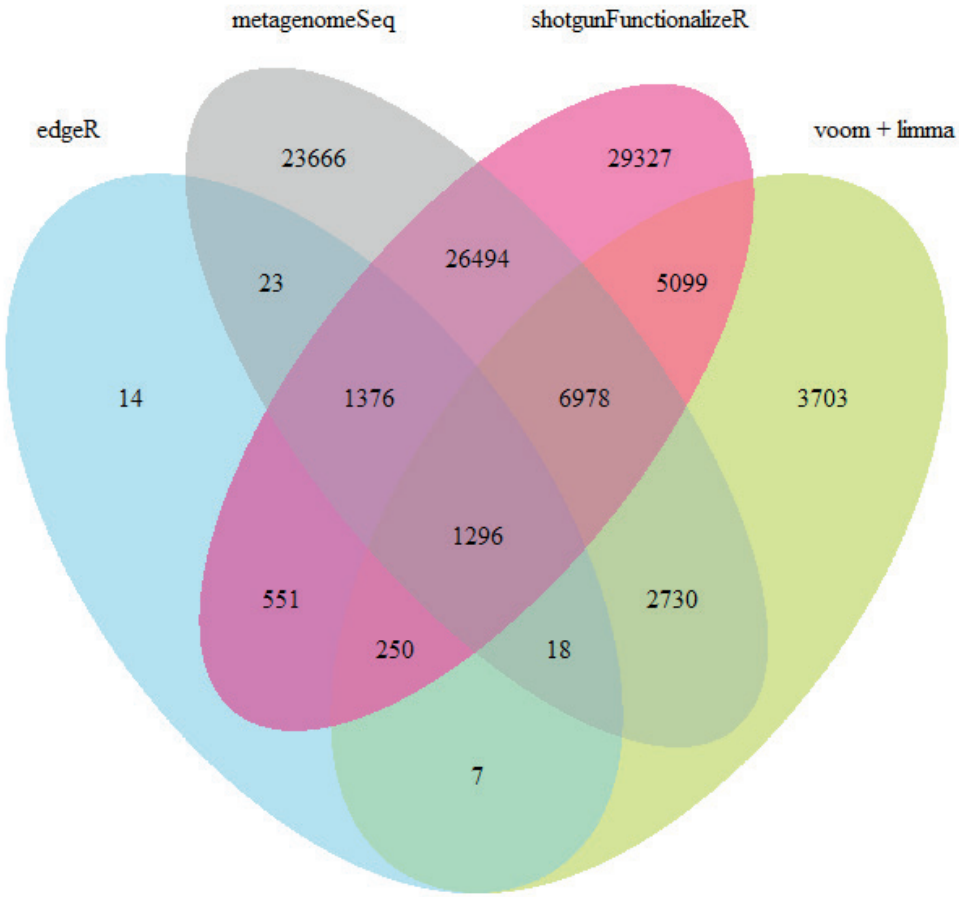


Figure 5.1: Venn diagram of significant associations identified by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR based on HELIUS 16S rRNA microbiome and FFQ data.

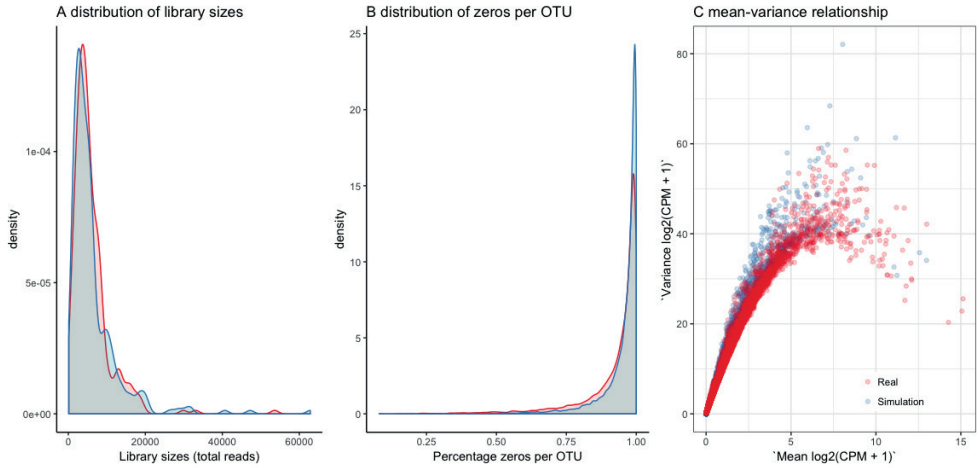


Figure 5.2: Comparison of simulated and real Human Microbiome Project stool 16S rRNA data. A. library size distributions B. distribution of percentage of zeros per OTU. C. mean-variance relationship. Every dot represents an OTU.

biome data with spiked-in associations between dietary variables and OTUs. We used a published FFQs (food frequency questionnaires) data as a template. To make sure our simulation framework can generate similar microbiome data as real ones, we compared our simulated data to the real HMP (Human Metabolome Project) stool 16S data. Our simulated microbiome data had similar distribution of library sizes and percentage of zeros per OTU, as well as similar mean-variance relationship (Figure 5.2). Our template FFQs data contained 214 dietary variables. In each simulation, we used one dietary variable. Therefore, in total we generated 214 simulated 16S rRNA data sets. Each data set contained 1000 subjects and had mean library size 50,000, and the same simulated data set was analyzed by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR. In our simulations, we observed large difference in results between the methods (Figure 5.3).

5.4.3 Method comparisons based on simulated data

Overall shotgunFunctionalizeR had both the highest true positive rate and the highest false positive rate (Figure 5.4). The median true positive rate of shotgunFunctionalizeR was (0.900), followed by metagenomeSeq (0.800), edgeR (0.624) and limma (0.519). Mean-

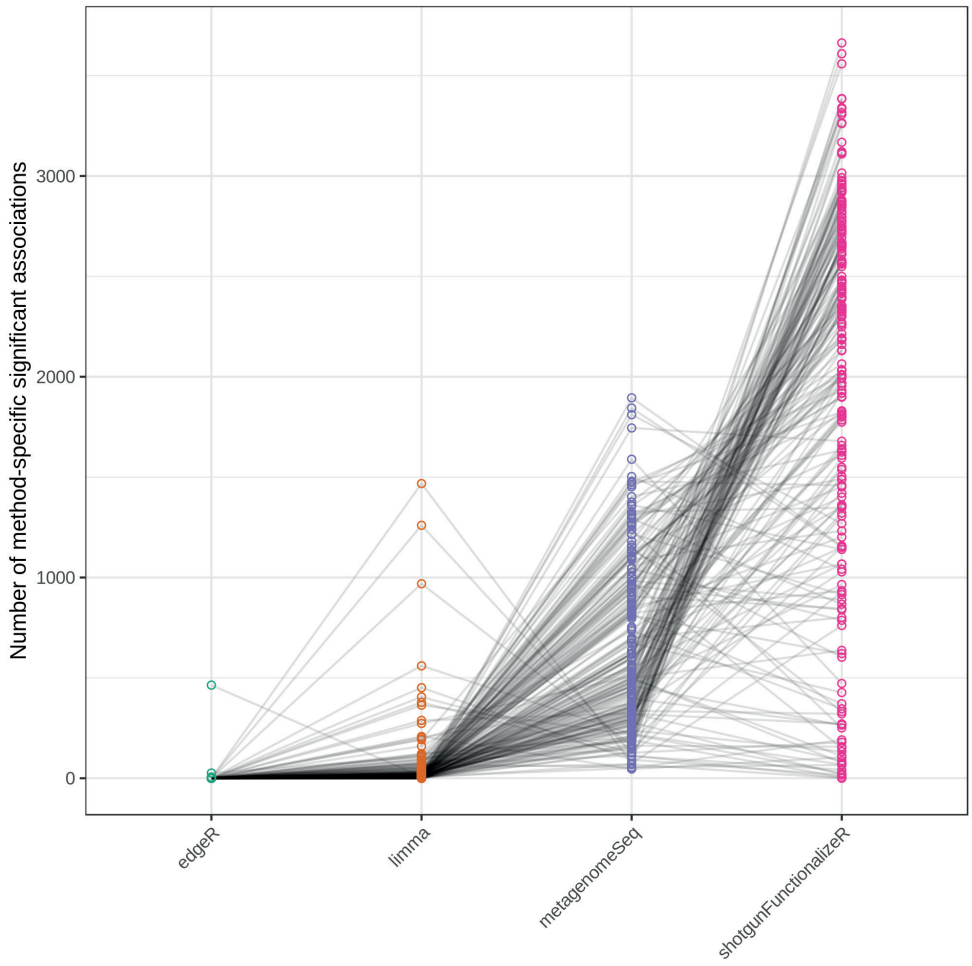


Figure 5.3: Every dot represents the number of significant associations identified only by the corresponding method. Each line represents a simulation, in which the same simulated data were analyzed by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR.

while the median false positive rate of shotgunFunctionalizeR, metagenomeSeq, limma, and edgeR were 0.716, 0.388, 0.125 and 0.0898, respectively. Depending on the simulated data set, the probability that a significant association is false (error probability) varied within each method over the 214 simulations (Figure 5.5A). With the same simulated data set, different methods showed different error probabilities (Figure 5.5B). Furthermore, the error probabilities in different methods were also influenced by the skewness of the distribution of the dietary variables (Figure 5.6). In the next step, we identified that 30% simulations in edgeR, 16% simulations in limma, 0.9% simulations in metagenomeSeq and 0% simulation in shotgunFunctionalizeR had error probabilities below 0.05 over the 214 simulations (Figure 5.7). However, when we focused on the significant associations that were identified by all four methods (we call them "overlap") in each simulation, we observed that 44% simulations had error probabilities below 0.05 over the 214 simulations (Figure 5.7).

5.5 Discussion

We learned from these relatively simple analyses that a key issue in the analysis of 16S rRNA microbiome data is the choice of the statistical method. Depending on the choice of statistical method, significant associations between dietary variables and microbial abundances varied dramatically. We observed that shotgunFunctionalizeR produced the largest number of unique significant associations, whereas most of the significant associations identified by edgeR were also identified by other methods. What really puzzled us is the relatively small number of significant associations identified by all methods. To find out which method we should choose for association studies, we developed a hierarchical model to simulate 16S rRNA data based on dietary variables with spiked-in associations. By comparing to the real HMP 16S microbiome data, we have shown that our simulation model can simulate realistic 16S rRNA microbiome data. Although in this work we focused on diet-microbe association analyses, our simulation framework can easily be adapted to simulate other scenarios.

Based on our simulation model, we simulated a large number of 16S microbiome data sets with sample size 1000 subjects and mean of sequencing depth 50,000. These settings were used to mimic the HELIUS data set. When we analyzed the simulated data sets

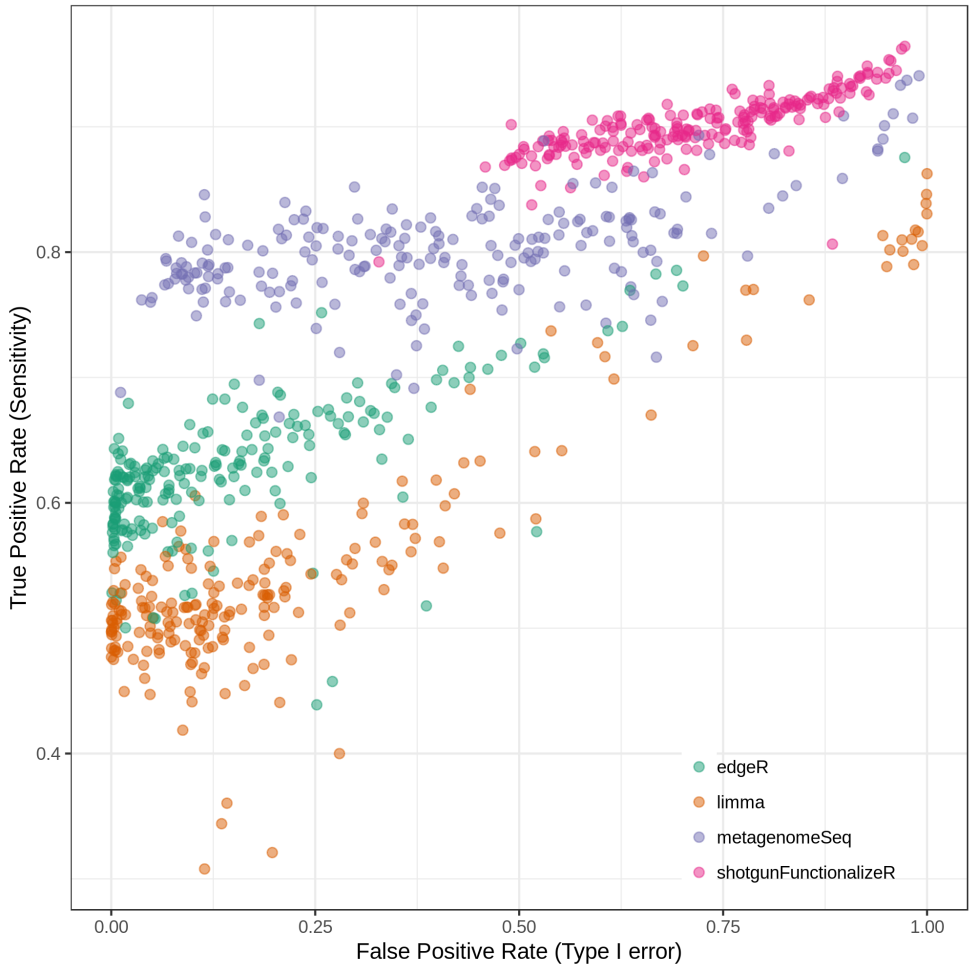


Figure 5.4: With each simulated data set, we calculated the performance of every method, in terms of true positive rate and false positive rate. Every dot represents a simulation.

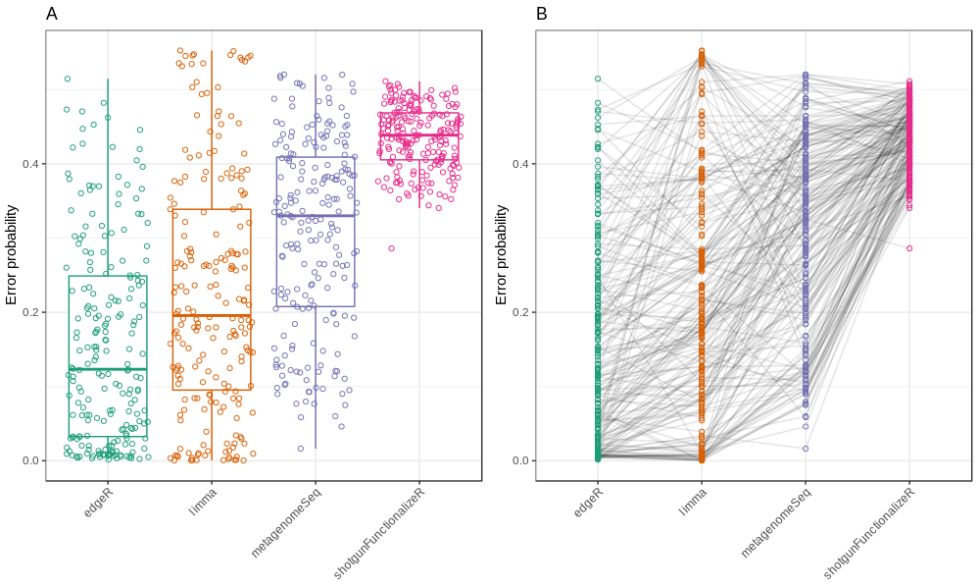


Figure 5.5: Every dot represents the probability that a significant association is false. Each line represents a simulation, in which the same simulated data were analyzed by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR. A: error probabilities vary a lot within each method depending on the simulated data set. B: With the same simulated data set, the error probabilities vary a lot across methods.

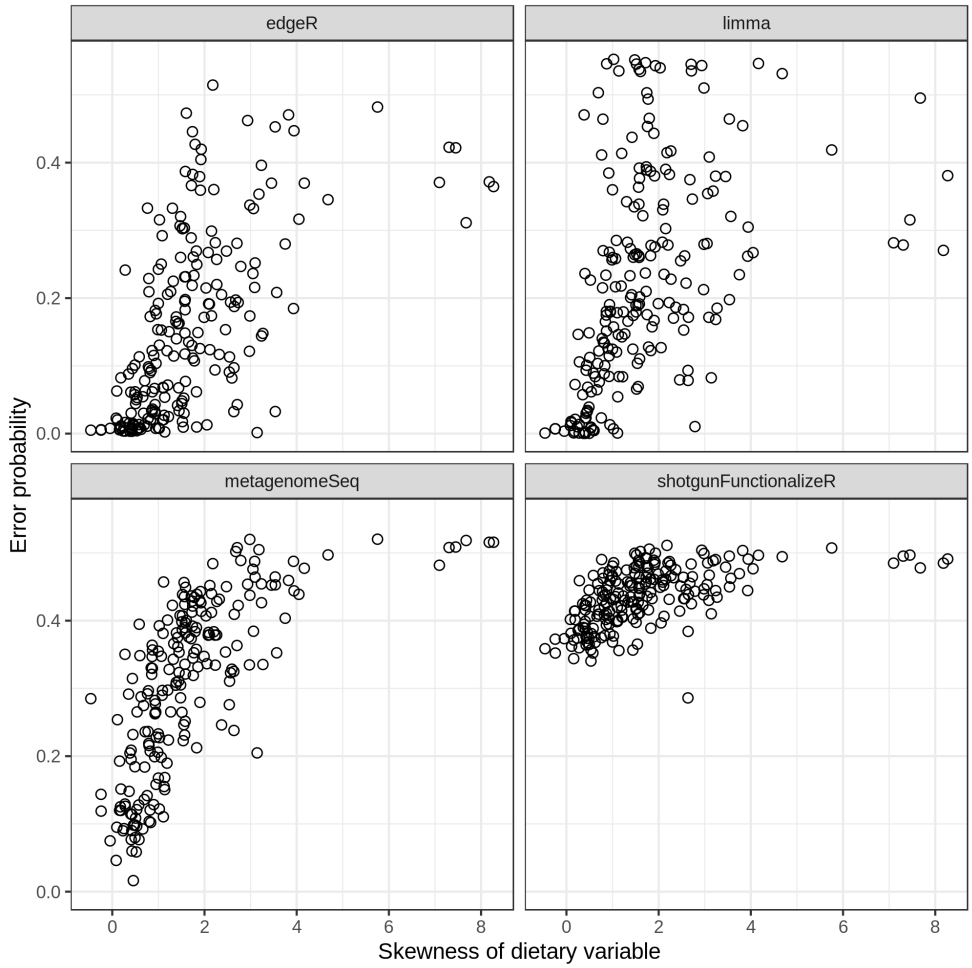


Figure 5.6: Skewness of predictor variable influences false positive rate. Every circle represents a simulation.

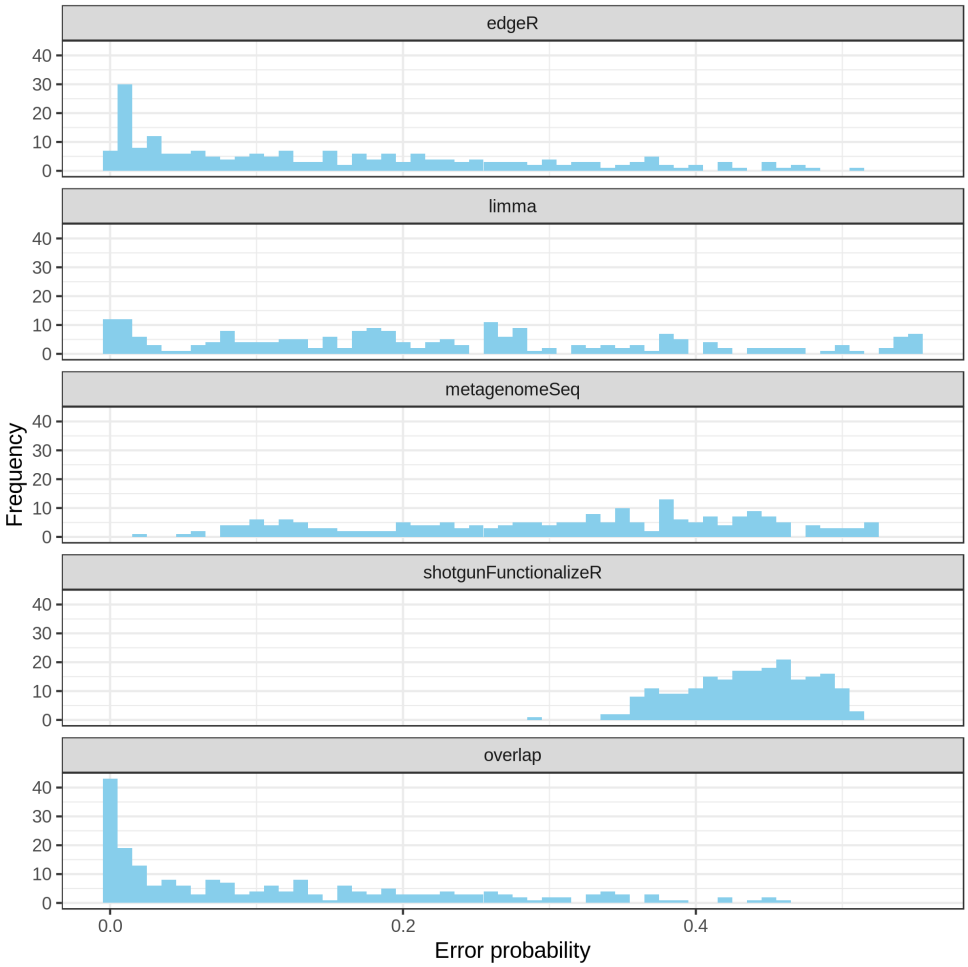


Figure 5.7: Distribution of probabilities that a significant association is false in edgeR, limma, metagenomeSeq and shotgunFunctionalizeR. The "overlap" refers to the distribution of error probabilities of significant associations identified by all four methods.

with edgeR, limma, metagenomeSeq and shotgunFunctionalizeR, we observed again large difference in number of significant associations between the methods. In general, we want our statistical method to detect as many as possible true positives, and as few as possible false positives. From our simulation studies, we learned that overall the most sensitive method (shotgunFunctionalizeR in this case) was likely to be the one with the most false positives. This phenomenon was observed in the differential abundance test scenario as well (Thorsen et al. 2016). Even though we set FDR as 0.05 in all our diet-microbe association analyses, our simulation results showed that control of FDR completely failed in shotgunFunctionalizeR, and rarely achieved in metagenomeSeq. On the other hand, edgeR and limma achieved FDR 0.05 in some cases. In the previous case-control simulations (Jonsson et al. 2016), metagenomeSeq and shotgunFunctionalizeR were shown to fail controlling false discovery rate at 0.05. However, edgeR was reported to be able to control false discovery rate at 0.05 (Jonsson et al. 2016). We think this is due to the fact that performing association analyses is more challenging than case-control comparisons because we cannot control both dependent and independent variables. Our further analysis showed that the skewness of the independent variable (e.g. dietary variable) influences the error probabilities in all methods. When the skewness of the dietary variable increased, the probability that a significant association is false also increased. When we focused on the significant diet-microbe associations that were identified by all four methods, we observed that more simulations had error probability below 0.05.

5.6 Conclusions

In summary, the choice of the statistical method is a key issue in the analysis of 16S rRNA microbiome data. No single method was optimal for diet-microbe association analyses. We recommend researchers to run multiple statistical models and focus on the significant associations identified by multiple methods. In this way, we expect to achieve better control of false discovery rate.

5.7 List of abbreviations

TP: true positive FP: false positive TN: true negative FN: false negative FDR: false discovery rate OTU: operational taxonomic unit HELIUS: HEalthy Life in an Urban Setting FFQs: food frequency questionnaires

5.8 Declarations

5.8.1 Ethics approval and consent to participate

The HELIUS study was complied with all relevant ethical regulations and in accordance with the Declaration of Helsinki (6th, 7th revisions); it was approved by the Academic Medical Center (AMC) Medical Ethics Committee and all participants provided written informed consent.

5.8.2 Consent for publication

Not applicable

5.8.3 Availability of data and material

The 16S rRNA gene sequences have been deposited at the European Genome-phenome Archive under study number EGAD00001004106. The FFQs data of this study are available from the study coordinator upon reasonable request. The 16S rRNA data and FFQs data, as well as the code used for simulation studies can be found at https://github.com/XiangZhangSC/HELIUS.

5.8.4 Competing interests

Not applicable

5.8.5 Funding

This work was supported by a personal ZONMW-VIDI grant 2013 [016.146.327] and a Dutch Heart Foundation CVON 2012 Grant (IN-CONTROL) [2012-03]. The funders had no role in the study design, the collection, analysis, and interpretation of data, the writing of the report, and the decision to submit the article for publication.

6

Forward individualized medicine from personal genomes to interactomes

This chapter has been published as Zhang X, Kuivenhoven JA and Groen AK (2015) Forward Individualized Medicine from Personal Genomes to Interactomes. Front. Physiol. 6:364.

6.1 Abstract

When considering the variation in the genome, transcriptome, proteome and metabolome, and their interaction with the environment, every individual can be rightfully considered as a unique biological entity. Individualized medicine promises to take this uniqueness into account to optimize disease treatment and thereby improve health benefits for every patient. The success of individualized medicine relies on a precise understanding of the genotype-phenotype relationship. Although omics technologies advance rapidly, there are several challenges that need to be overcome: Next generation sequencing can efficiently decipher genomic sequences, epigenetic changes, and transcriptomic variation in patients, but it does not automatically indicate how or whether the identified variation will cause pathological changes. This is likely due to the inability to account for 1) the consequences of gene-gene and gene-environment interactions, and 2) (post)transcriptional as well as (post)translational processes that eventually determine the concentration of key metabolites. The technologies to accurately measure changes in these latter layers are still under development, and such measurements in humans are also mainly restricted to blood and circulating cells. Despite these challenges, it is already possible to track dynamic changes in the human interactome in healthy and diseased states by using the integration of multi-omics data. In this review, we evaluate the potential value of current major bioinformatics and systems biology-based approaches, including genome wide association studies, epigenetics, gene regulatory and protein-protein interaction networks, and genome-scale metabolic modeling. Moreover, we address the question whether integrative analysis of personal multi-omics data will help understanding of personal genotype-phenotype relationships.

6.2 Introduction

Humans share the same genes but do not have identical DNA sequences. The latest 1000 Genomes Project reported over 84,000,000 single nucleotide polymorphisms (SNPs), 3,000,000 short insertions/deletions, and 60,000 structural variants in 2,504 subjects from 26 populations, by applying whole genome sequencing as well as exome sequencing and microarray genotyping technologies (1000 Genomes Project Consortium et al. 2015).

While there are large differences in the presence of both rare and common variants, it has been reported that every subject carries around 250 to 300 loss-of-function variants that lead gene products to having less or no function (1000 Genomes Project Consortium et al. 2010, 1000-Genomes-Project-Consortium:2012aa,UK10K-Consortium:2015aa). Nowadays, whole genome sequencing allows the determination of the entire DNA sequence of an individual, and the resulting genomic information is believed to enable prediction of disease risk and optimization of treatment outcome (Sadee 2011). In practice, predicting disease phenotypes from genetic sequences is extremely challenging because the genotype-phenotype relationship is far more complex than anticipated. A single gene can be associated with multiple disease phenotypes while a single disease phenotype can be caused by mutations in multiple genes (Barabási, Gulbahce, and Loscalzo 2011). Importantly, mutations do not have identical effects on individuals due to the individual variation in interaction between genes, proteins, metabolites and environmental factors (Barabási, Gulbahce, and Loscalzo 2011; Kathiresan and Srivastava 2012).

The complete set of (physical) interactions between molecules, such as genes, proteins and metabolites is known as the interactome (Cusick et al. 2005). In this review, we focus on the interactome in human cells. If we consider genome sequences as stills and phenotypes as a movie, then there must be a biological system which serves as a projector. It is indeed proposed that the interactome acts as the projector and eventually translates the phenotypic effects determined by both genotypes and environmental factors (Figure 6.1). Vidal et al. (Vidal, Cusick, and Barabási 2011; Emmert-Streib, Dehmer, and Haibe-Kains 2014) proposed that most disease phenotypes may be caused by the perturbation of the interactome, in which the products of disease genes were found to interact with each other and cluster as modules (Menche et al. 2015; Ghiassian, Menche, and Barabási 2015). These disease modules may overlap each other, explaining the shared associated genes and clinical symptoms of different diseases (Menche et al. 2015; Ghiassian, Menche, and Barabási 2015).

To understand the projector function of the interactome, one must capture all molecular components involved in cellular functions. With the rapid development of omics technologies, it is now possible to readily profile up to 19,797 protein-coding genes, 79,795 protein-coding transcripts, 30,057 proteins, and 4,229 metabolites (Psychogios et al. 2011; Harrow et al. 2012; Kim et al. 2014). Since individuality is present in the genomes, epigenomes,

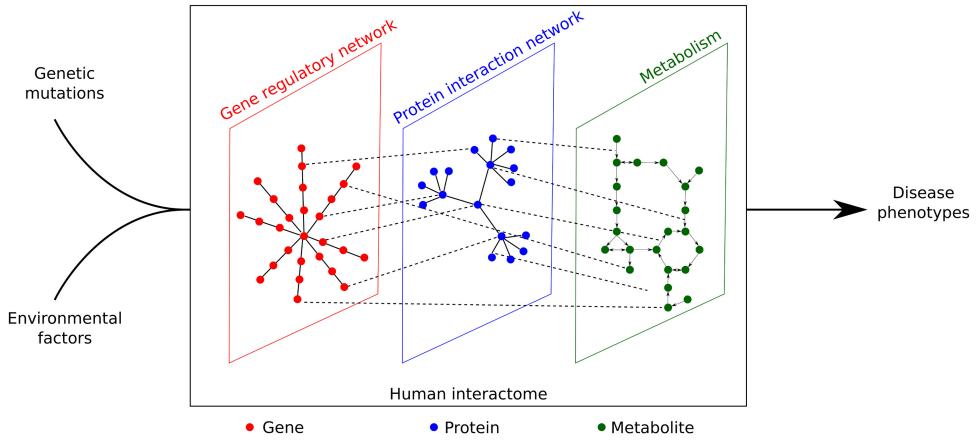


Figure 6.1: Genetic mutations and environmental effects can only lead to disease phenotypes through perturbation of the human interactome, which is a complex network constituted by gene regulatory network, protein interaction network, and metabolism.

transcriptomes, proteomes, and metabolomes, each cell type in every human subject will have a different interactome (Feinberg et al. 2010; Suhre et al. 2011; Yizhak et al. 2010b; Forler, Klein, and Klose 2014). In contrast to non-individualized medicine, personalized medicine attempts to address such subject-specific differences with respect to diagnosis and treatment (Topol 2014). This review aims to give an overview of bioinformatic and network modeling approaches that can be used to develop individualized medicine.

6.3 Genome-wide association studies, epigenetics and individualized medicine

Genome-wide association studies (GWAS) have identified a great number of common single nucleotide polymorphisms (SNPs) that are statistically associated with complex disease phenotypes. The National Human Genome Research Institute (NHGRI) GWAS catalog includes 1,751 curated publications of 11,912 SNPs (Welter et al. 2014). Besides disease-associated SNPs, GWAS also identified SNPs associated with drug efficacy and toxicity, fueling the development of pharmacogenomics and guiding individualized therapies (Sadee 2011; Crews et al. 2012; Low et al. 2014). The Pharmacogenomics Knowledgebase PharmGKB (Hewett et al. 2002; Altman 2007) is a literature-based database

which provides useful annotations on genes involved in pharmacokinetics (how the drug is absorbed, distributed, metabolized and eliminated) and pharmacodynamics (how the drug acts on its target and its mechanism of action). In the current release of PharmGKB, curated evidence for 1,073 human genes involved in drug response is present.

Epigenetics has been shown to play a key role in the crosstalk between environment and genome, pointing towards the notion that epigenetic marks might explain in part the role of the environment in disease development (Bjornsson, Fallin, and Feinberg 2004; Rivera and Ren 2013). Major epigenetic alterations include DNA methylation, histone modification, and chromatin remodeling (Rasool et al. 2015). A total number of 127 reference human epigenomes are available on the website of the Roadmap Epigenomics Project, including epigenetic landscapes of 111 primary cell and tissue types as well as 16 cell lines (Roadmap Epigenomics Consortium et al. 2015). Due to epigenetic modifications, cells can exhibit different phenotypes in response to various environmental factors, such as nutritional changes and oxidative stress. Feinberg (Feinberg 2007) defined this ability as phenotypic plasticity, whose abnormality is linked to diseases, such as cancers, neurodegenerative and autoimmune disorders (Howell et al. 2009). By integrating GWAS SNPs with epigenetic annotations, Farh et al. (Farh et al. 2015) identified that 90% of potentially causal variants of autoimmune diseases are non-coding and 60% map to enhancers of immune cells.

In general, information deriving from GWAS (Table 6.1) and epigenetics provide possible etiological pathways rather than the exact molecular mechanisms underlying diseases. Burke et al. (Burke and Korngiebel 2015) pointed out that although dramatic progress has been made in genomics research, there is still a gap between genomic knowledge and clinical application. To fill such gap, an accurate understanding of the genotype-phenotype relationship, which is hierarchically bridged by DNA, RNA, protein, metabolite and flux, must be developed (Figure 6.2). The integrative personal omics profile (iPOP) study (Chen et al. 2012) was the first example of individualized medicine attempting to overcome the gap by combining omics data sets. Over a 14-month period which also included two viral infections (HRV: human rhinovirus and RSV: respiratory syncytial virus), dr. Michael Snyder not only profiled his whole genome, but also the transcriptomes of his PBMCs (Peripheral Blood Mononuclear Cells) at 20 different time points, proteomes from PBMCs and serum across 14 time points, and metabolomes of his serum sampled 17 time

Table 6.1: Major SNP-trait association databases

Name	Link
NHGRI GWAS Catalog	www.genome.gov/gwastudies/
PharmGKB	http://www.pharmgkb.org/
GWASdb	http://jjwanglab.org/gwasdb
GWAS Central	http://www.gwascentral.org/
HuGE Navigator	http://www.hugenavigator.net/HuGENavigator/home.do
dbGaP	http://www.ncbi.nlm.nih.gov/gap
VaDE	http://bmi-tokai.jp/VaDE/

points, respectively. Integration of the data sets revealed the great potential of the individualized approach. In particular, the genetic variant information of dr. Snyder indicated that he is at risk for developing coronary artery disease, basal cell carcinoma, hypertryglyceridemia, and type 2 diabetes. At the same time, he was found carrying variants that are associated with response to glucose lowering drugs, rosigitazone and metformin. Interestingly, his time series measurements of transcriptome, proteome, and metabolome across healthy states, response to RSV infection, and recovery, enabled the authors to identify an alteration of the insulin signaling response following the RSV infection (Chen et al. 2012).

The iPOP study also provided us with some important insights on omics-based individualized medicine. First of all, as sequencing technologies vary considerably from each other due to sensitivity, accuracy, coverage and resolution, the measurements may contain systematic errors. Fortunately, since the human genome is constant over time, profiling with multiple DNA sequencing technologies is a way to improve the accuracy of genetic variant detection in an individual genome. As shown in the iPOP study (Chen et al. 2012), a genetic variant in the protein-coding genes can be trusted, if it is captured by the whole genome sequencing as well as whole exome sequencing. Same as above, we can also trust a genetic variant in the non protein-coding genes, if it is identified by different whole genome sequencing platforms. In contrast to the genome which is static, transcriptome, proteome, and metabolome are more dynamic and changes in their patterns represent the most valuable information for individualized medicine. To minimize systematic errors, the personal transcriptomes, proteomes, and metabolomes should be measured with standardized high-throughput methods at different time points and compared longitudinally. The longitudinal design also allows to perform statistical analysis with a single sample

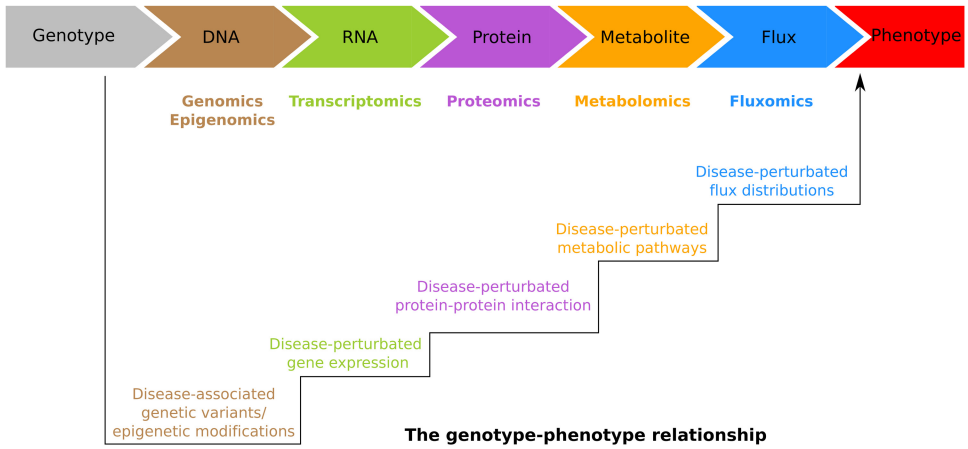


Figure 6.2: The genotype-phenotype relationship is hierarchically bridged by DNA, RNA, protein, metabolite and flux. These molecules are profiled in the genomics, epigenomics, transcriptomics, proteomics, metabolomics, and fluxomics, respectively. Bioinformatics and systems biology approaches try to translate these omics data sets into unified knowledge. In particular, from genomics and epigenomics, one attempts to identify the disease-associated genetic/epigenetic alterations. From transcriptomics, proteomics, metabolomics, and fluxomics, one aims to identify the genes, proteins, pathways, and the flux distributions involved in disease pathogenesis.

through applying well-established time-series data analysis techniques, such as Fourier spectral analysis and autocorrelation calculations (Chen et al. 2012). However, we have to admit that although the cost of sequencing technologies has dramatically decreased, sequencing with different platforms or multiple time points is unlikely to be performed for more than economic reasons only. In addition, the large volume of omics data sets will require substantial investments in data storage and management.

Topol (Topol 2014) rightfully indicated that individualized medicine needs translating large-scale omics data sets into useful knowledge. The approaches of omics data analysis can be roughly categorized as bioinformatics and network-based. Bioinformatics-based approaches often use statistical techniques to assess significant difference or association in the omics data. Their biological interpretation mainly relies on annotations in the community databases. Due to the chosen scope of this review, we are not going into details of these approaches. Network-based approaches, on the other hand, are mainly used to integrate multi-omics simultaneously and the network itself is subsequently used to explore biological insights. In general, network-based approaches first reconstruct biased or unbiased networks *in silico*, and then use the reconstructed network to interpret the omics data. A biased network indicates that prior biological knowledge is incorporated, whereas an unbiased network is purely data-driven.

Network-based approaches enable us to link genotype to phenotype, and vice versa. The constructed networks can be viewed as maps, in which we can locate GWAS results and improve our understanding the roles of genetic/epigenetic alterations in disease predisposition (Califano et al. 2012; Ghiassian, Menche, and Barabási 2015). At the same time, these maps can also help us tracking back molecular mechanisms of given clinical phenotypes. Like what has been shown by Bartel et al. (Bartel et al. 2015b), the "human blood metabolome-transcriptome interface", a network constructed based on the correlation between serum metabolomes and whole blood transcriptomes of 712 subjects, can identify active pathways/modules with concentrations of blood cholesterol and triglycerides. In the next sections, we focus on three types of network-based approaches, namely gene regulatory network, protein-protein interaction networks, and genome-scale metabolic modeling and discuss them in a schematic manner: i.e. 1) definition and generation; 2) usage and results; 3) strength and weakness. We also discuss their applicability for individualized medicine.

6.4 Gene regulatory networks

6.4.1 What are gene regulatory networks?

Thousands of gene products are produced from the human genome to support cell function and survival. The protein-coding genes can induce protein synthesis, whereas the non protein-coding genes encode noncoding RNAs (ncRNAs) as their gene products. Gene regulatory networks (GRNs) ensure proper levels of gene products present at the right time in the cell (Karlebach and Shamir 2008). In the GRN, nodes represent the genes and edges indicate the interactions between gene products.

6.4.2 How are GRNs generated?

Similar to gene coexpression networks, GRNs are statistically inferred from a large number of gene expression data sets. However, gene coexpression networks and GRNs are fundamentally different from each other. Pearson's correlation coefficient is used to infer coexpression networks, meaning that there is always a direct interaction for any pair of genes when their expressions are statistically correlated (Stuart et al. 2003). In contrast, GRNs are inferred mainly based on mutual information, which explicitly specifies direct or indirect interaction for each pair of genes. Mutual information defines how much information one random variable X provides about another random variable Y (Cover and Thomas 2006). For GRNs, the random variables refer to the gene expression levels. Almost all major algorithms developed for GRN inference are mutual information-based and include ARACNe (Algorithm for the Reconstruction of accurate Cellular Networks) (Basso et al. 2005; Margolin et al. 2006), CLR (Context Likelihood of Relatdeness) (Faith et al. 2007), MRNET (Meyer et al. 2007), RN (Relevance Network) (Butte and Kohane 2000), C3Net (Altay and Emmert-Streib 2010b), and BC3Net (de Matos Simoes and Emmert-Streib 2012). Different inference algorithms above were used to reconstruct human B cell GRNs and found the networks contained consistent biological information (Altay and Emmert-Streib 2010a; de Matos Simoes, Dehmer, and Emmert-Streib 2013). We refer readers to a recent review (Emmert-Streib, Dehmer, and Haibe-Kains 2014) for more general concepts of GRN inference and applications. In this review, we focus on ARACNe since it is the most widely used method. ARACNe makes use of two steps to

infer a genome-wide GRN (Basso et al. 2005). First, ARACNe assesses all the pair of genes by calculating their mutual information. Then, ARACNe discriminates whether the pair of genes are directly linked or separated by any other genes through applying a well-known property of mutual information called the data processing inequality (Basso et al. 2005; Cover and Thomas 2006).

6.4.3 What are GRNs used for?

The rationale of the GRN lies in the idea that genetic/epigenetic alterations contribute to disease phenotypes by inducing changes in a finite number of regulatory bottlenecks, i.e. transcription factors (TFs) (Lefebvre et al. 2010; Califano et al. 2012). ARACNe-inferred GRNs are used for identification of the crucial TFs (also called master regulators) that affect the transition from healthy to diseased states and vice versa. The identified master regulators then serve as starting points to search for the driver genetic/epigenetic alterations upstream.

6.4.4 What has come out?

Lefebvre et al. (Lefebvre et al. 2010) applied ARACNe to infer a human B-cell specific GRN from 254 B-cell microarray expression profiles representing 24 distinct phenotypes. The ARACNe-inferred B-cell GRN was subsequently used to identify *MYB* and *FOXM1* as the master regulators of B-cell proliferation. Similarly, an ARACNe-inferred glioblastoma GRN was created and used by Chen et al. (Chen et al. 2014) to identify two master regulators, *C/EBP*β and *C/EBP*δ that are known to be involved in mesenchymal subtype of glioblastoma patients (Carro et al. 2010). Furthermore, by combining the genetic variants from the same glioblastoma patients, the authors identified that *KLHL9* deletions are upstream of the two identified master regulators and act as driver mutations (Chen et al. 2014).

6.4.5 Strengths and weaknesses

One of the major advantages of ARACNe-inferred GRNs is that with whole genome microarray or total RNA sequencing, the entire genome can actually be included in the

ARACNe-inferred GRNs. Moreover, since it has been shown that the interactions inferred by the ARACNe algorithm are very likely to represent real biophysical and biochemical interactions (Basso et al. 2005; Lefebvre et al. 2010), ARACNe-inferred GRNs are suitable to explore all the possible interactions related to ncRNAs. This represents an important feature of ARACNe-inferred GRNs, as more or less 90% of the human genome is being transcribed, but only about 3% encodes protein. It is known that long noncoding RNAs (lncRNAs) can interact with DNA and proteins (Quinodoz and Guttman 2014), and some lncRNA interactions are related to human diseases. For example, Hirata et al. (Hirata et al. 2015) reported that interaction between lncRNA MALAT1 and histone-lysine N-methyltransferase EZH2 is involved in renal cell carcinoma.

The major drawback of ARACNe is that a large number (≥ 100) of gene expression profile data covering a broad range of phenotypes is required to infer the target GRNs (Basso et al. 2005; Margolin et al. 2006). This is indeed necessary to explore a significant range of gene expression dynamics in order to obtain adequate mutual information for inferring GRNs (Margolin et al. 2006). Obviously, in practice it is costly and time-consuming.

6.5 Protein-protein interaction networks

6.5.1 What are protein-protein interaction networks?

Proteins exert their function through interactions with other molecules (e.g. DNA, RNA, proteins, and metabolites). For instance, signal transduction is mediated through protein-protein interactions (PPIs), whereas gene expression (transcription factor-DNA) and metabolism (enzyme-substrate interaction) are mediated by protein-DNA and protein-metabolite interactions, respectively (Sevimoglu and Arga 2014). PPIs can also refer to formation of dimers, multi-protein complexes or supramolecular assemblies (e.g. actin filaments). Since some proteins are shared by different PPIs, individual PPIs are interconnected. In the PPI network, nodes represent genes whereas edges refer to physical interactions of the respective proteins.

Table 6.2: Primary sources of protein-protein interactions

Name	Link
HPRD	http://www.hprd.org/
IntAct	http://www.ebi.ac.uk/intact/
MINT	http://mint.bio.uniroma2.it/mint/Welcome.do
DIP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
BioGRID	http://thebiogrid.org/
PDB	http://www.rcsb.org/pdb/home/home.do

6.5.2 How are PPI networks generated?

There are three main resources of generic human PPI networks. The first resource is from the literature mining. We listed six primary databases (Table 6.2) that store and combine experimentally supported PPIs from small-scale studies. The second resource is derived from large-scale yeast-two-hybrid (Y2H) screening. In 2005, the first generation of Y2H-based human PPI network, HI-I-05, was introduced and included 2,700 high-quality binary PPIs among 1,705 proteins (Stelzl et al. 2005; Rual et al. 2005). In 2014, the second generation of Y2H-based human PPI network, HI-II-14, was released (Rolland et al. 2014). This time 13,944 PPIs were identified among 4,303 proteins. Both HI-I-05 and HI-II-14 can be downloaded (http://interactome.dfci.harvard.edu/H_sapiens/). In addition to the Y2H system, affinity-purification mass spectrometry (AP-MS) is also developed to profile PPIs in human cells (e.g. human HEK293T (Huttlin et al. 2015)). Compared to Y2H which is mainly used to identify binary interactions between two proteins, AP-MS is more focusing on deciphering the composition of protein complexes. The third resource of the human PPI network is the computational prediction, in which machine learning algorithms are applied to calculate the likelihood of interactions between two proteins based on the known interactions in the databases (Table 6.2). STRING (Search Tool for the Retrieval of Interacting Genes, http://string-db.org/) (Snel et al. 2000) is such a web-server including known and predicted protein interactions of over 2,000 organisms. In addition to STRING, databases, such as PIPs (http://www.compbio.dundee.ac.uk/www-pips/) (McDowall, Scott, and Barton 2009) and hPRINT (human Predicted Protein Interactome) (Elefsinioti et al. 2011) also predict PPIs without priori experimental evidence. The hPRINT results can be retrieved in STRING as well (Franceschini et al. 2013).

Human proteome studies have shown distinct proteome profiles in different cell and tis-

sue types (Kim et al. 2014; Uhlén et al. 2015). This makes it necessary to specify PPI networks in the target cell and tissue (Schaefer et al. 2013). TissueNet database (http://netbio.bgu.ac.il/tissuenet/) provides such context-specific PPI networks for 16 human tissues (Barshir et al. 2013). A generalized way to construct such context-specific PPI networks is introduced by Magger et al. (Magger et al. 2012), who developed a specific algorithm integrating context-specific gene expression data (proteomics or transcriptomics). Gene expression data are used to assess the probability of PPIs in the generic PPI network. If a gene is not expressed, the algorithm can either remove the gene from the generic PPI network or reduce the weight of the interactions associated with the gene.

6.5.2.1 What are PPI networks used for?

Human PPI networks are used to identify genes, proteins and subnetworks associated with diseases (Sevimoglu and Arga 2014). They are also used to systematically characterize PPI network perturbations associated with disease mutations. The PPI network perturbations include complete loss of gene products or alteration of PPI arrangement (Zhong et al. 2009; Sahni et al. 2013).

6.5.3 What has come out?

Goehler et al. (Goehler et al. 2004) generated a PPI network for Huntington's disease by using the Y2H. From there, they identified GIT1, a G protein-coupled receptor kinase-interacting protein, which directly interacts with huntingtin and turns out to enhance huntingtin aggregation. Based on the generic human PPI network derived from HPRD (Human Protein Reference Database) (Keshava Prasad et al. 2009), Jia et al. (Jia and Zhao 2014) focused on PPI subnetworks that contain multiple genes frequently mutated in lung adenocarcinoma and melanoma patients. The results showed that the driver mutations interrupted the PPIs that are involved in signaling pathways (e.g. EGF receptor signaling pathway) and biological processes (e.g. DNA repair systems) (Jia and Zhao 2014). Based on the Y2H protein interaction assays, Sahni et al. (Sahni et al. 2015) reported that common SNPs from healthy subjects rarely affected PPIs, but around 60% of human disease-associated missense mutations perturbed PPIs. Furthermore, they also

noticed that different mutations in the same gene can influence different PPIs.

6.5.4 Strengths and weaknesses

Unlike the ARACNe-inferred GRNs, in which the interactions are statistically inferred from the gene expression levels, PPI networks derived from the literature or Y2H screening are experimentally supported. Therefore, perturbations in PPI networks can be used with confidence to elucidate the molecular basis of diseases as described in the examples given above.

A weakness of the PPI networks is incomplete coverage. According to the up-to-date GENCODE release 23, there are 19,797 protein-coding genes in the human genome. The number of genes covered by the most comprehensive human PPI network, HI-II-14 (Rolland et al. 2014), is only 3,146 which suggests that there is still a long way to go. In addition, PPIs are often evaluated under unphysiological conditions, leading to false positive and negative PPIs included in generic PPI networks (Schaefer et al. 2013). Kuchaiev et al. (Kuchaiev et al. 2009) reported that the false positive and negative rate of Y2H could be as high as 64% and 71%, respectively.

6.6 Genome-scale metabolic models

6.6.1 What are genome-scale metabolic models?

Metabolites are implicated in maintenance of cellular functions and production of building blocks (e.g. purines and pyrimidines) for macromolecular biosynthesis. Computational biologists have reconstructed all metabolic reactions into one large network and name it "genome-scale metabolic model". GEMs and GSMMs are typically used as abbreviations in the literature.

6.6.2 How are GEMs generated?

In general, GEMs are constructed by using enzyme-mediated reactions, transporters and intermediary metabolites (Bordbar et al. 2014). The first landmark studies in this field

emerged in 2007 when Recon1 (Duarte et al. 2007) and EHMN (Edinburgh Human Metabolic Network) (Ma et al. 2007) were manually reconstructed based on genomic and experimental data in the literature. These two human metabolic networks were merged into the HMR (Human Metabolic Reaction) database (Agren et al. 2012). In 2010, a human hepatocyte-specific metabolic network, HepatoNet1, was reconstructed based on experimental evidence for presence of metabolic reactions in human hepatocytes (Gille et al. 2010). The experimental evidence was manually curated based on information from over 1,500 scientific articles. In 2013, the continuing development of Recon1, EHMN, and HepatoNet1 leads to the release of Recon2 (Thiele et al. 2013). A year later, another reconstruction of human hepatocyte-specific metabolic network, iHepatocytes2322, together with a new release of the Human Metabolic Reaction database, HMR2, were published (Mardinoglu et al. 2014).

Recon2 (Thiele et al. 2013) and HMR2 (Mardinoglu et al. 2014) represents all current knowledge of global human metabolism. Since different cell/tissue types may harbor synonymous enzymes to catalyze the same reaction and different metabolic pathways may result in the same product (Uhlén et al. 2015), it is important to reconstruct cell/tissue type specific GEMs to characterize the metabolism of target cells and tissues. For this purpose, algorithms, such as tINIT (task-driven Integrative Network Inference for Tissues) (Agren et al. 2014), GIMME (Gene Inactivity Moderated by Metabolism and Expression) (Becker and Palsson 2008), and mCADRE (metabolic Context-specificity Assessed by Deterministic Reaction Evaluation) (Wang et al. 2012) are used to generate cell/tissue type specific GEMs from the generic GEMs (e.g. Recon2 or HMR2). These algorithms use abundances of transcripts and proteins to estimate the probability of presence of enzymes in the generic GEMs. We refer readers to an excellent review (Machado and Herrgård 2014) for more details on the differences between the various algorithms.

6.6.3 What are GEMs used for?

Human GEMs, especially cell/tissue type specific GEMs, are mainly used as scaffolds to analyze transcriptomics data obtained from patient samples, in order to identify the metabolic pathways and metabolite biomarkers that are related to disease pathogenesis.

6.6.4 What has come out?

Using the tINIT algorithm with proteomics and transcriptomics data of human myocytes, Varemo et al. (Väremo et al. 2015) reconstructed a myocyte-specific GEM, iMyocytes2419, which made it possible to reveal that type 2 diabetes patients show extensive transcriptional changes in reactions involved in pyruvate oxidation, branched-chain amino acid catabolism, and tetrahydroflate metabolism. Mardinoglu et al. (Mardinoglu et al. 2014) applied iHepatocytes2322 and their previously developed Reporter Metabolite algorithm (Patil and Nielsen 2005) to analyze transcriptomics data of patients with non-alcoholic fatty liver disease, and identified that concentrations of chondroitin and heparan sulphates may represent novel biomarkers for diagnosing non-alcoholic steatohepatitis. Similar GEM-based analyses have been performed to study diseases such as, Alzheimer's disease (Lewis et al. 2010), obesity (Mardinoglu et al. 2013), and cancer (Agren et al. 2014; Yizhak et al. 2014).

6.6.5 Strengths and weaknesses

In our opinion, the major advantage of GEMs is that it allows to study global metabolic flux distributions. The rate of the metabolic reactions in a pathway (metabolic flux) is determined by many aspects, such as protein concentration, protein interaction (signal transduction), enzyme kinetics and metabolite concentrations (Winter and Krömer 2013). Therefore, metabolic fluxes can be considered as the ultimate outcome of cellular regulation at different levels (Nielsen 2003). When listing all the reactions as well as their corresponding flux values under a particular condition, one can construct a metabolic flux distribution that represents a particular cellular phenotype in detail.

Currently, ^{13}C stable isotope labeling is the most popular experimental method to measure *in vivo* fluxes (Blank and Ebert 2013). By performing ^{13}C fluxomic experiments, Murphy et al. (Murphy, Dang, and Young 2013) noticed that different levels of oncoprotein MYC can induce distinct metabolic flux distributions in P493-6 B cells. They showed that high MYC cells relied more heavily on amino acids and mitochondrial oxidative metabolism than low MYC cells. ^{13}C fluxomics also revealed distinct metabolic flux distributions in different cell lines. Niklas et al. (Niklas, Sandig, and Heinzle 2011) reported that human neuronal AGE1.HN cells had lower flux rates (around 2.3% of the glucose uptake flux)

in the pentose phosphate pathway than other cell lines, such as HEK-293 cells (15%) and hybridoma cells (20%). These ^{13}C fluxomic studies illustrate that various biological conditions can induce distinct metabolic flux distributions.

However, ^{13}C fluxomics cannot deliver us a complete picture of flux distributions in the metabolic network, since only a small number of reactions can be measured. Here, GEMs provide a means to estimate metabolic flux distributions under different conditions relying on a limited number of exchange fluxes, i.e. fluxes of substrates entering the cells and the fluxes of metabolites that are secreted from the cells. It is beyond the scope of this review to explain the related mathematical theory, but we recommend the article by Rossell et al. (Rossell et al. 2011), in which they formulated how to compute complete set of fluxes from the exchange fluxes.

Bordel et al. (Bordel, Agren, and Nielsen 2010) introduced a random sampling method which can calculate means and standard deviations for each flux in the GEM under different experimental conditions, when a limited number of measurements of exchange fluxes are given. By integrating changes in gene expression between different conditions, metabolic reactions can be classified as either transcriptionally regulated (significant changes in both flux and gene expression levels), post-transcriptionally regulated (significant changes in gene expression levels but not flux), or metabolically regulated (significant changes in flux but not gene expression levels). This random sampling method was applied together with the adipocyte-specific GEM, iAdipocytes1809, and helped identifying the fluxes of glucose uptake, fatty acids uptake, oxidative phosphorylation, mitochondrial and peroxisomal β-oxidation, fatty acid metabolism and tricarboxylic acid cycle as being differentially down regulated in obese subjects (Mardinoglu et al. 2013). Gavai et al. (Gavai et al. 2015) developed a novel algorithm called Lsei-FBA (Lesat-squares with equalities and inequalities Flux Balance Analysis), and identified the fluxes of glycolysis and oxygen uptake as being decreased in brains of Alzheimer's disease patients (29% and 46%, respectively) compared to healthy subjects. Similar to the random sampling method, Lsei-FBA also requires tissue-specific GEMs, and measurements of gene expression as well as exchange fluxes.

The second biggest advantage of GEMs is that up to now it is currently the only platform that can integrate genomics, transcriptomics, proteomics, metabolomics, and fluxomics data. Yizhak et al. (Yizhak et al. 2010a) integrated quantitative proteomics and

metabolomics with a GEM of the human erythrocyte, and predicted metabolic flux distributions in red blood cells. The flux distribution predictions were found to be consistent with the simulations made by a detailed kinetic model of human red blood cells. Bordbar et al. (Bordbar et al. 2012) analyzed transcriptomics, proteomics, and metabolomics data sets of LPS-stimulated RAW 264.7 cells with a GEM of the RAW 264.7 cell line, and identified a suppressive role for *de novo* nucleotide synthesis in macrophage activation.

Last but not the least, it has been shown by Uhlen et al. (Uhlén et al. 2015) that the minimum requirement of generating a cell/tissue type specific GEM is a single RNA sequencing profile.

Naturally, GEMs also have their limitations. First of all, although novel metabolite biomarkers for various diseases have been predicted by using cell/tissue type specific GEMs, few of them have been validated in humans, because of either technical limitation of measuring the metabolites in question or difficulty of accessing the patient materials. Secondly, since GEMs focus on metabolic enzyme-coding genes, reactions and pathways, GEMs cannot be used to study signal transduction pathways. Lastly, GEMs do not contain detailed kinetics of enzymes and produce metabolic flux distributions only under steady state conditions.

6.7 The future of individualized medicine

6.7.1 Role for GRNs

Regarding individualized medicine, longitudinal transcriptomics derived from cells/tissues of an individual including healthy and diseased states are the ideal resources to assemble an individualized GRN. Zoppoli et al. (Zoppoli, Morganella, and Ceccarelli 2010) introduced TimeDelay-ARACNe to infer GRNs specifically from time-course data. Such ARACNe-inferred GRN provides a personalized map, with which one can locate the genetic mutations identified in the one-dimensional genome sequences in a multi-dimensional network. By integrating gene differential expression information between healthy and diseased states, one can also identify the crucial transcription factors controlling the phenotype transition. Taken together with the network location information, one can make

the most of the personal genomic information and further prioritize the damaging effect of genetic mutations.

6.7.2 Role for PPI networks

PPI networks are proposed playing a role in buffering the impact of genetic mutations and environmental challenges (Forler, Klein, and Klose 2014; Garcia-Alonso et al. 2014). This opinion has been investigated by Garcia-Alonso et al. (Garcia-Alonso et al. 2014), who built up a human PPI network by merging generic PPI networks derived from three public databases (BioGRID (Stark et al. 2006), IntAct (Orchard et al. 2014), and MINT (Licata et al. 2012)). They used the reconstructed PPI network to study the effect of genetic variants predicted to be deleterious in the subjects participating in the 1000 Genomes Project, 252 healthy Spanish individuals, and 41 chronic lymphocytic leukemia patients. Interestingly, most of the potentially damaging genetic variants in healthy individuals were located in peripheral regions of the PPI network and did not really perturb the structure of the PPI network. However, when investigating the somatic variants that were predicted to be deleterious in chronic lymphocytic leukemia patients, they noticed that these mutations tended to be in internal regions of the PPI network. The above study indicates that PPI networks can help to identify whether genetic variants may be disrupting PPIs and hence may be important in explaining diseases.

6.7.3 Role for GEMs

GEMs have already been used successfully for individualized medicine. Argen et al. (Agren et al. 2014) reconstructed personalized GEMs for 6 hepatocellular carcinoma patients based on proteomics data, and used these models to identify potential anticancer drug targets for the individual patients. Yizhak et al. (Yizhak et al. 2014) reconstructed personalized GEMs for breast and lung cancer patients based on gene expression measurements obtained from biopsy samples. These personalized GEMs were used to predict the cancer cell growth rate, which was used to infer patient survival.

For successful individualized medicine, it should be realized that it is important to integrate information of cell/tissue type specific GEMs, in an attempt to capture whole-body

metabolism. Urine, plasma, and serum are the most common samples from human subjects for diagnostic purpose (Nicholson et al. 2012). Metabolic measurements based on these samples are the results of the crosstalk of many organs and can be regarded as serving the readouts of whole-body metabolism.

Bordbar et al. (Bordbar et al. 2011) build a multi-tissue GEM by integrating adipocyte, hepatocyte and myocyte-specific GEMs via a blood compartment. The assembled multi-tissue GEM was used to study the metabolic differences between non-type 2 diabetes obese and type 2 diabetes obese individuals. They reported that type 2 diabetes obese individuals lack activity in reactions catalyzed by lactate dehydrogenase, catalase and cysteine dioxygenase, comparing to the non-type 2 diabetes obese subjects. Besides integrating metabolism of different tissues and cells, the human gut microbiome is also considered important for whole-body metabolism (Mardinoglu and Nielsen 2015). Shoaie et al. (Shoaie et al. 2015) reconstructed five GEMs for five representative bacteria in the human gut, including *Bacteroides thetaiotanmicron*, *Eubacterium rectale*, *Bifidobacterium adolescentis*, *Faecalibacterium prausnitzii*, and *Ruminococcus bromii*. These GEMs were used to study 45 overweight and obese individuals who were subjected to an energy-restricted, high-protein diet intervention for 6 weeks. The authors reported that the diet intervention decreased the gut microbiota production of short chain fatty acids (acetate, butyrate, and propionate) and amino acids (e.g. alanine, proline and glycine etc.).

6.8 Concluding remarks

Due to the central role of the interactome in cellular functions, we think that the roadmap of individualized medicine is moving from human genomes to interactomes. However, construction of a complete human interactome is extremely complex and it might take at least another decade (Menche et al. 2015). This review shows that GRNs, PPI networks, GEMs can characterize part of the interactome in cells. Integrating different type of networks may contribute to better understanding of the interactome, and ultimately realizing true individualized medicine.

6.9 Disclosure/Conflict-of-Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6.10 Funding

This work was supported by grants CVON-Genius (CVON2011-19) and RESOLVE (FP7 305707).

7 General discussion and future perspectives

This dissertation focused on statistical analyses of high dimensional data in biomedical research. We attempted to show that statistical analysis is a construction process with a series of component decisions. From my perspective, components in statistical analyses include data, prior knowledge as well as statistical methods.

7.1 Data

Current biomedical research is data intensive. But what are data? Surprisingly this can be a question with multiple answers. For a clinician, data are spreadsheets in their SPSS workspace. For a biologist, data can be images of PCR or western blot, as well as figures in publications or slides. For machine learning engineers, data are anything with numeric values. In this dissertation, we refer to quantitative measurements as data. Particularly, we focused on high dimensional data generated by high-throughput technologies. I highlighted two aspects of these data that are important for any subsequent data analysis.

7.1.1 Technology underlying measurements matters

Until now, no single technology can measure all molecules. Genomics, epigenomics and transcriptomics analyse nucleic acids and rely on sequencing or array technologies. Proteomics and metabolomics are dependent on mass spectrometry or NMR technologies. Depending on technologies, measurements can have different physical meaning. For example, microarray uses fluorescence signal values to represent gene expression levels, and can have non-zero values even though a gene does not express (Zilliox and Irizarry 2007). Next generation sequencing is another popular technology for gene expression quantification. Measurements made by sequencing technologies represent the number of times each RNA is observed in each sample. Because gene expression values given by these two technologies have different physical meaning, different statistical models are used for data analysis.

More specifically, linear regression is often used to analyze microarray-based gene expression data, whereas generalized linear models (e.g. negative binomial regression) are used to analyze sequencing-based gene expression data.

In addition to different physical meaning, different technologies can give very different

values when we measure the same variable. For example, in chapter 3 the concentrations of circulating apolipoprotein B, LDL and HDL cholesterol were measured by clinical chemistry and the Nightingale metabolomics platform. Although the two technologies gave very similar HDL cholesterol measurements, a clear discrepancy in apoB and LDL cholesterol concentrations was observed. According to the LDL cholesterol concentration measured by clinical chemistry, all the subjects are hypercholesterolemic. However, based on the LDL cholesterol concentration made by the Nightingale metabolomics platform, none of them have hypercholesterolemia.

7.1.2 Batch effects and missing values

Sometimes biological samples were collected by different laboratories and processed by different people. In that case, we had to combine data from different sources. In chapter 4, we analyzed metabolomics data derived from subjects with European, Ghanaian and African Surinamese background. These participants were from two cohort studies with different times of sampling and measurement. Although all plasma samples of these participants were profiled by the same metabolomics platform and the data sets looked compatible with each other, we could not directly compare metabolite variables between ethnic groups. This was because values of metabolite concentrations were not only influenced by ethnic backgrounds, but also affected by batch effects including laboratory conditions and different personnel. If we ignore the batch effects, we could be easily misguided by the false positives. To bypass this hurdle, we ran regression analyses to evaluate relationships between HbA1c levels and circulating metabolites within each ethnic group and compared the associations across ethnic groups qualitatively.

In the same study, we applied a Bayesian approach when we ran the regression analyses. This was because the data set contained missing values. There were two types of missing values: 1) when the concentration of a metabolite was below the limit of detection, or 2) when values were rejected by the automatic sample and measurement quality control procedure in the Nightingale pipeline. The "standard" action in response to missing data is to delete them. However, this action will almost certainly bias the results of statistical analyses. Bayesian imputation is one of the ways to estimate missing values. In the Bayesian framework, data are what we observed, whereas parameters are what

we did not observe. Thus missing data were treated as parameters. Imputed values were constrained between zero and the minimal observed value if the missing values were below the limit of detection. We did not use this constraint for other type of missing values. I should mention that there are other valid multiple imputation algorithms. Personally I choose the Bayesian approach because it allows me to do missing value imputation and parameter estimation simultaneously.

7.2 Prior knowledge

We consider prior knowledge as extra "data". Prior knowledge is often well documented in the introduction and discussion sections of research manuscripts, but often ignored in the analysis of high-throughput omics data. Incorporation of prior knowledge can, however, provide better biological interpretation of statistical analysis outcomes (Reshetova et al. 2014). In this dissertation, I showed that prior knowledge can be introduced into the statistical analysis in various ways: 1) Prior knowledge was introduced in our analysis through biological databases. The KEGG pathway database and the genome-scale metabolic model (e.g. HMR2) provide us with information of the genes that participate in a particular pathway or that regulate the same metabolite. 2) Prior knowledge was also the basis to motivate the choice of a statistical method. In chapter 3, our prior knowledge that not all hypercholesterolemia cases were caused by known mutations in *LDLR*, *APOB* and *PCSK9*, indicated to us that at least two subgroups in a cohort of 119 hypercholesterolemic females should exist. We ran hierarchical clustering analysis to discover these subgroups using the metabolomics data. 3) Prior knowledge was introduced with prior distributions. In chapter 4, based on a preceding study (Dekker et al. 2015) we specified the prior distribution of HbA1c in Dutch population as a normal distribution with mean value approximately 40 mmol/mol and standard deviation about 6 mmol/mol. This prior was subsequently used as a basis of dealing with missing values. 4) Prior knowledge was used as a probabilistic model. In chapter 5, we constructed a probabilistic model to mimic the data generating process of human gut microbiome.

All these ways of introducing prior knowledge into statistical analyses have their limitations. Various biological databases contain inconsistent information, requiring experimental validations to evaluate the predictions based on the biological databases (Stobbe

et al. 2011). Result of hierarchical clustering analysis can be affected by the choice of dissimilarity measure, linkage statistic and decisions of cutting the dendrogram. Therefore, extra theory and data are needed to evaluate such clustering results. Bayesian approaches are often criticized for subjective prior distributions. In this thesis, we consider prior distributions as engineering components making Bayesian models running. Whenever we have prior information for parameters, we intend to assign informative prior distributions. However, usually we could not assign informative prior distributions for all the parameters. In chapter 4, we simply had no prior information for some parameters. For other parameters, it was difficult to translate prior information into prior distributions. In those scenarios, we used weakly informative priors, and applied prior predictive simulations to evaluate the choice of prior distributions. (Gelman, Simpson, and Betancourt 2017) demonstrated that the prior can often only be understood in the context of the likelihood. We confronted similar challenges when we constructed the generative model to simulate human gut microbiome in chapter 5. We compared simulated data to the real human gut microbiome data to make sure they had similar features.

7.3 Statistical methods

In this dissertation, we applied various statistical methods that were motivated by specific research questions, data availability and prior knowledge. In this section, I would like to point out that every statistical method has its own assumptions, and the choice of statistical methods can affect the outcomes of analyses.

7.3.1 Assumptions

Assumptions in statistical methods can be either implicit or explicit. Implicit assumptions are the ones implemented within a particular statistical software and often ignored by users who run statistical programs. For example, diverse methods have been developed to perform gene set enrichment analysis (Väremo, Nielsen, and Nookaew 2013b). Choosing a specific gene set enrichment analysis method automatically decides two things: 1) how to calculate the gene set summary statistics; 2) how to perform statistical inference. Similarly, the default settings of hierarchical clustering analysis often use the Euclidean

distance as the measure of similarity between each pair of samples, and complete linkage as dissimilarity quantify for each pair of groups. But default settings are not necessarily the proper settings for our purpose. We need to make our choices depending on the situations. In contrast to the implicit assumptions, the likelihood and prior distributions required by Bayesian regression models as well as the generative model for simulating human gut microbiome forced us to explicitly state all the assumptions.

7.3.2 Choice of statistical methods matters

Assumptions are part of statistical methods. In chapter 5, we used four methods that are based on Poisson, negative binomial, zero-inflated Gaussian distributions, as well as a weighted linear regression model to evaluate associations of nutrition intake with microbial abundances. We showed that choosing different statistical methods can lead to large differences in the outcomes of diet-microbe associations. Furthermore, a particular statistical method can generate a number of significant associations that are not identified by other statistical methods. "Which method should we choose?" is the immediate question from many scientists. In order to answer the question, we simulated a large number of microbiome data sets with known associations between microbial abundance and dietary intake. Based on the simulations, we realized that no statistical method was optimal for all microbiome data. In order to control false discovery rate, the best thing we can do is to run all eligible methods and focus on the results that are robust to the choice of statistical methods. "Why no statistical method is optimal?". This is a challenging question. "All models are wrong but some are useful." by George Box is taught in every school. Our measurements are noisy, and often not derived from any pure process assumed by statistical models. Therefore, all statistical models have to omit something, but what is omitted could be necessary for making useful and correct inferences.

7.4 Future perspective

The research examples in the dissertation focused on data analysis of single type omics, such as transcriptomics (chapter 2), metabolomics (chapter 3 and 4) and gut microbiome (chapter 5). These analyses of single omics data were limited to evaluate associations.

To reveal causal mechanistic information, integration across multiple omics data was suggested as the next step (Hasin, Seldin, and Lusis 2017). In chapter 6, we discussed three systems biology platforms for multi-omics integration. These systems biology platforms include gene regulatory networks, protein-protein interaction networks and genome-scale metabolic models. From my perspective, in addition to multi-omics data we also need to study biological systems under multiple conditions. Why? Different conditions can trigger different biological pathways in the same biological system. For example, plasma branched-amino acids (BCAAs) are well-known to be associated with insulin resistance. However, feeding animals with chow diet plus BCAAs did not cause insulin resistance (Newgard 2012). In contrast, feeding animals with high fat diet plus BCAAs did induce insulin resistance (Newgard 2012). A strategy to create such a compendium of conditions is to use chemical compounds (e.g. statin and LPS) with known biological targets and effects (O'Neill, Kishton, and Rathmell 2016).

Current statistical analysis in biomedical research relies heavily on methods such as statistical tests (e.g. t test), generalized linear models (e.g. linear regression) and multivariate methods (e.g. principal component analysis). These statistical techniques are very useful for discovery of relationships between variables but often ignore the underlying data generating processes. From my perspective, statistical models should contain more components motivated by biological knowledge or assumptions. Such statistical models will boost our understanding of biological mechanisms underlying a phenotype or disease. But we need to take into account the limitation of our knowledge. A fine example is the case of systems biology. Traditionally, systems biology uses ordinary differential equations to describe a biological system, such as glycolysis, pentose phosphate pathway and TCA cycle. However, this traditional systems biology modeling approach is often hampered by a large number of parameters with unknown values. As a result, an optimization algorithm is often used to estimate the parameters based on the experimental data. To take care of the uncertainties in parameters, a Bayesian approach was introduced (Vanlier et al. 2012). This is an important improvement although currently the model is computationally expensive and limited to a single biological pathway. It is a promising development, however, and I think the Bayesian approach should be further developed.

A second limitation of the traditional systems biology modeling approach is ignorance of heterogeneity between biological systems. To account for heterogeneity, the deterministic

differential equations model needs to be incorporated into a multi-level model or extended into a stochastic differential equations model (Székely and Burrage 2014). These extensions will make systems biology modeling even more computationally expensive. But the increased reality of the models make it worthwhile to consider this direction.

Summary

This dissertation discusses translating data obtained in biomedical research into knowledge through statistical analysis. To make the best inference based on collected data and prior knowledge, one needs to make a series of decisions during statistical analysis.

In chapter 1 we introduce high dimensional data such as transcriptomics, metabolomics and gut microbiome. We then highlight statistical analysis strategies such as permutation, quantification of similarities, Bayesian imputation and statistical simulation. In the end, we outline four research examples regarding inflammation, hypercholesterolemia, type 2 diabetes and microbe-diet associations.

In chapter 2 we discuss the first research example, in which gene expression data of peripheral blood mononuclear cells (PBMCs) challenged by four pathogenic agents were analyzed. By running gene set enrichment analysis in the context of either a human genome-scale metabolic model or a biological database, we identified metabolic biomarkers that discriminate different pathogenic stimulations.

In chapter 3 we discuss the second research example, in which plasma metabolomics data of 119 hypercholesterolemic females were analyzed. By running hierarchical clustering analysis, we showed that hypercholesterolemic subjects with and without known mutations had different plasma metabolomic profiles. Furthermore, we showed that the combination of metabolomics and genetic sequencing data can help us better understand the hypercholesterolemic cases without defined genetic mutations.

In chapter 4 we discuss the third research example, in which plasma metabolomics data of subjects with European, Ghanaian and African Surinamese background were analyzed. In

order to answer the question "why subjects with African background are more vulnerable to developing type 2 diabetes than subjects with European background?", we performed Bayesian regression analyses between HbA1c and plasma metabolites. We identified that the relationship between HbA1c and circulating acetoacetate were different in subjects with European and African background.

In chapter 5 we discuss the fourth research example, in which associations of gut microbiome with dietary intakes were analyzed. In this chapter, we showed that the choice of statistical methods can induce bias in significant diet-microbe associations. By performing a large number of simulation studies, we realized that no statistical method was optimal for all microbiome data. In order to achieve the targeted false discovery rate, the best we can do is to run multiple statistical analyses and focus on the significant results identified by multiple methods.

In chapter 6 we discuss the multi-omics integration, the next step to achieve a holistic picture of human phenotypes and disease. In this chapter, we discussed three systems biology platforms for the task. These platforms include gene regulatory networks, protein-protein interaction networks and genome-scale metabolic modeling.

In the last chapter (Chapter 7) we discuss "what are data in biomedical research?" and "how can we introduce prior knowledge into statistical analysis?". We then discuss the future direction of developments in statistical methods.

Nederlandse samenvatting

Dit proefschrift behandeld het vertalen van data uit biomedisch onderzoek naar kennis m.b.v. statistische analyse. Om de beste conclusie uit verzamelde data en bestaande kennis te verkrijgen, dienen er logische keuzes gemaakt worden tijdens de statistische analyse.

In hoofdstuk 1 introduceren we 'high dimensional data', zoals transcriptomics, metabolomics en intestinale microbioom. Vervolgens worden er statistische analyses uitgelicht, zoals permutation, quantification of similarities, Bayesian imputation and statistical stimulation. Tenslotte worden er vier onderzoeken besproken die betrekking hebben tot inflammatie, hypercholesterolemie, diabetes mellitus type 2 en microbe-dieetassociaties.

In hoofdstuk 2 bespreken we het eerste voorbeeld van onderzoek waarin de genexpressie van perifere bloed mononucleaire cellen (PBMC) onder invloed van 4 pathogenen geanalyseerd werd. Door middel van een genset verrijkingsanalyse in de context van ofwel een humaan metabool model op genoomschaal of een biologische database, vonden we metabole biomarkers die de verschillende pathogene stimuli onderscheiden.

In hoofdstuk 3 bespreken we een tweede voorbeeld van onderzoek waarin het plasma metaboloom van 119 vrouwen met hypercholesterolemie werd geanalyseerd. Door het uitvoeren van een hiërarchische clusteranalyse lieten we zien dat hypercholesterolemiepatiënten met en zonder een bekende mutaties verschillende profielen in plasma metaboloom hadden. Daarnaast lieten we zien dat de combinatie van metabolomics en genomics ons kan helpen om hypercholesterolemie zonder duidelijke genetische mutaties

beter te begrijpen.

In hoofdstuk 4 bespreken we een derde voorbeeld van onderzoek waarin het plasma metaboloom van mensen met een Europese, Ghanese en Afrikaanse achtergrond geanalyseerd werd. Om de vraag "waarom zijn mensen van Afrikaanse afkomst gevoeliger voor het ontwikkelen van type 2 diabetes dan mensen van Europese afkomst?" te beantwoorden, hebben we een Bayesiaanse regressieanalyse tussen HbA1c en plasmametabolieten gedaan. We stelden vast dat de relatie tussen HbA1c en circulerend acetoacetaat verschillend was tussen mensen van Afrikaanse en Europeaanse afkomst.

In hoofdstuk 5 bespreken we een vierde voorbeeld van onderzoek waarin de associatie tussen het intestinale microbioom en dieet werd geanalyseerd. In dit hoofdstuk laten we zien dat de keuze voor de statistische methode bias kan introduceren in significante dieet-microbioomassociaties. Door een groot aantal simulatiestudies uit te voeren, realiseerden we dat geen enkele statische methode optimaal is voor alle microbioomdata. Om de beoogde false discovery rate te behalen, is het aanbevolen om meerdere statistische analyses uit te voeren en te focussen op de significante resultaten die door meerdere methodes geïdentificeerd worden.

In hoofdstuk 6 bespreken we de integratie van multi-omics, de volgende stap in het verkrijgen van een holistisch beeld van humaan fenotype en ziekte. In dit hoofdstuk bediscussieerden we drie systeembiologie platformen voor deze taak. Deze platformen omvatten gen-regulatoire netwerken, eiwit-eiwitinteractie netwerken en metabole modellen op een genoomschaal.

In het laatste hoofdstuk, hoofdstuk 7, bespreken we de vragen "wat zijn data in biomedisch onderzoek?" en "hoe kunnen we bestaande kennis introduceren in een statistische analyse?". Vervolgens bediscussiëren we de toekomstige richting van ontwikkelingen in statistische methodes.

Portfolio

Name PhD student: Xiang Zhang

PhD period: 2013 October 1 - 2018 Febuary

Name PhD supervisor: Albert K. Groen; Aeiko H. Zwinderman

1. PhD training

Courses	Year	Workload (Hours/ECTS)
Epidemiology and Applied Statistics	2014	3.0
Data Integration for biologists	2014	1.5
Genetic Epidemiological Research and Data Analysis	2014	1.5
Scientific Writing A-Z, Long Track	2015	1.0
In silico life: constraint-based modelling at genome scale	2016	2.0

International conferences	Year	Workload (Hours/ECTS)
European Lipoprotein Club	2014	1.5
17th International Symposium on Athersclerosis	2015	6.0
86th European Atherosclerosis Society Congress	2018	6.0
Bayes@Lund2018	2018	1.0

2. Publications

Peer reviewed	Year
Forward individualized medicine from personal genomes to interactomes	2015
Identification of discriminating metabolic pathways and metabolites in human PBMCs stimulated by various pathogenic agents	2018
Use of plasma metabolomics to analyze phenotype-genotype relationships in young hypercholesterolemic females	2018

Authors

Identification of discriminating metabolic pathways and metabolites in human PBMCs stimulated by various pathogenic agents

Xiang Zhang[1,*], Adil Mardinoglu[2,3], Leo A.B. Joosten[4], Jan Albert Kuivenhoven[5], Yang Li[6], Mihai G. Netea[4,7], Albert K. Groen[1,8]

1. Department of Experimental Vascular Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
2. Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
3. Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden
4. Department of Internal Medicine, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands
5. Section Molecular Genetics, Department of Pediatrics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
6. Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
7. Department for Genomics & Immunoregulation Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany

8. Department of Laboratory Medicine, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Xiang Zhang did the analysis and wrote the manuscript; Leo A.B. Joosten recruited the participants and performed stimulation experiment; Adil Mardinoglu, Jan Albert Kuivenhoven, Yang Li, Mihai G. Netea and Albert K. Groen helped with manuscript writing.

Use of plasma metabolomics to analyze phenotype-genotype relationships in young hypercholesterolemic females

Xiang Zhang[1], Antoine Rimbert[2], Willem Balder[2,3], Aeiko Having Zwinderman[4], Jan Albert Kuivenhoven[2], Geesje Margaretha Dallinga-Thie[1,*], Albert Kornelis Groen[1,5,*]

1. Department of Experimental Vascular Medicine, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands
2. Department of Pediatrics, Section Molecular Genetics, University Medical Center Groningen, University of Groningen, Antonius Deusinglaan 1, Groningen, The Netherlands
3. Department of Cardiology, Jeroen Bosch Hospital, 's-Hertogenbosch, The Netherlands
4. Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands
5. Department of Pediatrics, University Medical Center Groningen, University of Groningen, Antonius Deusinglaan 1, Groningen, The Netherlands

- These authors contributed equally to the study

Xiang Zhang did the analysis of metabolomics data and wrote the manuscript; Antoine Rimbert and Willem Balder performed genetic analysis. Aeiko Having Zwinderman, Jan Albert Kuivenhoven, Geesje Margaretha Dallinga-Thie and Albert Kornelis Groen helped with manuscript writing.

Association of hemoglobin A1C with circulating metabolites in Dutch with European, African Surinamese and Ghanaian background

Xiang Zhang[1], Inge C.L. van den Munckhof[2], Joost H.W. Rutten[2], Mihai G. Netea[2,3,4], Albert K. Groen [1,5], Aeilko H. Zwinderman[6]

1. Departments of Experimental Vascular Medicine, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands
2. Department of Internal Medicine, Radboud university medical center, Nijmegen, The Netherlands
3. Radboud Center for Infectious Diseases, Radboud university medical Center, Nijmegen, The Netherlands
4. Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany
5. Department of Pediatrics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
6. Clinical Epidemiology, Biostatistics, and Bioinformatics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Xiang Zhang did the analysis of metabolomics data and wrote the manuscript; Inge C.L. van den Munckhof and Joost H.W. Rutten provided metabolomics data of 300-Obesity cohort. Aeilko H. Zwinderman provided metabolomics data of HELIUS cohort. Mihai G. Netea and Albert K. Groen helped with manuscript writing.

Statistical evaluation of diet-microbe associations

Xiang Zhang[1], Max Nieuwdorp[1,2], Albert K. Groen[1], Aeiko H. Zwinderman[3]

1. Amsterdam UMC, University of Amsterdam, Department of Experimental Vascular Medicine, Meibergdreef 9, Amsterdam, The Netherlands

2. Amsterdam UMC, University of Amsterdam, Department of Internal and Vascular Medicine, Meibergdreef 9, Amsterdam, The Netherlands
3. Amsterdam UMC, University of Amsterdam, Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Meibergdreef 9, Amsterdam, The Netherlands

Xiang Zhang did the analysis of metabolomics data and wrote the manuscript; Aeiko H. Zwinderman provided microbiome and diet data of HELIUS cohort. Max Nieuwdorp and Albert K. Groen helped with manuscript writing.

About the Author

Xiang Zhang was born on July 12th, 1988 in Chengdu, China. From 2007 to 2011 he did his bachelor at Sun Yat-sen University, Guangzhou, China and he studied biotechnology. In May 2011, he was awarded the Huygens Scholarship allowing him to come to the Netherlands and study biomedical sciences at University of Amsterdam. In July 2013, he obtained his master degree cum laude. In October 2013, he started his Ph.D. program at University Medical Center Groningen under the supervision of professor A.K. Groen and professor J.A. Kuivenhoven. There he mainly worked on omics data analysis and genome-scale metabolic modeling. In April 2016, he moved to Amsterdam and continued his Ph.D. work at Academic Medical Center under the supervision of professor A.K. Groen and professor A.H. Zwinderman. Here he mainly focused on statistical analysis of high dimensional omics data. He will continue his Postdoctoral research with professor A.K. Groen and A.H. Zwinderman, and he will focus on developing Bayesian methods for systems biology.

References

1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. 2010. "A map of human genome variation from population-scale sequencing." *Nature* 467 (7319): 1061–73. doi:10.1038/nature09534.

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, et al. 2015. "A global reference for human genetic variation." *Nature* 526 (7571): 68–74. doi:10.1038/nature15393.

Agren, Rasmus, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. 2012. "Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT." *PLoS Computational Biology* 8 (5): e1002518. doi:10.1371/journal.pcbi.1002518.

Agren, Rasmus, Adil Mardinoglu, Anna Asplund, Caroline Kampf, Mathias Uhlen, and Jens Nielsen. 2014. "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling." *Molecular Systems Biology* 10 (3): 721. doi:10.1002/msb.145122.

Altay, Gökmen, and Frank Emmert-Streib. 2010a. "Revealing differences in gene network inference algorithms on the network level by ensemble methods." *Bioinformatics (Oxford, England)* 26 (14): 1738–44. doi:10.1093/bioinformatics/btq259.

———. 2010b. "Inferring the conservative causal core of gene regulatory networks." *BMC Systems Biology* 4 (September): 132. doi:10.1186/1752-0509-4-132.

Altman, Russ B. 2007. "PharmGKB: a logical home for knowledge relating genotype to

drug response phenotype." *Nature Genetics* 39 (4): 426. doi:10.1038/ng0407-426.

Balder, Jan-Willem, Antoine Rimbert, Xiang Zhang, Martijn Viel, Roan Kanninga, Freerk van Dijk, Peter Lansberg, Richard Sinke, and Jan Albert Kuivenhoven. 2018. "Genetics, Lifestyle, and Low-Density Lipoprotein Cholesterol in Young and Apparently Healthy Women." *Circulation* 137 (8): 820–31. doi:10.1161/CIRCULATIONAHA.117.032479.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. "Network medicine: a network-based approach to human disease." *Nature Reviews. Genetics* 12 (1): 56–68. doi:10.1038/nrg2918.

Barshir, Ruth, Omer Basha, Amir Eluk, Ilan Y Smoly, Alexander Lan, and Esti Yeger-Lotem. 2013. "The TissueNet database of human tissue protein-protein interactions." *Nucleic Acids Research* 41 (Database issue): D841–4. doi:10.1093/nar/gks1198.

Bartel, Jörg, Jan Krumsiek, and Fabian J Theis. 2013. "Statistical methods for the analysis of high-throughput metabolomics data." *Computational and Structural Biotechnology Journal* 4: e201301009. doi:10.5936/csbj.201301009.

Bartel, Jörg, Jan Krumsiek, Katharina Schramm, Jerzy Adamski, Christian Gieger, Christian Herder, Maren Carstensen, et al. 2015a. "The Human Blood Metabolome-Transcriptome Interface." *PLoS Genetics* 11 (6): e1005274. doi:10.1371/journal.pgen.1005274.

———. 2015b. "The Human Blood Metabolome-Transcriptome Interface." *PLoS Genetics* 11 (6): e1005274. doi:10.1371/journal.pgen.1005274.

Bartuzi, Paulina, Daniel D Billadeau, Robert Favier, Shunxing Rong, Daphne Dekker, Alina Fedoseienko, Hille Fieten, et al. 2016. "CCC- and Wash-Mediated Endosomal Sorting of Ldlr Is Required for Normal Clearance of Circulating Ldl." *Nat Commun* 7 (March): 10961. doi:10.1038/ncomms10961.

Basso, Katia, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. 2005. "Reverse engineering of regulatory networks in human B cells." *Nature Genetics* 37 (4): 382–90. doi:10.1038/ng1532.

Becker, Scott A, and Bernhard O Palsson. 2008. "Context-specific metabolic networks are consistent with experiments." *PLoS Computational Biology* 4 (5): e1000082. doi:10.1371/journal.pcbi.1000082.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A

Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1). [Royal Statistical Society, Wiley]: 289–300. http://www.jstor.org/stable/2346101.

Beukers, M H, L H Dekker, E J de Boer, C W M Perenboom, S Meijboom, M Nicolaou, J H M de Vries, and H A M Brants. 2015. "Development of the HELIUS food frequency questionnaires: ethnic-specific questionnaires to assess the diet of a multiethnic population in The Netherlands." *European Journal of Clinical Nutrition* 69 (5): 579–84. doi:10.1038/ejcn.2014.180.

Bindraban, Navin R, Irene G M van Valkengoed, Gideon Mairuhu, Frits Holleman, Joost B L Hoekstra, Bob P J Michels, Richard P Koopmans, and Karien Stronks. 2008. "Prevalence of diabetes mellitus and the performance of a risk score among Hindustani Surinamese, African Surinamese and ethnic Dutch: a cross-sectional population-based study." *BMC Public Health* 8 (August): 271. doi:10.1186/1471-2458-8-271.

Bjornsson, Hans T, M Daniele Fallin, and Andrew P Feinberg. 2004. "An integrated epigenetic and genetic approach to common human disease." *Trends in Genetics : TIG* 20 (8): 350–8. doi:10.1016/j.tig.2004.06.009.

Blank, Lars M, and Birgitta E Ebert. 2013. "From measurement to implementation of metabolic fluxes." *Current Opinion in Biotechnology* 24 (1): 13–21. doi:10.1016/j.copbio.2012.10.019.

Blumenthal, Antje, Gayathri Nagalingam, Jennifer H Huch, Lara Walker, Gilles J Guillemin, George A Smythe, Sabine Ehrt, Warwick J Britton, and Bernadette M Saunders. 2012. "M. tuberculosis induces potent activation of IDO-1, but this is not essential for the immunological control of infection." *PloS One* 7 (5): e37314. doi:10.1371/journal.pone.0037314.

Bordbar, Aarash, Adam M Feist, Renata Usaite-Black, Joseph Woodcock, Bernhard O Palsson, and Iman Famili. 2011. "A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology." *BMC Systems Biology* 5 (October): 180. doi:10.1186/1752-0509-5-180.

Bordbar, Aarash, Monica L Mo, Ernesto S Nakayasu, Alexandra C Schrimpe-Rutledge, Young-Mo Kim, Thomas O Metz, Marcus B Jones, et al. 2012. "Model-driven multiomic data analysis elucidates metabolic immunomodulators of macrophage activation."

Molecular Systems Biology 8 (June): 558. doi:10.1038/msb.2012.21.

Bordbar, Aarash, Jonathan M Monk, Zachary A King, and Bernhard O Palsson. 2014. "Constraint-based models predict metabolic and associated cellular functions." *Nature Reviews. Genetics* 15 (2): 107–20. doi:10.1038/nrg3643.

Bordel, Sergio, Rasmus Agren, and Jens Nielsen. 2010. "Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes." *PLoS Computational Biology* 6 (7): e1000859. doi:10.1371/journal.pcbi.1000859.

Breuer, Karin, Amir K Foroushani, Matthew R Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L Winsor, Robert E W Hancock, Fiona S L Brinkman, and David J Lynn. 2013. "InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation." *Nucleic Acids Research* 41 (Database issue): D1228–33. doi:10.1093/nar/gks1147.

Brown, M S, and J L Goldstein. 1986. "A Receptor-Mediated Pathway for Cholesterol Homeostasis." *Science (New York, N.Y.)* 232 (4746): 34–47.

Brunham, Liam R, Janine K Kruit, Terry D Pape, Jenelle M Timmins, Anne Q Reuwer, Zainisha Vasanji, Brad J Marsh, et al. 2007. "Beta-cell ABCA1 influences insulin secretion, glucose homeostasis and response to thiazolidinedione treatment." *Nat Med* 13 (3): 340–47. doi:10.1038/nm1546.

Burke, Wylie, and Diane M Korngiebel. 2015. "Closing the gap between knowledge and clinical application: challenges for genomic translation." *PLoS Genetics* 11 (2): e1004978. doi:10.1371/journal.pgen.1004978.

Butte, A J, and I S Kohane. 2000. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418–29. http://www.ncbi.nlm.nih.gov/pubmed/10902190.

Califano, Andrea, Atul J Butte, Stephen Friend, Trey Ideker, and Eric Schadt. 2012. "Leveraging models of cell regulation and GWAS data in integrative network-based association studies." *Nature Genetics* 44 (8): 841–7. doi:10.1038/ng.2355.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan:

A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 1–32. doi:10.18637/jss.v076.i01.

Carro, Maria Stella, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, et al. 2010. "The transcriptional network for mesenchymal transformation of brain tumours." *Nature* 463 (7279): 318–25. doi:10.1038/nature08712.

Chen, James C, Mariano J Alvarez, Flaminia Talos, Harshil Dhruv, Gabrielle E Rieckhof, Archana Iyer, Kristin L Diefes, et al. 2014. "Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks." *Cell* 159 (2): 402–14. doi:10.1016/j.cell.2014.09.021.

Chen, Rui, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y K Lam, Rong Chen, Elana Miriami, et al. 2012. "Personal omics profiling reveals dynamic molecular and medical phenotypes." *Cell* 148 (6): 1293–1307. doi:10.1016/j.cell.2012.02.009.

Chen, Yunshun, Aaron T L Lun, and Gordon K Smyth. 2014. "Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR." In *Statistical Analysis of Next Generation Sequencing Data*, edited by Somnath Datta and Dan Nettleton, 51–74. Cham: Springer International Publishing. doi:10.1007/978-3-319-07212-8_3.

Cheng, Shih-Chin, Leo A B Joosten, and Mihai G Netea. 2014. "The interplay between central metabolism and innate immune responses." *Cytokine & Growth Factor Reviews* 25 (6): 707–13. doi:10.1016/j.cytogfr.2014.06.008.

Cover, Thomas M, and Joy A Thomas. 2006. *Elements of information theory {(2.} ed.)*. Wiley.

Crews, K R, J K Hicks, C-H Pui, M V Relling, and W E Evans. 2012. "Pharmacogenomics and individualized medicine: translating science into practice." *Clinical Pharmacology and Therapeutics* 92 (4): 467–75. doi:10.1038/clpt.2012.120.

Cusick, Michael E, Niels Klitgord, Marc Vidal, and David E Hill. 2005. "Interactome: gateway into systems biology." *Human Molecular Genetics* 14 Spec No (October): R171–81. doi:10.1093/hmg/ddi335.

de Matos Simoes, Ricardo, and Frank Emmert-Streib. 2012. "Bagging statistical network inference from large-scale gene expression data." *PloS One* 7 (3): e33624.

doi:10.1371/journal.pone.0033624.

de Matos Simoes, Ricardo, Matthias Dehmer, and Frank Emmert-Streib. 2013. "B-cell lymphoma gene regulatory networks: biological consistency among inference methods." *Frontiers in Genetics* 4: 281. doi:10.3389/fgene.2013.00281.

Dekker, Louise H, Rob M van Dam, Marieke B Snijder, Ron J G Peters, Jacqueline M Dekker, Jeanne H M de Vries, Evelien J de Boer, Matthias B Schulze, Karien Stronks, and Mary Nicolaou. 2015. "Comparable Dietary Patterns Describe Dietary Behavior across Ethnic Groups in the Netherlands, but Different Elements in the Diet Are Associated with Glycated Hemoglobin and Fasting Glucose Concentrations." *The Journal of Nutrition* 145 (8): 1884–91. doi:10.3945/jn.114.207472.

Dekker, Louise H, Marieke B Snijder, Marja H Beukers, Jeanne H M de Vries, Henny A M Brants, Evelien J de Boer, Rob M van Dam, Karien Stronks, and Mary Nicolaou. 2011. "A prospective cohort study of dietary patterns of non-western migrants in the Netherlands in relation to risk factors for cardiovascular diseases: HELIUS-Dietary Patterns." *BMC Public Health* 11: 441. doi:10.1186/1471-2458-11-441.

Dekkers, Koen F, Maarten van Iterson, Roderick C Slieker, Matthijs H Moed, Marc Jan Bonder, Michiel van Galen, Hailiang Mei, et al. 2016. "Blood Lipids Influence Dna Methylation in Circulating Cells." *Genome Biol* 17 (1): 138. doi:10.1186/s13059-016-1000-6.

Deschasaux, Mélanie, Kristien E Bouter, Andrei Prodan, Evgeni Levin, Albert K Groen, Hilde Herrema, Valentina Tremaroli, et al. 2018. "Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography." *Nature Medicine* 24 (10): 1526–31. doi:10.1038/s41591-018-0160-1.

Diabetes Prevention Program Research Group. 2015. "HbA1c as a predictor of diabetes and as an outcome in the diabetes prevention program: a randomized clinical trial." *Diabetes Care* 38 (1): 51–58. doi:10.2337/dc14-0886.

Du, Pan, Gilbert Feng, Warren Kibbe, and Simon Lin. 2016. *lumiHumanIDMapping: Illumina Identifier mapping for Human.* https://bioconductor.org/packages/release/data/annotation/html/lumiHumanIDMapping.html.

Du, Xian-Ming, Mi-Jurng Kim, Liming Hou, Wilfried Le Goff, M John Chapman, Miranda

Van Eck, Linda K Curtiss, et al. 2015. “HDL particle size is a critical determinant of ABCA1-mediated macrophage cellular cholesterol export.” *Circ Res* 116 (7): 1133–42. doi:10.1161/CIRCRESAHA.116.305485.

Duarte, Natalie C, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. 2007. “Global reconstruction of the human metabolic network based on genomic and bibliomic data.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (6): 1777–82. doi:10.1073/pnas.0610772104.

Durinck, Steffen, Paul T Spellman, Ewan Birney, and Wolfgang Huber. 2009. “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.” *Nature Protocols* 4 (8): 1184–91. doi:10.1038/nprot.2009.97.

Duvallet, Claire, Sean M Gibbons, Thomas Gurry, Rafael A Irizarry, and Eric J Alm. 2017. “Meta-analysis of gut microbiome studies identifies disease-specific and shared responses.” *Nature Communications* 8 (1): 1784. doi:10.1038/s41467-017-01973-8.

Eckardstein, Arnold von, and Rahel A Sibler. 2011. “Possible contributions of lipoproteins and cholesterol to the pathogenesis of diabetes mellitus type 2.” *Current Opinion in Lipidology* 22 (1): 26–32. doi:10.1097/MOL.0b013e3283412279.

Elefsinioti, Antigoni, Ömer Sinan Saraç, Anna Hegele, Conrad Plake, Nina C Hubner, Ina Poser, Mihail Sarov, et al. 2011. “Large-scale de novo prediction of physical protein-protein association.” *Molecular & Cellular Proteomics : MCP* 10 (11): M111.010629. doi:10.1074/mcp.M111.010629.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains. 2014. “Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks.” *Frontiers in Cell and Developmental Biology* 2: 38. doi:10.3389/fcell.2014.00038.

Everts, Bart, Eyal Amiel, Stanley Ching-Cheng Huang, Amber M Smith, Chih-Hao Chang, Wing Y Lam, Veronika Redmann, et al. 2014. “TLR-driven early glycolytic reprogramming via the kinases TBK1-IKKepsilon supports the anabolic demands of dendritic cell activation.” *Nature Immunology* 15 (4). Nature Research: 323–32.

doi:10.1038/ni.2833.

Faith, Jeremiah J, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. 2007. "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." *PLoS Biology* 5 (1): e8. doi:10.1371/journal.pbio.0050008.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, et al. 2015. "Genetic and epigenetic fine mapping of causal autoimmune disease variants." *Nature* 518 (7539): 337–43. doi:10.1038/nature13835.

Feinberg, Andrew P. 2007. "Phenotypic plasticity and the epigenetics of human disease." *Nature* 447 (7143): 433–40. doi:10.1038/nature05919.

Feinberg, Andrew P, Rafael A Irizarry, Delphine Fradin, Martin J Aryee, Peter Murakami, Thor Aspelund, Gudny Eiriksdottir, et al. 2010. "Personalized epigenomic signatures that are stable over time and covary with body mass index." *Science Translational Medicine* 2 (49): 49ra67. doi:10.1126/scitranslmed.3001262.

Ference, Brian A, Henry N Ginsberg, Ian Graham, Kausik K Ray, Chris J Packard, Eric Bruckert, Robert A Hegele, et al. 2017. "Low-Density Lipoproteins Cause Atherosclerotic Cardiovascular Disease. 1. Evidence from Genetic, Epidemiologic, and Clinical Studies. a Consensus Statement from the European Atherosclerosis Society Consensus Panel." *Eur Heart J* 38 (32): 2459–72. doi:10.1093/eurheartj/ehx144.

Festa, Andreas, Ken Williams, Anthony J G Hanley, James D Otvos, David C Goff, Lynne E Wagenknecht, and Steven M Haffner. 2005. "Nuclear magnetic resonance lipoprotein abnormalities in prediabetic subjects in the Insulin Resistance Atherosclerosis Study." *Circulation* 111 (25): 3465–72. doi:10.1161/CIRCULATIONAHA.104.512079.

Fischer, Krista, Johannes Kettunen, Peter Würtz, Toomas Haller, Aki S Havulinna, Antti J Kangas, Pasi Soininen, et al. 2014. "Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons." *PLoS Med* 11 (2): e1001606. doi:10.1371/journal.pmed.1001606.

Fizelova, Maria, Manna Miilunpohja, Antti J Kangas, Pasi Soininen, Johanna Kuusisto,

Mika Ala-Korpela, Markku Laakso, and Alena Stančáková. 2015. "Associations of multiple lipoprotein and apolipoprotein measures with worsening of glycemia and incident type 2 diabetes in 6607 non-diabetic Finnish men." *Atherosclerosis* 240 (1): 272–77. doi:10.1016/j.atherosclerosis.2015.03.034.

Forler, Stefanie, Oliver Klein, and Joachim Klose. 2014. "Individualized proteomics." *Journal of Proteomics* 107 (July): 56–61. doi:10.1016/j.jprot.2014.04.003.

Franceschini, Andrea, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, et al. 2013. "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." *Nucleic Acids Research* 41 (Database issue): D808–15. doi:10.1093/nar/gks1094.

Fryirs, Michelle A, Philip J Barter, Mathiyalagan Appavoo, Bernard E Tuch, Fatiha Tabet, Alison K Heather, and Kerry-Anne Rye. 2010. "Effects of high-density lipoproteins on pancreatic beta-cell insulin secretion." *Arterioscler Thromb Vasc Biol* 30 (8): 1642–8. doi:10.1161/ATVBAHA.110.207373.

Futema, Marta, Vincent Plagnol, KaWah Li, Ros A Whittall, H Andrew W Neil, Mary Seed, Simon Broome Consortium, et al. 2014. "Whole Exome Sequencing of Familial Hypercholesterolaemia Patients Negative for Ldlr/Apob/Pcsk9 Mutations." *Journal of Medical Genetics* 51 (8): 537–44. doi:10.1136/jmedgenet-2014-102405.

Garcia-Alonso, Luz, Jorge Jiménez-Almazán, Jose Carbonell-Caballero, Alicia Vela-Boza, Javier Santoyo-López, Guillermo Antiñolo, and Joaquin Dopazo. 2014. "The role of the interactome in the maintenance of deleterious variability in human populations." *Molecular Systems Biology* 10 (September): 752. doi:10.15252/msb.20145222.

Garvey, W Timothy, Soonho Kwon, Deyi Zheng, Sara Shaughnessy, Penny Wallace, Amy Hutto, Kimberly Pugh, Alicia J Jenkins, Richard L Klein, and Youlian Liao. 2003. "Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance." *Diabetes* 52 (2): 453–62.

Gautier, Thomas, David Masson, Miek C Jong, Linda Duverneuil, Naig Le Guern, Valérie Deckert, Jean-Paul Pais de Barros, et al. 2002. "Apolipoprotein Ci Deficiency Markedly Augments Plasma Lipoprotein Changes Mediated by Human Cholesteryl Ester Transfer Protein (Cetp) in Cetp Transgenic/Apoci-Knocked Out Mice." *J Biol Chem* 277 (35):

31354–63. doi:10.1074/jbc.M203151200.

Gavai, Anand K, Farahaniza Supandi, Hannes Hettling, Paul Murrell, Jack A M Leunissen, and Johannes H G M van Beek. 2015. "Using bioconductor package BiGGR for metabolic flux estimation based on gene expression changes in brain." *PloS One* 10 (3): e0119016. doi:10.1371/journal.pone.0119016.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics* 2 (4): 1360–83. doi:10.1214/08-AOAS191.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The prior can often only be understood in the context of the likelihood." *Entropy* 19 (10): 1–13. doi:10.3390/e19100555.

Ghiassian, Susan Dina, Jörg Menche, and Albert-László Barabási. 2015. "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome." *PLoS Computational Biology* 11 (4): e1004120. doi:10.1371/journal.pcbi.1004120.

Gille, Christoph, Christian Bölling, Andreas Hoppe, Sascha Bulik, Sabrina Hoffmann, Katrin Hübner, Anja Karlstädt, et al. 2010. "HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology." *Molecular Systems Biology* 6 (September): 411. doi:10.1038/msb.2010.62.

Goehler, Heike, Maciej Lalowski, Ulrich Stelzl, Stephanie Waelter, Martin Stroedicke, Uwe Worm, Anja Droege, et al. 2004. "A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease." *Molecular Cell* 15 (6): 853–65. doi:10.1016/j.molcel.2004.09.016.

Goeman, Jelle J, and Peter Bühlmann. 2007. "Analyzing gene expression data in terms of gene sets: methodological issues." *Bioinformatics (Oxford, England)* 23 (8): 980–7. doi:10.1093/bioinformatics/btm051.

Goldstein, Joseph L, and Michael S Brown. 2015. "A Century of Cholesterol and Coronaries: From Plaques to Genes to Statins." *Cell* 161 (1): 161–72. doi:10.1016/j.cell.2015.01.036.

Guasch-Ferré, Marta, Adela Hruby, Estefanía Toledo, Clary B Clish, Miguel A Martínez-

González, Jordi Salas-Salvadó, and Frank B Hu. 2016. "Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis." *Diabetes Care* 39 (5): 833–46. doi:10.2337/dc15-2251.

Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, et al. 2012. "GENCODE: the reference human genome annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74. doi:10.1101/gr.135350.111.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. 2017. "Multi-omics approaches to disease." *Genome Biology* 18 (1). Genome Biology: 83. doi:10.1186/s13059-017-1215-1.

Hellmuth, Christian, Franca Fabiana Kirchberg, Nina Lass, Ulrike Harder, Wolfgang Peissner, Berthold Koletzko, and Thomas Reinehr. 2016. "Tyrosine Is Associated with Insulin Resistance in Longitudinal Metabolomic Profiling of Obese Children." *J Diabetes Res* 2016: 2108909. doi:10.1155/2016/2108909.

Henderson, D C, and J J Rippin. 1995. "Stimulus-dependent production of cytokines and pterins by peripheral blood mononuclear cells." *Immunology Letters* 45 (1-2): 29–34. http://www.ncbi.nlm.nih.gov/pubmed/7622184.

Hewett, Micheal, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. 2002. "PharmGKB: the Pharmacogenetics Knowledge Base." *Nucleic Acids Research* 30 (1): 163–5. http://www.ncbi.nlm.nih.gov/pubmed/11752281 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC99138.

Hirata, Hiroshi, Yuji Hinoda, Varahram Shahryari, Guoren Deng, Koichi Nakajima, Z Laura Tabatabai, Nobuhisa Ishii, and Rajvir Dahiya. 2015. "Long Noncoding RNA MALAT1 Promotes Aggressive Renal Cell Carcinoma through Ezh2 and Interacts with miR-205." *Cancer Research* 75 (7): 1322–31. doi:10.1158/0008-5472.CAN-14-2931.

Holmes, Ian, Keith Harris, and Christopher Quince. 2012. "Dirichlet multinomial mixtures: generative models for microbial metagenomics." *PloS One* 7 (2): e30126. doi:10.1371/journal.pone.0030126.

Holmes, Michael V, Iona Y Millwood, Christiana Kartsonaki, Michael R Hill, Derrick A Bennett, Ruth Boxall, Yu Guo, et al. 2018. "Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke." *J Am Coll Cardiol* 71 (6): 620–32.

doi:10.1016/j.jacc.2017.12.006.

Hoppe, Andreas. 2012. "What mRNA Abundances Can Tell us about Metabolism." *Metabolites* 2 (3): 614–31. doi:10.3390/metabo2030614.

Hotamisligil, Gökhan S. 2017. "Inflammation, metaflammation and immunometabolic disorders." *Nature* 542 (7640): 177–85. doi:10.1038/nature21363.

Howell, Paul M, Suhu Liu, Suping Ren, Campbell Behlen, Oystein Fodstad, and Adam I Riker. 2009. "Epigenetics in human melanoma." *Cancer Control : Journal of the Moffitt Cancer Center* 16 (3): 200–218. doi:10.1177/107327480901600302.

Hu, Yan-Wei, Jun-Yao Yang, Xin Ma, Zhi-Ping Chen, Ya-Rong Hu, Jia-Yi Zhao, Shu-Fen Li, et al. 2014. "A lincRNA-Dynlrb2-2/Gpr119/Glp-1r/Abca1-Dependent Signal Transduction Pathway Is Essential for the Regulation of Cholesterol Homeostasis." *J Lipid Res* 55 (4): 681–97. doi:10.1194/jlr.M044669.

Huttlin, Edward L, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, et al. 2015. "The BioPlex Network: A Systematic Exploration of the Human Interactome." *Cell* 162 (2): 425–40. doi:10.1016/j.cell.2015.06.043.

Inouye, Michael, Johannes Kettunen, Pasi Soininen, Kaisa Silander, Samuli Ripatti, Linda S Kumpula, Eija Hämäläinen, et al. 2010. "Metabonomic, transcriptomic, and genomic variation of a population cohort." *Mol Syst Biol* 6 (December): 441. doi:10.1038/msb.2010.93.

Irani, Sara, Jahangir Iqbal, W James Antoni, Laraib Ijaz, and M Mahmood Hussain. 2018. "MicroRNA-30c Reduces Plasma Cholesterol in Homozygous Familial Hypercholesterolemic and Type 2 Diabetic Mouse Models." *J Lipid Res* 59 (1): 144–54. doi:10.1194/jlr.M081299.

Irizarry, Rafael a, Chi Wang, Yun Zhou, and Terence P Speed. 2009. "Gene set enrichment analysis made simple." *Statistical Methods in Medical Research* 18 (6): 565–75. doi:10.1177/0962280209351908.

Jia, Peilin, and Zhongming Zhao. 2014. "VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data." *PLoS Computational Biology* 10 (2): e1003460. doi:10.1371/journal.pcbi.1003460.

Johansson, Lennart F, Freerk van Dijk, Eddy N de Boer, Krista K van Dijk-Bos, Jan D

H Jongbloed, Annemieke H van der Hout, Helga Westers, et al. 2016. "CoNVaDING: Single Exon Variation Detection in Targeted Ngs Data." *Hum Mutat* 37 (5): 457–64. doi:10.1002/humu.22969.

Jonsson, Viktor, Tobias Österlund, Olle Nerman, and Erik Kristiansson. 2016. "Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics." *BMC Genomics* 17 (January): 78. doi:10.1186/s12864-016-2386-y.

Jovel, Juan, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, et al. 2016. "Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics." *Frontiers in Microbiology* 7: 459. doi:10.3389/fmicb.2016.00459.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2012. "KEGG for integration and interpretation of large-scale molecular data sets." *Nucleic Acids Research* 40 (Database issue): D109——14. doi:10.1093/nar/gkr988.

Karczewski, Konrad J., and Michael P. Snyder. 2018. "Integrative omics for health and disease." *Nature Reviews. Genetics* 19 (5). Nature Publishing Group: 299–310. doi:10.1038/nrg.2018.4.

Karlebach, Guy, and Ron Shamir. 2008. "Modelling and analysis of gene regulatory networks." *Nature Reviews. Molecular Cell Biology* 9 (10): 770–80. doi:10.1038/nrm2503.

Kathiresan, Sekar, and Deepak Srivastava. 2012. "Genetics of human cardiovascular disease." *Cell* 148 (6): 1242–57. doi:10.1016/j.cell.2012.03.001.

Keshava Prasad, T S, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, et al. 2009. "Human Protein Reference Database–2009 update." *Nucleic Acids Research* 37 (Database issue): D767–72. doi:10.1093/nar/gkn892.

Kettunen, Johannes, Michael V Holmes, Elias Allara, Olga Anufrieva, Pauli Ohukainen, Clare Oliver-Williams, Therese Tillin, et al. 2018. "Lipoprotein Signatures of Cholesteryl Ester Transfer Protein and Hmg-Coa Reductase Inhibition." *bioRxiv*. Cold Spring Harbor Laboratory. doi:10.1101/295394.

Khera, Amit V, Connor A Emdin, Isabel Drake, Pradeep Natarajan, Alexander G Bick, Nancy R Cook, Daniel I Chasman, et al. 2016. "Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease." *N Engl J Med* 375 (24): 2349–58.

doi:10.1056/NEJMoa1605086.

Khera, Amit V, Hong-Hee Won, Gina M Peloso, Kim S Lawson, Traci M Bartz, Xuan Deng, Elisabeth M van Leeuwen, et al. 2016. "Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients with Severe Hypercholesterolemia." *J Am Coll Cardiol* 67 (22): 2578–89. doi:10.1016/j.jacc.2016.03.520.

Kim, Min-Sik, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S Manda, Raghothama Chaerkady, Anil K Madugundu, et al. 2014. "A draft map of the human proteome." *Nature* 509 (7502): 575–81. doi:10.1038/nature13302.

Kleinnijenhuis, Johanneke, Marije Oosting, Leo A B Joosten, Mihai G Netea, and Reinout Van Crevel. 2011. "Innate immune recognition of Mycobacterium tuberculosis." *Clinical & Developmental Immunology* 2011: 405310. doi:10.1155/2011/405310.

KNOX, W E, and M LeMAY-KNOX. 1951. "The oxidation in liver of l-tyrosine to acetoacetate through p-hydroxyphenylpyruvate and homogentisic acid." *Biochem J* 49 (5): 686–93.

Kristiansson, Erik, Philip Hugenholtz, and Daniel Dalevi. 2009. "ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes." *Bioinformatics (Oxford, England)* 25 (20): 2737–8. doi:10.1093/bioinformatics/btp508.

Kruit, J K, P H C Kremer, L Dai, R Tang, P Ruddle, W de Haan, L R Brunham, C B Verchere, and M R Hayden. 2010. "Cholesterol efflux via ATP-binding cassette transporter A1 (ABCA1) and cholesterol uptake via the LDL receptor influences cholesterol-induced impairment of beta cell function in mice." *Diabetologia* 53 (6): 1110–9. doi:10.1007/s00125-010-1691-2.

Kruit, Janine K, Nadeeja Wijesekara, Jocelyn E Manning Fox, Xiao-Qing Dai, Liam R Brunham, Gavin J Searle, Garry P Morgan, et al. 2011. "Islet cholesterol accumulation due to loss of ABCA1 leads to impaired exocytosis of insulin granules." *Diabetes* 60 (12): 3186–96. doi:10.2337/db11-0081.

Kuchaiev, Oleksii, Marija Rasajski, Desmond J Higham, and Natasa Przulj. 2009. "Geometric de-noising of protein-protein interaction networks." *PLoS Computational Biology* 5 (8): e1000454. doi:10.1371/journal.pcbi.1000454.

Laarhoven, Arjan van, Sofiati Dian, Raúl Aguirre-Gamboa, Julian Avila-Pacheco, Isis

Ricaño-Ponce, Carolien Ruesen, Jessi Annisa, et al. 2018. "Cerebral tryptophan metabolism and outcome of tuberculous meningitis: an observational cohort study." *The Lancet. Infectious Diseases* 18 (5). Elsevier: 526–35. doi:10.1016/S1473-3099(18)30053-7.

Lachmandas, Ekta, Lily Boutens, Jacqueline M Ratter, Anneke Hijmans, Guido J Hooiveld, Leo A B Joosten, Richard J Rodenburg, et al. 2016. "Microbial stimulation of different Toll-like receptor signalling pathways induces diverse metabolic programmes in human monocytes." *Nature Microbiology* 2 (December). Nature Publishing Group: 16246. doi:10.1038/nmicrobiol.2016.246.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." *Genome Biology* 15 (2): R29. doi:10.1186/gb-2014-15-2-r29.

Leek, Jeffrey T, and John D Storey. 2007. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS Genetics* 3 (9): 1724–35. doi:10.1371/journal.pgen.0030161.

———. 2008. "A general framework for multiple testing dependence." *Proceedings of the National Academy of Sciences of the United States of America* 105 (48): 18718–23. doi:10.1073/pnas.0808709105.

Leek, Jeffrey T, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. 2012. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." *Bioinformatics (Oxford, England)* 28 (6): 882–83. doi:10.1093/bioinformatics/bts034.

Lefebvre, Celine, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, et al. 2010. "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers." *Molecular Systems Biology* 6 (June): 377. doi:10.1038/msb.2010.31.

Lewis, Nathan E, Gunnar Schramm, Aarash Bordbar, Jan Schellenberger, Michael P Andersen, Jeffrey K Cheng, Nilam Patel, et al. 2010. "Large-scale in silico modeling of metabolic interactions between cell types in the human brain." *Nature Biotechnology* 28 (12): 1279–85. doi:10.1038/nbt.1711.

Li, Yang, Marije Oosting, Patrick Deelen, Isis Ricaño-Ponce, Sanne Smeekens, Martin

Jaeger, Vasiliki Matzaraki, et al. 2016. “Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi.” *Nature Medicine* 22 (8): 952–60. doi:10.1038/nm.4139.

Licata, Luana, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, et al. 2012. “MINT, the molecular interaction database: 2012 update.” *Nucleic Acids Research* 40 (Database issue): D857–61. doi:10.1093/nar/gkr930.

Lin, Simon M, Pan Du, Wolfgang Huber, and Warren A Kibbe. 2008. “Model-based variance-stabilizing transformation for Illumina microarray data.” *Nucleic Acids Research* 36 (2): e11. doi:10.1093/nar/gkm1075.

Lorenzo, Andrea De, Juliana Duarte Lopes da Silva, Cinthia E James, Alexandre C Pereira, and Annie Seixas Bello Moreira. 2018. “Clinical, Anthropometric and Biochemical Characteristics of Patients with or Without Genetically Confirmed Familial Hypercholesterolemia.” *Arq Bras Cardiol* 110 (2): 119–23. doi:10.5935/abc.20180005.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology* 15 (12): 550. doi:10.1186/s13059-014-0550-8.

Low, Siew-Kee, Atsushi Takahashi, Taisei Mushiroda, and Michiaki Kubo. 2014. “Genome-wide association study: a useful tool to identify common genetic variants associated with drug toxicity and efficacy in cancer pharmacogenomics.” *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 20 (10): 2541–52. doi:10.1158/1078-0432.CCR-13-2755.

Lynch, Christopher J, and Sean H Adams. 2014. “Branched-chain amino acids in metabolic signalling and insulin resistance.” *Nat Rev Endocrinol* 10 (12): 723–36. doi:10.1038/nrendo.2014.171.

Lynch, Susan V, and Oluf Pedersen. 2016. “The Human Intestinal Microbiome in Health and Disease.” *The New England Journal of Medicine* 375 (24): 2369–79. doi:10.1056/NEJMra1600266.

Ma, Hongwu, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. 2007. “The Edinburgh human metabolic net-

work reconstruction and its functional analysis." *Molecular Systems Biology* 3: 135. doi:10.1038/msb4100177.

Machado, Daniel, and Markus Herrgård. 2014. "Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism." *PLoS Computational Biology* 10 (4): e1003580. doi:10.1371/journal.pcbi.1003580.

Mackey, Rachel H, Samia Mora, Alain G Bertoni, Christina L Wassel, Mercedes R Carnethon, Christopher T Sibley, and David C Goff Jr. 2015. "Lipoprotein particles and incident type 2 diabetes in the multi-ethnic study of atherosclerosis." *Diabetes Care* 38 (4): 628–36. doi:10.2337/dc14-0645.

Magger, Oded, Yedael Y Waldman, Eytan Ruppin, and Roded Sharan. 2012. "Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks." *PLoS Computational Biology* 8 (9): e1002690. doi:10.1371/journal.pcbi.1002690.

Mardinoglu, Adil, Rasmus Agren, Caroline Kampf, Anna Asplund, Intawat Nookaew, Peter Jacobson, Andrew J Walley, et al. 2013. "Integration of clinical data with a genome-scale metabolic model of the human adipocyte." *Molecular Systems Biology* 9 (649). Nature Publishing Group: 649. doi:10.1038/msb.2013.5.

Mardinoglu, Adil, Rasmus Agren, Caroline Kampf, Anna Asplund, Mathias Uhlen, and Jens Nielsen. 2014. "Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease." *Nature Communications* 5: 3083. doi:10.1038/ncomms4083.

Mardinoglu, Adil, and Jens Nielsen. 2015. "New paradigms for metabolic modeling of human cells." *Current Opinion in Biotechnology* 34 (August): 91–97. doi:10.1016/j.copbio.2014.12.013.

Mardinoglu, Adil, Francesco Gatto, and Jens Nielsen. 2013. "Genome-scale modeling of human metabolism - a systems biology approach." *Biotechnology Journal* 8 (9): 985–96. doi:10.1002/biot.201200275.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." *BMC Bioin-*

formatics 7 Suppl 1 (March): S7. doi:10.1186/1471-2105-7-S1-S7.

Marques-Pinheiro, Alice, Marie Marduel, Jean-Pierre Rabès, Martine Devillers, Ludovic Villéger, Delphine Allard, Jean Weissenbach, et al. 2010. "A Fourth Locus for Autosomal Dominant Hypercholesterolemia Maps at 16q22.1." *Eur J Hum Genet* 18 (11): 1236–42. doi:10.1038/ejhg.2010.94.

Maskarinec, Gertraud, Andrew Grandinetti, Grace Matsuura, Sangita Sharma, Marjorie Mau, Brian E Henderson, and Laurence N Kolonel. 2009. "Diabetes prevalence and body mass index differ by ethnicity: the Multiethnic Cohort." *Ethnicity & Disease* 19 (1): 49–55. http://www.ncbi.nlm.nih.gov/pubmed/19341163 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2702477.

Mayer, François L, Duncan Wilson, and Bernhard Hube. 2013. "Candida albicans pathogenicity mechanisms." *Virulence* 4 (2): 119–28. doi:10.4161/viru.22913.

McDowall, Mark D, Michelle S Scott, and Geoffrey J Barton. 2009. "PIPs: human protein-protein interaction prediction database." *Nucleic Acids Research* 37 (Database issue): D651–6. doi:10.1093/nar/gkn870.

McGettrick, Anne F, and Luke A J O'Neill. 2013. "How metabolism generates signals during innate immunity and inflammation." *The Journal of Biological Chemistry* 288 (32): 22893–8. doi:10.1074/jbc.R113.486464.

McMurdie, Paul J, and Susan Holmes. 2013. "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." *PLoS One* 8 (4): e61217. doi:10.1371/journal.pone.0061217.

———. 2014. "Waste not, want not: why rarefying microbiome data is inadmissible." *PLoS Comput Biol* 10 (4): e1003531. doi:10.1371/journal.pcbi.1003531.

Meeks, Karlijn A C, Karien Stronks, Erik J A J Beune, Adebowale Adeyemo, Peter Henneman, Marcel M A M Mannens, Mary Nicolaou, et al. 2015. "Prevalence of type 2 diabetes and its association with measures of body composition among African residents in the Netherlands–The HELIUS study." *Diabetes Res Clin Pract* 110 (2): 137–46. doi:10.1016/j.diabres.2015.09.017.

Menche, Jörg, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. 2015. "Disease networks. Uncovering

disease-disease relationships through the incomplete interactome." *Science (New York, N.Y.)* 347 (6224): 1257601. doi:10.1126/science.1257601.

Meyer, Patrick E, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. 2007. "Information-theoretic inference of large transcriptional regulatory networks." *EURASIP Journal on Bioinformatics & Systems Biology*, 79879. doi:10.1155/2007/79879.

Mills, Evanna, and Luke A J O'Neill. 2014. "Succinate: a metabolic signal in inflammation." *Trends in Cell Biology* 24 (5): 313–20. doi:10.1016/j.tcb.2013.11.008.

Mitsche, Matthew A, Jeffrey G McDonald, Helen H Hobbs, and Jonathan C Cohen. 2015. "Flux analysis of cholesterol biosynthesis in vivo reveals multiple tissue and cell-type specific pathways." *eLife* 4 (June): e07999. doi:10.7554/eLife.07999.

Moffett, John R, and Ma Aryan Namboodiri. 2003. "Tryptophan and the immune response." *Immunology and Cell Biology* 81 (4): 247–65. doi:10.1046/j.1440-1711.2003.t01-1-01177.x.

Munn, D H, E Shafizadeh, J T Attwood, I Bondarev, A Pashine, and A L Mellor. 1999. "Inhibition of T cell proliferation by macrophage tryptophan catabolism." *The Journal of Experimental Medicine* 189 (9): 1363–72. http://www.ncbi.nlm.nih.gov/pubmed/10224276 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2193062.

Murphy, Taylor A, Chi V Dang, and Jamey D Young. 2013. "Isotopically nonstationary 13C flux analysis of Myc-induced metabolic reprogramming in B-cells." *Metabolic Engineering* 15 (January): 206–17. doi:10.1016/j.ymben.2012.07.008.

Nagy, Csörsz, and Arvand Haschemi. 2015. "Time and Demand are Two Critical Dimensions of Immunometabolism: The Process of Macrophage Activation and the Pentose Phosphate Pathway." *Frontiers in Immunology* 6: 164. doi:10.3389/fimmu.2015.00164.

Nathan, D M, D E Singer, K Hurxthal, and J D Goodson. 1984. "The clinical information value of the glycosylated hemoglobin assay." *N Engl J Med* 310 (6): 341–46. doi:10.1056/NEJM198402093100602.

NCD Risk Factor Collaboration (NCD-RisC). 2016. "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants." *Lancet* 387 (10027): 1513–30. doi:10.1016/S0140-6736(16)00618-8.

Netea, Mihai G, Leo A B Joosten, Yang Li, Vinod Kumar, Marije Oosting, Sanne

Smeekens, Martin Jaeger, et al. 2016. “Understanding human immune function using the resources from the Human Functional Genomics Project.” *Nat Med* 22 (8): 831–33. doi:10.1038/nm.4140.

Netherlands Consortium, Genome of the. 2014. “Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population.” *Nature Genetics* 46 (8): 818–25. doi:10.1038/ng.3021.

Newgard, Christopher B. 2017. “Metabolomics and Metabolic Diseases: Where Do We Stand?” *Cell Metabolism* 25 (1): 43–56. doi:10.1016/j.cmet.2016.09.018.

Newgard, Christopher B, Jie An, James R Bain, Michael J Muehlbauer, Robert D Stevens, Lillian F Lien, Andrea M Haqq, et al. 2009. “A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance.” *Cell Metab* 9 (4): 311–26. doi:10.1016/j.cmet.2009.02.002.

Newgard, Christopher B. 2012. “Interplay between lipids and branched-chain amino acids in development of insulin resistance.” *Cell Metabolism* 15 (5). Elsevier Inc.: 606–14. doi:10.1016/j.cmet.2012.01.024.

Ngkelo, Anta, Koremu Meja, Mike Yeadon, Ian Adcock, and Paul A Kirkham. 2012. “LPS induced inflammatory responses in human peripheral blood mononuclear cells is mediated through NOX4 and Gi$$ dependent PI-3kinase signalling.” *Journal of Inflammation (London, England)* 9 (1): 1. doi:10.1186/1476-9255-9-1.

Nicholson, Jeremy K, Elaine Holmes, James M Kinross, Ara W Darzi, Zoltan Takats, and John C Lindon. 2012. “Metabolic phenotyping in clinical and surgical environments.” *Nature* 491 (7424): 384–92. doi:10.1038/nature11708.

Nielsen, Jens. 2003. “It is all about metabolic fluxes.” *Journal of Bacteriology* 185 (24): 7031–5. http://www.ncbi.nlm.nih.gov/pubmed/14645261 http://www.pubmedcentral.nih.gov/articlere

Niklas, Jens, Volker Sandig, and Elmar Heinzle. 2011. “Metabolite channeling and compartmentation in the human cell line AGE1.HN determined by 13C labeling experiments and 13C metabolic flux analysis.” *Journal of Bioscience and Bioengineering* 112 (6): 616–23. doi:10.1016/j.jbiosc.2011.07.021.

Oosting, Marije, Kathrin Buffen, Jos W M van der Meer, Mihai G Netea, and Leo A B Joosten. 2016. “Innate immunity networks during infection with Borrelia burgdorferi.”

Critical Reviews in Microbiology 42 (2): 233–44. doi:10.3109/1040841X.2014.929563.

Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, et al. 2014. "The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases." *Nucleic Acids Research* 42 (Database issue): D358–63. doi:10.1093/nar/gkt1115.

O'Brien, Edward J., Jonathan M. Monk, and Bernhard O. Palsson. 2015. "Using Genome-scale Models to Predict Biological Capabilities." *Cell* 161 (5). Elsevier Inc.: 971–87. doi:10.1016/j.cell.2015.05.019.

O'Neill, Luke A J, Rigel J Kishton, and Jeff Rathmell. 2016. "A guide to immunometabolism for immunologists." *Nature Reviews. Immunology* 16 (9): 553–65. doi:10.1038/nri.2016.70.

Packard, C J, and J Shepherd. 1997. "Lipoprotein Heterogeneity and Apolipoprotein B Metabolism." *Arterioscler Thromb Vasc Biol* 17 (12): 3542–56.

Paththinige, C S, N D Sirisena, and Vhw Dissanayake. 2017. "Genetic Determinants of Inherited Susceptibility to Hypercholesterolemia - a Comprehensive Literature Review." *Lipids Health Dis* 16 (1): 103. doi:10.1186/s12944-017-0488-4.

Patil, Kiran Raosaheb, and Jens Nielsen. 2005. "Uncovering transcriptional regulation of metabolism by using metabolic network topology." *Proceedings of the National Academy of Sciences of the United States of America* 102 (8): 2685–9. doi:10.1073/pnas.0406811102.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential abundance analysis for microbial marker-gene surveys." *Nature Methods* 10 (12): 1200–1202. doi:10.1038/nmeth.2658.

Pearce, Erika L, and Edward J Pearce. 2013. "Metabolic pathways in immune cell activation and quiescence." *Immunity* 38 (4): 633–43. doi:10.1016/j.immuni.2013.04.005.

Pearce, Erika L, Maya C Poffenberger, Chih-Hao Chang, and Russell G Jones. 2013. "Fueling immunity: insights into metabolism and lymphocyte function." *Science (New York, N.Y.)* 342 (6155): 1242454. doi:10.1126/science.1242454.

Phillips, Michael C. 2014. "Molecular mechanisms of cellular cholesterol efflux." *J Biol Chem* 289 (35): 24020–9. doi:10.1074/jbc.R114.583658.

Phipson, Belinda, and Gordon K Smyth. 2010. "Permutation P-values should never

be zero: calculating exact P-values when permutations are randomly drawn." *Statistical Applications in Genetics and Molecular Biology* 9: Article39. doi:10.2202/1544-6115.1585.

Pillois, Xavier, Thomas Gautier, Benjamin Bouillet, Jean-Paul Pais de Barros, Aline Jeannin, Bruno Vergès, Jacques Bonnet, and Laurent Lagrost. 2012. "Constitutive Inhibition of Plasma Cetp by Apolipoprotein C1 Is Blunted in Dyslipidemic Patients with Coronary Artery Disease." *J Lipid Res* 53 (6): 1200–1209. doi:10.1194/jlr.M022988.

Pornputtapong, Natapol, Intawat Nookaew, and Jens Nielsen. 2015. "Human metabolic atlas: an online resource for human metabolism." *Database : The Journal of Biological Databases and Curation* 2015: bav068. doi:10.1093/database/bav068.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. "FastTree 2–approximately maximum-likelihood trees for large alignments." *PloS One* 5 (3): e9490. doi:10.1371/journal.pone.0009490.

Psychogios, Nikolaos, David D Hau, Jun Peng, An Chi Guo, Rupasri Mandal, Souhaila Bouatra, Igor Sinelnikov, et al. 2011. "The human serum metabolome." *PloS One* 6 (2): e16957. doi:10.1371/journal.pone.0016957.

Quinodoz, Sofia, and Mitchell Guttman. 2014. "Long noncoding RNAs: an emerging link between gene regulation and nuclear organization." *Trends in Cell Biology* 24 (11): 651–63. doi:10.1016/j.tcb.2014.08.009.

Rasool, Mahmood, Arif Malik, Muhammad Imran Naseer, Abdul Manan, Shakeel Ansari, Irshad Begum, Mahmood Husain Qazi, et al. 2015. "The role of epigenetics in personalized medicine: challenges and opportunities." *BMC Medical Genomics* 8 Suppl 1 (January): S5. doi:10.1186/1755-8794-8-S1-S5.

Rebholz, Casey M, Bing Yu, Zihe Zheng, Patrick Chang, Adrienne Tin, Anna Köttgen, Lynne E Wagenknecht, Josef Coresh, Eric Boerwinkle, and Elizabeth Selvin. 2018. "Serum metabolomic profile of incident diabetes." *Diabetologia* 61 (5): 1046–54. doi:10.1007/s00125-018-4573-7.

Reshetova, Polina, Age K Smilde, Antoine H C van Kampen, and Johan A Westerhuis. 2014. "Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data." *BMC Systems Biology* 8 Suppl 2: S2. doi:10.1186/1752-0509-8-S2-

S2.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research* 43 (7): e47. doi:10.1093/nar/gkv007.

Rivera, Chloe M, and Bing Ren. 2013. "Mapping human epigenomes." *Cell* 155 (1): 39–55. doi:10.1016/j.cell.2013.09.011.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative analysis of 111 reference human epigenomes." *Nature* 518 (7539): 317–30. doi:10.1038/nature14248.

Roberts, Lee D, Albert Koulman, and Julian L Griffin. 2014. "Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome." *Lancet Diabetes Endocrinol* 2 (1): 65–75. doi:10.1016/S2213-8587(13)70143-8.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: a versatile open source tool for metagenomics." *PeerJ* 4: e2584. doi:10.7717/peerj.2584.

Rolland, Thomas, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, et al. 2014. "A proteome-scale map of the human interactome network." *Cell* 159 (5): 1212–26. doi:10.1016/j.cell.2014.10.050.

Rossell, Sergio, Christian Solem, Peter R Jensen, and Joseph J Heijnen. 2011. "Towards a quantitative prediction of the fluxome from the proteome." *Metabolic Engineering* 13 (3): 253–62. doi:10.1016/j.ymben.2011.01.010.

Rual, Jean-François, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, et al. 2005. "Towards a proteome-scale map of the human protein-protein interaction network." *Nature* 437 (7062): 1173–8. doi:10.1038/nature04209.

Sadee, Wolfgang. 2011. "Genomics and personalized medicine." *International Journal of*

Pharmaceutics 415 (1-2): 2–4. doi:10.1016/j.ijpharm.2011.04.048.

Sahni, Nidhi, Song Yi, Mikko Taipale, Juan I Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, et al. 2015. "Widespread macromolecular interaction perturbations in human genetic disorders." *Cell* 161 (3): 647–60. doi:10.1016/j.cell.2015.04.013.

Sahni, Nidhi, Song Yi, Quan Zhong, Noor Jailkhani, Benoit Charloteaux, Michael E Cusick, and Marc Vidal. 2013. "Edgotype: a fundamental link between genotype and phenotype." *Current Opinion in Genetics & Development* 23 (6): 649–57. doi:10.1016/j.gde.2013.11.002.

Schaefer, Martin H, Tiago J S Lopes, Nancy Mah, Jason E Shoemaker, Yukiko Matsuoka, Jean-Fred Fontaine, Caroline Louis-Jeune, et al. 2013. "Adding protein context to the human protein-protein interaction network to reveal meaningful interactions." *PLoS Computational Biology* 9 (1): e1002860. doi:10.1371/journal.pcbi.1002860.

Schiffer, Lucas, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B Dowd, Nicola Segata, and Levi Waldron. 2018. "HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor." *bioRxiv*, January. http://biorxiv.org/content/early/2018/08/29/299115.abstract.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." *Applied and Environmental Microbiology* 75 (23): 7537–41. doi:10.1128/AEM.01541-09.

Scholtens, Salome, Nynke Smidt, Morris A Swertz, Stephan J L Bakker, Aafje Dotinga, Judith M Vonk, Freerk van Dijk, et al. 2015. "Cohort Profile: LifeLines, a three-generation cohort study and biobank." *International Journal of Epidemiology* 44 (4): 1172–80. doi:10.1093/ije/dyu229.

Sevimoglu, Tuba, and Kazim Yalcin Arga. 2014. "The role of protein interaction networks in systems biomedicine." *Computational and Structural Biotechnology Journal* 11 (18): 22–27. doi:10.1016/j.csbj.2014.08.008.

Shoaie, Saeed, Pouyan Ghaffari, Petia Kovatcheva-Datchary, Adil Mardinoglu, Partho

Sen, Estelle Pujos-Guillot, Tomas de Wouters, et al. 2015. "Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome." *Cell Metabolism* 22 (2): 320–31. doi:10.1016/j.cmet.2015.07.001.

Sliz, Eeva, Johannes Kettunen, Michael V Holmes, Clare Oliver-Williams, Charles Boachie, Qin Wang, Minna Mannikko, et al. 2018. "Metabolomic Consequences of Genetic Inhibition of Pcsk9 Compared with Statin Treatment." *bioRxiv*. Cold Spring Harbor Laboratory. doi:10.1101/278861.

Smeekens, Sanne P, Aylwin Ng, Vinod Kumar, Melissa D Johnson, Theo S Plantinga, Cleo van Diemen, Peer Arts, et al. 2013. "Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans." *Nature Communications* 4: 1342. doi:10.1038/ncomms2343.

Smyth, Gordon K. 2004. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1): Article3. doi:10.2202/1544-6115.1027.

Snel, B, G Lehmann, P Bork, and M A Huynen. 2000. "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." *Nucleic Acids Research* 28 (18): 3442–4. http://www.ncbi.nlm.nih.gov/pubmed/10982861 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC110752.

Snijder, Marieke B, Henrike Galenkamp, Maria Prins, Eske M Derks, Ron J G Peters, Aeilko H Zwinderman, and Karien Stronks. 2017. "Cohort profile: the Healthy Life in an Urban Setting (HELIUS) study in Amsterdam, The Netherlands." *BMJ Open* 7 (12): e017873. doi:10.1136/bmjopen-2017-017873.

Soininen, Pasi, Antti J Kangas, Peter Würtz, Teemu Suna, and Mika Ala-Korpela. 2015. "Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics." *Circ Cardiovasc Genet* 8 (1): 192–206. doi:10.1161/CIRCGENETICS.114.000216.

Soutar, Anne K, and Rossi P Naoumova. 2007. "Mechanisms of Disease: Genetic Causes of Familial Hypercholesterolemia." *Nature Clinical Practice. Cardiovascular Medicine* 4 (4): 214–25. doi:10.1038/ncpcardio0836.

Spann, Nathanael J, Lana X Garmire, Jeffrey G McDonald, David S Myers, Stephen

B Milne, Norihito Shibata, Donna Reichart, et al. 2012. "Regulated accumulation of desmosterol integrates macrophage lipid metabolism and inflammatory responses." *Cell* 151 (1): 138–52. doi:10.1016/j.cell.2012.06.054.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. "BioGRID: a general repository for interaction datasets." *Nucleic Acids Research* 34 (Database issue): D535–9. doi:10.1093/nar/gkj109.

Stelzl, Ulrich, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, et al. 2005. "A human protein-protein interaction network: a resource for annotating the proteome." *Cell* 122 (6): 957–68. doi:10.1016/j.cell.2005.08.029.

Stobbe, Miranda D., Sander M. Houten, Gerbert A. Jansen, Antoine H C van Kampen, and Perry D. Moerland. 2011. "Critical assessment of human metabolic pathway databases: a stepping stone for future integration." *BMC Systems Biology* 5 (October): 165. doi:10.1186/1752-0509-5-165.

Stronks, Karien, Marieke B Snijder, Ron J G Peters, Maria Prins, Aart H Schene, and Aeilko H Zwinderman. 2013a. "Unravelling the impact of ethnicity on health in Europe: the HELIUS study." *BMC Public Health* 13 (April): 402. doi:10.1186/1471-2458-13-402.

———. 2013b. "Unravelling the impact of ethnicity on health in Europe: the HELIUS study." *BMC Public Health* 13 (April): 402. doi:10.1186/1471-2458-13-402.

Stuart, Joshua M, Eran Segal, Daphne Koller, and Stuart K Kim. 2003. "A gene-coexpression network for global discovery of conserved genetic modules." *Science (New York, N.Y.)* 302 (5643): 249–55. doi:10.1126/science.1087447.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.

Suhre, Karsten, So-Youn Shin, Ann-Kristin Petersen, Robert P Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, et al. 2011. "Human metabolic individuality in biomedical and pharmaceutical research." *Nature* 477 (7362): 54–60.

doi:10.1038/nature10354.

Székely, Tamás, and Kevin Burrage. 2014. "Stochastic simulation in systems biology." *Computational and Structural Biotechnology Journal* 12 (20-21). Elsevier B.V.: 14–25. doi:10.1016/j.csbj.2014.10.003.

Talmud, Philippa J, Fotios Drenos, Sonia Shah, Tina Shah, Jutta Palmen, Claudio Verzilli, Tom R Gaunt, et al. 2009. "Gene-Centric Association Signals for Lipids and Apolipoproteins Identified via the Humancvd Beadchip." *Am J Hum Genet* 85 (5): 628–42. doi:10.1016/j.ajhg.2009.10.014.

Talmud, Philippa J, Sonia Shah, Ros Whittall, Marta Futema, Philip Howard, Jackie A Cooper, Seamus C Harrison, et al. 2013. "Use of Low-Density Lipoprotein Cholesterol Gene Score to Distinguish Patients with Polygenic and Monogenic Familial Hypercholesterolaemia: A Case-Control Study." *Lancet (London, England)* 381 (9874): 1293–1301. doi:10.1016/S0140-6736(12)62127-8.

Tannahill, G M, A M Curtis, J Adamik, E M Palsson-McDermott, A F McGettrick, G Goel, C Frezza, et al. 2013. "Succinate is an inflammatory signal that induces IL-1β through HIF-1α." *Nature* 496 (7444): 238–42. doi:10.1038/nature11986.

Taylor, A, D Wang, K Patel, R Whittall, G Wood, M Farrer, R D G Neely, et al. 2010. "Mutation Detection Rate and Spectrum in Familial Hypercholesterolaemia Patients in the Uk Pilot Cascade Project." *Clinical Genetics* 77 (6): 572–80. doi:10.1111/j.1399-0004.2009.01356.x.

Teslovich, Tanya M, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, et al. 2010. "Biological, Clinical and Population Relevance of 95 Loci for Blood Lipids." *Nature* 466 (7307): 707–13. doi:10.1038/nature09270.

Thiele, Ines, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, et al. 2013. "A community-driven global reconstruction of human metabolism." *Nature Biotechnology* 31 (5): 419–25. doi:10.1038/nbt.2488.

Thomson, Scott J, Ara Askari, and David Bishop-Bailey. 2012. "Anti-inflammatory effects of epoxyeicosatrienoic acids." *International Journal of Vascular Medicine* 2012:

605101. doi:10.1155/2012/605101.

Thorsen, Jonathan, Asker Brejnrod, Martin Mortensen, Morten A Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. 2016. "Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies." *Microbiome* 4 (1): 62. doi:10.1186/s40168-016-0208-8.

Tillin, Therese, Alun D Hughes, Qin Wang, Peter Würtz, Mika Ala-Korpela, Naveed Sattar, Nita G Forouhi, et al. 2015. "Diabetes risk and amino acid profiles: cross-sectional and prospective analyses of ethnicity, amino acids and diabetes in a South Asian and European cohort from the SABRE (Southall And Brent REvisited) Study." *Diabetologia* 58 (5): 968–79. doi:10.1007/s00125-015-3517-8.

Topol, Eric J. 2014. "Individualized medicine from prewomb to tomb." *Cell* 157 (1): 241–53. doi:10.1016/j.cell.2014.02.012.

Turnbaugh, Peter J, Vanessa K Ridaura, Jeremiah J Faith, Federico E Rey, Rob Knight, and Jeffrey I Gordon. 2009. "The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice." *Science Translational Medicine* 1 (6): 6ra14. doi:10.1126/scitranslmed.3000322.

Uehara, Yoshinari, Thomas Engel, Zhengchen Li, Christian Goepfert, Stephan Rust, Xiaoqin Zhou, Claus Langer, et al. 2002. "Polyunsaturated fatty acids and acetoacetate downregulate the expression of the ATP-binding cassette transporter A1." *Diabetes* 51 (10): 2922–8. http://www.ncbi.nlm.nih.gov/pubmed/12351428.

Uhlén, Mathias, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-based map of the human proteome." *Science (New York, N.Y.)* 347 (6220): 1260419. doi:10.1126/science.1260419.

Ujcic-Voortman, Joanne K, Miranda T Schram, Monique A der Bruggen, Arnoud P Verhoeff, and Caroline A Baan. 2009. "Diabetes prevalence and risk factors among ethnic minorities." *Eur J Public Health* 19 (5): 511–15. doi:10.1093/eurpub/ckp096.

Valkengoed, Irene G M van, Carmen Argmann, Karen Ghauharali-van der Vlugt, Johannes M F G Aerts, Lizzy M Brewster, R J G Peters, Frédéric M Vaz, and Riekelt H Houtkooper. 2017. "Ethnic differences in metabolite signatures and type 2 diabetes: a

nested case-control analysis among people of South Asian, African and European origin." *Nutrition & Diabetes* 7 (12): 300. doi:10.1038/s41387-017-0003-z.

Vanlier, J, C A Tiemann, P A J Hilbers, and N A W van Riel. 2012. "An integrated strategy for prediction uncertainty analysis." *Bioinformatics (Oxford, England)* 28 (8): 1130–5. doi:10.1093/bioinformatics/bts088.

Väremo, Leif, Jens Nielsen, and Intawat Nookaew. 2013a. "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods." *Nucleic Acids Research* 41 (8): 4378–91. doi:10.1093/nar/gkt111.

———. 2013b. "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods." *Nucleic Acids Research* 41 (8): 4378–91. doi:10.1093/nar/gkt111.

Väremo, Leif, Camilla Scheele, Christa Broholm, Adil Mardinoglu, Caroline Kampf, Anna Asplund, Intawat Nookaew, Mathias Uhlén, Bente Klarlund Pedersen, and Jens Nielsen. 2015. "Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes." *Cell Reports* 11 (6): 921–33. doi:10.1016/j.celrep.2015.04.010.

Vergeer, Menno, Liam R Brunham, Joris Koetsveld, Janine K Kruit, C Bruce Verchere, John J P Kastelein, Michael R Hayden, and Erik S G Stroes. 2010. "Carriers of loss-of-function mutations in ABCA1 display pancreatic beta-cell dysfunction." *Diabetes Care* 33 (4): 869–74. doi:10.2337/dc09-1562.

Vermeulen, E, K Stronks, M Visser, I A Brouwer, M B Snijder, R J T Mocking, E M Derks, A H Schene, and M Nicolaou. 2017. "Dietary pattern derived by reduced rank regression and depressive symptoms in a multi-ethnic population: the HELIUS study." *European Journal of Clinical Nutrition* 71 (8): 987–94. doi:10.1038/ejcn.2017.61.

Vidal, Marc, Michael E Cusick, and Albert-László Barabási. 2011. "Interactome networks and human disease." *Cell* 144 (6): 986–98. doi:10.1016/j.cell.2011.02.016.

Wang, J, A Stančáková, P Soininen, A J Kangas, J Paananen, J Kuusisto, M Ala-Korpela, and M Laakso. 2012. "Lipoprotein subclass profiles in individuals with varying degrees of glucose tolerance: a population-based study of 9399 Finnish men." *J Intern Med* 272

(6): 562–72. doi:10.1111/j.1365-2796.2012.02562.x.

Wang, Jian, Jacqueline S Dron, Matthew R Ban, John F Robinson, Adam D McIntyre, Maher Alazzam, Pei Jun Zhao, et al. 2016. "Polygenic Versus Monogenic Causes of Hypercholesterolemia Ascertained Clinically." *Arterioscler Thromb Vasc Biol* 36 (12): 2439–45. doi:10.1161/ATVBAHA.116.308027.

Wang, Thomas J, Martin G Larson, Ramachandran S Vasan, Susan Cheng, Eugene P Rhee, Elizabeth McCabe, Gregory D Lewis, et al. 2011. "Metabolite profiles and the risk of developing diabetes." *Nature Medicine* 17 (4): 448–53. doi:10.1038/nm.2307.

Wang, Thomas J, Debby Ngo, Nikolaos Psychogios, Andre Dejam, Martin G Larson, Ramachandran S Vasan, Anahita Ghorbani, et al. 2013. "2-Aminoadipic acid is a biomarker for diabetes risk." *J Clin Invest* 123 (10): 4309–17. doi:10.1172/JCI64801.

Waters, Kevin M, Daniel O Stram, Mohamed T Hassanein, Lo\"\ic Le Marchand, Lynne R Wilkens, Gertraud Maskarinec, Kristine R Monroe, et al. 2010. "Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups." *PLoS Genet* 6 (8). doi:10.1371/journal.pgen.1001078.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic Acids Research* 42 (Database issue): D1001–6. doi:10.1093/nar/gkt1229.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Winter, Gal, and Jens O Krömer. 2013. "Fluxomics - connecting 'omics analysis and phenotypes." *Environmental Microbiology* 15 (7): 1901–16. doi:10.1111/1462-2920.12064.

Wu, Gary D, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, et al. 2011. "Linking long-term dietary patterns with gut microbial enterotypes." *Science (New York, N.Y.)* 334 (6052): 105–8. doi:10.1126/science.1208344.

Würtz, Peter, Aki S Havulinna, Pasi Soininen, Tuulia Tynkkynen, David Prieto-Merino, Therese Tillin, Anahita Ghorbani, et al. 2015. "Metabolite Profiling and Cardiovascular Event Risk: A Prospective Study of 3 Population-Based Cohorts." *Circulation* 131 (9):

774–85. doi:10.1161/CIRCULATIONAHA.114.013116.

Würtz, Peter, Antti J. Kangas, Pasi Soininen, Debbie A. Lawlor, George Davey Smith, and Mika Ala-Korpela. 2017. "Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies." *American Journal of Epidemiology* 186 (9): 1084–96. doi:10.1093/aje/kwx016.

Würtz, Peter, Ville-Petteri Mäkinen, Pasi Soininen, Antti J Kangas, Taru Tukiainen, Johannes Kettunen, Markku J Savolainen, et al. 2012. "Metabolic signatures of insulin resistance in 7,098 young adults." *Diabetes* 61 (6): 1372–80. doi:10.2337/db11-1355.

Würtz, Peter, Pasi Soininen, Antti J Kangas, Tapani Rönnemaa, Terho Lehtimäki, Mika Kähönen, Jorma S Viikari, Olli T Raitakari, and Mika Ala-Korpela. 2013. "Branched-chain and aromatic amino acids are predictors of insulin resistance in young adults." *Diabetes Care* 36 (3): 648–55. doi:10.2337/dc12-0895.

Würtz, Peter, Mika Tiainen, Ville-Petteri Mäkinen, Antti J Kangas, Pasi Soininen, Juha Saltevo, Sirkka Keinänen-Kiukaanniemi, et al. 2012. "Circulating metabolite predictors of glycemia in middle-aged men and women." *Diabetes Care* 35 (8): 1749–56. doi:10.2337/dc11-1838.

Würtz, Peter, Qin Wang, Pasi Soininen, Antti J Kangas, Ghazaleh Fatemifar, Tuulia Tynkkynen, Mika Tiainen, et al. 2016. "Metabolomic Profiling of Statin Use and Genetic Inhibition of Hmg-Coa Reductase." *J Am Coll Cardiol* 67 (10): 1200–1210. doi:10.1016/j.jacc.2015.12.060.

Xia, Yinglin, and Jun Sun. 2017. "Hypothesis Testing and Statistical Analysis of Microbiome." *Genes & Diseases* 4 (3): 138–48. doi:10.1016/j.gendis.2017.06.001.

Yizhak, Keren, Tomer Benyamini, Wolfram Liebermeister, Eytan Ruppin, and Tomer Shlomi. 2010a. "Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model." *Bioinformatics* 26 (12): i255–60. doi:10.1093/bioinformatics/btq183.

———. 2010b. "Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model." *Bioinformatics (Oxford, England)* 26 (12): i255–60. doi:10.1093/bioinformatics/btq183.

Yizhak, Keren, Edoardo Gaude, Sylvia Le Dévédec, Yedael Y Waldman, Gideon Y Stein,

Bob van de Water, Christian Frezza, and Eytan Ruppin. 2014. “Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer.” *eLife* 3 (November). doi:10.7554/eLife.03641.

York, Autumn G, Kevin J Williams, Joseph P Argus, Quan D Zhou, Gurpreet Brar, Laurent Vergnes, Elizabeth E Gray, et al. 2015. “Limiting Cholesterol Biosynthetic Flux Spontaneously Engages Type I IFN Signaling.” *Cell* 163 (7): 1716–29. doi:10.1016/j.cell.2015.11.045.

Zhang, Xinyan, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi. 2017. “Negative binomial mixed models for analyzing microbiome count data.” *BMC Bioinformatics* 18 (1): 4. doi:10.1186/s12859-016-1441-7.

Zhong, Quan, Nicolas Simonis, Qian-Ru Li, Benoit Charloteaux, Fabien Heuze, Niels Klitgord, Stanley Tam, et al. 2009. “Edgetic perturbation models of human inherited disorders.” *Molecular Systems Biology* 5: 321. doi:10.1038/msb.2009.80.

Zilliox, Michael J, and Rafael A Irizarry. 2007. “A gene expression bar code for microarray data.” *Nature Methods* 4 (11): 911–3. doi:10.1038/nmeth1102.

Zoppoli, Pietro, Sandro Morganella, and Michele Ceccarelli. 2010. “TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach.” *BMC Bioinformatics* 11 (March): 154. doi:10.1186/1471-2105-11-154.