Dept. of Biomedical Engineering and Computational Science

Metabolomics meets genetics

– from an NMR metabolomics platform to the genetic architecture of serum metabolites

Taru Tukiainen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 16/2012

Metabolomics meets genetics – from an NMR metabolomics platform to the genetic architecture of serum metabolites

Taru Tukiainen

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium F239 at the Aalto University School of Science (Espoo, Finland) on the 2nd of March 2012 at 12 noon.

Aalto University
School of Science
Dept. of Biomedical Engineering and Computational Science

Supervisor
Prof. Kimmo Kaski

Instructors
Prof. Mika Ala-Korpela, University of Oulu, Finland
Adj. Prof. Samuli Ripatti, Institute for Molecular Medicine Finland, Finland

Preliminary examiners
Adj. Prof. Vesa Olkkonen, Minerva Foundation Institute for Medical Research, Finland
Prof. Erik Ingelsson, Karolinska Institutet, Sweden

Opponent
Prof. Thomas Illig, Hannover Medical School, Germany

Aalto University publication series
DOCTORAL DISSERTATIONS 16/2012

ISBN 978-952-60-4509-2 (printed)
ISBN 978-952-60-4510-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2012

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

441 697
Printed matter

Author
Taru Tukiainen

Name of the doctoral dissertation
Metabolomics meets genetics – from an NMR metabolomics platform to the genetic architecture of serum metabolites

Publisher School of Science

Unit Department of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 16/2012

Field of research Computational Systems Biology

Manuscript submitted 13 December 2011 **Manuscript revised** 23 January 2012

Date of the defence 2 March 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Metabolomics is a recently emerged field of science studying metabolites and how their levels change with biological perturbations. A key requirement for metabolomics analyses is a technology that can capture a multitude of metabolite information in a single measurement. As many of the available platforms have lacked automation in the metabolomics experimentation, including the data analysis and handling, the measurements have been costly and time-consuming, and thus metabolomics data had not been widely applied in large-scale studies. Metabolomics profiling, however, has great potential to provide further biological knowledge by, for example, elucidating in detail the mechanisms and pathways underlying disease.

The first two publications of this thesis present a high-throughput proton nuclear magnetic resonance (NMR) –based serum metabolomics platform designed to facilitate the use of metabolomics data in large biomedical studies. The platform allows the highly-automated metabolomics profiling of tens of thousands of samples per year in a cost-effective manner and with the implemented models more than a hundred metabolites, including lipoprotein subclasses, other lipids and small molecules, can be quantified from the serum NMR data. The metabolomics profiling provided by the NMR-based platform has gained wide interest; the platform has run non-stop since it was set up in late 2008 as many Finnish and international cohorts have had their samples measured and used the data in several publications.

In the two other publications included in this thesis, the quantitative metabolite data obtained through the platform was combined with detailed data on genetic variants in more than 8000 Finnish individuals. This unique data set was used a) to comprehensively characterize, in terms of metabolite and genetic associations, the genomic regions known to associate with blood lipid levels, and b) to dissect genetic components associated with the changes in the metabolite levels. A wealth of biological information was uncovered in these studies including new metabolic associations for the known genetic regions and several new genetic regions associated with the metabolites. These findings can help to understand the links between the genes and clinical conditions.

Together the results of this thesis show how detailed metabolomics data greatly complements the conventional laboratory measurements and support the use of this data in biomedical studies as means to provide valuable biological knowledge.

Keywords metabolomics, NMR, genetics, SNP, lipoprotein subclasses

ISBN (printed) 978-952-60-4509-2 **ISBN (pdf)** 978-952-60-4510-8

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 146 **The dissertation can be read at** http://lib.tkk.fi/Diss/

Tekijä
Taru Tukiainen

Väitöskirjan nimi
Metabolomiikka kohtaa genetiikan – NMR-pohjaisesta metabolomiikkaprotokollasta seerumin metaboliittitasojen geneettiseen taustaan

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 16/2012

Tutkimusala Laskennallinen systeemibiologia

Käsikirjoituksen pvm 13.12.2011 **Korjatun käsikirjoituksen pvm** 23.01.2012

Väitöspäivä 02.03.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Metabolomiikka on tieteenala, joka tutkii metaboliitteja ja kuinka niiden tasot muuttuvat biologisten häiriötekijöiden vaikutuksesta. Metabolomiikassa pyritään siihen, että kaikki, tai ainakin mahdollisimman iso osa, näytteen metaboliiteista saadaan määritettyä samanaikaisesti. Täten tärkeä osa metabolomiikkatutkimusta on menetelmä, joka pystyy mittaamaan suuren määrän metaboliitti-informaatiota kerralla. Iso ongelma metabolomiikka-tutkimuksissa näihin on kuitenkin ollut, että käytössä olevien laitteiden mittauskapasitetti ja automaation mahdollisuudet, myös datan käsittelyssä ja tulkinnassa, ovat olleet rajallisia. Täten mittaukset ovat olleet usein kalliita ja aikaavieviä, ja siksi harvat isot tutkimukset ovat käyttäneet metabolomiikkatietoa. Tutkimukset ovat kuitenkin osoittaneet, että metaboliitti-informaation käyttäminen auttaa muun muassa havainnollistamaan tatutimekanismeja.

Osa tästä väitöskirjasta käsittelee NMR-spektroskopiaan pohjautuvaa metabolomiikkaprotokollaa, joka kehiteltiin edesauttamaan metabolomiikkatietoa käyttöä epidemiologisissa ja kliinisissä sovelluksissa, jotka vaativat suurehkoja näytemääriä. Protokolla on pitkälti automatisoitu, ja siten sitä käyttäen isojen näytejoukkojen mittaaminen sujuu nopeasti ja kustannustehokkaasti. Lisäksi, protokollan osana on kehitetty kvantitointimalleja, joilla seerumin NMR-spektreistä saadaan tarkkaa tietoa yli sadasta metaboliitista. Protokolla pystytettiin vuoden 2008 lopussa, ja sen jälkeen sillä on mitattu kymmeniätuhansia näytteitä ja sillä kerättyä tietoa on käyetty useissa julkaisuissa.

Tähän väitöskirjaan sisältyy myös kaksi muuta julkaisua, joissa protokollalla mitattua metabolomiikkatietoa yhdistettiin geneettiseen informaatioon. Näihin tutkimuksiin oli käyettävissä ainutlaatuinen aineisto: yli kahdeksan tuhatta suomalaista viidestä eri tutkimusaineistosta, joista kustakin oli määritetty sekä yli sata seerumin metaboliittia että koko genomin kattavasti yhden nukleotidin muutoksia. Tätä tietoa käyttäen selvitettiin, mitkä genomin muutokset assosioituvat metaboliittitasojen muutoksiin ja karakterisoitiin tarkemmin genomin alueita, joiden on aiemmin havaittu vaikuttavan veren rasvapitoisuuksiin. Näissä tutkimuksista saatiin paljon uutta metabolista ja geneettistä tietoa, joka voi edesauttaa tarkentamaan niitä mekanismeja ja metabolisia reittejä, jotka yhdistävät geenejä ja kliinisiä tiloja.

Yhdessä tämän väitöskirjan löydökset näyttävät, kuinka käyttämällä metabolomiikkapohjaista tietoa saadaan kerättyä uutta biologista informaatiota, myös verrattuna perinteisesti käytettyihin laboratoriomittauksiin, ja täten havainnot tukevat tarkan metaboliitti-informaation käyttöä biolääketieteellisissä tutkimuksissa.

Avainsanat metabolomiikka, NMR, genetiikka, SNP, lipoproteiinialaluokat

ISBN (painettu) 978-952-60-4509-2 **ISBN (pdf)** 978-952-60-4510-8

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 146 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

Preface

Little did I anticipate what an eventful period lay ahead when I started to pursue a doctor's degree. The occasional detours and sometimes surprising turns during the past four years shaped this thesis so that it resembles very little the first research plans. However, looking back now, I am happy to see it all turned out much better than I ever dared to dream of. As amazing as it is to see the nearly final thesis in my hands as a record of the research conducted, I feel some of the greatest outcomes are less tangible: I have had the pleasure to meet numerous exceptional people and most importantly learn much about science, myself and life in science.

The work for this thesis was mostly conducted in the Department of Biomedical Engineering and Computational Science (BECS) at the Aalto University School of Science (former Helsinki University of Technology, HUT) and at the Institute for Molecular Medicine Finland (FIMM). I wish to acknowledge Professor Jouko Lampinen, the head of BECS, and Professor Olli Kallioniemi, the director of FIMM, for providing the excellent research facilities.

My work was financially supported by Academy of Finland Centre of Excellence program, the BioSHaRE project, Instrumentarium Science Foundation, Finnish Cultural Foundation, Aalto University School of Science and Technology Research Training Scholarship, Finnish Concordia Fund, Emil Aaltonen Foundation and Finnish Foundation for Technology Promotion, all of which are gratefully acknowledged.

Professor Thomas Illig is warmly thanked for accepting the invitation and finding the time to act as the Opponent of my defense. Professor Erik Ingelsson and Adjunct Professor Vesa Olkkonen are acknowledged for their valuable comments and words of encouragement as the preliminary examiners of my thesis. I also wish to thank my supervising professor Kimmo Kaski for, throughout these years, believing in me and my skills and giving me the academic freedom to follow my own path.

My deepest gratitude goes to my instructors Professor Mika Ala-Korpela from University of Oulu and Adjunct Professor Samuli Ripatti from FIMM. Through Mika and Samuli's collaboration I was given the possibility to engage myself with the multidisciplinary research topic. I am thankful for them for trusting me with such a fascinating project. Mika, you are an exceptional scientist and your

catching enthusiasm helped to overcome some of the moments of despair encountered on the way. I am also grateful for you for opening many exciting avenues for research. Samuli, your drive, expertise, support and rationality have been invaluable to the project and my thesis. Thank you so much for accepting me as a member of your group.

I warmly acknowledge Professors Aarno Palotie and Markus Perola and Adjunct Professor Matti Jauhianen for their great ideas and encouragement along the way. Aarno's many years' experience in human genetics has been of enormous value to the project. Thanks to Markus, Tartu will always have a place in my heart. Matti's limitless knowledge on lipids never ceases to amaze me.

Dr Johannes Kettunen has been one of the closest collaborators and a kind of a mentor to me during the past year and a half. It has been a great pleasure – and fun – to work with someone from whom I have learned numerous things about various aspects of science. Antti Kangas is another close colleague and has contributed to the work in many ways. Thank you, Antti, for your huge input on the computational and visual side of the work but also for the valuable peer support and friendship. A big thank you also goes to Drs Ville-Petteri Mäkinen and Peter Würtz for all the help with the many small and large things. Dr Pasi Soininen is warmly acknowledged for all the NMR expertise and always having the patience to answer my more or less relevant NMR-related questions.

During the course of the thesis I spent an adventurous year in London working in Professor Marjo-Riitta Järvelin's group in Imperial College at the Department of Epidemiology and Biostatistics. I wish to thank Marjo-Riitta, whose ever-excited nature I continue to admire, for the educational experience, friendliness and the opportunity to work in Imperial. Dr Paul O'Reilly's help in guiding my first steps into genetic analyses is also warmly acknowledged.

Much of the work conducted for this thesis would not have been possible without the collaborations with the excellent study cohorts. I am grateful to all the personnel of the Northern Finland Birth Cohort 1966, The Cardiovascular Risk in Young Finns Study, Helsinki Birth Cohort Study, The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study, The Finnish twin registry, The Health 2000 cohort and the Kuopio cognitive decline study sample. Sincere thank you also to all the volunteers who have taken the time to participate in the cohorts' collections. I also wish to express my gratitude to all other co-authors and collaborators for their input to the work.

The help of all the secretaries, IT support and other personnel is greatly acknowledged. Thank you Laura Pyysalo, Sari Kivikko, Huei-yi Shen, FIMM IT guys and many others at BECS, Imperial and FIMM.

Spending time at different labs during the course of this thesis has given me the joy to work with and meet loads of wonderful people. Warm thanks for the fun times organizing and attending the NMR symposia and for the Friday afternoon group meetings spiced up with coffee, pulla and anecdotes to the former and present members of our Computational Medicine research group: Aino, Antti, Jaakko, Janne, Johanna, Lauri, Linda, Niina, Niko, Pasi S., Pasi J., Peter, Tomi

and Ville. Beatriz, Christine, Erika, Katerina, Paul, Penny and everyone else from Imperial are thanked for making my stay in London so enjoyable. A big thanks to Alfredo, Ansku, Antti, Diana, Elisabeth, Emmi, Hanski, Heidi, Helena, Himanshu, Ida, Jaakko, Jarkko, Johannes, Karola, Kati, Liisa, Mari, Marjis, Marine, Mikko, Minna, Minttu, Olli, Outi, Pekka, Peter, Perttu, Pietari, PP, Päivi, Saana, Tarja, Tea, Tiia, Teppo, Tero, Verneri, Virpi, Will, and the many others from FIMM and THL whom I may have forgotten to mention, for the warm and creative work environment and the fun times in the office, on coffee (and other beverage) breaks, at meetings and at all the social outings.

My heartfelt thanks go to my dear old friends Anja, Hanna, Jenni, Jonna, Maiju and Marjo. Thanks for your lively company and for understanding that I have sometimes prioritized my thesis over spending time with you. Big thanks also to all the friends met during the nearly ten years spent at HUT and Aalto for all the warm memories from Otaniemi.

I am ever grateful to my dear family, parents Aila and Ensio and sister Tiina. Thank you for all the love and care and for teaching me never to give up too easily. Without you three as my role models I would not have achieved any this. And Pauli, your unconditional love and support mean the world to me. You light up my life.

Helsinki, January 31, 2012
Taru Tukiainen

Contents

List of original publications

This thesis consists of an overview and the following Publications, which are referred to in the text by their Roman numerals.

I. Tukiainen T, Tynkkynen T, Mäkinen VP, Jylänki P, Kangas A, Hokkanen J, Vehtari A, Gröhn O, Hallikainen M, Soininen H, Kivipelto M, Groop PH, Kaski K, Laatikainen R, Soininen P, Pirttilä T, Ala-Korpela M. A multi-metabolite analysis of serum by ^{1}H NMR spectroscopy: Early systemic signs of Alzheimer's disease. Biochem Biophys Res Commun. 2008;375(3):356-61.

II. Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, Järvelin MR, Kähönen M, Lehtimäki T, Viikari J, Raitakari OT, Savolainen MJ, Ala-Korpela M. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. Analyst. 2009;134(9):1781-5.

III. Tukiainen T*, Kettunen J*, Kangas AJ, Lyytikäinen LP, Soininen P, Sarin AP, Tikkanen E, O'Reilly PF, Savolainen MJ, Kaski K, Pouta A, Jula A, Lehtimäki T, Kähönen M, Viikari J, Taskinen MR, Jauhiainen M, Eriksson JG, Raitakari O, Salomaa V, Järvelin MR, Perola M, Palotie A, Ala-Korpela M, Ripatti S. Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. Hum Mol Genet. *In press*.

IV. Kettunen J*, Tukiainen T*, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikäinen LP, Kangas AJ, Soininen P, Würtz P, Silander K, Dick DM, Rose RJ, Savolainen MJ, Viikari J, Kähönen M, Lehtimäki T, Pietiläinen KH, Inouye M, McCarthy MI, Jula A, Eriksson J, Raitakari OT, Salomaa V, Kaprio J, Järvelin MR, Peltonen L, Perola M, Freimer NB, Ala-Korpela M, Palotie A, Ripatti S. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet. *In press*.

* These authors contributed equally to the work

Author's contribution

For Publication I the author performed the data analysis and wrote the manuscript with Prof Mika Ala-Korpela. In Publication II the author's contribution was in analyzing the NMR data. The author also reviewed and commented the manuscript. The author performed the data analyses for Publications III and IV with Dr Johannes Kettunen, wrote the manuscript for Publication III and reviewed and edited the manuscript of Publication IV.

1. Introduction

The concept of using the characteristics of biological fluids, such as blood or urine, as markers of disease dates back thousands of years yet similar approach still largely serves as the basis for the modern day risk evaluation and diagnostics of metabolic conditions; While in ancient China ants were attracted to urine samples with high amount of glucose, nowadays the presence of diabetes is confirmed by assessing the blood glucose levels using targeted assays. The justification for charting circulating or urine metabolites, molecules that are intermediates or products of metabolism, is that the metabolite levels reflect the whole of the various biological processes of the human body. Perturbations of the metabolic homeostasis, induced for example by disease, can be observed as changes in the metabolite levels.

Aside from the conventional approach of measuring one or a couple of interesting metabolites as indicators of disease or increased risk for disease, a systems-level approach of simultaneously detecting a wider range of metabolites and using this comprehensive molecular level data to provide further insight into the perturbation is increasingly gaining ground. This field of study focusing on the detection and analysis of ideally all the metabolites from a sample is called metabolomics[1]. The emergence of metabolomics during the last decades has been largely due to the adaptation of a number of techniques from analytical chemistry, such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) that allow the simultaneous detection of a multitude of molecules, to biomedical purposes for profiling metabolites in biofluids.

The potential uses of this multivariate metabolomics data are various; the metabolite profiles are used to classify samples, elucidate the metabolic pathways affected or identify novel biomarkers in relation to disease, genetic differences or environmental factors. Although many have provided promising results, a major shortcoming of most metabolomics studies addressing biomedical problems to date has been the limited sample size; as typical numbers of individuals in the studies has been less than a hundred, the studies have lacked statistical power and been prone to false findings.

[1] There are two names for this area of science, metabolomics and metabonomics, both having similar definitions. The term metabolomics is used throughout this thesis in order to be consistent and as this is the more commonly used term in the field.

Major grounds preventing the wide-scale facilitation of metabolomics data has been the lack of high-throughput platforms; with limited capacity and little automation metabolomics measurements have often been costly and time-consuming. Moreover, normalization, handling and interpretation of the complex metabolomics data involving hundreds of peaks is rarely straightforward, and thus difficult to automate.

This thesis presents a high-throughput proton (^{1}H) NMR –based serum metabolomics platform designed to facilitate the use of metabolomics data in clinical and epidemiological settings (Publications I and II). The platform takes advantage of the latest laboratory and spectroscopic equipment to enable automated sample preparation and measurement, thus providing means to process tens of thousands of samples per year (Publication II). The experimentation was optimized to detect the majority of the metabolite information available from a serum sample by NMR (Publication I). Additionally, the platform includes efficient data handling modules, and models for automated quantification of more than a hundred metabolite measures have been implemented.

Subsequently, moving from the technology into applying it, two studies that combined metabolomics measurements with genetic information were conducted (Publications III and IV). With a unique set-up of over 8000 individuals from five population-based Finnish cohorts with both dense genotype and serum metabolomics data genetic variants contributing to the population-level variation in the traits were uncovered (Publication IV). Additionally, by studying a set of mono- and dizygotic Finnish twins the heritabilities of the metabolomics traits were determined (Publication IV). In Publication III the detailed metabolomics and genetic data were used to further characterize the genetic regions previously reported to associate with blood lipids.

The dual structure of this thesis, i.e., methods development and applications, is reflected in the presentation of the theory and results in this summary part of the thesis. The second chapter introduces the reader to metabolomics and provides the theoretical background to understand the NMR metabolomics platform (Publications I and II) presented in Chapter 3. The next three chapters give an overview to metabolites and metabolism, genetics and the materials and analysis methods applied in Publications III and IV. Chapter 7 summarizes the results from the two last publications and, finally, in Chapter 8 conclusions are drawn and future prospects contemplated.

2. Metabolomics

Metabolomics is the study of the metabolome, i.e, the complete set of metabolites found in a sample. The metabolite content and levels of the metabolites are assessed in order to provide a global view to the physiology of the system. Metabolomics is a broad term encompassing a wide range of techniques, samples and research aims and the field has expanded rapidly during the past decade; the applications range from plant science[1], animal studies[2] and toxicology[3] to human disease diagnostics[4] and risk prediction[5]. Reviewing the whole area of science including the various technologies and applications is beyond the scope of this thesis, thus this chapter focuses on presenting the key concepts of metabolomics, with special focus on serum proton (^{1}H) nuclear magnetic resonance (NMR) metabolomics, and some of the achievements in the field with respect to human health and disease.

2.1 Human metabolome

In analogy to genome, transcriptome and proteome, metabolome defines the whole of metabolites in a sample (Figure 1). However, compared to the other 'omes, the metabolome is less well specified. The human metabolome database[6] currently lists around 8000 metabolites found in the various biofluids or tissues of human body. The metabolites represent a wide range of chemical and physical properties – from larger hydrophobic lipid molecules to relatively small water-soluble sugars – and concentrations – from abundant metabolites with millimolar concentrations to those present in only nanomolar quantities.

The metabolite composition reflects the biological state of the system summing up the contribution of the genotype, gene expression and protein expression (Figure 1) but is also influenced by environmental factors. Therefore, among the various 'omes, metabolome is the most reflective of the phenotype. As disturbance of the metabolic homeostasis, e.g., by genetic modifications or disease processes, causes alterations in the metabolite levels, the system-wide metabolite profiles can be used for hypothesis-free observations of the influence of and pathways underlying these perturbations. Variation in the levels of metabolites that arises from sources largely independent of the perturbation can confound the studies of the metabolome. For instance, metabolome varies in time and shows influences of

diet[7, 8], gender[9-12] and exercise[13]. However, a substantial amount of the variation is genetically determined[14-20] (Publication IV).

The number and identity of metabolites differs considerably between different tissues and biofluids, and efforts are underway to characterize, using various technologies, the detectable metabolomes of the clinically important biofluids[21, 22]. Blood plasma, serum and urine are the most commonly used targets in metabolomics as the sampling is minimally or non-invasive and many studies typically collect this data. Depending on the research questions more exotic samples, including cerebrospinal fluid, saliva or faecal extracts, can be more appropriate targets of for the studies.

2.1.1 Serum metabolome

Blood is a body fluid carrying various substances to and from cells. It composes of two parts: blood cells, including white and red blood cells and platelets, and plasma, in which the substances, e.g., metabolites like amino acids and lipoproteins, are dissolved. Serum is similar to plasma but the collection procedures differ so that serum does not contain fibrinogens. As serum and plasma reflect the whole systemic metabolism, these biofluids are appealing targets for metabolomics experiments and especially suitable for studies of vascular and systemic diseases. Human serum metabolome database[21] currently catalogues over 4200 metabolites detected from human serum.

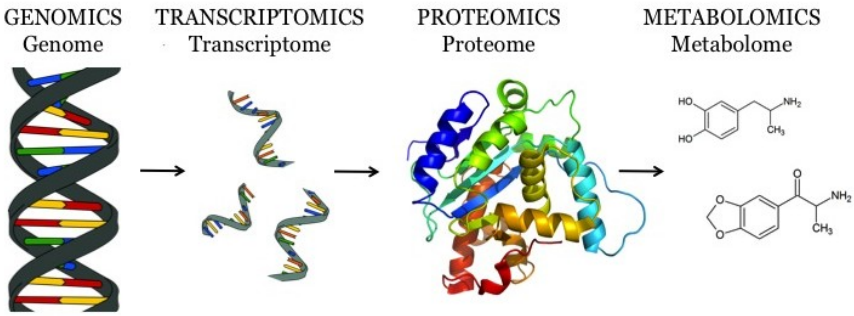

Figure 1. *The 'omics cascade. Genomics studies genomes, i.e., the hereditary information, transcriptomics the gene expression, proteomics the protein content and metabolomics the metabolome. The figure is put together by the author from separate figures taken from Wikimedia.*

2.2 Metabolomics measurement technologies

Metabolomics represents the end point of the 'omics cascade (Figure 1) and like other 'omics sciences, it is a large-scale study; A primary aim in metabolomics is the comprehensive measurement of ideally the whole metabolite content of a sample. Therefore, a key requirement for metabolomics analyses is a technology that enables the simultaneous measurement of a multitude of metabolites. For this purpose, analytical chemistry technologies have been adopted to metabolomics research. However, due to the wide range of chemical and physical

properties of metabolites, there exists no single technology that can observe the whole metabolome. For example, single metabolomics analyses can detect only a fraction of the compounds of the serum metabolome[21]. However, the metabolite information captured from a serum sample with the different analyses overlap only partially[21], and thus better coverage of the metabolome is obtained using several approaches.

There are two key technologies widely applied in metabolomics studies for detecting the metabolites: ^{1}H NMR and MS[23]. Both technologies produce spectra/chromatograms where the positions and sizes of the peaks carry the metabolite information, however, the basic principles and properties of the two technologies differ leading to different advantages and disadvantages and, thus, to the complementary nature of the technologies. A summary of the properties of the two technologies are given in Table 1 and briefly discussed below.

NMR detects protons, hydrogen ions, which are present basically in all biological molecules, based on their magnetic properties. These properties depend on the chemical environment of the proton, i.e., the other atoms surrounding the protons in the molecules. Hence, different molecules produce characteristic spectral shapes. MS, however, detects the molecules based on their mass and charge. This measurement requires the ionization of the sample in order to produce charged particles, and therefore, unlike in NMR, the sample cannot be recovered for further analyses after an MS experiment.

MS analyses are often coupled with a separation step of the molecules in the sample, e.g., liquid or gas chromatography, prior the measurement to enhance the detection of the metabolites. Moreover, MS experiments typically require a sample preprocessing step, in which the sample components that can interfere with the analysis, e.g., salts and proteins, are removed. The sample pretreatment, however, can be chosen so to target the analyses to metabolites of specific types. E.g., targeted analysis of serum lipid species, lipidomics, can be considered as a subfield of metabolomics. NMR, however, requires little or no sample preprocessing and no separation step of the molecules is used. Therefore, NMR is an unbiased method as in principle all hydrogen-containing metabolites with adequate concentration can be detected. Also, the simplicity of an NMR experiment makes the analyses highly reproducible and cost-effective.

A major advantage of MS is that it offers great sensitivity in detecting molecules also with minute concentrations. Therefore, compared to MS, NMR is insensitive as it detects only medium and high abundance metabolites. In general, MS-based metabolomics approaches allow for a more comprehensive coverage of the metabolome. However, in certain cases NMR performs better. For example, MS methods cannot distinguish between isoleucine and leucine as the two metabolites have the same mass, however, these amino acids produce distinct NMR spectra. Additionally, NMR allows for the measurement of lipoproteins.

The basics of ^{1}H NMR are discussed in more detail below.

Table 1. *The basics properties of the MS and NMR technologies used in metabolomics studies of biofluids. Data from Griffin et al.[24], Dettmer et al.[25], Issaq et al[26]. and Bictash et al.[27].*

	MS	**NMR**
Number and types of metabolites measured	From tens to thousands, depending on the sample preparation and separation methods various kinds of metabolites can be observed (aqueous metabolites including amino acids, glucose and other small molecules, and a wide range of lipid species including glycerophospholipids and acylcarnitine lipids), can detect metabolites with minute concentrations (down to nanomolar quantities)	From tens to around a hundred, in principle detects all hydrogen-containing metabolites but only those of medium and high-abundance in the sample, the observed metabolites depend on the biofluid studied but often include energy-metabolism related substances like glucose, lactate, pyruvate, creatinine and some amino acids (lipoproteins can be measured from serum/plasma)
Metabolites detected by	Mass-to-charge ratio of the ionized metabolites, separation methods used to enhance the detection	The magnetic properties of hydrogen ions, protons, in the molecules, which depend on the chemical environment of the proton
Measurement time	From few minutes to up to an hour (depending on the separation method and MS technique used)	Around ten minutes
Sample preparation	Often a requirement, the matrix (e.g. proteins) has to be removed to avoid this interfering with the analyses by, for example, blocking the separation column, the choice of preparation technique depends on whether the analyses are targeted or aiming for a more global profiling and on which separation method is used, for example, fatty acids and amino acids require derivatization prior gas-chromatography separation	Little requirements: with the latest spectrometers only buffer needs to be added to the sample prior analysis, sample preparation can be used for, e.g., removing the proteins or targeting the analyses on lipids
Sample separation	The metabolites of the sample are often separated by using, for example, gas or liquid chromatography prior the measurement, the separation method is chosen to suit the metabolites the analyses are targeted to or multiple separation methods can be used for a better coverage of the metabolome	Not applied
Metabolite identification	With the help of available libraries or internal standards	Libraries are used, nearly all of the signals can be annotated to metabolites
Advantages	Extremely sensitive and therefore allows the measurement of thousands of metabolites, various sample preparation and separation methods enable the measurement of different parts of the metabolome	Non-destructive to the sample, non-selective as all hydrogen-containing metabolites with adequate quantity can be detected, little or no sample preparation is required, low-cost and highly reproducible analyses
Downsides	Requires sample preparation and separation and material/metabolite information may be lost during these steps, analyses often targeted to metabolites of specific properties, some of the peaks in the spectra are not annotated, may be costly	Insensitive as detects only metabolites with medium to high concentrations, considerable peak overlap in spectra that complicates metabolite quantification

2.2.1 Proton NMR spectroscopy

Proton NMR spectroscopy is a technique for detecting chemical and physical properties of hydrogen-containing molecules. The technique takes advantage of the phenomenon called nuclear magnetic resonance, i.e., the behaviour of atoms when exposed to electromagnetic radiation in a magnetic field. The most common use of ^{1}H NMR spectroscopy is in organic chemistry as a tool to determine the structure of chemical compounds, yet it is increasingly applied in biomedicine.

Physical background
Certain atomic nuclei, atomic isotopes that have an odd number of protons, neutrons or both, have a property called spin, a magnetic moment, that makes the nuclei sensitive to external magnetic fields. One of the types of nuclei with spin is the hydrogen isotope ^{1}H, called proton, most of the hydrogen atoms consisting of this isotope. As hydrogen is present in nearly all naturally occurring compounds, NMR spectroscopy targeted in detecting protons is an attractive approach for biomedical studies.

In NMR spectroscopy, the sample containing the sensitive isotopes is exposed to an external magnetic field, which causes the nuclei to align to different energy states, in the simplest case to two states, either with or against the magnetic field. However, the nuclei are not distributed evenly to the energy states but when further exposed to electromagnetic radiation some of the nuclei absorb energy and move to the state with higher energy. Once the radiation subsides, the nuclei gradually return to their original state releasing electromagnetic radiation, a free induction decay (FID), which the NMR machine detects.

Not all of the nuclei in a molecule react in the same way, but the chemical environment of the atom, the number of electrons surrounding the proton, determines the frequency of the radiation needed to excite the nucleus, the resonance frequency of the nucleus. The nuclei with more electrons around them require less energy to be excited and thus resonate at a lower frequency. The more electronegative atoms or functional groups in proximity of the nuclei, the less shielding from electrons the nuclei get as the electrons are drawn towards the electronegative atoms. Thus, the frequencies of the electromagnetic radiation released when the nuclei relax contain information on the chemical structures of the molecules and can be used for structure elucidation (used in organic chemistry) or for molecule identification.

^{1}H NMR spectra
After a Fourier transform of the FID signal, the output of an NMR spectrometer is a spectrum of peaks of various sizes and shapes positioned along a frequency axis according to the chemical shifts, i.e., the resonance frequencies scaled by the magnetic field strength of the NMR spectrometer, of each unique nucleus. As a single molecule usually has protons in more than one type of chemical environment, each molecule manifests several peaks. The characteristics of the peaks, including the position, intensity and multiplicity, in an NMR spectrum

contain a wealth of information on the underlying molecules. For molecule identification not only the chemical shifts of the peaks pointing to specific functional groups in the chemical neighbourhood of the proton are of importance but several signals in a peak, i.e., the multiplicity of the peak, further help to determine the structure by reflecting the interactions of nearby protons. The peak areas are proportional to the number of protons contributing to each signal, and therefore the intensity values can be used for molecule quantification.

The ^{1}H NMR spectrum of a biological sample containing multiple compounds, such as serum, is complex with hundreds of peaks. As an example, typical NMR spectra of serum and urine are shown in Figure 2. As the chemical shift scale is limited, peak overlap is significant in many parts of the spectrum, therefore complicating the identification and quantification of the metabolites, although the redundancy arising from one molecule manifesting several peaks may help to overcome this issue. Further complexity is introduced by the small shifts in peak positions occurring from sample to sample.

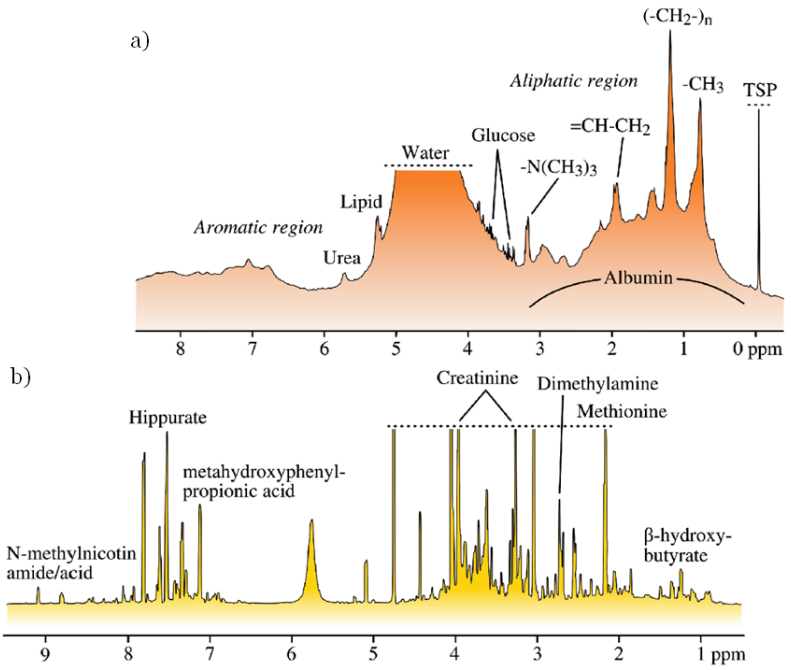

Figure 2. *^{1}H NMR spectra of two biofluids obtained with standard settings: a) serum, b) urine. Some of the peaks are annotated to metabolites. The figure is modified from the thesis of Ville-Petteri Mäkinen with permission.*

The first studies NMR of plasma and serum were conducted in late 1970's and early 1980's[28, 29]. Since then, the knowledge of the origins of the peaks has increased greatly and nowadays extensive catalogues exists where the observed peaks at specific chemical shifts with specific shapes are assigned to molecules[21, 30]. NMR has the appealing feature that it detects the lipoprotein profile; A typical ^{1}H NMR spectrum of plasma and serum contains broad signals arising from various types of lipoproteins (signals annotated -CH_3, $(-CH_2)_n$, =CH-CH_2, -

$N(CH_3)_3$ in Figure 2a). The concept of using ^1H NMR to quantify lipoproteins and lipoprotein subclasses was invented in 1990's[31, 32], and as this offers a more efficient method over the conventionally used ultracentrifugation, applying NMR-based lipoprotein profiling is increasingly being applied in biomedical research (e.g., Chasman et al.[33], Hodge et al.[34]). The broad lipoprotein signals, however, hamper the detection of the less-abundant molecules. These metabolites can be observed either by removing the proteins from the sample or by using different NMR pulse sequences[35, 36], which is the more often applied approach.

2.3 Metabolic fingerprinting vs. quantitative metabolomics

The outputs from the metabolomics measurements are information-rich, complex spectra where thousands of peaks represent the metabolites detected (e.g. Figure 2). This metabolite data can be used as such, as a whole, to identify typical patterns of changes in the raw metabolite profile that relate to biological perturbations, and the metabolites contributing to these patterns can subsequently be identified. Alternatively, the metabolomics spectra can be used as a source of individual metabolite data, and once identified and quantified the levels of the metabolites can be used in analyses either individually or in combination. Both approaches have their benefits and limitations.

The former use of the data, so called metabolic fingerprinting, is the approach the first metabolomics studies took and many still apply (e.g., Jung et al.[37]). Using the whole spectra retains all the metabolite information, however, the analyses of the complex spectra require statistical methods capable of handling the multivariate data, such as principal component analysis or partial least squares. The interpretation of the results from these analyses is rarely straightforward and the link to the underlying biology can be lost.

The latter approach, quantitative metabolomics, is becoming more popular, as this approach allows also for the univariate analyses, analyses conventionally applied in epidemiological and clinical studies. However, the quantification and identification of the individual metabolites from the spectral data is a challenging task due to the multitude of often overlapping and shifting peaks and is thus commonly done manually or semimanually. Another downside is that not all peaks can be annotated to specific metabolites, especially with MS data.

2.4 Applications of metabolomics in biomedical research

The applications of metabolomics are numerous, as this area of science has exploded during the past decade. Thus only examples of the studies and findings from studies of human disease and metabolism are presented below.

One of the first human metabolomics studies gave high hopes, reporting that the NMR profiles of blood plasma predicted not only the presence but also the severity of coronary heart disease[38]. These claims, however, were not replicated in

a study from another group using a different set of patients[39], raising the possibility that the original findings were mainly due to a highly selected patient sample[40]. More successful applications to cardiovascular disease and other metabolic conditions have been published since.

Wang et al. identified a novel pathway underlying atherosclerosis using unbiased gas chromatography (GC) – MS[41]. The three peaks that showed differences between the cases and controls in two cohorts (75 cases and 75 controls in total) were identified as choline, trimethylamine N-oxide and betaine, and the levels of these metabolites predicted cardiovascular disease risk also in an independent clinical cohort. Unlike in many metabolomics studies up to date, the group went further to elucidate the underlying mechanisms in animal models. The metabolites were confirmed to be derived from dietary phosphatidylcholine and to associate with the risk of atherosclerosis through a pathway involving gut flora that promoted the formation of macrophage foam cells, thus highlighting the role of both dietary regulation and manipulation of the microbial composition in the treatment of cardiovascular disease. Gut microbiota was also pointed out as one of the contributing factors to the differences in urinary ^{1}H NMR profiles between different population samples in a large study of over 4,600 individuals from China, Japan, UK and USA[42]. The authors also identified potential urinary biomarkers for blood pressure.

Aside from aiming at separation between two groups, the metabolomics data in combination with suitable multivariate statistics can be used to stratify individuals into metabolically more accurate groupings than the clinical definitions allow. Self-organizing map analysis of serum NMR data from 4,309 young adults revealed that there was no single metabolic phenotype underlying high carotid intima-media thickness, a surrogate marker of cardiovascular disease, but the condition was described by three distinct profiles, varying levels of blood lipids contributing most to these profiles[43]. Another study applying the same methodology in 613 type I diabetics uncovered the biochemical background and the complex interactions between various diabetic complications pinpointing the limitations of conventional definitions in classifying the high-risk individuals[10].

The above studies analyzed the metabolomics data as a whole, however recently the focus has shifted on quantitative metabolomics. Wang et al. identified novel biomarkers for type 2 diabetes (T2D) using targeted liquid chromatography (LC) – tandem mass spectrometry (MS/MS) to profile 61 metabolites from serum samples of 189 pairs of future diabetics and matched controls[5]. Five branched-chain (isoleucine, leucine and valine) and aromatic amino acids (tyrosine and phenylalanine) were found to predict the development of diabetes adding predictive value to the traditional risk factors, and the findings were validated in an independent large lower-risk cohort.

Intriguingly, a few other recent studies have also pinpointed the role of circulating amino acids in metabolic disorders. Newgard et al. assayed a number of conventional metabolites as well as metabolites with MS/MS and GC-MS in

obese and lean individuals (N = 74 and N = 67, respectively)[44]. Among other metabolites, nine amino acids had significantly different concentrations between the groups. Principal components analysis of the metabolites showed that a component mainly consisting of variance from branched-chain amino acids (BCAA; valine, isoleucine and leucine) contributed significantly to the separation between the lean and obese and also associated with insulin resistance. Further investigation in an animal model showed that BCAA supplementation with high fat diet resulted in insulin resistance although the animals gained less weight than those on high fat diet without supplementation.

In another study from the same group the potential of serum metabolites to discriminate between coronary artery disease (CAD) patients and healthy controls was evaluated[45]. In total of 69 metabolites were measured with MS from test (174 cases, 174 controls) and replication (140 cases, 140 controls) sets. Two principal components consisting of branched-chain amino acid and urea cycle metabolites associated significantly with CAD. Furthermore, in a small pilot-like study that combined metabolomics data from three platforms to cover a larger range of serum metabolites, Suhre et al. identified the branched-chain amino acids as one of the metabolite groups deregulated in diabetes[46].

As metabolite levels only provide a partial view to the whole of biological processes in human body, some studies have analyzed metabolomics data in concert with data from other ‘omics platforms in order to provide a systems biology view. Inouye et al. performed the first study that combined serum NMR metabolomic, transcriptomic and genetic data[47]. A module of co-expressing genes that are key components of inflammation and allergy associated with multiple metabolites. Causal network inferred using the genotype data revealed that the expression of the module was reactive to the levels of some of the metabolites and also that metabolite levels were reactive to other metabolites. The use of metabolomics data in genome-wide association studies is covered in Chapter 5.4.2.

To summarize, the metabolomics studies have uncovered a number of new potential biomarkers and provided hypotheses for the mechanisms underlying metabolic conditions therefore clearly supporting the use of the hypothesis-free metabolomics approach in further studies. Many of the discoveries have been made with surprisingly limited sample sizes. As metabolomics data from large well-characterized cohorts becomes available many more discriminating or predictive metabolites can be expected to be found. A major challenge, as in all biomedical research, is to translate the findings into interventions and treatments or tools for risk evaluation and diagnostics. One step on the way is to gain more knowledge on the variability of the discovered metabolites or metabolite profiles due to genetic, gender, age and environmental differences.

3. ^{1}H NMR serum metabolomics platform

This chapter presents the set-up for a serum ^{1}H NMR metabolomics platform, reported in Publications I and II, designed to provide high-throughput metabolite data for clinical and epidemiological studies in a cost-effective manner; Publication I presents the three molecular window approach applied to obtain a large quantity of various molecular data from a single serum sample and Publication II describes the protocol from NMR spectroscopy details and experiment flow to the data handling and analyses. The platform development has been a group effort and builds on the expertise of people from various fields, including NMR analytics and bioinformatics, and on the previous work from the group members[10, 48-50]. The author's contribution to the platform has been on the spectra interpretation and analysis, thus the platform is summarized in this chapter from this perspective.

3.1 ^{1}H NMR of serum - Three molecular windows (Publication I)

The NMR metabolomics platform is designed for the measurement of the metabolite profile of blood serum, a primary body fluid reflecting the biochemistry of the whole system and thus suitable for the study of systemic and vascular complications. The molecular variety in serum is wide ranging from large macromolecules like lipoprotein particles and albumin with millimolar concentrations to small analytes with minute quantities and thus not detectable by NMR. The signals from the most abundant macromolecules dominate the NMR spectrum acquired with the standard settings and although these provide valuable information on the lipoprotein particle profile the broad signals hamper the identification and quantification of the smaller molecules.

To facilitate the detection of as much of the metabolite content as possible from a single serum sample, the NMR experiments were targeted to three molecular windows; Two NMR experiments are run for the native serum sample utilizing different pulse sequences (Table 2) to acquire a variety of molecular information, and the third analysis is acquired from serum lipid extracts providing data on the

individual lipid molecules. Figure 3 shows typical spectra from the three molecular windows with majority of the peaks annotated to metabolites.

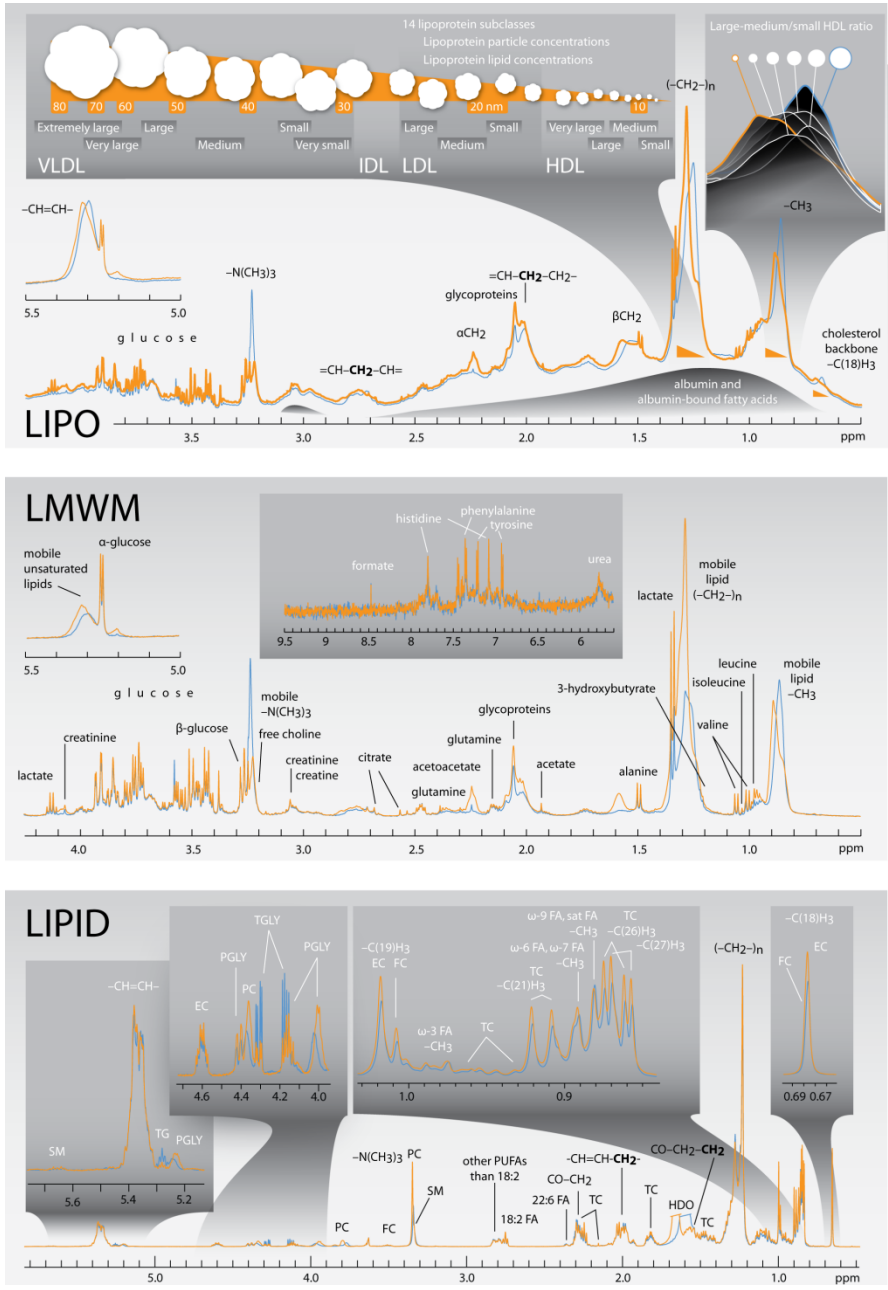

Figure 3. *The three molecular windows of the NMR platform with majority of the peaks annotated to the metabolites or functional groups giving rise to the signals. LIPO and LMWM windows are acquired from a native serum sample and contain information on the lipoproteins and small molecules, respectively. LIPID window is acquired from serum lipid extracts and thus shows signals arising from, e.g., fatty acids. The figure is courtesy of Antti J. Kangas.*

The lipoprotein lipids (LIPO) window represents a conventional ^{1}H NMR spectrum of serum with broad overlapping signals arising mainly from lipoprotein particles and albumin and a few sharp peaks from smaller molecules superimposed on the round shapes (the top section in Figure 3). The size and shape of some of the broad peaks (methyl, $-CH_3$, at 0.8 ppm and methylene, $-CH_2-$, at 1.3 ppm) in the spectrum reflect the lipoprotein particle distribution in detail; these signals are superpositions of the signals arising from the lipoprotein subclasses. Therefore various lipoprotein subclass particles and measures can be derived from these peaks with mathematical modelling. The zoom-ins in Figure 3 show how the signals from the individual subclasses contribute to the observed peak shapes.

Further peaks from the smaller molecules are revealed when a pulse sequence that suppresses the majority of the signals from lipoproteins (Table 2) is applied. Thus, the low-molecular-weight metabolites (LMWM) window features a number of sharp signals that can be assigned to tens of different molecules (the middle section in Figure 3). For example, the spectrum features signals from various amino acids including alanine (peaks at 1.5 ppm), valine (at 1.0 ppm) and glutamine (at 2.5 ppm), and other small molecules, e.g., creatinine (at 3.1 and 4.1 ppm). Also some residual signals from lipoproteins are present (mobile lipids at 0.8, 1.3 and 5.3 ppm). In combination LIPO and LMWM windows are likely to contain most of the information on the molecules observable by ^{1}H NMR of native serum.

The lipid extraction procedure breaks down the lipoprotein particle structure uncovering the various lipid and fatty acid species within the particles. The signals in the NMR spectrum acquired from the lipid extracts (LIPID window, the bottom section Figure 3) arise from, e.g., various fatty acid (FA) groups including ω-3 and ω-6 FAs (e.g., at ~0.95 ppm), and other types of lipids including sphingomyelin (e.g., at 3.3 ppm), phosphatidylcholine (e.g., at 3.4 ppm) and free and esterified cholesterol (e.g., at 0.6 ppm). In addition, some detailed molecular characteristics can be derived from the signals including the average number of double bonds in a FA chain and a measure for the average FA chain length.

3.2 Experiment flow (Publication II and other data)

To ensure the high-throughput and minimize experimental variation the sample preparation and analysis have been highly optimized utilizing the latest robotics-controlled laboratory tools and NMR spectrometer components. As each sample undergoes the same protocol, the results also from different study sets should be directly comparable in terms of the NMR experimentation. Figure 4 illustrates the structure of the platform and the analysis flow of the metabolomics experiments.

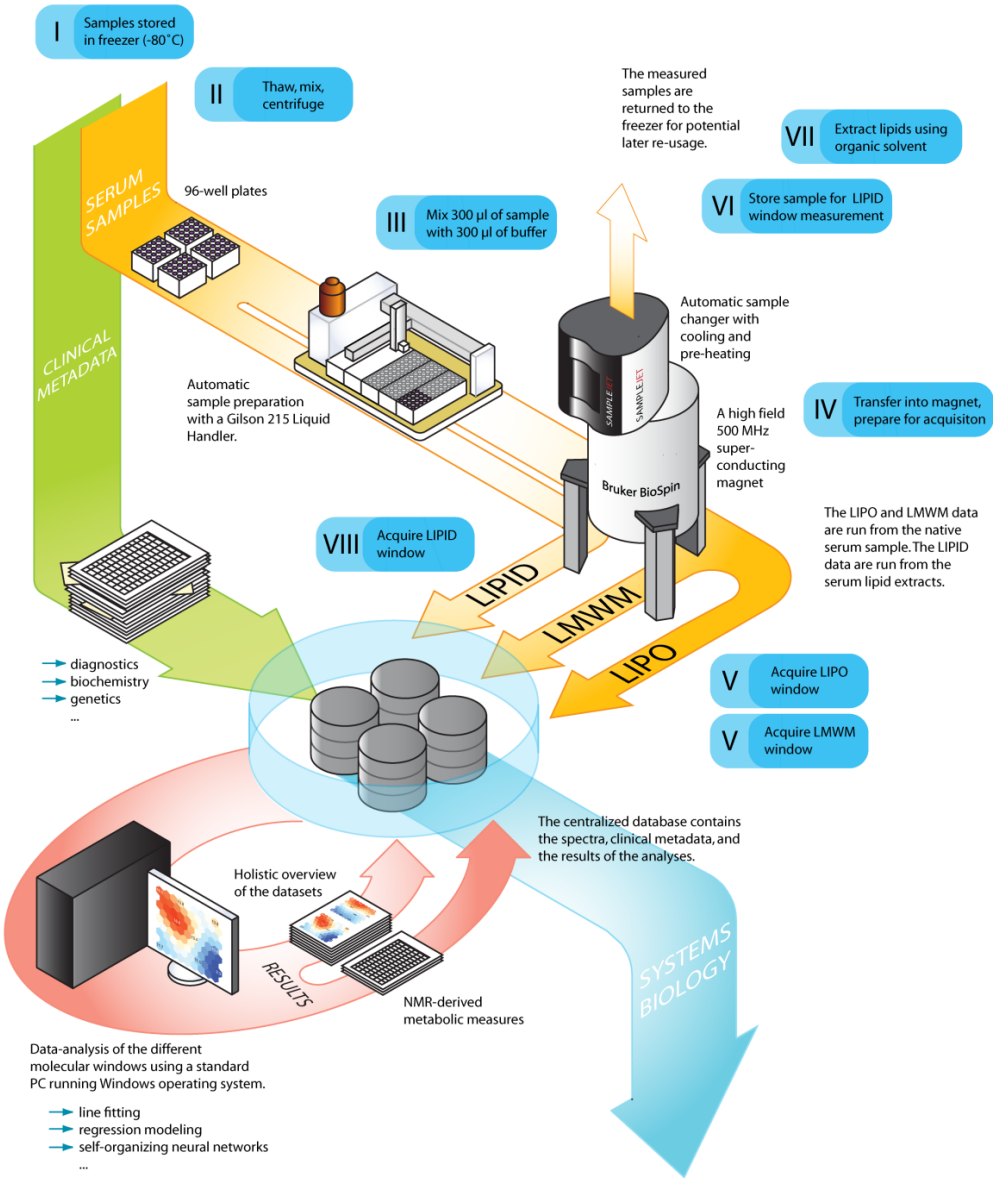

Figure 4. *An illustration of the components and the analysis flow in the metabolomics platform. The roman numerals refer to the order of the steps in sample preparation and measurement. The original figure and text are courtesy of Antti J. Kangas and Pasi Soininen and modified with permission.*

Table 2. *The experimental parameters for the three molecular windows used in the NMR measurements.*

	LIPO	**LMWM**	**LIPID**
Data points (k)	80	64	64
Transients	8	24	32
Acquired after	4 dummy scans	4 dummy scans	4 dummy scans
Excitation pulse	Automatically calibrated	Automatically calibrated	Fixed length
Water peak suppression	Bruker noesypresat: presaturation during relaxation delay (irradiation field of 25 Hz), 10 ms mixing time and a spoil gradient	Presaturation during relaxation delay (irradiation field of 25 Hz)	-
Filter	-	78 ms T_2-filter with a fixed echo delay of 403 μs	-
Acquisition time (s)	2.7	3.3	3.3
Relaxation delay (s)	3	3	3
Measurement temperature (˚C)	37	37	22
Preprocessing of FID			
Zero-filled to (k)	128	128	128
Line broadening (Hz)	1	1	0.5

The NMR experimentation of the LIPO and LMWM windows requires little sample preparation; The serum samples stored in a freezer at -80°C are fist thawed in a refrigerator (+4°C) overnight, then mixed and centrifuged (3400 × g) for two minutes, and finally 300 μl of serum is mixed with 300 μl of sodium phosphate buffer. A liquid handler (Gilson Liquid Handler 215) automatically prepares the samples directly to NMR tubes (outer diameter 5 mm) in a process of slowly mixing and aspirating the serum and buffer to avoid sample foaming. The sample preparation process for 96 samples takes approximately 2 hours.

The prepared samples in 96-tube racks are inserted to a robotic sample changer (Bruker SampleJet) mounted on top of a Bruker NMR spectrometer. The sample changer can hold up to five well plates at a time (480 samples) and it is equipped with a cooling unit to keep the samples at a refrigerator temperature (+6°C) while awaiting for the measurement. Additionally, the SampleJet includes a pre-heating unit to warm the samples to a physiological temperature (+37.5°C; 0.5°C of heat is lost during the sample transfer into the spectrometer) just before the NMR experiment, thus minimizing the time needed for temperature stabilization inside the magnet.

The ^{1}H NMR experimentation is performed with a Bruker AVANCE III spectrometer operating at 500.36 Hz (11.74 T) and dedicated to the metabolomics measurements. The spectrometer is equipped with the latest components, including an inverse selective probehead, that enable high sensitivity in the analysis and thus facilitate metabolite quantification. The total time required for LIPO and LMWM measurements and spectra preprocessing is less than 9 minutes per sample, while with the previous manual experimentation used in Publication I

the corresponding time was about 30 minutes. The experimental details for LIPO and LMWM windows are given in Table 2.

As NMR experimentation is non-destructive to the sample, the sample used for LIPO and LMWM measurements can be stored for later use, including the lipid extraction and LIPID window measurement. Currently, the lipid extraction protocol is done manually as specified in Publication I, but automation of the procedure is in progress. The manual extraction, however, poses no bottleneck time-wise as the throughput per week (500 samples) meets the capacity of the NMR spectrometer. The experimental details for the LIPID window are in Table 2 and the measurement lasts approximately 10 minutes.

An essential part of the platform is the storage of the vast amounts of data generated. The data from the experiments as well as other metadata are stored to a dedicated server to a centralised data base from where it can be easily accessed and analysed with the analysis tools incorporated or other software.

3.3 Data-analysis example: Self-organizing map

One of the data-analysis and visualization tools incorporated to the platform is self-organizing map (SOM). SOM is an unsupervised pattern recognition method that projects the multidimensional metabolomics data onto a two-dimensional map, the metabolically similar samples residing close to each other and dissimilar further apart. With the statistical colourings implemented (Mäkinen et al.[10], Publication I), SOM provides a powerful approach for simultaneous visualization and comparison of various metabolic (NMR) and clinical (external) features and their relationships. The SOM-analysis does not require metabolite quantification but can use the NMR spectra as an input and thus is a suitable first-stage analysis approach.

In Publication I the SOM approach was used in the analysis of the metabolic characteristics of mild cognitive impairment (MCI), a transitional stage between normal cognition and dementia. The NMR spectra data from the three molecular windows from 180 serum samples were analysed. Although the study sample was limited, the metabolic features captured in the serum NMR spectra, i.e., systemic metabolism, were reflective of the cognitive decline and resulted in a significant cluster of the MCI samples. Selected signals were quantified from the spectra to further analyze the contribution of the metabolites to the observed distribution of the samples on the SOM. A metabolic characteristic that coincided remarkably well with the high proportion of MCI cases on the map was a low relative amount of ω-3 fatty acids. This finding provided another perspective on the role of polyunsaturated fatty acids in dementia development[51].

With the platform running the metabolomics measurement of considerably larger, epidemiological, data sets is now possible. In Publication II the metabolomics data (LIPO and LMWM) from 4470 serum samples were analysed

with SOM to exemplify how the data-driven metabolic phenotyping with NMR metabolomics data reflects metabolic syndrome and its components.

3.4 Metabolite quantification

To facilitate quantitative metabolomics and the use of NMR as an alternative to standard laboratory assays, models for automated metabolite quantification from the NMR spectra have been developed. As these models are yet to be published, detailed data on the models and their performance cannot be presented in this thesis.

The deconvolution of lipoprotein measures from the largely overlapping and broad peaks of the LIPO window is not straightforward. For the LIPO window we have implemented models for 14 different lipoprotein subclasses including their lipid and particle concentrations (Table 3 presents the subclasses, particle sizes and measures), as well as other serum lipoprotein and lipid measures, e.g. serum triglycerides. In total 90 measures are quantified from the LIPO window.

The quantification bases on regression models similar to those presented by Vehtari et al[50] as we have observed that regression models perform better than the line fitting-based approaches in the lipoprotein quantification. All models were calibrated via NMR-independent measures from high-performance liquid chromatography (HPLC) and cross-validated to evaluate the performance. The average r^2 for the measures from HPLC and NMR is 0.75 (SD = 0.14) with 73% of the models having $r^2 > 0.7$. Utilizing a similar approach Bruker has recently developed a commercial package for lipoprotein subclass quantification. As an example of the accuracy of the lipoprotein measure quantification from NMR spectra a comparison between the enzymatically measured serum triglycerides and the NMR measured triglycerides is given in Figure 5.

Due to the sharper peaks and less extensive overlap in the LMWM and LIPID windows the method of choice for metabolite and lipid signal quantification from LMWM and LIPID spectra, respectively, was initially iterative lineshape fitting analysis (Figure 6) done using software from PERCH Solutions Ltd. However, in the latest quantification protocol the metabolites also from these windows are quantified using regression models, which perform more efficiently time-wise. The metabolites were identified using multidimensional NMR spectra and literature references of chemical shifts and peak shapes[6]. As material loss may occur during the lipid extraction procedure, the data from LIPID window is scaled with the ratio of the total cholesterol signals from LIPO and LIPID windows. This approach assumes the loss of material is equal for all lipid species. In total 22 metabolites are currently quantified from the LMWM spectra and 15 from the LIPID spectra. As an example of the accuracy of the small molecule quantification the creatinine levels measured via conventional laboratory techniques and NMR from the same serum samples are compared in Figure 5.

Altogether 117 metabolites are currently quantified from serum samples using the NMR-based platform. The metabolites and their roles in human metabolism are overviewed in the next chapter.

Table 3. *The lipoprotein subclasses and lipid components quantified from the NMR spectra. The components that can be reliably quantified for each subclass are marked with an x.*

Lipoprotein subclass	Average diameter (nm)	TC	FC	CE	PL	TG	TL	[P]
Chylomicrons and extremely large VLDL	75 upwards	-	-	-	x	x	x	x
Very large VLDL	64	-	-	-	x	x	x	x
Large VLDL	53.6	x	x	x	x	x	x	x
Medium VLDL	44.5	x	x	x	x	x	x	x
Small VLDL	36.8	x	x	-	x	x	x	x
Very small VLDL	31.3	-	-	-	x	x	x	x
IDL	28.6	x	x	-	x	x	x	x
Large LDL	25.5	x	x	x	x	-	x	x
Medium LDL	23.0	x	-	x	x	-	x	x
Small LDL	18.7	x	-	-	-	-	x	x
Large HDL	14.3	x	x	x	x	x	x	x
Medium HDL	12.1	x	x	x	x	-	x	x
Small HDL	10.9	x	x	x	x	-	x	x
Very small HDL	8.7	-	-	-	-	x	x	x

TC, total cholesterol; FC, free cholesterol; CE, cholesterol esters; PL, phospholipids; TG, triglycerides; TL, total lipids; [P], particle concentration.

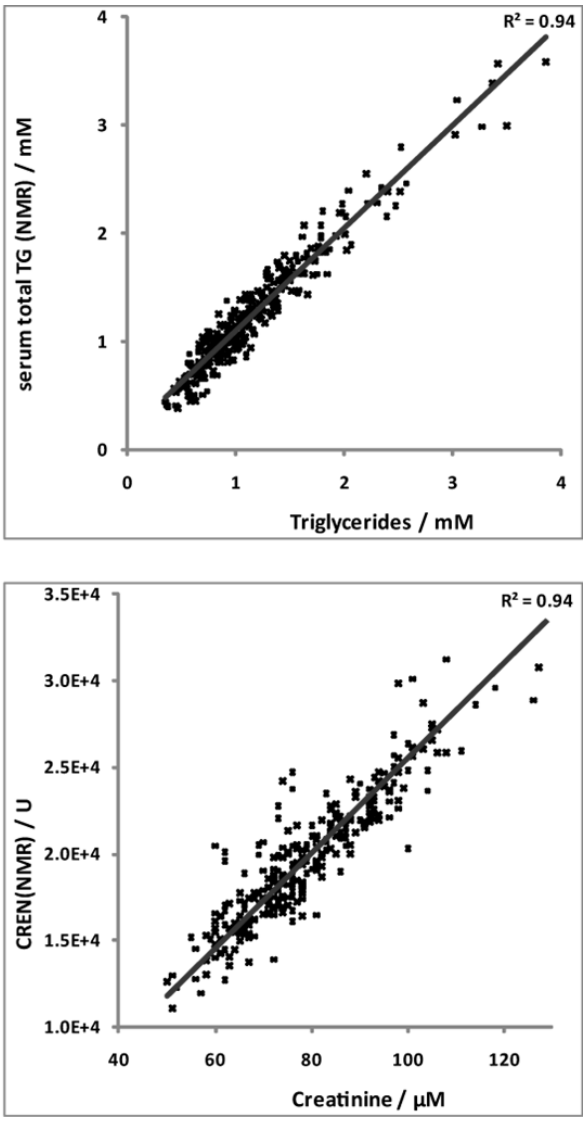

Figure 5. *Comparison of the NMR measured triglycerides and creatinine to the corresponding measures obtained with conventional laboratory assays. The figure is courtesy of Pasi Soininen.*

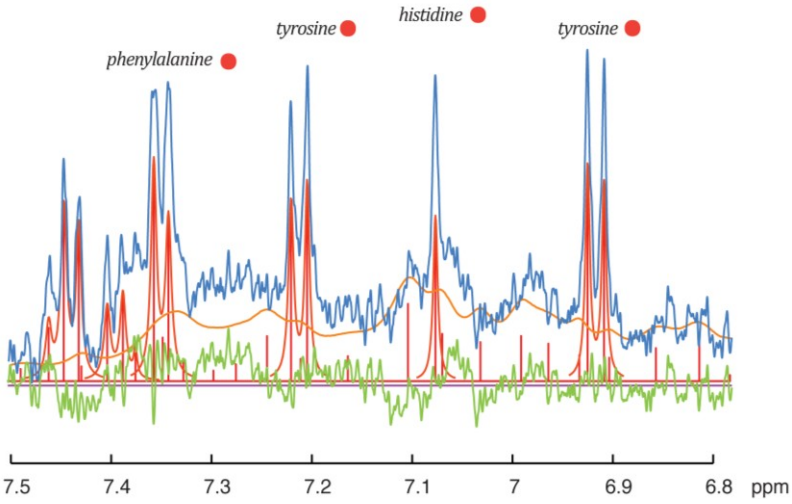

Figure 6. *An example of the line shape fitting procedure. The figure is courtesy of Antti J. Kangas.*

3.5 Discussion

Over the past decade many research groups have set up mainly NMR or MS-based metabolomics protocols and commercial platforms have also emerged. The metabolomics data from NMR overlaps only partly with the metabolomics data from MS-based platforms and thus the two approaches provide complementary information. The metabolites unique to NMR are mainly arising from the LIPO window, as MS-based methods typically are not capable of detecting lipoproteins. However, with MS a multitude of other detailed molecular information can be captured.

In general, NMR-based approaches are less expensive compared to various MS metabolomics approaches largely due to the minimal sample preparation and measurement time required in NMR experimentation. However, methods are in constant development and MS profiling of over 250 serum metabolites can be achieved in 24 minutes[14]. In terms of the throughput and automated analyses the serum NMR platform presented here is to the author's best knowledge unique. For example, another NMR-based metabolomics platform designed for large-scale metabolic profiling offers the measurement of 48 plasma or serum samples or 120 urine specimens per day[27, 52] but currently offers automated or semiatomated quantification of only a small number of metabolites[27]. A commercial Chenomx platform uses their patented Targeted Profiling technique for metabolite quantification from NMR data promising rapid and accurate analyses, however, no information on the details of this method was available on the company website (http://www.chenomx.com/, accessed 24.11.2011).

NMR-based metabolite profiling can be used for all kinds of fluids and also tissues. This platform was optimized for serum, mainly due to the group's interest in vascular problems, this biofluid providing a source of relevant metabolite information to these studies. Plasma is a fluid similar to serum, but contains

clotting factors, and could have served as an alternative for the target biofluid. It has been reported, using MS-based metabolomics, that plasma measures of metabolites are less variable than the corresponding measures from serum, but also that the metabolite concentrations are higher in serum [53]. We have noticed that more metabolites can be quantified from serum NMR spectra compared to plasma. Work is ongoing to adjust the platform and metabolite quantification models also for plasma samples.

NMR analyses are highly reproducible, however, some variability can be introduced during sample storage and handling. Too high sample storage temperature and an increasing number of freeze-thaw cycles can have an effect on NMR signals, especially those arising from lipoproteins[54-57]. In our analyses we have observed that samples stored in -80°C produce a highly similar spectrum and result in similar quantification values for metabolites even after years of storage. Repeated freezing and thawing of the sample appears to be a more of a critical factor, however, significant effects begin to appear only after several freeze-thaw cycles. Furthermore, as the sample preparation is automated, with the exception of the lipid extraction, the variation introduced in this step is minimal.

Quantification of further metabolites could be facilitated to some extent using a spectrometer with stronger field strength or using longer acquisition times. The NMR analyses were, however, optimized for the set-up presented in this chapter to allow for cost-effective and high-throughput metabolomics analyses. Thus, the platform has provided metabolite profiles and quantitative data on more than a hundred metabolites for several population-based studies. The platform has been running for roughly three years, and several published papers have utilized the data from the platform[43, 47, 58-65].

4. Metabolic context of the NMR measured metabolites

The NMR metabolomics platform presented in the previous chapter currently allows for the quantification of in total 117 metabolites from three NMR spectra acquired from a fasting serum sample or the lipid extracts of the sample. The quantified metabolites, which are summarized in Table 4, are not selected based on suspected role in cardiometabolic diseases but rather represent the most abundant serum metabolites and thus observable by NMR. The metabolomics protocol was utilized in Publications III and IV to obtain quantitative metabolite data. This chapter gives a brief overview to the quantified metabolites and the primary pathways they are involved in and connected by in the context of human metabolism. As the greatest proportion of the quantitative metabolite information is lipoprotein-related and as Publication III largely focuses on lipoproteins, a major emphasis in this chapter is on these metabolites and their metabolism.

Table 4. *The metabolites quantified via the NMR platform grouped together according to the metabolite type and/or metabolic pathways the metabolites are involved in, i.e, not necessarily according to the molecular windows the metabolites are quantified from. Only the particle class for each lipoprotein subclass is given in the table; the lipid components quantified for each subclass are given in Table 3 and Appendix I, which lists all the metabolite measures, including the derived measures and metabolite ratios, analyzed in Publications III and IV.*

Lipoprotein measures	**Lipids and related metabolites**	**Glycolysis and citric acid cycle metabolites**	**Ketone bodies**
Extremely large VLDL	Total fatty acids	Citrate	3-hydroxybutyrate
Very large VLDL	Polyunsaturated fatty acids	Glucose	Acetoacetate
Large VLDL	ω-3 fatty acids	Lactate	**Waste products**
Medium VLDL	ω-6 fatty acids	Pyruvate	Urea
Small VLDL	ω-9 and saturated fatty acids	**Amino acids**	Creatinine
Very small VLDL	Docosahexaenoic acid	Alanine	**Other metabolites**
IDL	Linoleic acid	Glutamine	Glycoproteins
Large LDL	Total cholesterol	Histidine	Acetate
Medium LDL	Free cholesterol	Isoleucine	
Small LDL	Esterified cholesterol	Leucine	
Very large HDL	Total triglycerides	Phenylalanine	
Large HDL	Total phosphoglycerides	Tyrosine	
Medium HDL	Total cholines	Valine	
Small HDL	Phosphatidylcholines		
TC	Sphingomyelins		
TG	3 signals from mobile lipids		
LDL-C	Albumin		
HDL-C	Glycerol		

VLDL, very-low-density lipoproteins; IDL, intermediate-density lipoproteins; LDL, low-density lipoproteins; HDL, high-density lipoproteins; LDL-C, LDL cholesterol; HDL-C, HDL cholesterol.

4.1 Lipoproteins

Lipids, including cholesterol, triglycerides and phospholipids, are essential components for many molecular reactions and cell membrane composition and serve as an important source of energy. Due to their poor solubility to blood, lipids are transported in lipoprotein particles, which are complexes of lipids and lipid-binding proteins, apolipoproteins[2]. Thus, lipoproteins are key players in human metabolism and abnormal lipoprotein levels, dyslipidemia, is an acknowledged risk factor for cardiovascular disease.

[2] The levels of apolipoproteins B and A1 can be estimated from the quantified NMR lipids using the so-called extended Friedewald formula[152]

4.1.1 Composition and classification

Lipoproteins are spherical particles that have a lipid-rich core consisting mainly of hydrophobic lipids, i.e., esterified cholesterol and triglycerides, and a hydrophilic surface layer of mainly unesterified cholesterol, phospholipids and apolipoproteins. Lipoproteins are a heterogeneous group of particles with different particle size, density and both lipid and apolipoprotein compositions. Each of these characteristics can be used to categorize the particles. A commonly applied criterion is the density, by which lipoproteins can be classified into five distinct groups: chylomicrons (CM), very-low-density lipoproteins (VLDL), intermediate-density lipoproteins (IDL), low-density-lipoproteins (LDL) and high-density lipoproteins (HDL); the less dense the particles the larger they are in size and compose of more lipids than protein. The basic properties of these classes are given in Table 5 and illustrated in Figure 7a.

The lipoprotein classes HDL, LDL and VLDL are heterogeneous composites and can further be subdivided into more detailed lipoprotein subclasses that further differ in size and composition. Lipoprotein subclasses have become of interest as different subclasses may exhibit different functions[66, 67]. The role of lipoprotein subclasses as well as the lipoprotein particle number has been debated especially with regards to cardiovascular risk[67-70]. The NMR platform quantifies 14 different lipoprotein subclasses and a number of lipid components for each particle type (Tables 3 and 4).

Table 5. *Density, diameter, main apolipoprotein constituents and weight percentages for the lipid components for the main lipoprotein fractions.*

Particle	Density (g/ml)	Particle diameter (nm)	Main apos	%TG	%C	%PL	%PROT
CM	< 0.95	100-500	apoB	84	8	7	< 2
VLDL	0.95-1.006	30-80	apoB	50	22	18	10
IDL	1.006-1.019	25-50	apoB	31	29	22	18
LDL	1.019-1.063	18-28	apoA	4	50	21	25
HDL	> 1.063	5-15	apoA	8	30	29	33

Apo, apolipoprotein; TG, triglyceride; C, cholesterol; PL, phospholipid; PROT, protein. Modified from Biochemistry 2nd Edition, 1995, Garret & Grisham.

The lipoprotein metabolism is a complex interplay of the various lipoprotein particles, lipid transfer proteins, cell surface receptors and enzymes. Briefly, the functions of the five lipoprotein classes are the following (Figure 7b): The CM, VLDL and IDL particles transfer dietary (CM) or internally synthesized (VLDL and IDL) triglycerides to peripheral tissues, and the LDL and HDL particles function in maintaining the cholesterol homeostasis by transporting cholesterol to (LDL) and from (HDL) tissues. The following paragraphs and Figure 8 describe these processes in more detail.

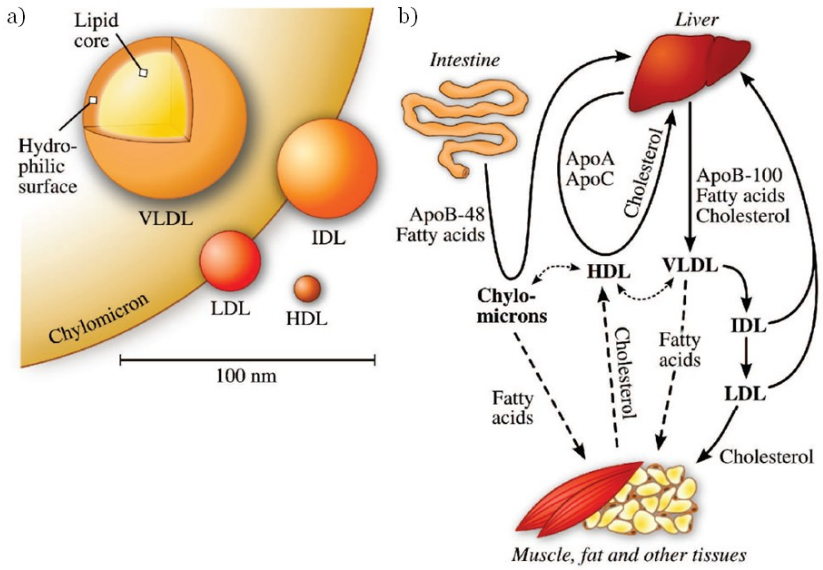

Figure 7. *Illustrations of a) the lipoprotein particle composition and the five main classes of lipoproteins, and b) the primary processes in lipoprotein metabolism. The figure is modified from the thesis of Ville-Petteri Mäkinen with permission.*

4.1.2 Lipoprotein metabolism

Chylomicrons are formed in the epithelial cells of the small intestine from the dietary lipids, a majority of which are triglycerides, and apolipoprotein components, apolipoprotein B-48 being specific for CM particles. After secretion to the blood stream the CM particles rapidly transport the triglycerides to extrahepatic tissues, e.g., muscle and adipose tissue, where lipoprotein lipase (LPL) first hydrolyzes the triglycerides before taking the fatty acids in the tissue. The delipidated CMs, chylomicron remnants, consisting mainly of apolipoproteins and cholesterol, are taken up by the liver mainly through LDL receptor (LDLR) or LDLR related proteins (LRP)[71]. Chylomicrons are only present in human circulation in the postprandial state, thus lipoprotein measures from fasting blood do not include CMs.

The triglycerides synthesized in liver are packed to VLDL particles, which, like CMs, transport triglycerides to peripheral tissues, but are considerably smaller in size and contain apolipoprotein B-100. The secretion of VLDL particles to the bloodstream starts the delipidation cascade, in which, with depleting TG content, VLDL particles gradually turn into IDL and finally to LDL particles that contain only small amounts of TG. The triglycerides from VLDL and IDL particles are internalized through the actions of LPL[72] and hepatic lipase (HL) the latter mainly hydrolyzing triglycerides from the smaller particles (IDL as well as LDL and HDL)[73]. The end products of the delipidation cascade, LDL particles, have little triglycerides and the major lipid component is cholesterol esters. The LDL particles are internalized through the LDLR on the cell surface to peripheral

tissues and the liver, where the particles are broken down to obtain the cholesterol.

The metabolism of HDL particles is more complex than of the other lipoprotein classes involving multiple steps, and describing these processes in all detail is beyond the scope of this thesis. Therefore only the primary steps and mediators are touched upon. The pre-forms of HDL, so called pre-beta HDL particles or nascent HDL, are formed when apoA-1, the primary apolipoprotein component of HDL, first acquires phospholipids and free cholesterol from liver and peripheral tissues through ATP-binding cassette transporter A1 (ABCA1)[74]. The discoidal pre-beta HDL particle matures first to a small spherical HDL particle and subsequently to a larger HDL, when lecithin-cholesterol acyltransferase (LCAT) converts the free cholesterol in the HDL particle to cholesterol ester that is moved to the core of the particle resulting in a concentration gradient allowing further free cholesterol to flow into the particle[75]. Also, ATP-binding cassette transporters G1 (ABCG1) and G5 (ABCG5) facilitate the flow of free cholesterol and further phospholipids from peripheral tissues to the various HDL particles[76]. The cholesterol esters mainly from the mature large HDL particles are donated to liver through scavenger receptor class B member 1 (SR-BI) protein and the lipid-depleted HDL particles again enter the HDL maturation process.

HDL particles undergo constant remodelling and interact with other lipoprotein particles through the actions of lipases and enzymes like cholesterol ester transfer protein (CETP), phospholipase transfer protein (PLTP) and LCAT. CETP transfers cholesterol esters from HDL particles to TG-rich lipoproteins concomitantly transferring TG to HDL, resulting in smaller TG-enriched HDL particles. PLTP transfers phospholipids between VLDL and HDL and also between different HDL particles. It also modulates HDL particle size[77].

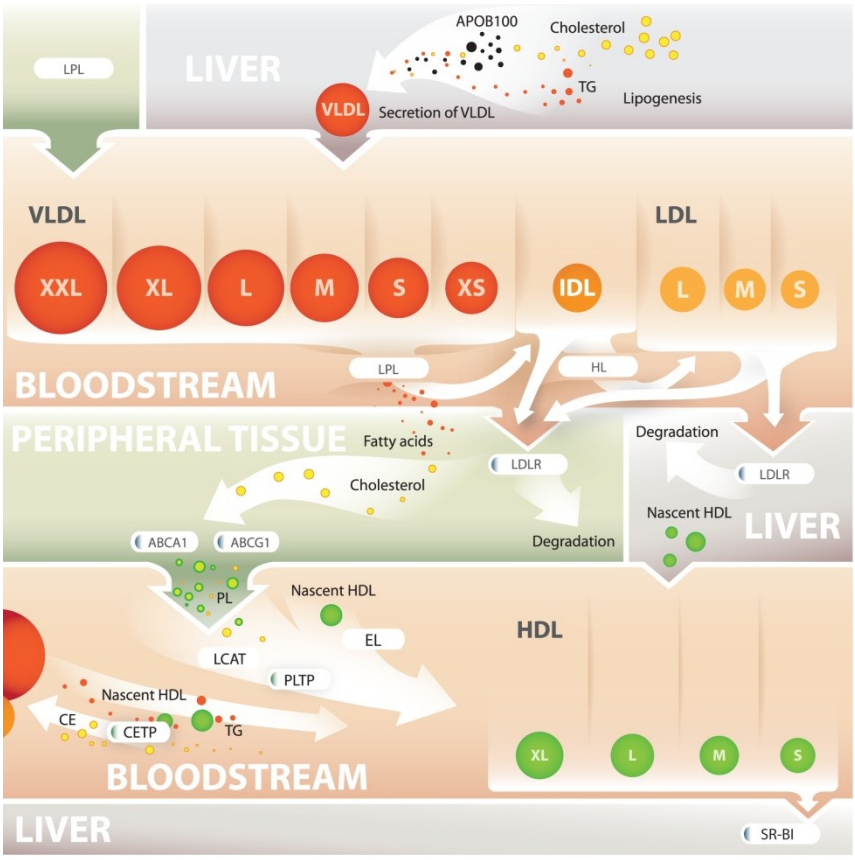

Figure 8. *A schematic illustration of the endogenous lipoprotein metabolism. The figure is modified from Figure 3 of Publication III.*

4.1.3 Lipoprotein measurements

There are various methods for measuring lipoproteins that isolate the particles based on their different properties. The methods include ultracentrifugation, precipitation, electrophoresis, chromatography, enzymatic methods and NMR. The gold standard for lipoprotein subclass isolation is ultracentrifugation that separates different classes based on their density. As many of the measurement methods are tedious (e.g., ultacentrifugation[78]), the lipoproteins are rarely characterized comprehensively for clinical purposes, but only levels of TC, TG, HDL-C and LDL-C are regularly assessed with enzymatic laboratory assays. LDL-C is occasionally determined through a mathematical formula, Friedewald formula, which estimates the amount of LDL-C from the levels of the three other lipid measures. The measurement of the lipoprotein subclasses with NMR contrasts favourably to the other available methods, as the quantification of the lipoprotein data can be achieved in a single measurement lasting less than 10 minutes[36].

4.2 Other quantified metabolites

In addition to lipoproteins, the metabolomics platform allows the quantification of a number of other metabolic intermediates. Many of these metabolites play multiple roles and are involved in various metabolic pathways, thus only the primary catabolic and anabolic processes of the metabolites are touched upon below.

4.2.1 Lipids and related metabolites

The lipid extraction procedure breaks down the lipoprotein particle structure and thus enables the detection of the aggregates of the lipid components carried in the particles. This data includes information on various types of fatty acids, for example, the total amounts of ω-3 and ω-6 FA. Two specific fatty acid species are quantified; linoleic acid (LA), an ω-6 FA essential for humans, i.e., it must be acquired through diet, that acts as a precursor in the synthesis of other ω-6 FA, and docosahexaenoic acid (DHA), an ω-3 FA, synthesized from α-linolenic acid. The metabolic pathways of ω-3 and ω-6 FA share the same enzymes, e.g., fatty acid desaturases 1 and 2.

Fatty acids are transported in the circulation in lipoprotein particles in the form of triglycerides or free fatty acids are bound to albumin, and are stored in cells as triglycerides. A triglyceride consists of three fatty acids attached to a glycerol molecule. Phosphoglycerides are structurally similar to triglycerides but one of the fatty acids is substituted by a phosphorylated (amino) alcohol, e.g., choline is the alcohol moiety of phosphatidylcholines. Sphingomyelins, another type of phospholipid quantified via the platform, have a fatty acid and phosphoryl choline attached to a sphingosine backbone.

Lipids serve as an important source of energy in human body. The first step in the catabolism of lipids is the hydrolysis of triglycerides to fatty acids and glycerol. The released fatty acids are oxidized to acetyl-coenzyme A (acetyl-CoA), a key substance in energy transfer, that enters the citric acid cycle. Acetyl-CoA can also be a substrate in the synthesis of fatty acids, lipogenesis. The freed glycerol is used in glycolysis, i.e. the breakdown of glucose, or gluconeogenesis, the regeneration of glucose, or in the synthesis of fatty acids.

4.2.2 Glycolysis, citric adid cycle and ketone body metabolites

Glucose is a central source of energy in human metabolism and is stored in cells in the form of glycogen. In glycolysis glucose is degraded in a set of reactions into pyruvate. Pyruvate is involved in various metabolic processes. In addition to being the end product of glycolysis, pyruvate can serve as a substrate for gluconeogenesis. The non-essential amino acid alanine can be generated from pyruvate and the glucogenic amino acids, including alanine, can be degraded to pyruvate. Under low-oxygen conditions, e.g., in skeletal muscle during exercise, pyruvate is converted to lactate, which is subsequently transported to liver where

it is converted to glucose. Normally, however, pyruvate is transported to mitochondria where it is oxidized to acetyl-CoA. Acetyl-CoA is also produced in the β-oxidation of fatty acids and from the breakdown of ketogenic amino acids including isoleucine, leucine, phenylalanine and tyrosine.

Acetyl-CoA enters the citric acid cycle, where it is first converted to citrate. Citrate is further converted to other citric acid cycle intermediates or transported to cytosol where it is retransformed to acetyl-CoA and further used, e.g., in fatty acid synthesis. Conditions in which gluconeogenesis is increased, for example in untreated diabetes or after prolonged reduced energy intake, the citric acid cycle slows down as the produced oxaloacetate is used for gluconeogenesis. Under these conditions acetyl-CoA is increasingly converted in hepatocytes to ketone bodies, acetone, acetoacetate and 3-hydroxybutyrate. Acetoacetate and 3-hydroxybutyrate are transported in blood to extrahepatic tissues where they are used as fuel, i.e., oxidized to acetyl-CoA.

4.2.3 Amino acids

Amino acids are building blocks of proteins but serve also as a source of energy and are substrates for, e.g., neurotransmitter synthesis. The NMR platform currently quantifies eight amino acids: the essential amino acids histidine, isoleucine, leucine, phenylalanine and valine and the non-essential amino acids alanine, glutamine and tyrosine. Alanine is synthesized from pyruvate, tyrosine from phenylalanine and glutamine from glutamate, however, the *de novo* synthesis of tyrosine and glutamine can be insufficient in young or during illness, and therefore these amino acids are conditionally essential. Tyrosine is a precursor of dopamine, norepinephrine, epinephrine and histidine the precursor of histamine.

All of the amino acids can serve as an energy source when their carbon skeletons are degraded into acyl-CoA, acyl-CoA derivatives, pyruvate or the citric acid cycle intermediates and further converted to either ketone bodies (ketogenic amino acids) or glucose (glucogenic amino acids) via gluconeogenesis. Most of the catabolism of amino acids takes place in the liver, however, isoleucine, leucine and valine, the branced-chain amino acids (BCAA), are degraded in extrahepatic tissues such as muscle and adipose tissue, where they are converted to acyl-CoA derivatives further used to produce energy. Alanine and glutamine are the two most abundant amino acids in circulation and both transport amino groups, which are produced upon amino acid breakdown and can be toxic, away from tissues for degradation as urea. Alanine has a central role in the glucose-alanine cycle, a transport system of amino groups from muscle to liver and subsequently for degradation.

4.2.4 Waste products and other small molecules

The amino groups transported to the liver are converted to urea in the urea cycle, and the urea is subsequently transported to kidney for excretion. Creatinine is formed when creatine, an energy substrate, breaks down in muscle, therefore the amount of creatinine produced depends on an individual's muscle mass. Like urea, creatinine is a waste product and excreted though the kidneys. Thus, elevated levels of these metabolites can indicate kidney dysfunction.

Other NMR measured metabolites include acetate and glycoproteins, the measure for the latter mainly comprising of α1-acid glycoprotein. Acetate is a building block for a variety of metabolites, and α1-acid glycoprotein is an acute phase protein, a marker of low-grade inflammation.

4.3 Metabolite levels in health and disease

Maintaining metabolic homeostasis, i.e., controlling the metabolite levels and fluxes, is essential for the functioning of the human body and is a complex process. One of the key players involved are enzymes that convert metabolites to others, e.g., phenylalanine hydroxylase that converts phenylalanine to tyrosine, and transporters that control the intake and excretion of metabolites in human body in order to maintain homeostasis. As the metabolic pathways are highly interconnected a dysfunction, i.e., disturbance of homeostasis, in a certain part of the metabolism may be reflected as fluctuations also in the levels of metabolites not directly involved.

Part of the enzyme and transporter activity is genetically determined; genomic variation may result, e.g., in changes in gene expression or in dysfunctional protein products and thus in altered metabolite levels. The variation can have severe consequences causing inherited metabolic disease. For example, mutation that renders phenylalanine dehydroxylase dysfunctional leads to elevated phenylalanine levels and may cause, e.g., mental retardation. These conditions are, however, rare and more common genetic variation in the same genes likely leads to smaller-scale fluctuation in the metabolite levels. For instance, genome-wide association studies have identified that common variants affecting genes encoding for an aromatic amino acid transporter and glutaminase, an enzyme that degrades glutamine to glutamate, are associated with variation in the levels of tyrosine and glutamine, respectively (Suhre et al.[14] and Publication IV). In addition, variants in several genes encoding for enzymes and transporters involved in lipoprotein metabolism have been identified to associate with blood lipid and lipoprotein levels at the population level (Teslovich et al.[79], Publications III and IV).

Metabolic homeostasis is also disturbed by various other conditions, e.g., diseases like type 2 diabetes that have a substantial environmental component in addition to the genetic preponderance. For example, the insensitivity to insulin in

type 2 diabetes leads to the dysregulation of glucose levels and can also be reflected in other metabolic parameters including the levels of the ketone bodies.

Blood levels of some of the metabolites quantified by the NMR platform are regularly used in the clinic for disease diagnosis or risk evaluation, and recent studies have suggested potential role for the metabolites as biomarkers. Table 6 summarizes the most established biomarkers among the quantified metabolites.

Table 6. *A summary of the metabolites quantified via the NMR platform that are used or suggested as markers for metabolic disorders.*

Metabolite	**Marker for**
α1-acid glycoprotein	Inflammation
Acetoacetate, 3-hydroxybutyrate	Diabetic ketoacidocis
Albumin	Fluid balance, liver and kidney function
Amino acids	Rare metabolic disorders
Glucose	Diabetes diagnosis and monitoring
Lactate	Acid-base balance, e.g., in diabetes
Lipoproteins (e.g. LDL-C and HDL-C)	Risk markers for coronary artery disease
Urea, creatinine	Kidney function
BCAAs, tyrosine, phenylalanine	Suggested markers for the risk of diabetes[5]
DHA, tyrosine, glutamine	Suggested markers for incident high intima-media thickness[80]
Sphingomyelin	Suggested marker for kidney disease in type 1 diabetes[81]

5. Genetics

Genetics is a discipline in biology studying heredity, the genetic transmission of characteristics from parents to offspring. Genetics is a broad area of science that covers studies from the molecular basis of inheritance, e.g., gene structure and function or the organization and information in the genome sequence, to the genetic differences within and between populations or between species. This chapter focuses on the structure and variation of human genome and the use of the variation in genome-wide association analyses to dissect the genetic components contributing to the variance of complex traits.

5.1 Structure and variation of human genome

Human genome consists of a double helix of deoxyribonucleic acid (DNA) where the order of the four different nucleotide bases, adenine (A), thymine (T), guanine (G) and cytosine (C), contains the "genetic code". Most of the human genome is packed into 23 chromosome pairs, 22 of which are autosomal and one sex-determining, that are stored in the nuclei of cells. Additionally, a small amount of genetic material, a circular DNA molecule, mitochondrial DNA, is stored in cell organelles called mitochondria. A chromosome consists of a long string of DNA double helix, paired strands of DNA, bound around proteins called histones.

There are approximately 3 billion base pairs in the 23 human chromosomes. Only a small minority of this sequence, 1.5%, is known to code for proteins corresponding to between 20,000 and 25,000 genes. With the enormous amount of base pairs in the human genome, it is no wonder that no two human beings are fully identical in their genetic make-up. However, all human genomes are 99.9% identical, and between any two individuals only 0.1% of the genome varies. Most of this variation has no biological effect, and therefore only a small proportion of the genome together with environmental factors contributes to the phenotypic differences between individuals. The variable sites in the genome are used in genetic mapping to identify DNA regions contributing to phenotypic differences.

There are two main mechanisms introducing variation to the human genome, recombination occurring during meiosis and mutations taking place at any time in any cell. An individual inherits half of the genetic material from the mother and a half from the father, one of each chromosome from each parent. The

chromosomes, however, do not pass on to the offspring as such, but the maternal and paternal genetic materials are shuffled during meiosis in a process called homologous recombination. Recombination events occur approximately once every one hundred million base pairs, thus, recombination can take place on average over 30 times per chromosome per meiosis.

Additional variation to the genome is introduced by spontaneous mutations that can occur due to errors in DNA replication, or can be caused, for example, by exposure to radiation or mutagens. Additionally malfunction in the recombination system can cause structural mutations. If a mutation occurs in the germline, it is passed on to the next generation. The simplest type of mutation is the substitution of a single nucleotide by another. The average mutation rate for a base is ~2.5 $\times 10^{-8}$ per meiosis[82]. In addition, parts of DNA sequence varying in length from a single nucleotide to larger chunks can be inserted (insertion), duplicated (duplication) or deleted (deletion) from a genomic sequence including the exchange of genetic material between different chromosomes (translocation) (Figure 9).

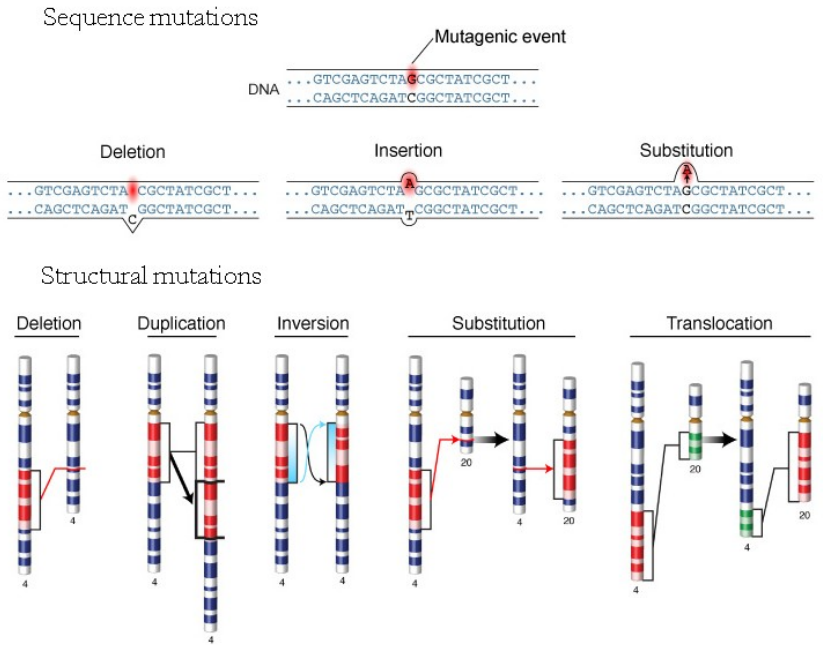

Figure 9. *Different types of mutations in human genome. Modified from* www.genome.gov/glossary.

5.1.1 Single nucleotide polymorphisms

A mutation of a single nucleotide occurring in a population is called a single nucleotide polymorphism (SNP). These genomic variants are typically found every 100 to 300 bases along the DNA sequence and make up approximately 90% of human genetic variation[83]. Due to their abundance, SNPs have become a major target in genetics research. Although the substitution of a base by another is a

minor change in the genome, and many of the SNPs likely have no biological effect, SNPs that change the amino acid sequence may have large impacts on the protein products.

The 1000 Genomes project aims at providing a comprehensive resource on the genomic variation, including SNPs that have minor allele frequencies of at least 1% in a population, by sequencing, initially 1000, individuals from various populations. The pilot phase of the study with less than 200 individuals sequenced reported approximately 15 million SNPs[83], and with the completion of sequencing of 1094 individuals over 38 million SNPs are now catalogued (June 2011 release). The next goal of the project is to sequence 2500 individuals from 25 populations.

The genotypes of SNPs can be determined individually by direct genotyping but often SNP information is collected using commercial Illumina or Affymetrix SNP arrays that nowadays provide information on up to five million variants across the genome. The population cohorts studied in Publications III and IV were genotyped using SNP arrays from Illumina that captured 370, 610 or 670 thousand variants selected to tag a large proportion of the common variation in European populations. As the number of SNPs in the genome is considerably larger, the SNP set is often augmented to include a larger number of variants by imputation (see Chapter 5.3).

5.2 Linkage disequilibrium

A variant in the parental genome may be passed on to next generations. As the number of recombination events in a generation is rather small, the offspring not only inherit the genomic variant but a long stretch of DNA around it. Over many generations several recombination events take place, the DNA gets more and more mixed and therefore the pieces of DNA sequences the individuals descended from a common ancestor share are getting shorter. However, even apparently unrelated individuals share recognizable stretches of DNA. These regions of chromosomes that have not been broken up by recombination are called haplotypes.

The non-random co-inheritance of alleles from two or more loci is called linkage disequilibrium (LD): two loci inherited together are in complete LD and typically the further apart the loci are, the weaker is the LD between them. LD occurs as haplotype blocks, i.e., regions of the genome with little recombination events (recombination coldspots) between two sites of more frequent recombination (recombination hotspots). As humans are a rather young species, most of the SNP variation in any current human population comes from the variation present in the ancestral human population, and therefore for most parts of the chromosomes only a limited number of common haplotypes exists. However, the extent of LD varies from population to population and depends on the population history[84, 85]. The range of LD is the shortest among the Africans while in Finland where the

population history includes multiple bottlenecks, subsequent isolation and rapid expansion[86] leading to limited number of ancestral haplotypes (Figure 10), the LD is among the most extensive[87].

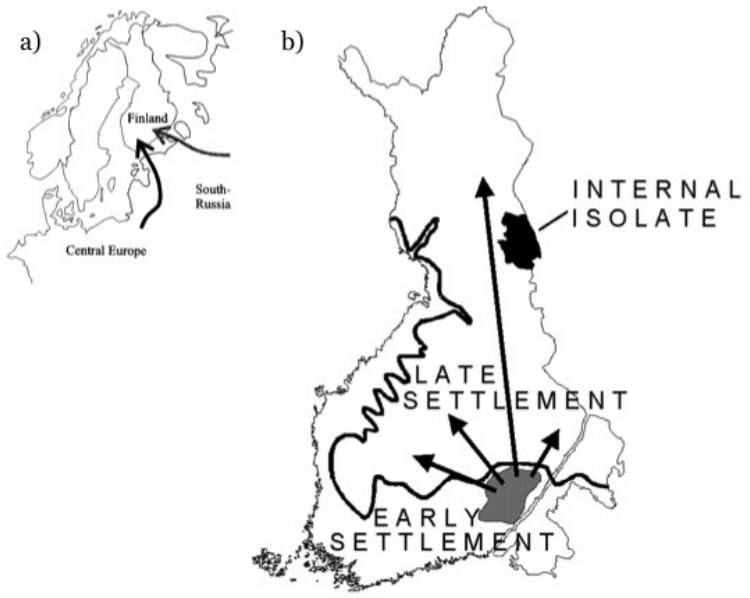

Figure 10. *The inhabitation of Finland. a) The two migratory waves to Finland, 4000 years ago from east and 2000 years ago from south. b) The habitation was largely concentrated on the coastal area (the early settlement) until the 16th century when the internal migration movement started from a small southeastern area of Finland to the uninhabited parts of the country (late settlement) and resulted in isolated rural populations, including the internal isolate of Kuusamo (shaded black), that have remained surprisingly stable over time. The population of Finland grew rapidly from the 250,000 inhabitants on the 18th century to its present 5 million. The figure is modified from Peltonen et al.*[88] *and Varilo et al.*[86].

The correlation of the alleles within a haplotype block is utilized in association studies. The variation from a SNP predisposing to a specific phenotype can be detected via another SNP in LD with the causal variant (genome-wide association studies). Also, LD can be used to predict the genotypes of SNPs that are not directly observed using the information from the genotyped variants (imputation).

In 2003 the International HapMap Consortium started building a haplotype map of the entire human genome, which shows the LD structure across chromosomes in various populations and predicts which markers are inherited together and thus facilitates the use of SNP data in genomics studies. The first HapMap catalogue contained information on roughly one million common (population frequency > 5%) SNPs[85], the following Phase II increased the number to 3.1 million[89] and Phase III catalogued further 1.6 million SNPs[90]. Now data from the 1000 Genomes project provides a similar resource to a larger number of variants[83].

5.3 Genotype imputation

SNP arrays genotype only a small set of the known variants, however due to LD the genotypes for the untyped SNPs can be predicted by utilizing the limited set of observed SNP data. Genotype imputation describes this process of filling in the missing genotypes. Successful imputation requires a detailed reference panel of haplotypes, i.e., a dense data set of genotyped markers in another study set that includes data on the markers being imputed, e.g., 1000 Genomes reference haplotypes potentially complemented with population-specific haplotype sets[91]. In addition information on the recombination rates in the genomic regions is required.

Genotype imputation has become standard practice in genetics within the last decade, since genome-wide association studies (GWASs) often require the meta-analysis of tens of cohorts that are often genotyped on different platforms. The cohorts can be combined when they all are imputed to include the same set of SNPs. Aside facilitating meta-analyses genotype imputation can aid in fine-mapping analyses by providing a denser set of SNPs[92]. Many imputation methods also allow the imputation of sporadic missing SNPs from the genotype chip data, variants that are not in the reference panel or genetic variation other than SNPs, for example copy number variations[92, 93].

Figure 11 presents a summary of the workflow in genotype imputation. A cohort of individuals has been genotyped with a SNP array that leaves a large number of variants untyped (Step 1 in Figure 11). In imputation the essential goal is to fill in this missing data. Most of the imputation algorithms first phase each individual at the observed SNPs, i.e., each genotype is resolved into its two haplotypes (combinations of alleles; paternal and maternal alleles), statistically or by using the reference set of haplotype information at the typed SNPs. The resulting phased haplotypes are then considered as mosaics of the different reference haplotypes (Step 2 in Figure 11) by looking for perfect or nearly perfect matches between the two sets. Assuming that the stretches of haplotypes that match the observed SNPs also match the untyped SNPs in the study cohort, the missing genotypes are selected from the matching mosaic haplotypes (Step 3 in Figure 11).

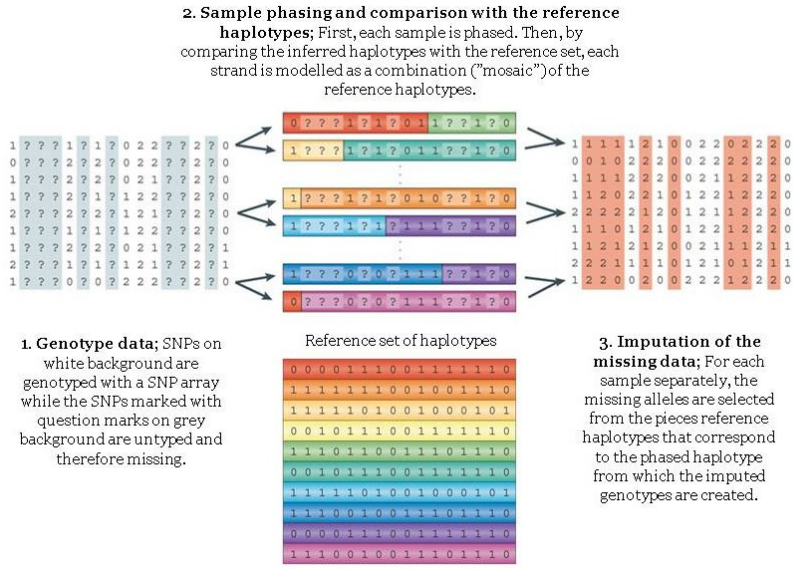

Figure 11. *Illustration of the workflow of genotype imputation. Modified from Marchini et al.*[92].

As with all prediction procedures there is uncertainty in the results of genotype imputation. The genotypes cannot be phased with 100% certainty nor is matching the pieces of reference haplotypes error-free. Therefore, the imputation algorithms, many of which are Markov chain Monte Carlo based algorithms[94], present the imputed genotypes as probability distributions, thus allowing this uncertainty to be accounted for in the subsequent analyses.

5.4 Genome-wide association analyses

Genome-wide association study (GWAS) is an analysis approach to identify the variants, typically SNPs, in human genome that associate with the trait of interest by studying a large number of genetic markers in thousands of unrelated individuals. The traits studied vary from diseases, e.g., type II diabetes or lupus, to anthropometric and biochemical traits, e.g., body-mass index (BMI) and blood lipid levels, respectively. The common denominator for the studied traits is that they are expected to be complex traits, i.e., multiple genes in addition to environmental factors contribute to trait variance, in contrast to Mendelian disorders, for which the underlying gene has often been identified previously with other means of genetic research.

The field has taken giant leaps since the publication of the first GWAS in 2005[95] that studied the association of 116,204 SNPs to age-related macular degeneration in 96 cases and 50 controls. The National Human Genome Research Institute GWAS Catalog currently (2011 2nd quarter) lists the significant results from altogether 1,449 published GWAS for 237 different traits (available at www.genome.gov/GWAStudies, accessed 6.11.2011). The low-hanging fruits, i.e.,

the variants with large effects, were quickly detected for a variety of traits (including FTO for body mass index[96]) with reasonable sample sizes. The current-day association studies of the extensively studied traits like BMI and blood lipid levels require more than 100,000 individuals to detect new variants, since the effect sizes of these SNPs are minute. Few single cohorts include such sample sizes and therefore most GWASs are multi-centre efforts.

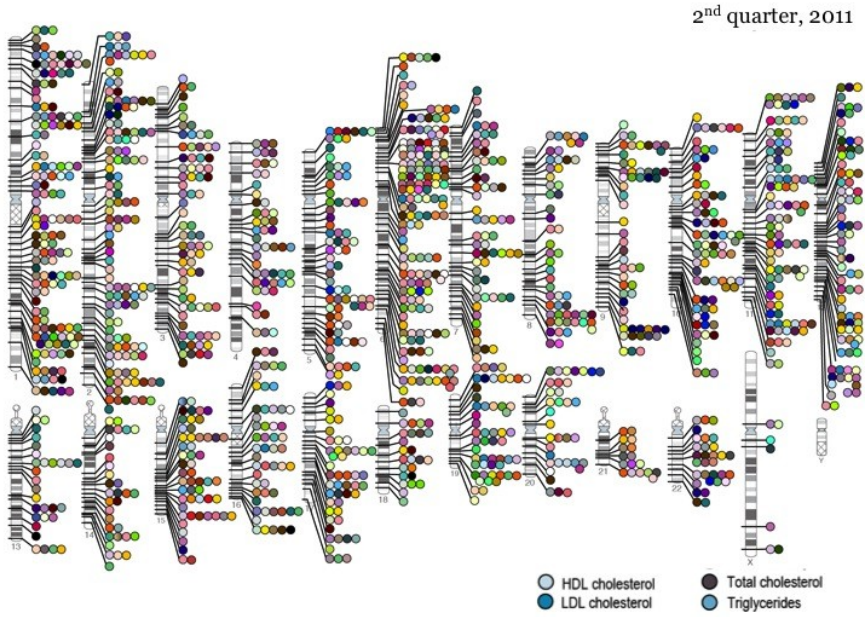

Figure 12. *Illustration of all published and catalogued genome-wide association studies (N = 1,449) and their findings by chromosome as of the June 2011. The circles show the positions of the significant ($P \leq 5 \times 10^{-8}$) associations in the genome and the different colors illustrate the various traits (N = 237) studied. Key for the four blood lipids (TC, LDL-C, HDL-C and TG) is given. The figure is modified from http://www.genome.gov/GWAStudies/.*

The association between a SNP and the trait is most often tested with simple regression, depending on the type of trait studied, either with logistic (for dichotomous traits) or linear (quantitative traits) regression. Regression is a simplistic approach but considering the amount of SNPs tested in thousands of individuals, the computational burden is considerable. The traits studied are often adjusted with a few major covariates that significantly correlate with the trait, e.g., gender and age. Additionally, it is often necessary to adjust for population stratification, i.e., the systematic differences in allele frequencies between individuals from different geographical origins. As the current GWASs include multiple cohorts that are analyzed separately, the results are combined in a meta-analysis. Testing more than a million variants in a GWAS causes a multiple-testing issue, thus a stringent *P*-value threshold needs to be used to rule out false positives. The widely accepted significance threshold, often referred to as genome-wide significance, is 5×10^{-8}. This threshold is derived based on the

assumption that there are roughly a million independent variants in the human genome[97].

5.4.1 GWAS of blood lipids

As abnormal levels of the blood lipid parameters of TC, TG, LDL-C and HDL-C are major risk factors for coronary artery disease and the measurements are readily available for many cohorts, these lipids are among the most extensively studied traits in GWASs. Tens of studies have investigated the genetic variants contributing to the variance of the four blood lipids with sample sizes of the studies increasing year by year[79, 98-119]. The latest published GWAS of the four lipids, by Teslovich et al., which included over 100,000 individuals from 46 population-based cohorts, identified 95 loci associated with the traits[79]. Fifty-nine of the reported loci were novel and the 36 other loci included all the loci identified in previous GWASs of the same traits. Table 7 summarizes the results from this study.

Many of the loci identified by Teslovich et al. harbor genes that have previously been linked to Mendelian lipid disorders, are targets for hyperlipidemia drugs, have a known function in lipid metabolism or are good functional candidates, including *APOB* (encoding for apolipoprotein B), *LIPC* (hepatic lipase) and *PLTP* (phospholipid transfer protein). However, many of the loci have not been previously implicated in lipoprotein metabolism. Further studies are required to identify the causal variants and affected genes and to elucidate the role and relevance of these genes for human metabolism. For example, an extensive functional study using human cohorts, hepatocytes and knockout mice was carried out to identify *SORT1* as the causative gene and the related pathway underlying the association of a SNP in chromosome 1p13 with both LDL-C and myocardial infarction[120].

Table 7. *The 95 loci identified in the GWAS of Teslovich et al. to associate with one or more of the conventional blood lipids.*

Locus	Lead trait	Effect size	*P*-value	Locus	Lead trait	Effect size	*P*-value
LDLRAP1	TC	−1.22	4×10^{-11}	*CYP26A1*	TG	−2.28	2×10^{-8}
PABPC4	HDL	−0.48	4×10^{-10}	*GPAM*	TC	1.14	2×10^{-10}
PCSK9	LDL	2.01	2×10^{-28}	*AMPD3*	HDL	−0.41	5×10^{-8}
ANGPTL3	TG	−4.94	9×10^{-43}	*SPTY2D1*	TC	−1.04	3×10^{-8}
EVI5	TC	−1.18	3×10^{-8}	*LRP4*	HDL	0.78	3×10^{-18}
SORT1	LDL	−5.65	1×10^{-170}	*FADS1-2-3*	TG	3.82	5×10^{-24}
ZNF648	HDL	−0.47	3×10^{-10}	*APOA1*	TG	16.95	7×10^{-240}
MOSC1	TC	−1.39	6×10^{-13}	*UBASH3B*	TC	0.97	2×10^{-10}
GALNT2	HDL	−0.61	4×10^{-21}	*ST3GAL4*	LDL	1.95	1×10^{-15}
IRF2BP2	TC	−1.36	5×10^{-14}	*PDE3A*	HDL	0.4	4×10^{-8}
APOB	LDL	4.05	4×10^{-114}	*LRP1*	TG	−2.70	4×10^{-10}
GCKR	TG	8.76	6×10^{-133}	*MVK*	HDL	−0.44	7×10^{-15}
ABCG5/8	LDL	2.75	2×10^{-47}	*BRAP*	TC	−0.96	7×10^{-12}
RAB3GAP1	TC	1.25	2×10^{-8}	*HNF1A*	TC	1.42	1×10^{-14}
COBLL1	TG	−2.01	2×10^{-10}	*SBNO1*	HDL	0.86	7×10^{-9}
IRS1	HDL	0.46	3×10^{-9}	*ZNF664*	HDL	0.44	3×10^{-10}
RAF1	TC	−1.42	4×10^{-9}	*SCARB1*	HDL	0.61	3×10^{-14}
MSL2L1	TG	−2.22	3×10^{-8}	*NYNRIN*	LDL	1.14	5×10^{-11}
KLHL8	TG	−2.25	9×10^{-12}	*CAPN3*	TG	7	2×10^{-8}
SLC39A8	HDL	−0.84	7×10^{-11}	*FRMD5*	TG	5.13	2×10^{-11}
ARL15	HDL	−0.49	5×10^{-8}	*LIPC*	HDL	1.45	3×10^{-96}
MAP3K1	TG	2.57	1×10^{-10}	*LACTB*	HDL	−0.39	9×10^{-9}
HMGCR	TC	2.84	9×10^{-47}	*CTF1*	TG	−2.13	3×10^{-8}
TIMD4	TC	−1.98	7×10^{-28}	*CETP*	HDL	3.39	7×10^{-380}
MYLIP	LDL	−1.43	1×10^{-11}	*LCAT*	HDL	1.27	8×10^{-33}
HFE	LDL	−2.22	6×10^{-10}	*HPR*	TC	2.34	3×10^{-24}
HLA	TC	2.31	4×10^{-19}	*CMIP*	HDL	−0.45	2×10^{-11}
C6orf106	TC	−1.86	5×10^{-11}	*STARD3*	HDL	−0.48	1×10^{-13}
FRK	TC	−1.18	2×10^{-10}	*OSBPL7*	LDL	0.78	2×10^{-8}
CITED2	HDL	−0.39	3×10^{-8}	*ABCA8*	HDL	−0.42	2×10^{-10}
LPA	LDL	−0.56	2×10^{-17}	*PGS1*	HDL	−0.39	8×10^{-9}
DNAH11	TC	1.43	9×10^{-10}	*LIPG*	HDL	−1.31	3×10^{-49}
NPC1L1	TC	2.01	3×10^{-11}	*MC4R*	HDL	−0.42	7×10^{-9}
TYW1B	TG	−7.91	1×10^{-9}	*ANGPTL4*	HDL	−0.45	3×10^{-8}
MLXIPL	TG	−9.32	6×10^{-58}	*LDLR*	LDL	−6.99	4×10^{-117}
KLF14	HDL	0.59	1×10^{-15}	*LOC55908*	HDL	−0.64	3×10^{-9}
PPP1R3B	HDL	−1.21	6×10^{-25}	*CILP2*	TC	−4.74	3×10^{-38}
PINX1	TG	2.01	1×10^{-8}	*APOE*	LDL	7.14	9×10^{-147}
NAT2	TG	2.85	5×10^{-14}	*FLJ36070*	TC	1.27	2×10^{-10}
LPL	TG	−13.64	2×10^{-115}	*LILRA3*	HDL	0.83	4×10^{-16}
CYP7A1	TC	1.23	2×10^{-12}	*ERGIC3*	TC	−1.19	4×10^{-10}
TRPS1	HDL	−0.44	6×10^{-11}	*MAFB*	TC	−1.38	6×10^{-11}
TRIB1	TG	−5.64	3×10^{-55}	*TOP1*	LDL	1.39	4×10^{-19}
PLEC1	LDL	1.4	4×10^{-13}	*HNF4A*	HDL	−1.88	1×10^{-15}
TTC39B	HDL	−0.65	3×10^{-12}	*PLTP*	HDL	−0.93	2×10^{-22}
ABCA1	HDL	−0.94	2×10^{-33}	*UBE2L3*	HDL	−0.46	1×10^{-8}

ABO	LDL	2.24	6×10^{-13}	*PLA2G6*	TG	−1.54	4×10^{-8}
JMJD1C	TG	−2.38	3×10^{-12}				

Locus, the candidate gene or the nearest gene of the associated SNP as reported by Teslovich et al.[79]; Lead trait, the most associated trait; Effect size, the effect size of the association with respect to the minor allele in units of mg/dl.

In addition to the above studies, Chasman et al. performed a GWAS on 17 NMR-derived lipoprotein measures (the concentrations of eight lipoprotein subclasses, the total particles for four lipoprotein classes, mean particle sizes for three lipoprotein classes, and the estimates for HDL-C and TG) and five conventional lipoprotein or apolipoprotein measures[33]. In total 31 loci were associated with one or more of the 22 lipoprotein measures in the primary association analysis. At the time the study was published seven of the associated loci were novel. Four loci were later discovered in the GWAS of total lipid measures by Teslovich et al.[79], however with a significantly larger study sample, thus pointing to the reduced biological variance resulting from the use of detailed lipoprotein measures. Table 3 summarizes the three novel loci that all associated to HDL measures. Further support for the use of NMR-based lipoprotein phenotyping was provided in the study from Kaess et al., where the authors identified further and/or stronger associations when HDL particle size was included in the analyses in addition to HDL-C[121]. In addition, in a recent paper by Petersen et al.[122] the associations of the 95 known lipid loci to 15 NMR-derived lipoprotein subclasses were investigated, and the authors found that the associations to the subclass measures strengthened the associations compared to conventional lipids.

Table 8. *The three novel loci found by Chasman et al.[33] that have not been identified in GWASs of the main lipid fractions.*

Locus	Chr	SNP	Trait	Alleles/MAF	Effect size	*P*-value
PCCB, STAG1	3	rs3856637	HDL small	G/A/0.28	0.37	1×10^{-8}
ASCL1, PAH	12	rs10778213	HDL-C by NMR	G/A/0.47	-0.79	2×10^{-8}
		rs1818702	HDL total	A/G/0.29	-0.42	9×10^{-10}
PRKAR1A, WIPI1	17	rs2909207	HDL medium	A/G/0.22	0.09	1×10^{-8}

Locus, the candidate gene or nearest gene of the associated SNP as reported by Chasman et al.; Chr, chromosome; Trait, the associated trait; Alleles/MAF, the minor and major alleles and frequency of the minor allele; Effect size, the effect size of the association with respect to the minor allele.

5.4.2 GWAS and metabolomics

Metabolomics data for samples large enough to conduct genome-wide analyses with are just becoming available and only a handful of studies assessing the genetic variants associated with the metabolite traits from serum or urine measured with either MS or NMR have been conducted thus far. Three genome-wide association studies have focused the analyses on MS-based serum metabolites[14-16, 46], while two other studies investigated the NMR-measured metabolites from urine and/or plasma[17, 18]. The top panel in Table 9 summarizes

these studies. All but one of the studies have also analysed ratios of metabolites motivated by the findings that these serve as proxies for enzymatic activity and thus may increase the strength of association[16, 123]. Additionally, one study has investigated the genetic components of more targeted sets of metabolites, i.e., sphingolipids[124] (middle section of Table 9).

Together the studies have identified more than 40 loci that associate with metabolite levels or ratios of metabolites. For many loci there are plausible candidate genes, which have functions that match to the metabolites associated with the variants in the locus. For example, several genes known to function in fatty acid beta-oxidation are associated with the levels of carnitines[15], essential components for lipid metabolism. Additionally, in the cases where the nearby genes have no known role in metabolism, the associated metabolite traits may provide further insight into the functions of genes in the region. As an example, inspired by the association of a locus with serum levels of carnitine, Suhre et al experimentally validated the gene in the locus, *SLC16A9*, to function as a carnitine efflux transporter[14]. Many of the identified loci have been previously linked to clinical outcomes, including *NAT2* to coronary artery disease and *GCKR* to diabetes. Thus, the metabolite associations uncovered for these loci may provide bases for new pathophysiological hypotheses and help to identify new pharmacological targets.

Although the sample sizes in the studies are considerably smaller, only up to 3000 individuals for the untargeted metabolomics, than the current-day association studies of complex traits that involve over hundred thousand individuals, a wealth of genetic loci have been identified. Many of the traits have not been studied before, which may explain the discovery rate to some extent. However, this is also likely due to the metabolomics traits being closer to the actual pathways the genes act on, thus the strengths of association are larger than for the clinical phenotypes. This assumption is supported by the large effect size most of the loci show; while a single SNP typically explains less than a percent of the variation of a clinical phenotype, such as TC, the proportions explained of variance of the metabolite levels are up to 30%.

Many of the metabolomics-GWA studies have also investigated the associations of the variants to ratios of the metabolites. This approach is justified as many loci show significantly stronger associations to the ratios. In these cases the associated variants may directly affect the enzymes converting one metabolite to another and therefore using a ratio of the metabolites reduces the biological variance.

Testing a large number of traits simultaneously - including the metabolite ratios the number of traits tested is more than 30,000 - causes a multiple testing issue. Thus, most studies have applied a stringent *P*-value threshold to avoid false positive findings. Additionally, as sample sizes have been limited, some studies have focused the analyses on variants that have minor allele frequencies of at least 10% to avoid false findings.

Table 9. *The published genome-wide association studies that have used metabolomics data. The top panel lists the studies that have applied untargeted metabolomics analyses and the middle panel the study that has targeted the analyses on lipid species. In the bottom panel details of a study that investigated sexual dimorphisms in the metabolite associations is given.*

Study, year	Sample	Metabolomics platform	N	Met	SNPs	Loci
Untargeted metabolomics						
Suhre et al.[14], 2011	Serum	2 separate UHPLC/MS/MS2 injections, 1 GC/MS injection + Metabolon library	1,768 (KORA F4) + 1,052 (Twins UK)	295 + > 37,000 ratios	655,658 / 534,665	37
Illig et al.[15], 2010	Serum	ESI-MS/MS, Biocrates AbsoluteIDQ	1,809 (First stage: 1,029 (KORA F4); Second stage: 780 (KORA F4)); Replication: 422 (Twins UK)	163 + 26,406 ratios	517,480	9
Gieger et al.[16], 2008	Serum	ESI-MS/MS at Biocrates Life Sciences AG	284 (KORA F3)	363 + 201 ratios	187,454	3
Nicholson et al. [18], 2011	Plasma and urine	(FIA-MS Biocrates +) in-house NMR	142 (MolTWIN); Replication: 69 (MolOBB)	526 peaks from NMR	2,541,644	3
Suhre et al.[17], 2010	Urine	NMR 400 MHz, Chenomx NMR Suite 6.1	862 (SHIP-0 discovery); Replication: 870 (SHIP-0 females) + 992 (KORA F4); Verification: 170 (SHIP-1 verification)	59 + 1661 ratios	645,249	5
Targeted metabolomics						
Hicks et al.[124], 2009	Plasma	ESI-MS/MS	4,400 from four cohorts (ERF, MICROS, NSPHS, ORCADES, VIS)	33 lipids + 43 matched ratios	318,237	5
Sex-specific analyses						
Mittelstrass et al.[9], 2011	Serum	ESI-MS/MS, Biocrates AbsoluteIDQ	3,061 (1452 males, 1552 females, KORA F4) + 377 (197 males, 180 females, KORA F3)	131	651,596	1

Sample, the biofluid used in the metabolomics analyses; N, the number of individuals in the study (cohort is given in parenthesis); Met, the number (and type) of metabolites studied; SNPs, the number of SNPs studied; Loci, the number of significantly associated loci.

Although gender and age are shown to have an effect on the metabolite profiles of urine[11, 12] and serum/plasma[10, 21, 125], only two (Suhre et al.[14] and Nicholson et al.[18]) of the GWASs on the untargeted metabolomics traits have adjusted the analyses for either or both of these confounding factors. Mittelstrass et al. comprehensively analysed the gender differences in MS-measured metabolites and identified significant effects for 78% of the studied traits[9]. In the same study sex-specific genome-wide analysis was performed and one locus was found to have significantly different effects between the genders (bottom panel in Table 9).

In addition to these studies where metabolomics data has been applied in the discovery of new loci, some studies have used NMR metabolomics measurements to provide further insight into the biology underlying the loci identified to associate with other metabolic parameters, i.e., liver enzymes[58] and blood pressure[59].

5.5 Heritability

In order for a trait to be genetically determined, it has to be heritable, i.e., genetic factors in addition to environmental influences contribute to the trait variance. The observed variance in a trait (V_T) can be decomposed to variance from both genetic (V_G) and environmental factors (V_E) :

$$V_T = V_G + V_E$$

Heritability determines the proportion of the trait variance that is due to genetic variation:

$$H^2 = \frac{V_G}{V_T}$$

This is the definition of broad sense heritability that takes into account the variance of both additive (V_A) genetic effects, reflecting the effects of individual alleles, and dominant (V_D) genetic effects, reflecting the allelic interactions, which compose the total genetic variance. The narrow sense heritability determines the proportion the additive genetic variance from the trait variance:

$$h^2 = \frac{V_A}{V_T}$$

Similarly the environmental component of variance can be further decomposed to contributing factors, for example, to environmental variance common to siblings and that arising from unique environmental influences.

Heritability can be estimated by comparing the observed and expected resemblance between relatives, e.g., families or twins. A common way to assess the heritability is to study monozygotic (MZ) and dizygotic (DZ) twin pairs, who share 100% or 50% of their genomes, respectively. A crude estimate for heritability can be calculated from the difference in the intraclass correlation (ICC) between the MZ and DZ twins:

$$h^2 = 2(ICC_{MZ} - ICC_{DZ})$$

Twin studies, however, often apply more detailed models for estimating the heritability. These models take into account the different combinations of genetic and environmental sources of variation, e.g., additive and dominant genetic components and shared and unique environmental influences.

5.5.1 Heritability estimates of blood lipids

In family studies the heritability estimates for blood lipids have been shown to range from 0.39 to 0.62 for TC, from 0.35 to 0.83 for HDL-C, from 0.24 to 0.50 for LDL-C and from 0.20 to 0.55 for TG[20, 126]. Some twin studies have reported somewhat higher estimates (up to 0.81 for TC, 0.76 for HDL-C, 0.79 LDL-C and 0.75 for TG)[127, 128] but these studies have included data sets of young twins that have less environmental variance and thus the estimates should be evaluated in that context. Also, gender and race differences have been reported[128, 129]. Few studies have assessed the heritabilities of lipoprotein subclasses. The heritability estimates for LDL particle size have varied from 0.26 to 0.60[127, 130, 131] and for HDL from 0.25 to 0.56[127, 132]. The heritability estimates for five HDL subclasses in a Finnish twin sample ranged from 0.46 to 0.63[127].

5.5.2 Heritability of metabolomics measures

The large number of loci identified for the blood and urine metabolites pinpoints the genetic components underlying the metabolite level regulation. However, thus far, few studies have assessed the extent of the inherited proportion in the variation in metabolites.

Shah et al. studied the heritabilities of metabolites measured with MS in comparison to the heritabilities of conventional metabolites linked to cardiovascular risk in 117 individuals from eight families burdened with premature coronary artery disease[20]. Many of the metabolites showed higher degree of heritability than the conventional risk factors; High heritabilities were identified for several amino acids, arginine showing an exceptionally high heritability (h^2=0.80), some free fatty acids and acylcarnitine species. However, as the study sample was limited these heritability estimates had large standard errors.

Nicholson et al. used a longitudinal twin study design to decompose the variation in plasma and urine NMR spectra peaks into familial, individual-environmental and longitudinally unstable components[19]. Due to the small sample of only 77 twin pairs, the group could not assess the heritability, but

determined the familiality that combines the genetic and common environmental variation. Peaks in NMR spectra of plasma showed a greater degree of familiality, and less between-visit variability compared to urine NMR. However, for both biofluids much of the variation in the peaks was due to the stable, i.e., genetic and environmental, components, and less due to short-term fluctuations in the metabolite profile, thus highlighting the potential of NMR-based metabolites as biomarkers. The highest familialities were identified for plasma creatinine (77%) and urine trimethylamine (92%), and the findings were mostly consistent with the heritability estimates from the previous study by Shah et al.

5.6 Mapping gene expression

The genetic markers identified in association studies rarely provide much detail of the underlying genes and mechanisms. Further understanding on how the associated loci contribute to the trait variance can be gained by studying the associations of the SNPs to the variation in gene expression. The loci that regulate gene expression are called expression quantitative trait loci (eQTL).

The genetic markers identified in association studies often map in loci where there is no evident functional candidate gene, e.g., there may be several genes in the region or the associated locus may be in a gene desert. Therefore GWASs are often complemented with an eQTL analysis to further interpret the results. For example, the largest GWAS on blood lipids up to date investigated the correlations of the lead SNPs of the 95 associated loci and the gene transcripts from liver, omental fat and subcutaneous fat located within 500 kb of the variant, i.e., *cis*-eQTL, and identified significant eQTLs for 32 loci. Some of the associated transcripts can be remote from the SNPs. E.g., a variant that correlates with the expression of *PPP1R3B* lies nearly 200 kb from the gene[79]. In addition to identifying functional candidates, the gene expression data can be used to provide insight into which of the associated SNPs in the region are more likely to be tagging the causative variant.

Gene expression varies considerably between tissues, and thus the identified eQTLs depend on the choice of tissue for the study. For example, 30 % of eQTLs are shared among lymphoblastoid cell line, skin and fat tissues, and a high proportion of these show significant differences in the magnitudes of effects between the different tissues[133]. Ideally transcripts from multiple tissues would be included in the eQTL analyses, however, well-characterized data sets of hundreds of individuals with both genome-wide SNP data and genome-wide transcripts from several tissues are rare. For eQTL studies of variants associated with lipid-related traits, transcripts from liver seem the ideal target, as liver is a major player in lipid metabolism.

6. Materials and methods

The metabolomics platform presented in Chapter 3 provided a foundation for the two other studies presented in thesis, i.e., Publications III and IV. However, in addition to the means to assess the metabolites, the thorough investigation of the genetic components underlying the metabolite levels performed in these studies required, for example, study cohorts, genotyping, imputation and a number of other analysis methods. This chapter presents the materials and methods used in Publications III and IV.

6.1 Study subjects

More detailed descriptions of the cohorts can be found in the original publications (III and IV) and references therein. All studies were approved by local ethics committees and participants provided informed consent.

6.1.1 The Northern Finland Birth Cohort 1966 (III, IV)

The Northern Finland Birth Cohort 1966 (NFBC1966) is a longitudinal birth cohort following mothers and their children born in 1966 in the Oulu and Lapland provinces of Finland and comprises of 12068 deliveries and 12231 children.[134] The data collection began prenatally and the offspring were followed-up at the ages of 6 months, 14 years and at the age of 31. 6007 cohort members still living in Northern Finland or in Helsinki region attended the latest assessment. The participants went through a medical examination and provided fasting blood samples that were used for metabolomics experiments and DNA extraction.

6.1.2 The Cardiovascular Risk in Young Finns Study (III, IV)

The Cardiovascular Risk in Young Finns Study (YF) is an ongoing multi-centre follow-up study initiated in 1980 to study the cardiovascular risk from childhood to adulthood.[135] 3596 children and adolescents from five Finnish university cities and their rural surroundings aged 3, 6, 9, 12, 15 and 18 years attended the baseline examination. The cohort was followed-up at three year intervals between 1980 and 1992, and in 2001 and 2007. Blood samples for DNA extraction were donated at the 2001 follow-up when the participants were 24–39 years of age. The

fasting blood samples for the metabolomics measurements used in the Publications III and IV were taken at the latest, 27-year assessment, the participants being 30–45 years at the time.

6.1.3 Helsinki Birth Cohort Study (III, IV)

The Helsinki Birth Cohort Study (HBCS) comprises of 8760 men and women born 1934-1944 in Helsinki, Finland.[136] Growth, socioeconomic aspects and general health data of the participants has been abstracted from birth records, child welfare clinic and school health records and linked to national health care registers. A subset of 2500 randomly selected individuals from the cohorts took part in a clinical examination and gave fasting blood samples between 2001 and 2004.

6.1.4 Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (III, IV)

The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study was collected in 2007 as an extension to the National FINRISK Study 2007 survey. The DILGOM sample includes 5025 Finnish individuals 25–74 years of age, who participated in a thorough clinical examination including measurements for fasting glucose, answered a dietary questionnaire and provided fasting blood samples.

6.1.5 The Health 2000 GenMets sample (III, IV)

The Health 2000 GenMets (H2000) sample is a subset of 2212 Finnish individuals including metabolic syndrome cases and their matched controls from the Health 2000 survey collected in 2000.[137] Participants provided fasting blood samples, underwent a health examination and answered questions concerning, e.g., their health status, living conditions and employment.

6.1.6 Finnish twin registry (IV)

The twin sample used for heritability estimates in Publication IV is a subset of 507 monozygotic and 826 dizygotic twin brothers and sisters who participated either the FinnTwin-12 (FT12, http://wiki.helsinki.fi/display/twineng/Finntwin12) or FinnTwin-16 (FT16, http://wiki.helsinki.fi/display/twineng/Finntwin16) cohort study. The cohorts are population-based longitudinal studies following twins born in 1983-1987 (FT12) and 1975-1979 (FT16). The twins of the younger study were contacted the year they turned 11 and were followed-up at the ages of 14, 17 and ~22. The participants of the older cohort were initially approached in the 1-2 months following the twins' 16th birthday and follow-up assessments were made when the twins were 17, 18.5 and ~25 years of age. Each visit included comprehensive survey of health, personality and social relationships. Fasting blood samples for DNA extraction and metabolomics experiments were taken at the last assessment.

6.2 Genotypes and imputation

The cohorts were genotyped using commercial Illumina HumanHap SNP arrays: 370k array for NFBC1966, 610k for DILGOM and H2000 and custom generated 670k array for YF and HBCS. Also, a subset of the twins was genotyped using HumanHap 670k array. Quality control was performed for each study separately using the following criteria: DNA samples and markers that had genotype failures in > 5% of samples or markers, respectively, were removed. In addition, if the data indicated excessive genome-wide heterozygosity (indicating sample contamination) or gender discrepancies, these individuals were removed as well as were closely related individuals.

The cleaned genotypes were augmented by imputation using IMPUTE software[93] and a reference panel that included the 1000 Genomes reference (low-coverage pilot release from March 2010), HapMap3 reference (release #2 from Feb 2009) which further included an additional Finnish imputation reference in HapMap3 depth. After imputation, the SNP set included in total 7.7 million genotyped or imputed polymorphic markers.

In order to assess the quality of the imputation to the 1000 Genomes reference, the imputed genotypes for 316 markers that showed genome wide significance in Publication IV were compared with directly genotyped SNPs from Cardiometabochip, which was available for the DILGOM study sample (Figure 13). As the concordance of the genotypes between the imputed and genotyped SNPs was high (94% of the SNPs had $r^2 > 0.8$) the imputation was accurate for the reported SNPs. Imputation accuracy also seemed not the vary depending on the allele frequency (1% < MAF < 50%).

6.3 Gene expression profiling

Leukocyte gene expression data was collected for a subset of the DILGOM cohort (N = 585). PAXgene Blood RNA System (PreAnalytiX GMbH, Hombrechtikon, Switzerland) was used to obtain stabilized total RNA using the protocol as recommended by the manufacturer. 750 ng of biotinylated cRNA produced from total RNA with Ambion Illumina TotalPrep RNA Amplification Kit (Applied Biosystems, Foster City, CA, USA) was hybridized onto Illumina HumanHT-12 Expression BeadChips (Illumina Inc., San Diego, CA, USA), using a standard protocol.

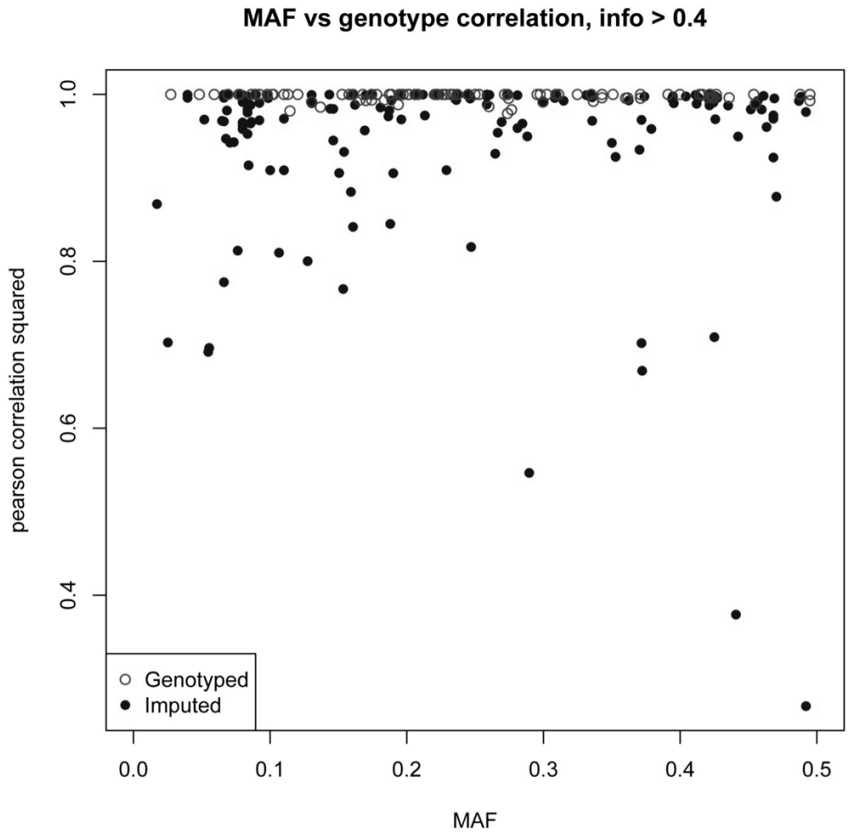

Figure 13. *The correlation between the genotypes of 316 SNPs imputed from the 1000 Genomes reference panel and genotyped with Cardiometabochip as a function of the minor allele frequency (MAF). The figure is from the Supplementary Material of Publication IV.*

6.4 Metabolite and enzymatic lipid measurements

The serum ^{1}H NMR metabolomics platform presented in Chapter 3 was applied to measure 117 metabolites from fasting serum samples including 80 lipoprotein measures, 15 lipid measures and 22 small molecules from the six cohorts. The measurements were targeted to three molecular windows, LIPO, LMWM and LIPID. The details on the platform and measurements can be found in Chapter 3.

The blood levels of TC, TG, LDL-C and HDL-C were also determined for the individuals in each cohort using standard enzymatic assays.

6.4.1 Metabolite transformations and corrections

Before applying any corrections or transformations to the metabolite traits individuals who were pregnant or using lipid-lowering medication were excluded. Also those individuals who had not fasted before blood sampling were removed from further analyses, although, the metabolite quantification protocol automatically excludes samples that fall too far from the training set maximum or minimum (sample is excluded if the whole quantifiable area in the spectrum is

10% or a single spectral point 10% in LIPO or 40% in LIPID and LMWM windows below or above the maximum and minimum values, respectively) as outliers. The phenotypes were subsequently corrected for gender, age (not NFBC1966 where all the individuals were of the same age) and ten first principal components to account for the population substructure[138, 139]. The covariate adjustment was performed in R using regression. Inspection of the raw metabolite traits revealed skewness in the distributions. Therefore, to ensure normal distribution, the residuals from the regression were normalized to mean 0 and standard deviation 1 using inverse normal transformation.

In addition to the 117 directly measured metabolites additional 99 targeted ratios of the metabolites or other derived measures were calculated. The number of ratios was limited to a predefined set of biologically interesting ones to avoid exhaustive computational load. The metabolite ratios were calculated from the unadjusted metabolite data. Extreme outliers, i.e., values more than ±4 standard deviations from the phenotype mean, were excluded. The raw ratios were then corrected and transformed as above.

6.5 Association testing

6.5.1 Association analyses

To test the associations between the SNPs and the metabolite traits the phenotypes were correlated against the genotype data assuming an additive genetic model:

$$y = \alpha + x\beta + \epsilon$$

where y is the trait, α a constant, x the genotypes of the SNP {0,1,2}, β the regression coefficient, i.e., the effect size, and ϵ is the prediction error of the model. The model assumptions are the following: 1) the relationship of the SNP and the mean value of the trait is linear, i.e., each copy of the variant allele increases the mean trait value with a constant amount, 2) the trait is normally distributed, and 3) the error terms are normally distributed and independent of each other and the values of x.

The association analyses were performed for each cohort separately using SNPTEST software[94] (version 2.1.1) or R[140].

6.5.2 Meta-analysis of the cohorts

The cohorts were combined in a fixed-effects inverse variance meta-analysis using META software[141] or GWAMA software[142]. SNPs had to pass the following criteria to be included in further inspection: the SNP had to have a result in all five cohorts, it had to be imputed with good quality (imputation info > 0.4) and no heterogeneity in the effect sizes between cohorts was allowed (*P*-value for Q statistics < 1×10^{-5}).

Genomic inflation factors, lambdas, illustrate the deviation of the *P*-value distribution of a phenotype from the expected distribution, uniform distribution [0,1]. A value over 1.05 usually indicates population stratification. In Publication IV the lambdas for the 216 metabolite traits were between 0.99 (for the ratio of glucose and pyruvate) and 1.06 (for the ratio of alanine and citrate). Although the lambda values gave no convincing evidence of population stratification, the analyses were further corrected with these values. Also, a stringent *P*-value threshold was adopted, 2.31×10^{-10}, which is the genome-wide significance level corrected for the 216 traits tested.

6.5.3 Conditional association analyses

Conditional association analyses were performed in Publications III and IV to identify further independent signals (III, IV) or to confirm the independence of the associated SNPs from previously reported signals (IV) by using genotypes of the SNPs as covariates. In Publication III the association of the variants in each locus was conditioned first on the genotype of the previously reported lead variant of the locus and then, if significant associations ($P < 5 \times 10^{-8}$), remained recursively adding the most significantly associated SNP to the covariates. In Publication IV the analyses were conditioned on the identified lead variant of each locus and the significance level applied was $P < 2.31 \times 10^{-10}$. The conditional association analyses were performed for all metabolite traits in analysis windows of 1Mb flanking regions of the previously reported (III) or identified (IV) lead variant and the independent variants if these were identified. In Publication IV, further conditional analyses were performed in two loci to verify the independence of the observed association from a previously reported association signal from a nearby locus but further than 1Mb away by conditioning the analyses on the reported lead variant the locus.

The conditional association analyses were performed for each cohort separately using linear regression in the SNPTEST software[94] (version 2.2.0) and the cohorts were combined in a fixed-effects meta-analysis using METAL[141] as described above.

6.5.4 Proportion of variance explained

In Publication IV the proportion of trait variance explained by the associated SNPs was assessed in the twin sample, which was independent of the discovery cohorts. As part of the twin sample was stratified by alcohol consumption, one pair of these twins was randomly chosen for the analyses, thus resulting in a random population sample of 436 individuals. Gene scores were built from the dosages of the 33 significantly associated SNPs so that for each trait all the SNPs that showed a nominal genome-wide significant association with the trait in the meta-analysis were included in the gene score.

6.5.5 *Cis*-eQTL analysis

The SNPs, which had significant association with one or more of the metabolite traits in Publication III or IV, were correlated against the leukocyte gene expression in a subset of the DILGOM cohort (N = 585). All expression probes within 1 Mb of the SNP were tested for correlation with the SNP's allele dosage using Spearman rank correlation in R. The level of significance used was $P < 9 \times 10^{-7}$.

6.6 Heritability estimates

The heritablilities for the metabolite traits were estimated in the twin sample (N = 561 pairs). Before calculating the heritability estimates the metabolite measures were corrected for age and gender and the residuals were inverse normal transformed in R[140]. The heritability of each metabolite measure was estimated using standard modelling methods utilizing the specialized "OpenMx" package in R. For each metabolomics phenotype, models estimating the hypothetical combinations of the different genetic and environmental sources of influence (ACE, ADE, AE, CE and E, where A is the additive genetic influence, C is the shared environmental influence, D is the dominance genetic influence, and E is the unique environmental influence) were built and tested against a saturated model, where no inference on the underlying architecture of the phenotype is assumed. The simplest genetic model that fitted the data best was chosen by comparing the fit statistics (likelihood ratio test and Akaike's Information Criterion) of the hierarchically nested, hypothetical models against that of the saturated model.

6.7 Other statistical and visualization methods

6.7.1 *P*-gain

P-gain is a statistic calculated to evaluate the difference in the associations of a SNP to two traits, i.e., *P*-gain is the ratio of the two *P*-values. This statistic was applied in Publication III to formally quantify the gain from using the metabolomics measures over the enzymatic lipids. A *P*-gain, here the ratio of the *P*-values of the SNP association to the lead enzymatic lipid and the lead metabolomics trait, over 47 was considered significant. The significance level of 47 was derived from the number of principal components of the full metabolomics data set explaining over 99% of variance. In Publication IV *P*-gain was used to determine whether a ratio of two metabolites has better power of association than the individual metabolites. A *P*-gain was calculated for all associated ratios for all associated SNPs by taking the minimum of the two *P*-values of association of the individual metabolites of the ratio and dividing this by

the *P*-value of the ratio. If the *P*-gain was over 1, then the ratio was considered to provide information beyond the individual metabolites.

6.7.2 Heat map visualization

In Publication III the associations of the SNPs in the lipid loci to the metabolites were visualised using a heat map to enable the simultaneous comparison of the associations. The beta coefficients determined the colour scale of the map. For the heat map containing the associations of all the lead SNPs of the 95 lipid loci with the lipoprotein subclass measures, the SNPs were ordered based on their beta coefficients using multidimensional scaling with the apoB-related lipoprotein measures and HDL measures contributing with equal weight to the ordering. The visualization was done in Matlab programming environment.

7. NMR metabolomics meets genetics

This chapter presents the results from two studies (Publications III and IV), in which that metabolite information obtained via the metabolomics platform presented in Chapter 3 was combined with genetic data; Publication III utilized the quantified metabolite data and a dense map of variants to further characterize the known lipid loci[79] both by phenotype and by genotype. In Publication IV genome-wide scan of the metabolite traits was conducted and, by including a twin data set, the heritabilities of the metabolite traits were estimated and the proportion the found variants explain of the variance of the metabolite traits was determined.

The data set used in the studies comprises of in total 8330 Finnish individuals from five population-based cohorts with both genotype and metabolite data. Table 10 gives the basic cohort characteristics. The genotype data was augmented by imputation to the 1000 Genomes reference panel yielding a dense genotype marker set of 7.7 million SNPs across the genome. The metabolites were quantified from the NMR spectra automatically as described in Chapter 3. Altogether quantitative data was available for 117 metabolites (80 from LIPO, 22 from LMWM and 15 form LIPID windows). Additionally, a set of 99 interesting derived measures, including selected ratios of the metabolites, were calculated based on existing biological knowledge. The studied metabolites and derived metabolite measures with the abbreviations used are given in Appendix I.

Table 10. *The basic characteristics of the five population-based Finnish cohorts used in Publications III and IV.*

Study		**N**	**Mean age, years**	**% Female**
NFBC1966	Northern Finland Birth Cohort 1966	4703	31 ± 0	51%
YF	The Cardiovascular Risk in Young Finns Study	1904	37.7 ± 5.0	54%
HBCS	Helsinki Birth Cohort Study	708	61.3 ± 2.9	60%
H2000	Health 2000 GenMets Study	572	55.8 ± 7.3	57%
DILGOM	The Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic Syndrome	443	50.1 ± 13.5	56%

7.1 Metabolic and genetic characterization of the known lipid loci (Publication III)

In Publication III the metabolomics and genotype data were used to further uncover the metabolic and genetic architecture of the 95 loci identified through associations with enzymatic measures of TC, TG, LDL-C and HDL-C[79]. As suggested in another study[33], some of the lipid genes can have specific effects on certain types of lipoprotein particles. Therefore, we hypothesized that the variants could be more strongly associated with specific lipoprotein subclass measures compared to the aggregate enzymatic lipid measures. On the other hand, some of the lipid genes have showed associations to a wide range of traits, thus we investigated the associations across the whole metabolomics panel. Furthermore, a recent study showed for seven LDL-C associated genes how a more detailed coverage of the SNPs in the associated regions than the genotyping arrays provide leads to the discovery of stronger and further variants[143]. Thus, we utilized the detailed marker map obtained via imputation to the 1000 Genomes reference panel and investigated the associations of all the variants within the lipid loci to the metabolite traits, and also performed conditional analyses to formally search for multiple independent variants.

7.1.1 Detailed metabolic characterization of the lipid loci

We first investigated the associations of the 102 lead SNPs reported for the 95 loci to the wide range of metabolomics traits (216 traits) and the four enzymatic lipid measures. Twenty-two of the previously reported lead SNPs showed significant associations ($P < 5 \times 10^{-8}$). The associated loci and traits are given in Table 11. For only six of the lead SNPs (bottom panel in Table 11) the strongest association was to an enzymatically measured lipid, but for 16 SNPs a more detailed lipoprotein measure or a ratio of the metabolites was the most associated trait. Had only enzymatic traits been studied, only 16 significantly associated SNPs would have been found.

We assessed the increase in the strength of association from using the detailed metabolite traits from the metabolomics platform over the enzymatically determined lipids by calculating the *P*-gain statistic (see Materials and Methods for details). For a majority of the associated loci, i.e., 13 SNPs (top panel in Table 11), there was a significant *P*-gain from using the metabolomics phenotypes. In line with this observation, the proportion of trait variance the associated SNPs explained was considerably higher for the metabolite traits (median 0.66%) than for the enzymatic lipids (median 0.39%). The highest *P*-gain was observed for the *FADS1-2-3* locus, which encodes for the fatty acid desaturase genes; the strongest association was to a specific ratio of polyunsaturated lipids and a single SNP explained 15.41% of the variance of this trait.

Table 11. *The 22 lead SNPs that associated significantly with the metabolite measures or enzymatic lipids.*

Locus	Chr	Lead SNP	Lead trait	*P*-value	Variance explained	*P*-gain
Significant P-gain						
*ANGPTL3**	1	rs2131925	Val/Serum-TG	4.01×10^{-12}	0.59%	2.69×10^{4}
*GALNT2**	1	rs4846914	M-HDL-L/S-HDL-L	1.67×10^{-12}	0.66%	7.95×10^{6}
APOB	2	rs1042034	XS-VLDL-TG	9.80×10^{-18}	0.89%	3.80×10^{8}
MLXIPL	7	rs17145738	VLDL-D	1.77×10^{-12}	0.62%	1.02×10^{3}
LPL	8	rs12678919	Val/Serum-TG	5.81×10^{-13}	0.63%	4.47×10^{3}
ABCA1	9	rs1883025	Free-C/Est-C	3.04×10^{-11}	0.57%	6.07×10^{1}
FADS1-2-3	11	rs174546	LA/PUFA	4.77×10^{-268}	15.41%	2.32×10^{259}
APOA1	11	rs964184	Val/Serum-TG	1.09×10^{-26}	1.38%	2.73×10^{2}
LIPC	15	rs1532085	XL-HDL-TG	5.52×10^{-72}	3.97%	4.67×10^{55}
*HPR**	16	rs2000999	Gp/Tot-C	1.47×10^{-13}	0.67%	8.67×10^{8}
*CILP2**	19	rs10401969	MobCH	5.21×10^{-9}	0.42%	1.10×10^{3}
*APOE**	19	rs439401	XS-VLDL-TG	3.12×10^{-9}	0.43%	7.31×10^{2}
*PLTP**	20	rs6065906	L-HDL-L/M-HDL-L	1.29×10^{-24}	1.27%	4.97×10^{22}
Small P-gain						
GCKR	2	rs1260326	Ala/Gln	1.20×10^{-18}	1.03%	1.24
PPP1R3B	8	rs9987289	IDL-C	3.20×10^{-9}	0.42%	3.93
APOE	19	rs4420638	S-LDL-L	3.38×10^{-23}	2.15%	2.76
No P-gain						
SORT1	1	rs629301	LDL-C-lab	3.68×10^{-15}	0.73%	-
APOB	2	rs1367117	LDL-C-lab	4.50×10^{-12}	0.60%	-
HMGCR	5	rs12916	LDL-C-lab	1.30×10^{-10}	0.50%	-
CETP	16	rs3764261	HDL-C-lab	6.32×10^{-49}	2.52%	-
LDLR	19	rs6511720	LDL-C-lab	6.28×10^{-22}	1.12%	-
HNF4A	20	rs1800961	HDL-C-lab	1.03×10^{-8}	0.37%	-

*Locus, the candidate gene or the nearest gene associated to the lead SNP as reported in Teslovich et al.[79]; Chr, chromosome; Lead SNP, the most associated variant of the locus in Teslovich et al. [79]; Lead trait, the trait most associated with the lead SNP; Variance explained, the proportion of the variance of the lead trait the lead SNP explains; P-gain, the value of the P-gain statistic used to evaluate the gain from using the more detailed metabolite measures (See Materials and methods), value > 47 was considered significant; *, the lead SNP was not associated with the enzymatic lipids. The key for the metabolite trait abbreviations is given in Appendix I.*

The reported lead SNPs of the loci associated with up to 70 metabolite measures. Unsurprisingly, as the lipoprotein subclasses are correlated and connected in the lipoprotein cascade, most SNPs showed associations to several lipoprotein subclass measures. The subclass associations were, in general, in line with the observations from the original study, e.g., the SNP in *LDLR* locus that was previously reported to associate with LDL-C associated to all LDL subclass measures. However, for some loci the association profiles were specific to certain subclass particle types or measures, and the enzymatic lipid measures appeared not to fully describe these associations. For example, both *PLTP* and *LIPC* associated with very large, large and small HDL particles, but the sign of

association was the opposite between the larger and smaller HDLs (See Figure 14 for a visualization of the subclass associations for these loci). Thus, the association to total HDL-C was insignificant (*PLTP*) or considerably weaker (*LIPC*) than the subclass associations. The subclass associations of both loci are in line with the known functions of the encoded proteins; e.g., one of the roles of PLTP is to modulate HDL particle size.

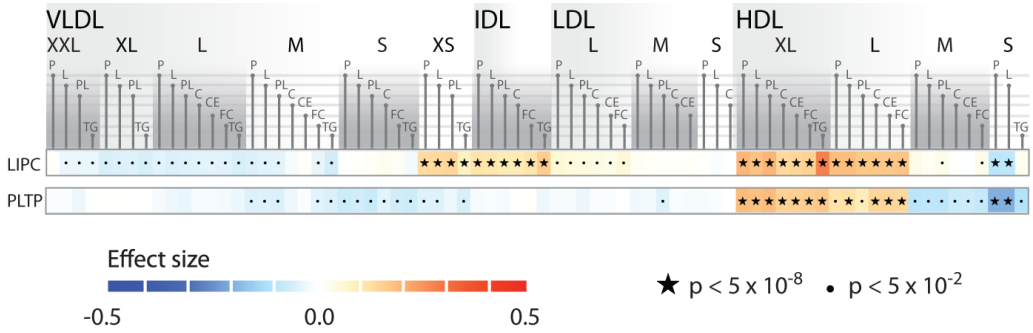

Figure 14. *A heat map visualization of the associations of the reported lead SNPs in PLTP (rs6065906) and LIPC (rs1532085) loci with the lipoprotein subclass measures. The colouring represents the effect sizes of the associations, and the significance of an association is indicated either with a star (genome-wide significance) or a dot (nominal significance). The effect size is in units of standard deviations and shown in respect to the A and C alleles for LIPC and PLTP SNPs, respectively.*

The associations of the lipid loci were not limited to the lipoprotein measures, but some loci showed associations also with other metabolites, including *GCKR*, a well-established susceptibility locus for T2D[144] that encodes for glucokinase regulatory protein. In addition the to the associations with VLDL and fatty acid measures *GCKR* associated with several amino acids, small molecules and their ratios. For example, the SNP was the most associated with the ratio of alanine and glutamine. Table 12 lists the significant small molecule associations of the *GCKR* SNP.

Table 12. *The significant small molecule associations of the lead SNP (rs1260326) in GCKR locus. The effect size is in units of standard deviations and given in respect to the T allele.*

Metabolite	*P*-value	Effect size
Amino acids & small molecules		
Alanine	1.41×10^{-17}	0.138
Isoleucine	2.93×10^{-14}	0.123
Leucine	9.96×10^{-10}	0.099
Pyruvate	1.53×10^{-10}	0.104
Glycoproteins	3.72×10^{-10}	0.101
Metabolite ratios		
Alanine/Glutamine	1.20×10^{-18}	0.15
Isoleucine/Glucose	6.49×10^{-17}	0.135
TG/Glucose	9.76×10^{-16}	0.13
Isoleucine/Phenlylalanine	1.18×10^{-15}	0.13
Glutamine/Isoleucine	2.22×10^{-14}	-0.13
Glucose/Pyruvate	1.27×10^{-13}	-0.12

7.1.2 Genetic and metabolic architecture of the lipid loci

Next the associations of all the variants within the 95 lipid loci (440,870 SNPs), i.e., within 1Mb of the reported lead SNP of the locus, to the metabolomics and enzymatic measures were studied in order to identify alternative variants to the reported lead SNPs that would show stronger associations to the metabolites or lipids in the homogeneous Finnish population. In addition, conditional analyses were performed to identify statistically independent variants in the loci that could increase the proportion of trait variance the loci explain.

In total 31 of the lipid loci showed significant associations, and for 27 loci a variant other than the previously reported lead SNP was the most associated SNP. Thus, in only four loci (*CETP*, *HNF4A*, *APOA1* and *GCKR*) the reported lead SNP showed the strongest associations. For eleven of the associated regions (*LDLRAP1*, *PCSK9*, *ABCG5/8*, *C6orf106*, *ABO*, *LRP4*, *LRP1*, *HNF1A*, *SCARB1*, *LCAT* and *LIPG*) the original lead SNP was not associated with any of the metabolite traits in our data.

The new variants increased the variance explained of the traits compared to the reported lead SNPs (median 0.81% vs. median 0.42%). However, as the regional association plots of the *LIPC* and *PLTP* loci show (Figure 15), the greatest increase in the explained variance arises from using the detailed metabolite traits.

Some of regions' most associated variants were weakly correlated with the reported lead SNPs, and therefore these associations may arise from a signal different to the original lipid association. For example, the strongest associations in *HPR* and *LRP1* loci point to the neighbouring amino acid loci (Publication IV and Suhre et al.[14]). However, the most associated variant in *APOE* locus is weakly correlated with the previous lead SNP but was also found in another study to be the variant in the region most associated to LDL-C[143].

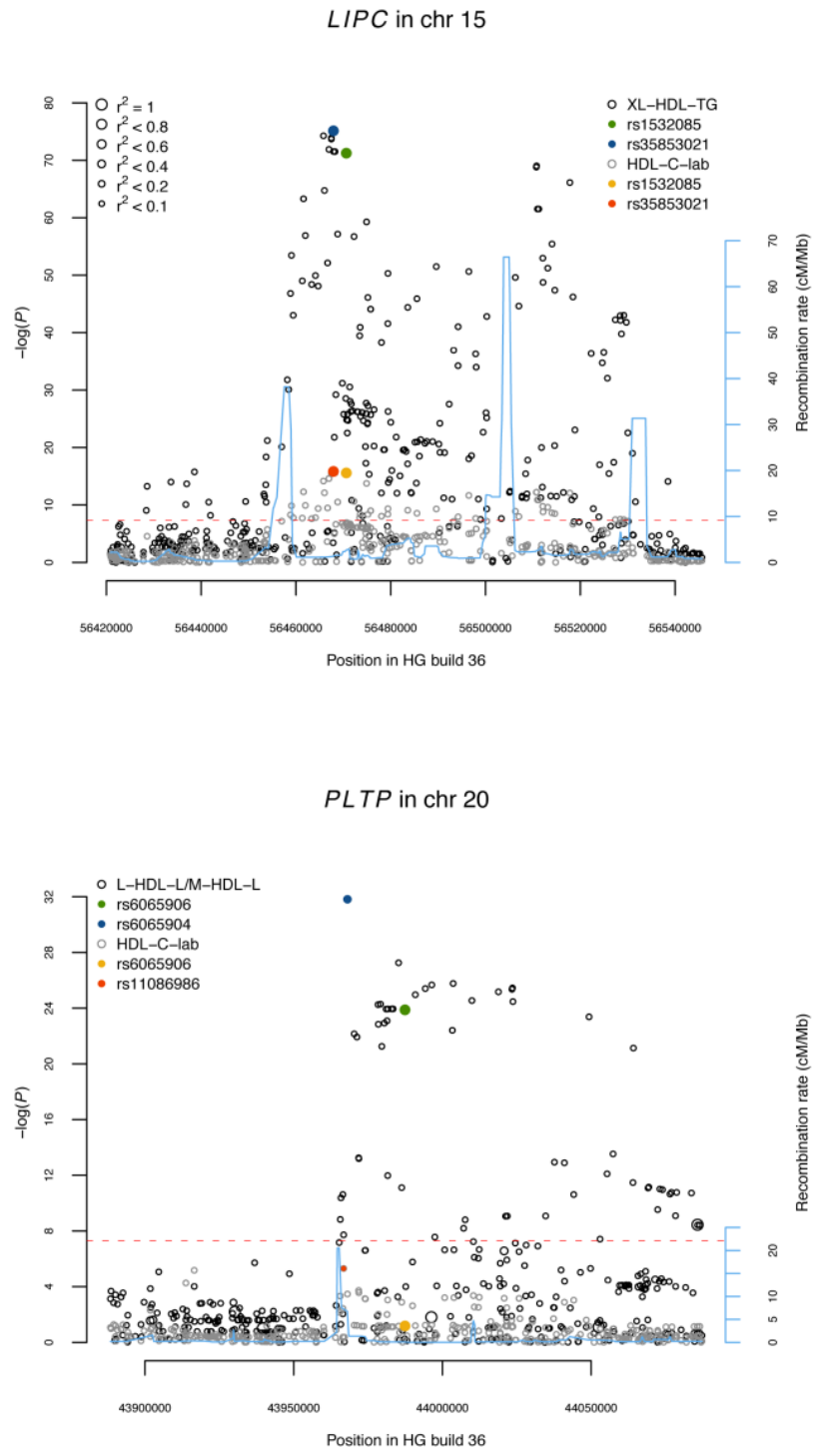

Figure 15. *The regional association plots of the LIPC and PLTP loci the previous and new lead variants highlighted. Associations for both the lead enzymatic trait and the lead metabolomics trait are shown. The plots are from the Supplementary Material of Publication III.*

In order to formally test the independence of the variants within the loci, conditional association analyses were performed (See Materials and Methods for details). These revealed twelve loci (*PCSK9*, *APOB*, *TYW1B*, *FADS1-2-3*, *LRP1*, *LIPC*, *HPR*, *CETP*, *LOC55908*, *CILP2*, *APOE* and *PLTP)* that harboured two or more independently associated variants; *APOB* and *APOE* loci harboured three and *LIPC* and *HPR* four independent SNPs.

Interestingly, the study of the associations of the independent and associated variants across the wide metabolite panel revealed that the variants in *APOB* show distinct differences in their association profiles; two variants associated mostly with IDL and LDL particles while the two other SNPs associated with the larger apoB-lipoproteins (Figure 16). The first two variants reside in the 5' end of the *APOB* gene that encodes the MTP-binding domain, and the latter two in the other end of the gene near the LDLR-binding domain.

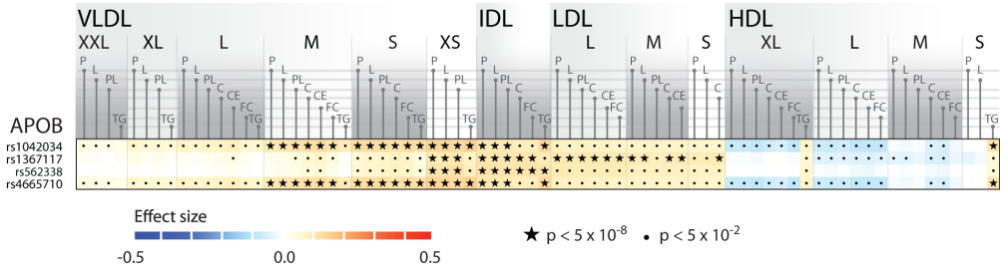

Figure 16. *The associations of the three independent and the region's most associated variant in the APOB locus to the lipoprotein subclass measures. The colouring represents the effect sizes of the associations, and the significance of an association is indicated either with a star (genome-wide significance) or a dot (nominal significance). The effect size is in units of standard deviations and shown in respect to the alleles that associated with an increase of the most associated trait. The figure is modified from Figure 2 of Publication III.*

7.1.3 Discussion

Taken together, the thorough metabolic and genetic characterization utilizing NMR metabolomics data and a dense map of genotyped and imputed variants from the 1000 Genomes reference panel identified significant associations for 31 known lipid loci. Table 13 summarizes the lead associations (both variant and phenotype) for these loci. Interestingly, only for one of the 31 loci, *HNF4A*, no further information in terms of associated SNPs or traits beyond those reported in the original paper could be provided, thus demonstrating the power of the combined use of a wide metabolite panel and a dense marker set.

We found that many loci demonstrated stronger associations to more detailed measures of metabolism than the enzymatic lipid measures; the proportion of variance explained was considerably larger for the lead metabolite traits than for the enzymatic lipids. This points to that these loci may have more specific metabolic roles, which the aggregate lipid measures can capture only partially.

For example, with the detailed lipoprotein subclass variables we showed the heterogeneous association patterns of *LIPC* and *PLTP* to the HDL measures. These results support the previous findings that refined lipoprotein profiling often results in stronger associations than the use of the conventional enzymatic lipid measures[33, 121].

An increase in the explained variance when lipoprotein subclasses are used instead of the conventional lipids was also observed in the study by Petersen et al.[122] published simultaneously with our paper (Publication III). The authors took a similar approach and studied the associations of the lead variants in the 95 lipid loci to 15 NMR-derived lipoprotein subclasses in 1,791 individuals. Due to the smaller sample size a smaller number or significant loci was identified in this study, but, for example, the findings regarding *PLTP* and *LIPC* are mostly consistent with our results, although the subclass definitions differ between the studies.

Studying a wide range of metabolites, including lipoproteins and lipids as well as small molecules, enabled us to comprehensively profile the potential gene effects on various parts of metabolism. Although the studied loci have been identified to associate with lipid levels, our analysis showed that some loci have associations to metabolites not known to directly relate to lipoprotein metabolism. The uncovered associations of *GCKR* with the various amino acids and other small molecules underpin the benefits of the hypothesis-free profiling approach and may provide bases for new hypotheses for the pathways linking *GCKR* and the number of metabolic conditions the locus has been previously associated with.

The reported lead SNP was the most significantly associated variant in only four loci. The large number of alternative region specific stronger variants identified shows the benefits of a detailed map of variants. Using Finnish population-based cohorts provided us a homogeneous study sample. However, as Finns have unique LD patterns, the associations and the identified variants may not be fully accurate as such in other European populations. However, at least one of the new region specific strongest variants has been found in another study to associate with lipids[143]. Many loci were found to harbour multiple independently associated markers that together explained a considerably larger proportion of trait variance, a finding which is in line with the observations from a recent study[143].

Table 13. *A summary of the lead variants and lead traits for the 31 significantly associated lipid loci.*

Locus	Chr	SNP	Trait	*P*-value	Variance explained	Correlation of SNPs
Lead SNP and lead trait remains						
*CETP**	16	rs3764261	HDL-C-lab	6.32×10^{-49}	2.51%	1.000
HNF4A	20	rs1800961	HDL-C-lab	1.03×10^{-8}	0.37%	1.000
Lead SNP but new trait						
GCKR	2	rs1260326	Ala/Gln	1.20×10^{-18}	1.03%	1.000
APOA1	11	rs964184	Val/Serum-TG	1.09×10^{-26}	1.38%	1.000
Enzymatic trait but new SNP						
*PCSK9**	1	1-55892749	LDL-C-lab	1.34×10^{-21}	1.46%	0.024
LDLRAP1	1	rs35346083	LDL-C-lab	4.95×10^{-8}	0.40%	0.929
SORT1	1	rs660240	LDL-C-lab	1.94×10^{-15}	0.75%	0.990
ABCG5/8	2	rs6756629	LDL-C-lab	1.88×10^{-10}	0.47%	0.036
HMGCR	5	rs7703051	LDL-C-lab	3.86×10^{-11}	0.52%	0.931
ABO	9	rs11244035	LDL-C-lab	3.87×10^{-9}	0.49%	0.270
LRP4	11	rs3758673	HDL-C-lab	7.99×10^{-11}	0.49%	0.306
SCARB1	12	12-123911977	HDL-C-lab	1.68×10^{-8}	0.49%	0.046
LCAT	16	16-66448807	HDL-C-lab	2.34×10^{-8}	0.40%	0.652
LDLR	19	19-11058749	LDL-C-lab	4.01×10^{-24}	1.45%	0.881
*APOE**	19	rs7412	LDL-C-lab	2.75×10^{-64}	4.62%	0.048; 0.011
New trait and new SNP						
ANGPTL3	1	rs1168029	MobCH	1.18×10^{-13}	0.71%	0.884
GALNT2	1	rs11122454	M-HDL-L/S-HDL-L	8.36×10^{-13}	0.71%	0.907
*APOB**	2	rs4665710	XS-VLDL-TG	9.17×10^{-18}	0.89%	1.000; 0.134
C6orf106	6	6-34803686	M-HDL-CE	2.26×10^{-8}	0.81%	0.109; 0.064
MLXIPL	7	rs1324787	VLDL-D	4.78×10^{-14}	0.83%	0.673
PPP1R3B	8	rs983309	IDL-C	2.43×10^{-9}	0.42%	0.901
LPL	8	8-19956650	M-VLDL-PL	2.28×10^{-15}	0.90%	0.774
ABCA1	9	rs2575876	Tot-C/Est-C	1.63×10^{-11}	0.58%	0.988
*FADS1-2-3**	11	rs174547	LA/PUFA	1.31×10^{-269}	15.72%	0.986
*LRP1**	12	rs2638315	Gln/Glc	2.38×10^{-36}	2.42%	0.001
HNF1A	12	rs58706475	Tyr	2.07×10^{-8}	0.51%	0.135
*LIPC**	15	rs35853021	XL-HDL-TG	7.11×10^{-76}	4.46%	0.813
*HPR**	16	rs4788815	Phe/Tyr	7.37×10^{-18}	0.98%	0.038
LIPG	18	rs7228085	XL-HDL-TG	4.34×10^{-11}	0.59%	0.164
*CILP2**	19	rs17216588	MobCH	1.04×10^{-9}	0.45%	0.912
*PLTP**	20	rs6065904	L-HDL-L/M-HDL-L	1.49×10^{-32}	1.77%	0.598

*Locus, the candidate gene or the nearest gene associated to the lead SNP as reported in Teslovich et al.[79]; Chr, chromosome; Lead SNP and lead trait, the most associated SNP-trait pair in the locus; Variance explained, the proportion of the variance of the lead trait the lead SNP explains; *; the locus harboured multiple independently associated markers. The key for the metabolite trait abbreviations is given in Appendix I.*

We studied the lipid loci using NMR-based profiling of serum metabolites and many of the studied 117 metabolites and the 99 derived measures were lipoprotein related. This capability of NMR to profile lipoproteins in detail was of particular use as the focus was on lipid loci. However, as some of the loci were identified to have associations beyond lipoprotein measures, studies using metabolite data

from other metabolomics platforms providing complementing information seem justified to further uncover the biological processes underlying the loci.

To conclude, this study uncovered a complex metabolic and genetic architecture underlying the known lipid loci. As lipoproteins are key players in various metabolic conditions, the found associations may be utilized to provide hypotheses for further studies to better understand the variety and interconnections of the metabolic processes involved.

7.2 Genome-wide scan of the metabolomics traits (Publication IV)

While the study presented in Publication III focused on a predefined set of loci, in Publication IV a fully hypothesis-free approach in terms of the genetic associations was taken: all the 216 assayed metabolites and derived variables were correlated against the full set of 7.7 million markers distributed across the genome. In addition to the five population-based cohorts studied in Publication III, a data set of 561 Finnish twin pairs (221 monozygotic, 340 dizygotic, aged 22-25) with genotype and metabolomics data was added to the analyses to enable the estimation of the heritabilities of the metabolite traits.

The handful of the previous metabolomics-GWASs have provided valuable information on the genetic variants contributing to metabolite variation. However, as only some of the metabolites assayed using the NMR metabolomics platform overlap with the previous studies, we hypothesized new genetic information could be uncovered. Also, as our data set included more individuals and variants than the previous metabolomics-GWASs, further variants could be identified due to the increase in power and the more comprehensive coverage of the genetic variation. Furthermore, little data existed on the heritabilites of the metabolomics traits. Thus, the estimates obtained studying the twin sample would provide a resource to evaluate the overall proportion of trait variance the genetic factors explain.

7.2.1 Heritability estimates of the metabolomics traits

The small molecules showed in general lower heritability estimates (lowest for histidine, 0.23; highest for glutamine, 0.55) than the lipids (lowest for DHA, 0.48; highest for LA, 0.62) or lipoproteins (lowest for the concentration of chylomicrons and extremely large VLDL particles (XXL-VLDL-P), 0.50; highest for the free cholesterol in large HDL, 0.76). Taking into account also the derived metabolite measures, the trait showing the highest heritability (0.79) was the ratio of total lipids in large and medium HDL particles. Several measures of very large and large HDL particles as well as the mean diameter of HDL particles showed exceptionally high heritabilities (above 0.70), and > 40% of the metabolites had estimates of heritability above 0.60.

The high heritability estimates for the lipoprotein subclasses motivated a comparison of the heritabilities of the enzymatic lipid measures and the

corresponding measures derived from NMR data. The heritability estimates calculated in a subset of the twins (N = 256) were similar for TC, LDL-C and HDL-C, but were slightly different for TG the heritability estimate for the NMR-measure being higher (0.68) that the enzymatic one (0.55).

7.2.2 Genome-wide association analysis

The genome-wide association analysis identified 31 loci showing significant associations with a total of 180 metabolomics traits. A stringent *P*-value threshold (2.31×10^{-10}, genome-wide significance corrected for the 216 traits tested) was adopted to prevent false positive findings. Thirteen of the associated loci, seven of which had not been previously associated with other metabolic phenotypes, showed associations with amino acids or other small molecules. The remaining 18 loci demonstrated associations to lipoprotein or lipid measures; three novel and eleven known loci showed the strongest association to a lipoprotein measure and one novel and three known loci associated primarily to other NMR-derived lipid measures. Figure 17 summarizes the findings and places those on the primary pathways of human metabolism. The associated loci are briefly described below.

Loci associated with small molecules
The seven novel and six known loci that showed associations to small molecules are summarized in Table 14 with their lead associations, significant eQTLs and potential candidate genes / reported candidates and traits. Six of the novel loci associated with amino acid measures; five of the loci associated with measures of the amino acids recently shown to predict T2D[5], and one with glutamine. In addition one locus was associated with citrate levels. Only one locus showed a significant eQTL in leukocytes: the lead SNP in the citrate locus in chromosome 22 had an eQTL with *CLTCL1*. Interestingly, this locus harbours a plausible candidate gene, *SLC25A1*, which encodes for a citrate transporter. The genes near two other loci also have functions that closely match the associated metabolites. The SNP in chromosome 2 that associated with the ratio of alanine and valine is in the first intron of *SLC1A4*, a neutral amino acid transporter. The variant associated with the ratio of phenylalanine and tyrosine maps 25 kb upstream of *TAT*. *TAT* encodes for tyrosine aminotransferase, an enzyme that catalyzes the conversion of tyrosine to hydroxyphenylpuryvate, and mutations in this gene have been shown to cause type 2 tyrosinemia [OMIM 276600] the symptoms of which include intellectual disability, keratitis, painful palmoplantar hyperkeratosis, and elevated serum tyrosine levels.

Among the known associations are two previously reported gluocse loci, which associated in our data with the NMR-derived glucose, and two recently reported amino acid loci, *SLC16A10* and *GLS2*, for which we report associations to similar metabolite measures. In addition, *GCKR* locus shown to associate, e.g., to TG, showed the strongest association to the ratio of alanine and glutamine but also associated to various lipoprotein, especially VLDL, measures. Finally, we provided further biological evidence for a locus in chromosome 6, which was recently

associated with bradykinin, while we showed an association to histidine and related ratios.

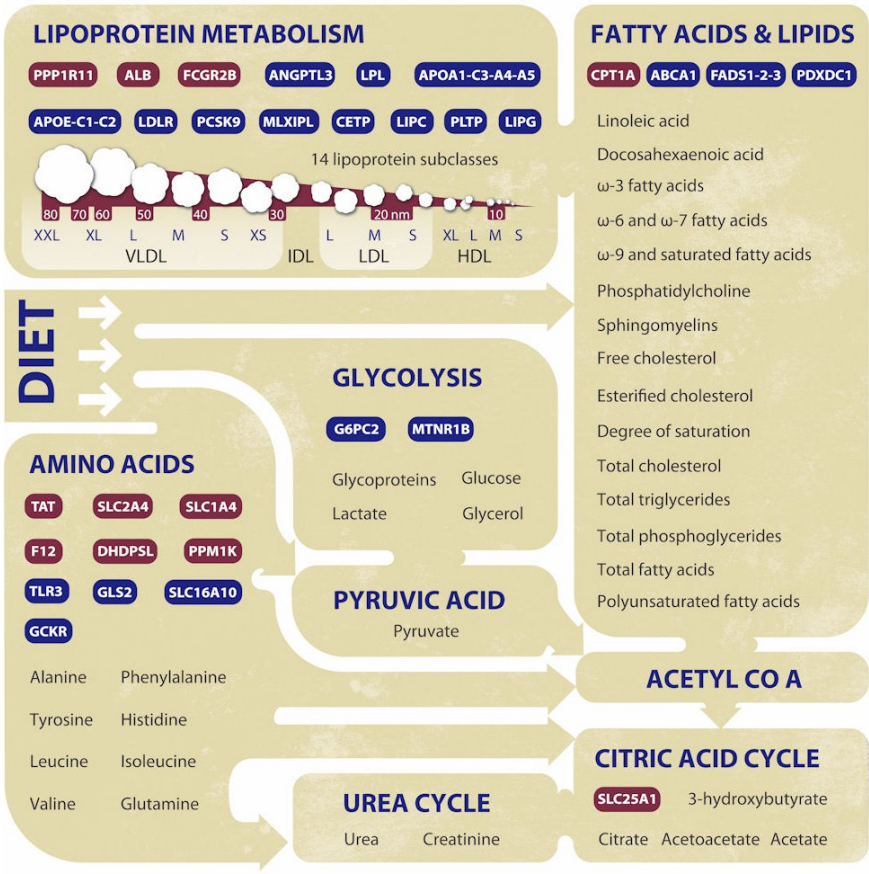

Figure 17. *A summary of the identified 31 loci and studied metabolites in the context of the primary pathways of human metabolism. New loci are highlighted in red and loci that were found in previous GWAS are marked with blue colour. The figure is modified from Publication IV.*

Table 14. *The loci that associated with amino acids or other small molecules. The top panel presents the novel loci and the bottom panel the known loci.*

SNP	Chr	Lead trait	*P*-value	Effect size	eQTL	Candidate (trait)
Novel loci						
rs2160387	2	Ala/Val	2.62×10^{-22}	-0.17	-	*SLC1A4*
rs1440581	4	Fischer's ratio	1.96×10^{-16}	0.13	-	*PPM1K*
rs2545801	5	Phe	8.70×10^{-11}	-0.12	-	*F12*
rs2297644	10	Gln/His	1.23×10^{-12}	0.15	-	*DHDPSL*
rs4788815	16	Phe/Tyr	1.54×10^{-17}	0.15	-	*TAT*
17-7083575*	17	Fischer's ratio	2.64×10^{-14}	-0.51	-	*SLC2A4*
rs807669	22	Citrate	3.30×10^{-16}	-0.14	*CLTCL1*	*SLC25A1*
Known loci						
rs1260326	2	Ala/Gln	2.59×10^{-18}	-0.15	-	*GCKR*, TG[100]
rs560887	2	Glc	2.19×10^{-17}	0.15	-	*G6PC2*, Glucose[145]
rs4241816	4	His/Val	5.58×10^{-13}	0.12	-	*KLKB1*, Bradykinin[14]
rs6900341	6	Ala/Tyr	3.68×10^{-15}	0.13	-	*SLC16A10*, Isoleucine/Tyrosine[14]
rs10830963	11	Glc	3.19×10^{-11}	0.14	-	*MTNR1B*, Glucose[146]
rs2638315	12	Gln/Glc	2.43×10^{-35}	-0.29	*SPRYD4*	*GLS2*, Glutamine[14]

*SNP, the most associated SNP in the region; Chr, chromosome; Lead trait, the most associated trait; Effect size, the beta coefficient in units of standard deviations; eQTL, a significant association of the lead SNP with the expression levels the shown gene; Candidate, a potential candidate gene, for the known loci the reported associated trait is given with a reference to the study; *, the locus would not have been identified if genotypes imputed HapMap II reference panel been used instead to imputing to the 1000 Genomes reference.*

Loci associated with lipoproteins and lipids

In total eighteen loci associated with lipoprotein and lipid measures (Table 15). Four of these have not been previously reported to associate with metabolic traits. These regions include a locus associated with a specific HDL particle cholesterol measure, one showing associations to a range of lipoprotein measures and serum albumin, one associated with VLDL measures, and a locus associated with a ratio of polyunsaturated fatty acids.

A single SNP in chromosome 1 associated with the cholesterol ester content of very large HDL particles. This and other variants in the locus were significant eQTL SNPs for two genes, the alpha and beta subunits of Fc fragment of IgG, low affinity II, receptor (*FCGR2A* and *FCGR2B*), that play roles in activating immune response. *FCGR2B* has been shown to modify the risk of atherosclerosis in mice[147, 148].

Locus close to *ALB* in chromosome 4 associated with albumin and several apoB-related lipoproteins, cholesterol and sphingomyelin. Previous reports have shown the association of rare variants in *ALB* gene to analbuminemia, hypercholesterolemia and hyperlipidemia[149-151]. In line with these observations but, interestingly, in contrast to the mostly positive correlations between albumin and the lipid metabolites, in our data the same allele that associated with an increase in albumin levels associated with a decrease in the lipid measures.

A variant in class I MHC locus associated with XXL-VLDL-P and other lipoprotein measures. A potential candidate gene is *PPP1R11*, an inhibitor of PP1, a highly-conserved serine/threonine phosphatase with a central role in glycogen metabolism and maintaining blood glucose levels.

A SNP within six Mb of *FADS* (fatty acid desaturase) gene cluster in chromosome 11, which is known to associate with lipoproteins and lipids, associated with the ratio of linoleic acid to other polyunsaturated fatty acids (LA/PUFA) (rs17610395; $P = 7.6 \times 10^{-12}$). This SNP was, however, confirmed to be independent of the previously identified marker in the FADS locus. Matching to the associated metabolite trait, a ratio of polyunsaturated fatty acids, the variant is a non-synonymous SNP (Ala275Thr) located in *CPT1A*, a gene encoding carnitine palmitoyltransferase IA, a liver-expressed enzyme involved in long-chain fatty acid oxidation. Rare mutations of this gene cause CPT I deficiency, an autosomal recessive metabolic disorder of long-chain fatty acid oxidation [OMIM 255120].

Our data replicated the associations of 14 previously reported lipid or lipoprotein related loci to a similar phenotype. Twelve loci previously associated with TG, LDL-C or HDL-C show associations in our data to similar phenotypes. In addition, in line with reported associations, the *FADS* locus associated with fatty acid measures. And finally, *PDXDC1* locus that was recently shown to associate with eicosatrienoylglycerophospholipids associated with linoleic acid, which can be converted to an eicosanoid precursor.

7.2.3 The proportion of variance explained

The proportions of variance of the metabolite traits the significantly associated variants together explain was studied in the twin cohort that was independent of the discovery cohorts. The associated variants explained up to 9.5% of trait variance for the metabolites (highest for the concentration of IDL particles (IDL-P) and for the total lipid content of IDL) and a considerably larger proportion, up to 25%, for the derived measures (highest for LA/PUFA). The corresponding proportions of the heritable variance that was explained were 14.5% for IDL-P and 40.4% for LA/PUFA. The exceptionally high proportion of explained variance for LA/PUFA is largely driven by a single common SNP in *FADS* locus: each risk allele resulted in a 0.57 SD increase in the fatty acid ratio. The box plot in Figure 18 illustrates the change in LA-PUFA ratio with the increasing number of effect alleles in each of the five cohorts.

Table 15. *The loci that associated with NMR-measures of lipoproteins or lipids. Top panel presents the novel loci and bottom panel the known loci.*

SNP	Chr	Lead trait	*P*-value	Effect size	eQTL	Candidate (trait)
Novel loci						
1-159807481*	1	XL-HDL-CE	1.21×10^{-10}	0.19	*FCGR2B, FCGR2A*	*FCGR2B, FCGR2A*
4-73541429	4	Albumin	4.84×10^{-18}	-0.51	-	*ALB*
rs6917603*	6	XXL-VLDL-P	2.81×10^{-29}	-0.24	-	*PPP1R11*
rs17610395*	11	LA/PUFA	7.57×10^{-12}	0.17	-	*CPT1A*
Known loci						
1-55889093	1	L-LDL-FC	1.10×10^{-19}	-0.59	-	*PCSK9*, LDL[100]
rs1168029	1	MobCH	2.66×10^{-13}	0.13	-	*ANGPTL3*, TG[100]
rs13247874	7	VLDL-D	8.43×10^{-14}	-0.16	-	*MLXIPL*, TG[100]
8-19956650	8	M-VLDL-PL	4.26×10^{-15}	-0.22	-	*LPL*,TG[100]
rs2575876	9	Tot-C/Est-C	1.63×10^{-11}	-0.14	-	*ABCA1*, HDL[100]
rs174547	11	LA/PUFA	8.02×10^{-262}	0.57	-	*FADS1-2-3*, *PC*[16]
rs651821	11	Val/Serum-TG	7.98×10^{-20}	0.27	-	*APOA1-C3-A4-A5*, LDL[99]
rs35853021	15	XL-HDL-TG	7.11×10^{-76}	0.31	-	*LIPC*, HDL[100]
rs11075253	16	LA/PUFA	4.98×10^{-15}	-0.14	-	*PDXC1*, metabolism of C20:2 and C20:3 fatty acids[14]
rs3764261	16	HDL-C	1.23×10^{-36}	0.22	-	*CETP*, HDL[101]
rs7228085*	18	XL-HDL-TG	6.70×10^{-11}	0.11	-	*LIPG*, HDL[100]
rs55791371	19	M-LDL-C/M-LDL-PL	8.21×10^{-17}	-0.26	-	*LDLR*, LDL[100]
rs7412	19	L-LDL-FC	2.52×10^{-58}	-0.75	-	*APOE-C1-C2*, LDL[100]
rs6065904	20	L-HDL-L/M-HDL-L	2.29×10^{-31}	-0.22	*PLTP*	*PLTP*, TG[101]

See Table 14 for key.

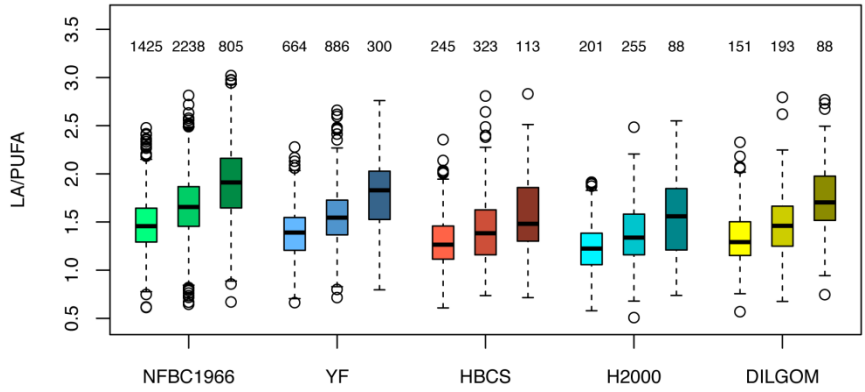

Figure 18. *Box plot of the lead SNP in FADS locus (rs174547) and its effect on the LA-PUFA ratio in the five study cohorts. The numbers above the boxes indicate the number of individuals with each genotype (TT/TC/CC) in each cohort.*

7.2.4 Discussion

This study presented the results from the largest (in terms of sample size and studied variants) metabolomics GWAS conducted thus far. The study also was, to the author's best knowledge, the first GWAS utilizing blood NMR metabolomics with quantitative data from more than a few metabolites[18]. Additionally, this study was among the first and thus far the best powered to assess the heritabilities of metabolomics traits.

Altogether 31 loci were identified to associate with the metabolite traits, including eleven novel loci. The study uncovered loci for three of the five amino acids, i.e., phenylalanine, tyrosine and valine, the levels of which were recently shown to predict the development of T2D[5], and thus the findings may lead to better understanding of the biochemical pathways involved in the pathogenesis of T2D. Mutations in the candidate genes of three of the novel loci have been linked with metabolic abnormalities; *TAT* to type 2 tyrosinemia, *CPT1A* CPT 1 deficiency and *ALB* to analbuminemia and dyslipidemia. We identified common variants in these loci that likely result in a similar but less severe phenotype. In addition, *FCGR2B* that associated with the cholesterol ester content of very large HDL particles, has been linked with atherosclerosis in mouse studies[147, 148]. This finding may help further studies to pinpoint the pathways and mechanisms linking HDL metabolism and atherosclerosis.

Some of the measured metabolites and traits similar to these have been studied in previous GWASs; small molecules and lipid species that have some overlap with the NMR lipids have been studied in previous metabolomics GWASs and conventional lipid levels have been extensively studied by large GWAS consortia. Nevertheless, we uncovered new loci. This may be due to several factors. A detailed dissection of the lipoprotein components revealed the very specific lipoprotein associations in *FCGR2B* and *PPP1R11*, which could thus not be detected with the aggregate enzymatic lipid measures potentially not even with considerably larger sample sizes. In comparison to other GWASs assessing the genetic components underlying the levels of the smaller circulating metabolites, e.g., amino acids, our larger sample may have led to the discovery of further loci. In addition, we applied a more detailed panel of markers than the previous studies, and in fact, four of the novel loci (marked with * in Tables 14 and 15) would not have been identified had only genotypes imputed to HapMap II reference panel been used.

Applying imputation requires the use of stringent quality control filters. Thus, we limited the investigation to markers with frequencies of > 1% in population, good imputation quality and coherent effects in all five cohorts. In nine of the eleven loci the most strongly associated variant was imputed. For seven of the loci there was a significant association also to a directly genotyped variant, and for the two other the imputation was validated by genotyping.

We showed that a great proportion of the variance in the NMR measured metabolite traits is accounted for by genetic factors. Our heritablitity estimates for

the small molecules were mostly consistent with the estimates from two previous metabolomics studies that have, however, been less well powered (Table 16). The observed high heritabilities may in part be explained by the accurate NMR measurement of the metabolites in contrast, e.g., to clinical lipid parameters that are composite measures of lipids carried in various lipoprotein particles.

Table 16. *A Comparison of the heritability estimates from Publication IV to the estimates from two previous studies. Nicholson et al. estimated the familiality of the traits, i.e., the combined contribution of genetic and common environmental effects, thus the corresponding estimates from Publication IV show proportion variance explained by both genetic and shared environmental components if the final model selected included these components.*

Metabolite	Nicholson et al. (N = 144)[19]	Shah et al. (N = 117 / 8 families)[20]	Publication IV
Creatinine	77%	–	41% A + 17% C
Tyrosine	~70%	38% A	39% A
Histidine	~70%	35% A	23% A + 18% C
Glucose	~60%	47% A	25% A + 24% C
Citrate	~62%	39% A	54% A
Glycoprotein acetyls	~60%	–	53% A
Alanine	~60%	55% A	30% A + 19% C
Leucine	~55%	–	52% A
Glycerol	~52%	33% A	33% A + 25% C
Valine	~50%	44% A	45% A
Acetate	~50%	–	30% C
3-hydroxybutyrate	41%	51% A	53% A
Lactate	~35%	–	25% A + 20% C
Isoleucine	~33%	–	51% A
Albumin	~33%	–	39% A + 16% C
Acetoacetate	~30%	–	50% A
Glutamine	~25%	–	55% A
Pyruvate	~15%	–	52% A

A, additive genetic influence; C shared environmental influence. The table is modified from Alfredo Ortega-Alonso with permission.

The NMR metabolomics platform enabled the measurements of detailed metabolite traits in large cohorts providing an enhancement to the conventional clinical measures often available in these sample sizes. The 216 metabolite measures assessed in this study provide a broad view to human metabolism, but represent only a small amount of the various circulating metabolic species found in the human body. Other techniques, including MS, measure mainly other components of the metabolome. GWASs applying these metabolomics measures have provided valuable insight into the genetic control of metabolism[14-16] and the results we reported complement these findings.

8. Conclusions and future prospects

The field of biomedicine has taken giant leaps during the past decade. Technologies that enable capturing information from various layers of biology in high-throughput manner have emerged and thus have given rise to 'omics sciences. Assessing the variation in the genome with SNP arrays has become routine practice, and this data has been extensively utilized to uncover the genetic underpinnings of various traits. However, in the other end of the 'omics cascade, the limited availability of technologies that enable the high-throughput measurement of the metabolite information has held back the use of metabolomics data in large studies. The metabolomics platform presented in this thesis answers to this call by providing a means to capture a wealth of metabolite information cost-effectively using NMR spectroscopy. Since the set up of the metabolomics platform in late 2008 tens of thousands of samples have undergone the same NMR experimentation providing a considerable resource of metabolomics data for utilization in clinical and epidemiological studies.

The high throughput of the platform provided a basis for the two applications presented in this thesis. Utilizing a unique data set of altogether 8330 Finnish individuals with both metabolomics and genotype data the genetic components underlying the quantified serum metabolites were elucidated in a genome-wide association analysis and additionally the detailed metabolite and genotype information was used to further characterize the known lipid loci. A substantial amount of novel biological information was uncovered due to the enhanced metabolic profiling thus showing the utility of metabolomics measurements as more accurate descriptors of metabolism over the conventional clinical assays.

We are, however, only in the beginning of the path of utilizing metabolomics in combination with genetics and further biological information awaits to be discovered. As we and others[14] have observed, a considerably greater numbers of loci than the reported ones are associated with the metabolite levels, but due to the rather moderate sample sizes, in the context of GWAS, that were available for the conducted studies, a wealth of association signals still reside above the applied *P*-value threshold. Therefore, an obvious next step approach, in line with other GWASs of complex traits, is to boost the analysis power by including further cohorts with genotypes and the metabolite data obtained through the NMR metabolomics platform. The 117 metabolites assayed through the metabolomics

platform cover only a minor part of the extensive serum metabolome. Additional metabolite information, however, resides in the spectra and, as the quantification models are under constant development, further metabolite data is expected to be reliably extracted.

The detailed map of 7.7 million variants obtained by imputing the genotypes to the 1000 Genomes reference panel provided a most extensive coverage of the genetic variation and led to discoveries not approachable by using the older imputation references. Recent updates from the 1000 Genomes project have increased the number of available genetic variants to 38 million therefore offering an even more comprehensive map of variants and likely leading to further findings. With the improvements in genotyping technologies, exome or even whole-genome sequence data for large cohorts will become available in the near future, leading to an immense set of genetic information and providing a unique resource to comprehensively investigate the effects of genome variation.

As the data sets are getting larger in terms of individuals, genetic variants and also the metabolites assayed, a considerable challenge for the analyses will, and already is, posed by the available computational capacity. For example, an association analysis of all the metabolite measures from the NMR platform, including all possible combinations of metabolite ratios, i.e., over 9000 phenotypes, in the five Finnish cohorts totalling to 8330 individuals with the genotypes imputed to the newest 1000 Genomes reference panel, thus including 38 million variants, requires roughly 12 million CPU hours and 200 TB of disk space to store the data, therefore posing also a major financial burden.

A major motivation for the research conducted in this thesis was to provide more understanding on the complex biological pathways of human metabolism. Dissecting genetic variants affecting the metabolite levels not only helps to understand the differences in metabolic capacities between individuals, potentially of use in individualized therapy, and sheds light on the gene functions, but, especially in terms of the acknowledged biomarker metabolites, may help to elucidate the pathways and mechanisms involved in disease. Thus, while the discovery of the genetic underpinnings of the metabolic complexity continues, the existing findings, especially those bearing potential clinical significance by being linked to a clinical endpoint, should be characterized further to translate these into therapies and interventions.

Bibliography

1. Ward, J.L., Baker, J.M., Llewellyn, A.M., Hawkins, N.D., Beale, M.H. (2011). Metabolomic analysis of Arabidopsis reveals hemiterpenoid glycosides as products of a nitrate ion-regulated, carbon flux overflow. Proc. Natl. Acad. Sci. U. S. A. 108, 10762-10767.
2. Long, J.Z., Cisar, J.S., Milliken, D., Niessen, S., Wang, C., Trauger, S.A., Siuzdak, G., Cravatt, B.F. (2011). Metabolomics annotates ABHD3 as a physiologic regulator of medium-chain phospholipids. Nat. Chem. Biol. 7, 763-765.
3. Shima, N., Miyawaki, I., Bando, K., Horie, H., Zaitsu, K., Katagi, M., Bamba, T., Tsuchihashi, H., Fukusaki, E. (2011). Influences of methamphetamine-induced acute intoxication on urinary and plasma metabolic profiles in the rat. Toxicology 287, 29-37.
4. Bjerrum, J.T., Nielsen, O.H., Hao, F., Tang, H., Nicholson, J.K., Wang, Y., Olsen, J. (2010). Metabonomics in ulcerative colitis: diagnostics, biomarker identification, and insight into the pathophysiology. J. Proteome Res. 9, 954-962.
5. Wang, T.J., Larson, M.G., Vasan, R.S., Cheng, S., Rhee, E.P., McCabe, E., Lewis, G.D., Fox, C.S., Jacques, P.F., Fernandez, C. et al. (2011). Metabolite profiles and the risk of developing diabetes. Nat. Med.
6. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S. et al. (2007). HMDB: the Human Metabolome Database. Nucleic Acids Res. 35, D521-6.
7. Lenz, E.M., Bright, J., Wilson, I.D., Hughes, A., Morrisson, J., Lindberg, H., Lockton, A. (2004). Metabonomics, dietary influences and cultural differences: a 1H NMR-based study of urine samples obtained from healthy British and Swedish subjects. J. Pharm. Biomed. Anal. 36, 841-849.
8. Rezzi, S., Ramadan, Z., Martin, F.P., Fay, L.B., van Bladeren, P., Lindon, J.C., Nicholson, J.K., Kochhar, S. (2007). Human metabolic phenotypes link directly to specific dietary preferences in healthy individuals. J. Proteome Res. 6, 4469-4477.
9. Mittelstrass, K., Ried, J.S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J. et al. (2011). Discovery of sexual dimorphisms in metabolic and genetic biomarkers. PLoS Genet. 7, e1002215.
10. Makinen, V.P., Soininen, P., Forsblom, C., Parkkonen, M., Ingman, P., Kaski, K., Groop, P.H., FinnDiane Study Group, Ala-Korpela, M. (2008). 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. Mol. Syst. Biol. 4, 167.
11. Psihogios, N.G., Gazi, I.F., Elisaf, M.S., Seferiadis, K.I., Bairaktari, E.T. (2008). Gender-related and age-related urinalysis of healthy subjects by NMR-based metabonomics. NMR Biomed. 21, 195-207.
12. Slupsky, C.M., Rankin, K.N., Wagner, J., Fu, H., Chang, D., Weljie, A.M., Saude, E.J., Lix, B., Adamko, D.J., Shah, S. et al. (2007). Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Anal. Chem. 79, 6995-7004.
13. Lewis, G.D., Farrell, L., Wood, M.J., Martinovic, M., Arany, Z., Rowe, G.C., Souza, A., Cheng, S., McCabe, E.L., Yang, E. et al. (2010). Metabolic signatures of exercise in human plasma. Sci. Transl. Med. 2, 33ra37.
14. Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wagele, B., Altmaier, E., CARDIoGRAM, Deloukas, P., Erdmann, J. et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature 477, 54-60.
15. Illig, T., Gieger, C., Zhai, G., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmuller, G., Kato, B.S., Mewes, H.W. et al. (2010). A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42, 137-141.
16. Gieger, C., Geistlinger, L., Altmaier, E., Hrabe de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J. et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 4, e1000282.

17. Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D. et al. (2011). A genome-wide association study of metabolic traits in human urine. Nat. Genet. 43, 565-569.
18. Nicholson, G., Rantalainen, M., Li, J.V., Maher, A.D., Malmodin, D., Ahmadi, K.R., Faber, J.H., Barrett, A., Min, J.L., Rayner, N.W. et al. (2011). A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet. 7, e1002270.
19. Nicholson, G., Rantalainen, M., Maher, A.D., Li, J.V., Malmodin, D., Ahmadi, K.R., Faber, J.H., Hallgrimsdottir, I.B., Barrett, A., Toft, H. et al. (2011). Human metabolic profiles are stably controlled by genetic and environmental variation. Mol. Syst. Biol. 7, 525.
20. Shah, S.H., Hauser, E.R., Bain, J.R., Muehlbauer, M.J., Haynes, C., Stevens, R.D., Wenner, B.R., Dowdy, Z.E., Granger, C.B., Ginsburg, G.S. et al. (2009). High heritability of metabolomic profiles in families burdened with premature cardiovascular disease. Mol. Syst. Biol. 5, 258.
21. Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B. et al. (2011). The human serum metabolome. PLoS One 6, e16957.
22. Wishart, D.S., Lewis, M.J., Morrissey, J.A., Flegel, M.D., Jeroncic, K., Xiong, Y., Cheng, D., Eisner, R., Gautam, B., Tzur, D. et al. (2008). The human cerebrospinal fluid metabolome. J. Chromatogr. B. Analyt Technol. Biomed. Life. Sci. 871, 164-173.
23. Dunn, W.B., Broadhurst, D.I., Atherton, H.J., Goodacre, R., Griffin, J.L. (2011). Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem. Soc. Rev. 40, 387-426.
24. Griffin, J.L., Atherton, H., Shockcor, J., Atzori, L. (2011). Metabolomics as a tool for cardiac research. Nat. Rev. Cardiol. 8, 630-643.
25. Dettmer, K., Aronov, P.A., Hammock, B.D. (2007). Mass spectrometry-based metabolomics. Mass Spectrom. Rev. 26, 51-78.
26. Issaq, H.J., Van, Q.N., Waybright, T.J., Muschik, G.M., Veenstra, T.D. (2009). Analytical and statistical approaches to metabolomics research. J. Sep. Sci. 32, 2183-2199.
27. Bictash, M., Ebbels, T.M., Chan, Q., Loo, R.L., Yap, I.K., Brown, I.J., de Iorio, M., Daviglus, M.L., Holmes, E., Stamler, J. et al. (2010). Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. J. Clin. Epidemiol. 63, 970-979.
28. Brown, F.F., Campbell, I.D., Kuchel, P.W., Rabenstein, D.C. (1977). Human erythrocyte metabolism studies by 1H spin echo NMR. FEBS Lett. 82, 12-16.
29. Nicholson, J.K., Buckingham, M.J., Sadler, P.J. (1983). High resolution 1H n.m.r. studies of vertebrate blood and plasma. Biochem. J. 211, 605-615.
30. Nicholson, J.K., Foxall, P.J., Spraul, M., Farrant, R.D., Lindon, J.C. (1995). 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. Anal. Chem. 67, 793-811.
31. Hiltunen, Y., Ala-Korpela, M., Jokisaari, J., Eskelinen, S., Kiviniitty, K., Savolainen, M., Kesaniemi, Y.A. (1991). A lineshape fitting model for 1H NMR spectra of human blood plasma. Magn. Reson. Med. 21, 222-232.
32. Otvos, J.D., Mora, S., Shalaurova, I., Greenland, P., Mackey, R.H., Goff, D.C.,Jr. (2011). Clinical implications of discordance between low-density lipoprotein cholesterol and particle number. J. Clin. Lipidol. 5, 105-113.
33. Chasman, D.I., Pare, G., Mora, S., Hopewell, J.C., Peloso, G., Clarke, R., Cupples, L.A., Hamsten, A., Kathiresan, S., Malarstig, A. et al. (2009). Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. PLoS Genet. 5, e1000730.
34. Hodge, A.M., Jenkins, A.J., English, D.R., O'Dea, K., Giles, G.G. (2011). NMR-determined lipoprotein subclass profile is associated with dietary composition and body size. Nutr. Metab. Cardiovasc. Dis. 21, 603-609.
35. Tukiainen, T., Tynkkynen, T., Makinen, V.P., Jylanki, P., Kangas, A., Hokkanen, J., Vehtari, A., Grohn, O., Hallikainen, M., Soininen, H. et al. (2008). A multi-metabolite analysis of serum by 1H NMR spectroscopy: early systemic signs of Alzheimer's disease. Biochem. Biophys. Res. Commun. 375, 356-361.
36. Soininen, P., Kangas, A.J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Järvelin, M.R., Kähönen, M., Lehtimäki, T., Viikari, J. et al. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. Analyst 134, 1781-1785.
37. Jung, J.Y., Lee, H.S., Kang, D.G., Kim, N.S., Cha, M.H., Bang, O.S., Ryu do, H., Hwang, G.S. (2011). 1H-NMR-based metabolomics study of cerebral infarction. Stroke 42, 1282-1288.
38. Brindle, J.T., Antti, H., Holmes, E., Tranter, G., Nicholson, J.K., Bethell, H.W., Clarke, S., Schofield, P.M., McKilligin, E., Mosedale, D.E. et al. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. Nat. Med. 8, 1439-1444.

39. Kirschenlohr, H.L., Griffin, J.L., Clarke, S.C., Rhydwen, R., Grace, A.A., Schofield, P.M., Brindle, K.M., Metcalfe, J.C. (2006). Proton NMR analysis of plasma is a weak predictor of coronary artery disease. Nat. Med. 12, 705-710.
40. Ala-Korpela, M. (2008). Critical evaluation of 1H NMR metabonomics of serum as a methodology for disease risk assessment and diagnostics. Clin. Chem. Lab. Med. 46, 27-42.
41. Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.M. et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472, 57-63.
42. Holmes, E., Loo, R.L., Stamler, J., Bictash, M., Yap, I.K., Chan, Q., Ebbels, T., De Iorio, M., Brown, I.J., Veselkov, K.A. et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453, 396-400.
43. Wurtz, P., Soininen, P., Kangas, A.J., Makinen, V.P., Groop, P.H., Savolainen, M.J., Juonala, M., Viikari, J.S., Kahonen, M., Lehtimaki, T. et al. (2011). Characterization of systemic metabolic phenotypes associated with subclinical atherosclerosis. Mol. Biosyst 7, 385-393.
44. Newgard, C.B., An, J., Bain, J.R., Muehlbauer, M.J., Stevens, R.D., Lien, L.F., Haqq, A.M., Shah, S.H., Arlotto, M., Slentz, C.A. 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, 311-326.
45. Shah, S.H., Bain, J.R., Muehlbauer, M.J., Stevens, R.D., Crosslin, D.R., Haynes, C., Dungan, J., Newby, L.K., Hauser, E.R., Ginsburg, G.S. et al. (2010). Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ. Cardiovasc. Genet. 3, 207-214.
46. Suhre, K., Meisinger, C., Doring, A., Altmaier, E., Belcredi, P., Gieger, C., Chang, D., Milburn, M.V., Gall, W.E., Weinberger, K.M. et al. (2010). Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS One 5, e13953.
47. Inouye, M., Kettunen, J., Soininen, P., Silander, K., Ripatti, S., Kumpula, L.S., Hämäläinen, E., Jousilahti, P., Kangas, A.J., Männistö, S. et al. (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. Mol. Syst. Biol. 6, 441.
48. Ala-Korpela, M., Korhonen, A., Keisala, J., Horkko, S., Korpi, P., Ingman, L.P., Jokisaari, J., Savolainen, M.J., Kesaniemi, Y.A. (1994). 1H NMR-based absolute quantitation of human lipoproteins and their lipid contents directly from plasma. J. Lipid Res. 35, 2292-2304.
49. Ala-Korpela, M. (1995). 1H NMR spectroscopy of human blood plasma. Prog Nucl Magn Reson Spectrosc 27, 475-554.
50. Vehtari, A., Makinen, V.P., Soininen, P., Ingman, P., Makela, S.M., Savolainen, M.J., Hannuksela, M.L., Kaski, K., Ala-Korpela, M. (2007). A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data. BMC Bioinformatics 8 Suppl 2, S8.
51. Samieri, C., Feart, C., Letenneur, L., Dartigues, J.F., Peres, K., Auriacombe, S., Peuchant, E., Delcourt, C., Barberger-Gateau, P. (2008). Low plasma eicosapentaenoic acid and depressive symptomatology are independent predictors of dementia risk. Am. J. Clin. Nutr. 88, 714-721.
52. Beckonert, O., Keun, H.C., Ebbels, T.M., Bundy, J., Holmes, E., Lindon, J.C., Nicholson, J.K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat. Protoc. 2, 2692-2703.
53. Yu, Z., Kastenmuller, G., He, Y., Belcredi, P., Moller, G., Prehn, C., Mendes, J., Wahl, S., Roemisch-Margl, W., Ceglarek, U. et al. (2011). Differences between human plasma and serum metabolite profiles. PLoS One 6, e21230.
54. Bell, J.D., Brown, J.C., Norman, R.E., Sadler, P.J., Newell, D.R. (1988). Factors affecting 1H NMR spectra of blood plasma: cancer, diet and freezing. NMR Biomed. 1, 90-94.
55. Teahan, O., Gamble, S., Holmes, E., Waxman, J., Nicholson, J.K., Bevan, C., Keun, H.C. (2006). Impact of analytical bias in metabonomic studies of human blood serum and plasma. Anal. Chem. 78, 4307-4318.
56. Bernini, P., Bertini, I., Luchinat, C., Nincheri, P., Staderini, S., Turano, P. (2011). Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. J. Biomol. NMR 49, 231-243.
57. Zivkovic, A.M., Wiest, M.M., Nguyen, U.T., Davis, R., Watkins, S.M., German, J.B. (2009). Effects of sample handling and storage on quantitative lipid analysis in human serum. Metabolomics 5, 507-516.
58. Chambers, J.C., Zhang, W., Sehmi, J., Li, X., Wass, M.N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S.E. et al. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. 43, 1131-1138.
59. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C. et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478, 103-109.

60. Surakka, I., Isaacs, A., Karssen, L.C., Laurila, P.P., Middelberg, R.P., Tikkanen, E., Ried, J.S., Lamina, C., Mangino, M., Igl, W. et al. (2011). A Genome-Wide Screen for Interactions Reveals a New Locus on 4p15 Modifying the Effect of Waist-to-Hip Ratio on Total Cholesterol. PLoS Genet. 7, e1002333.
61. Vanhala, M., Kumpula, L.S., Soininen, P., Kangas, A.J., Ala-Korpela, M., Kautiainen, H., Mäntyselkä, P., Saltevo, J. (2011). High serum adiponectin is associated with favorable lipoprotein subclass profile in 6.4-year follow-up. Eur. J. Endocrinol. 164, 549-552.
62. Koskinen, J., Magnussen, C.G., Wurtz, P., Soininen, P., Kangas, A.J., Viikari, J.S., Kahonen, M., Loo, B.M., Jula, A., Ahotupa, M. et al. (2011). Apolipoprotein B, oxidized low-density lipoprotein, and LDL particle size in predicting the incidence of metabolic syndrome: the Cardiovascular Risk in Young Finns study. Eur. J. Cardiovasc. Prev. Rehabil.
63. Valcarcel, B., Wurtz, P., Seich al Basatena, N.K., Tukiainen, T., Kangas, A.J., Soininen, P., Jarvelin, M.R., Ala-Korpela, M., Ebbels, T.M., de Iorio, M. (2011). A differential network approach to exploring differences between biological states: an application to prediabetes. PLoS One 6, e24702.
64. Stancakova, A., Paananen, J., Soininen, P., Kangas, A.J., Bonnycastle, L.L., Morken, M.A., Collins, F.S., Jackson, A.U., Boehnke, M.L., Kuusisto, J. et al. (2011). Effects of 34 risk loci for type 2 diabetes or hyperglycemia on lipoprotein subclasses and their composition in 6,580 nondiabetic Finnish men. Diabetes 60, 1608-1616.
65. Haas, B.E., Weissglas-Volkov, D., Aguilar-Salinas, C.A., Nikkola, E., Vergnes, L., Cruz-Bautista, I., Riba, L., Stancakova, A., Kuusisto, J., Soininen, P. et al. (2011). Evidence of how rs7575840 influences apolipoprotein B-containing lipid particles. Arterioscler. Thromb. Vasc. Biol. 31, 1201-1207.
66. Asztalos, B.F., Tani, M., Schaefer, E.J. (2011). Metabolic and functional relevance of HDL subspecies. Curr. Opin. Lipidol. 22, 176-185.
67. Krauss, R.M. (2010). Lipoprotein subfractions and cardiovascular disease risk. Curr. Opin. Lipidol. 21, 305-311.
68. Mora, S., Otvos, J.D., Rifai, N., Rosenson, R.S., Buring, J.E., Ridker, P.M. (2009). Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident cardiovascular disease in women. Circulation 119, 931-939.
69. Ala-Korpela, M., Soininen, P., Savolainen, M.J. (2009). Letter by Ala-Korpela et al regarding article, "Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident cardiovascular disease in women". Circulation 120, e149; author reply e150.
70. Prado, K.B., Shugg, S., Backstrand, J.R. (2011). Low-density lipoprotein particle number predicts coronary artery calcification in asymptomatic adults at intermediate risk of cardiovascular disease. J. Clin. Lipidol. 5, 408-413.
71. Redgrave, T.G. (2004). Chylomicron metabolism. Biochem. Soc. Trans. 32, 79-82.
72. Wang, H., Eckel, R.H. (2009). Lipoprotein lipase: from gene to obesity. Am. J. Physiol. Endocrinol. Metab. 297, E271-88.
73. Zambon, A., Bertocco, S., Vitturi, N., Polentarutti, V., Vianello, D., Crepaldi, G. (2003). Relevance of hepatic lipase to the metabolism of triacylglycerol-rich lipoproteins. Biochem. Soc. Trans. 31, 1070-1074.
74. Lawn, R.M., Wade, D.P., Garvin, M.R., Wang, X., Schwartz, K., Porter, J.G., Seilhamer, J.J., Vaughan, A.M., Oram, J.F. (1999). The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway. J. Clin. Invest. 104, R25-31.
75. Rye, K.A., Bursill, C.A., Lambert, G., Tabet, F., Barter, P.J. (2009). The metabolism and anti-atherogenic properties of HDL. J. Lipid Res. 50 Suppl, S195-200.
76. Wang, N., Lan, D., Chen, W., Matsuura, F., Tall, A.R. (2004). ATP-binding cassette transporters G1 and G4 mediate cellular cholesterol efflux to high-density lipoproteins. Proc. Natl. Acad. Sci. U. S. A. 101, 9774-9779.
77. Korhonen, A., Jauhiainen, M., Ehnholm, C., Kovanen, P.T., Ala-Korpela, M. (1998). Remodeling of HDL by phospholipid transfer protein: demonstration of particle fusion by 1H NMR spectroscopy. Biochem. Biophys. Res. Commun. 249, 910-916.
78. Griffin, B.A., Caslake, M.J., Yip, B., Tait, G.W., Packard, C.J., Shepherd, J. (1990). Rapid isolation of low density lipoprotein (LDL) subfractions from plasma by density gradient ultracentrifugation. Atherosclerosis 83, 59-67.
79. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713.
80. Würtz, P., Raiko, J.R., Magnussen, C.G., Soininen, P., Kangas, A.J., Tynkkynen, T., Thomson, R., Laatikainen, R., Savolainen, M.J., Laurikka, J. et al. (2012). High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. European Heart Journal In press.
81. Mäkinen, V., Tynkkynen, T., Soininen, P., Forsblom, C., Peltola, T., Kangas, A., Groop, P., Ala-Korpela, M. Sphingomyelin is associated with kidney disease in type 1 diabetes (The FinnDiane Study). Metabolomics, 1-7.

82. Nachman, M.W., Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297-304.
83. 1000 Genomes Project Consortium, Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
84. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. et al. (2001). Linkage disequilibrium in the human genome. Nature 411, 199-204.
85. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.
86. Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger, J.D., Peltonen, L. (2003). The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum. Mol. Genet. 12, 51-59.
87. Service, S., DeYoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A. et al. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat. Genet. 38, 556-560.
88. Peltonen, L., Jalanko, A., Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. Hum. Mol. Genet. 8, 1913-1923.
89. International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.
90. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F. et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.
91. Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., Salomaa, V., Daly, M., Palotie, A., Peltonen, L. et al. (2010). Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res. 20, 1344-1351.
92. Marchini, J., Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499-511.
93. Howie, B.N., Donnelly, P., Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529.
94. Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906-913.
95. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T. et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science 308, 385-389.
96. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W. et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science (New York, N. Y 316, 889-94.
97. Storey, J.D., Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. 100, 9440-9445.
98. Demirkan, A., Amin, N., Isaacs, A., Jarvelin, M.R., Whitfield, J.B., Wichmann, H.E., Kyvik, K.O., Rudan, I., Gieger, C., Hicks, A.A. et al. (2011). Genetic architecture of circulating lipid levels. Eur. J. Hum. Genet. 19, 813-819.
99. Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R.J., Marcano, A.C., Hajat, C. et al. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. Am. J. Hum. Genet. 82, 139-149.
100. Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M. et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. 40, 161-169.
101. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S. et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 40, 189-197.
102. Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., Song, K., Yuan, X., Johnson, T., Ashford, S. et al. (2008). LDL-cholesterol concentrations: a genome-wide association study. Lancet 371, 483-491.
103. Burkhardt, R., Kenny, E.E., Lowe, J.K., Birkeland, A., Josowitz, R., Noel, M., Salit, J., Maller, J.B., Pe'er, I., Daly, M.J. et al. (2008). Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. Arterioscler. Thromb. Vasc. Biol. 28, 2078-2084.

104. Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C.G., Zaitlen, N.A., Varilo, T., Kaakinen, M. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat. Genet. 41, 35-46.
105. Kathiresan, S., Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., Schadt, E.E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T. et al. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet. 41, 56-65.
106. Aulchenko, Y.S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I.M., Pramstaller, P.P., Penninx, B.W., Janssens, A.C., Wilson, J.F., Spector, T. et al. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat. Genet. 41, 47-55.
107. Hiura, Y., Shen, C.S., Kokubo, Y., Okamura, T., Morisaki, T., Tomoike, H., Yoshida, T., Sakamoto, H., Goto, Y., Nonogi, H. et al. (2009). Identification of genetic markers associated with high-density lipoprotein-cholesterol by genome-wide screening in a Japanese population: the Suita study. Circ. J. 73, 1119-1126.
108. Waterworth, D.M., Ricketts, S.L., Song, K., Chen, L., Zhao, J.H., Ripatti, S., Aulchenko, Y.S., Zhang, W., Yuan, X., Lim, N. et al. (2010). Genetic variants influencing circulating lipid levels and risk of coronary artery disease. Arterioscler. Thromb. Vasc. Biol. 30, 2264-2276.
109. Shen, H., Damcott, C.M., Rampersaud, E., Pollin, T.I., Horenstein, R.B., McArdle, P.F., Peyser, P.A., Bielak, L.F., Post, W.S., Chang, Y.P. et al. (2010). Familial defective apolipoprotein B-100 and increased low-density lipoprotein cholesterol and coronary artery calcification in the old order amish. Arch. Intern. Med. 170, 1850-1855.
110. Ridker, P.M., Pare, G., Parker, A.N., Zee, R.Y., Miletich, J.P., Chasman, D.I. (2009). Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study. Circ. Cardiovasc. Genet. 2, 26-33.
111. Heid, I.M., Boes, E., Muller, M., Kollerits, B., Lamina, C., Coassin, S., Gieger, C., Doring, A., Klopp, N., Frikke-Schmidt, R. et al. (2008). Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. Circ. Cardiovasc. Genet. 1, 10-20.
112. Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gomez Perez, F.J., Frazer, K.A., Elliott, P., Scott, J. et al. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat. Genet. 40, 149-151.
113. Chasman, D.I., Pare, G., Zee, R.Y., Parker, A.N., Cook, N.R., Buring, J.E., Kwiatkowski, D.J., Rose, L.M., Smith, J.D., Williams, P.T. et al. (2008). Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. Circ. Cardiovasc. Genet. 1, 21-30.
114. Pollin, T.I., Damcott, C.M., Shen, H., Ott, S.H., Shelton, J., Horenstein, R.B., Post, W., McLenithan, J.C., Bielak, L.F., Peyser, P.A. et al. (2008). A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322, 1702-1705.
115. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University,and Novartis Institutes of BioMedical Research, Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N. et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316, 1331-1336.
116. Kim, Y.J., Go, M.J., Hu, C., Hong, C.B., Kim, Y.K., Lee, J.Y., Hwang, J.Y., Oh, J.H., Kim, D.J., Kim, N.H. et al. (2011). Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. Nat. Genet. 43, 990-995.
117. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. 42, 210-215.
118. Ma, L., Yang, J., Runesha, H.B., Tanaka, T., Ferrucci, L., Bandinelli, S., Da, Y. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. BMC Med. Genet. 11, 55.
119. Igl, W., Johansson, A., Wilson, J.F., Wild, S.H., Polasek, O., Hayward, C., Vitart, V., Hastie, N., Rudan, P., Gnewuch, C. et al. (2010). Modeling of environmental effects in genome-wide association studies identifies SLC2A2 and HP as novel loci influencing serum cholesterol levels. PLoS Genet. 6, e1000798.
120. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M. et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714-719.
121. Kaess, B.M., Tomaszewski, M., Braund, P.S., Stark, K., Rafelt, S., Fischer, M., Hardwick, R., Nelson, C.P., Debiec, R., Huber, F. et al. (2011). Large-scale candidate gene analysis of HDL particle features. PLoS One 6, e14529.
122. Petersen, A., Stark, K., Musameh, M.D., Nelson, C.P., Römisch-Margl, W., Kremer, W., Raffler, J., Krug, S., Skurk, T., Rist, M.J. et al. (2011). Genetic Associations with Lipoprotein Subfractions Provide Information on their Biological Nature. Human Molecular Genetics.

123. Altmaier, E., Ramsay, S.L., Graber, A., Mewes, H.W., Weinberger, K.M., Suhre, K. (2008). Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication. Endocrinology 149, 3478-3489.
124. Hicks, A.A., Pramstaller, P.P., Johansson, A., Vitart, V., Rudan, I., Ugocsai, P., Aulchenko, Y., Franklin, C.S., Liebisch, G., Erdmann, J. et al. (2009). Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet. 5, e1000672.
125. Wang, X., Magkos, F., Mittendorfer, B. (2011). Sex differences in lipid and lipoprotein metabolism: it's not just about sex hormones. J. Clin. Endocrinol. Metab. 96, 885-893.
126. Isaacs, A., Sayed-Tabatabaei, F.A., Aulchenko, Y.S., Zillikens, M.C., Sijbrands, E.J., Schut, A.F., Rutten, W.P., Pols, H.A., Witteman, J.C., Oostra, B.A. et al. (2007). Heritabilities, apolipoprotein E, and effects of inbreeding on plasma lipids in a genetically isolated population: the Erasmus Rucphen Family Study. Eur. J. Epidemiol. 22, 99-105.
127. Pietilainen, K.H., Soderlund, S., Rissanen, A., Nakanishi, S., Jauhiainen, M., Taskinen, M.R., Kaprio, J. (2009). HDL subspecies in young adult twins: heritability and impact of overweight. Obesity (Silver Spring) 17, 1208-1214.
128. Iliadou, A., Snieder, H., Wang, X., Treiber, F.A., Davis, C.L. (2005). Heritabilities of lipids in young European American and African American twins. Twin Res. Hum. Genet. 8, 492-498.
129. Fenger, M., Benyamin, B., Schousboe, K., Sorensen, T.I., Kyvik, K.O. (2007). Variance decomposition of apolipoproteins and lipids in Danish twins. Atherosclerosis 191, 40-47.
130. Bosse, Y., Perusse, L., Vohl, M.C. (2004). Genetics of LDL particle heterogeneity: from genetic epidemiology to DNA-based variations. J. Lipid Res. 45, 1008-1026.
131. Austin, M.A., Newman, B., Selby, J.V., Edwards, K., Mayer, E.J., Krauss, R.M. (1993). Genetics of LDL subclass phenotypes in women twins. Concordance, heritability, and commingling analysis. Arterioscler. Thromb. 13, 687-695.
132. Lamon-Fava, S., Jimenez, D., Christian, J.C., Fabsitz, R.R., Reed, T., Carmelli, D., Castelli, W.P., Ordovas, J.M., Wilson, P.W., Schaefer, E.J. (1991). The NHLBI Twin Study: heritability of apolipoprotein A-I, B, and low density lipoprotein subclasses and concordance for lipoprotein(a). Atherosclerosis 91, 97-106.
133. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K. et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 7, e1002003.
134. Rantakallio, P. (1969). Groups at risk in low birth weight infants and perinatal mortality. Acta Paediatr. Scand. 193, Suppl 193:1+.
135. Raitakari, O.T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J. et al. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. Int. J. Epidemiol. 37, 1220-1226.
136. Eriksson, J.G. (2007). Epidemiology, genes and the environment: lessons learned from the Helsinki Birth Cohort Study. J. Intern. Med. 261, 418-425.
137. Perttilä, J., Merikanto, K., Naukkarinen, J., Surakka, I., Martin, N.W., Tanhuanpää, K., Grimard, V., Taskinen, M.R., Thiele, C., Salomaa, V. et al. (2009). OSBPL10, a novel candidate gene for high triglyceride trait in dyslipidemic Finnish subjects, regulates cellular lipid metabolism. J. Mol. Med. 87, 825-835.
138. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38, 904-9.
139. Patterson, N., Price, A.L., Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. 2, e190.
140. R Development Core Team. (2007). R: A Language and Environment for Statistical Computing (Vienna, Austria.
141. Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L., Berrettini, W., Knouff, C.W., Yuan, X., Waeber, G. et al. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. 42, 436-440.
142. Magi, R., Morris, A.P. (2010). GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288.
143. Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H.M., Jackson, A.U., Piras, M.G., Usala, G., Maninchedda, G., Sassu, A. et al. (2011). Fine mapping of five Loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. 7, e1002198.
144. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L. et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat. Genet. 42, 105-116.
145. Bouatia-Naji, N., Rocheleau, G., Van Lommel, L., Lemaire, K., Schuit, F., Cavalcanti-Proenca, C., Marchand, M., Hartikainen, A.L., Sovio, U., De Graeve, F. et al. (2008). A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. Science 320, 1085-1088.

146. Prokopenko, I., Langenberg, C., Florez, J.C., Saxena, R., Soranzo, N., Thorleifsson, G., Loos, R.J., Manning, A.K., Jackson, A.U., Aulchenko, Y. et al. (2009). Variants in MTNR1B influence fasting glucose levels. Nat. Genet. 41, 77-81.
147. Zhao, M., Wigren, M., Duner, P., Kolbus, D., Olofsson, K.E., Bjorkbacka, H., Nilsson, J., Fredrikson, G.N. (2010). FcgammaRIIB inhibits the development of atherosclerosis in low-density lipoprotein receptor-deficient mice. J. Immunol. 184, 2253-2260.
148. Hernandez-Vargas, P., Ortiz-Munoz, G., Lopez-Franco, O., Suzuki, Y., Gallego-Delgado, J., Sanjuan, G., Lazaro, A., Lopez-Parra, V., Ortega, L., Egido, J. et al. (2006). Fcgamma receptor deficiency confers protection against atherosclerosis in apolipoprotein E knockout mice. Circ. Res. 99, 1188-1196.
149. Shin, Y., Vaziri, N.D., Willekes, N., Kim, C.H., Joles, J.A. (2005). Effects of gender on hepatic HMG-CoA reductase, cholesterol 7alpha-hydroxylase, and LDL receptor in hereditary analbuminemia. Am. J. Physiol. Endocrinol. Metab. 289, E993-8.
150. Rosipal, S., Debreova, M., Rosipal, R. (2006). A speculation about hypercholesterolemia in congenital analbuminemia. Am. J. Med. 119, 181-182.
151. Koot, B.G., Houwen, R., Pot, D.J., Nauta, J. (2004). Congenital analbuminaemia: biochemical and clinical implications. A case report and literature review. Eur. J. Pediatr. 163, 664-670.
152. Niemi, J., Mäkinen, V.P., Heikkonen, J., Tenkanen, L., Hiltunen, Y., Hannuksela, M.L., Jauhiainen, M., Forsblom, C., Taskinen, M.R., Kesäniemi, Y.A. et al. (2009). Estimation of VLDL, IDL, LDL, HDL2, apoA-I, and apoB from the Friedewald inputs--apoB and IDL, but not LDL, are associated with mortality in type 1 diabetes. Ann. Med. 41, 451-461.

List of abbreviations

[P]	Particle concentration
^{1}H	Proton
A	Adenine
A	Additive genetic influence
ABCA1	ATP-binding cassette transporter A1
ABCG1	ATP-binding cassette transporter G1
ABCG5	ATP-binding cassette transporter G5
acetyl-CoA	Acetyl-coenzyme A
Apo	Apolipoprotein
APOB	Apolipoprotein B (gene)
BCAA	Branched-chain amino acid
C	Cytosine
C	Shared environmental influence
CAD	Coronary artery disease
CE	Cholesterol esters
CETP	Cholesterol ester transfer protein
CM	Chylomicrons
D	Dominance genetic influence
DHA	Docosahexaeonic acid
DILGOM	The Dietary, Lifestyle, and Genetic Determinants of Obesity and Metabolic Syndrome
DNA	Deoxyribonucleic acid
DZ	Dizygotic
E	Unique environmental influence
eQTL	Expression quantitative trait loci
FA	Fatty acid
FC	Free cholesterol
FID	Free induction decay
FT12	FinnTwin-12
FT16	FinnTwin-16
G	Guanine
GC	Gas cromatography
GWAS	Genome-wide association study
H2000	The Health 2000 GenMets
HBCS	Helsinki Birth Cohort Study
HDL	High-density lipoproteins
HL	Hepatic lipase
HPLC	High-performance liquid chromatography
ICC	Intraclass correlation
IDL	Intermediate-density lipoproteins

LA	Linoleic acid
LC	Liquid chromatography
LCAT	Lecithin-cholesterol acyltransferase
LD	Linkage disequilibrium
LDL	Low-density lipoproteins
LDLR	LDL receptor
LIPC	Hepatic lipase (gene)
LIPID	Lipid extracts
LIPO	Lipoprotein lipids
LMWM	Low-molecular-weight
LPL	Lipoprotein lipase
LRP	LDLR related proteins
MCI	Mild cognitive impairment
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MZ	Monozygotic
NFBC1966	Northern Finland Birth Cohort 1966
NMR	nuclear magnetic resonance
PL	Phospholipids
PLTP	Phospholipase transfer protein
PLTP	Phospholipase transfer protein (gene)
PROT	Protein
SNP	Single nucleotide polymorphism
SOM	Self-organizing map
SR-BI	Scavenger receptor class B member 1
T	Thymine
T2D	Type 2 diabetes
TC	Total cholesterol
TG	Triglycerides
TL	Total lipids
VLDL	Very-low-density lipoproteins
YF	The Cardiovascular Risk in Young Finns Study

Appendix I

Abbreviations and full names for the metabolites and derived traits studied in Publications III and IV.

Abbreviation	Full metabolite name
AcAcO	Acetoacetate
AcO	Acetate
Ala	Alanine
Alb	Albumin
bOHBuO	3-hydroxybutyrate
Cit	Citrate
Crea	Creatinine
DHA	22:6, docosahexaenoic acid (DHA)
Est-C	Esterified cholesterol
FAw3	Omega-3 fatty acids
Faw67	Omega-6 and -7 fatty acids
Faw9S	Omega-9 and saturated fatty acids
Free-C	Free cholesterol
Glc	Glucose
Gln	Glutamine
Glol	Glycerol
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein
HDL-C	Total cholesterol in HDL (from NMR)
HDL-C-lab	Enzymatically measured HDL-C
His	Histidine
IDL-C	Total cholesterol in IDL
IDL-FC	Free cholesterol in IDL
IDL-L	Total lipids in IDL
IDL-P	Concentration of IDL particles
IDL-PL	Phospholipids in IDL
IDL-TG	Triglycerides in IDL
Ile	Isoleucine
LA	18:2, linoleic acid (LA)
Lac	Lactate
LDL-C	Total cholesterol in LDL
LDL-C-lab	Enzymatically measured LDL-C
Leu	Leucine
L-HDL-C	Total cholesterol in large HDL
L-HDL-CE	Cholesterol esters in large HDL

L-HDL-FC	Free cholesterol in large HDL
L-HDL-L	Total lipids in large HDL
L-HDL-P	Concentration of large HDL particles
L-HDL-PL	Phospholipids in large HDL
L-LDL-C	Total cholesterol in large LDL
L-LDL-CE	Cholesterol esters in large LDL
L-LDL-FC	Free cholesterol in large LDL
L-LDL-L	Total lipids in large LDL
L-LDL-P	Concentration of large LDL particles
L-LDL-PL	Phospholipids in large LDL
L-VLDL-C	Total cholesterol in large VLDL
L-VLDL-CE	Cholesterol esters in large VLDL
L-VLDL-FC	Free cholesterol in large VLDL
L-VLDL-L	Total lipids in large VLDL
L-VLDL-P	Concentration of large VLDL particles
L-VLDL-PL	Phospholipids in large VLDL
L-VLDL-TG	Triglycerides in large VLDL
M-HDL-C	Total cholesterol in medium HDL
M-HDL-CE	Cholesterol esters in medium HDL
M-HDL-FC	Free cholesterol in medium HDL
M-HDL-L	Total lipids in medium HDL
M-HDL-P	Concentration of medium HDL particles
M-HDL-PL	Phospholipids in medium HDL
M-LDL-C	Total cholesterol in medium LDL
M-LDL-CE	Cholesterol esters in medium LDL
M-LDL-L	Total lipids in medium LDL
M-LDL-P	Concentration of medium LDL particles
M-LDL-PL	Phospholipids in medium LDL
MobCH	Double bond protons of mobile lipids
MobCH2	CH_2 groups of mobile lipids
MobCH3	CH_3 groups of mobile lipids
M-VLDL-C	Total cholesterol in medium VLDL
M-VLDL-CE	Cholesterol esters in medium VLDL
M-VLDL-FC	Free cholesterol in medium VLDL
M-VLDL-L	Total lipids in medium VLDL
M-VLDL-P	Concentration of medium VLDL particles
M-VLDL-PL	Phospholipids in medium VLDL
M-VLDL-TG	Triglycerides in medium VLDL
PC	Phosphatidylcholine and other cholines
Phe	Phenylalanine
PUFA	Other polyunsaturated fatty acids than 18:2
Pyr	Pyruvate
Serum-C	Serum total cholesterol (from NMR)
Serum-TG	Serum total triglycerides (from NMR)
S-HDL-L	Total lipids in small HDL
S-HDL-P	Concentration of small HDL particles
S-HDL-TG	Triglycerides in small HDL
S-LDL-C	Total cholesterol in small LDL
S-LDL-L	Total lipids in small LDL
S-LDL-P	Concentration of small LDL particles

SM	Sphingomyelins
S-VLDL-C	Total cholesterol in small VLDL
S-VLDL-FC	Free cholesterol in small VLDL
S-VLDL-L	Total lipids in small VLDL
S-VLDL-P	Concentration of small VLDL particles
S-VLDL-PL	Phospholipids in small VLDL
S-VLDL-TG	Triglycerides in small VLDL
TC	Enzymatically measured total cholesterol
TG	Enzymatically measured total triglycerides
Tot-C	Total cholesterol
Tot-CH	Total cholines (and other N-trimethyl compounds)
Tot-FA	Total fatty acids
Tot-PG	Total phosphoglycerides
Tot-TG	Total triglycerides
Tyr	Tyrosine
Urea	Urea
Val	Valine
VLDL-TG	Triglycerides in VLDL
XL-HDL-C	Total cholesterol in very large HDL
XL-HDL-CE	Cholesterol esters in very large HDL
XL-HDL-FC	Free cholesterol in very large HDL
XL-HDL-L	Total lipids in very large HDL
XL-HDL-P	Concentration of very large HDL particles
XL-HDL-PL	Phospholipids in very large HDL
XL-HDL-TG	Triglycerides in very large HDL
XL-VLDL-L	Total lipids in very large VLDL
XL-VLDL-P	Concentration of very large VLDL particles
XL-VLDL-PL	Phospholipids in very large VLDL
XL-VLDL-TG	Triglycerides in very large VLDL
XS-VLDL-L	Total lipids in very small VLDL
XS-VLDL-P	Concentration of very small VLDL particles
XS-VLDL-PL	Phospholipids in very small VLDL
XS-VLDL-TG	Triglycerides in very small VLDL
XXL-VLDL-L	Total lipids in chylomicrons and extremely large VLDL
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL
Abbreviation	**Full derived measure name**
AcO/AcAcO	Acetate to acetoacetate ratio
Ala/Cit	Alanine to citrate ratio
Ala/Glc	Alanine to glucose ratio
Ala/Gln	Alanine to glutamine ratio
Ala/His	Alanine to histidine ratio
Ala/Ile	Alanine to isoleucine ratio
Ala/Leu	Alanine to leucine ratio
Ala/Phe	Alanine to phenylalnine ratio
Ala/Pyr	Alanine to pyruvate ratio
Ala/Tyr	Alanine to tyrosine ratio
Ala/Val	Alanine to valine ratio
ApoA1	Apolipoprotein A-I (Lipido)

ApoB	Apolipoprotein B (Lipido)
ApoB/ApoA1	Apolipoprotein B by apolipoprotein A-I (Lipido)
BCAAs	Total branched chain amino acids; Val+Leu+Ile
Bis/DB	Ratio of bisallylic groups to double bonds
Bis/FA	Ratio of bisallylic groups to total fatty acids
bOHBuO/AcAO	3-hydroxybutyrate to acetoacetate ratio
bOHBuO/AcO	3-hydroxybutyrate to acetate ratio
CH2/DB	Average number of methylene groups per a double bond
CH2/FA	Average number of methylene groups in a fatty acid chain
Crea/Alb	Creatinine to albumin ratio
DB/FA	Average number of double bonds in a fatty acid chain
DHA/FAw3	Docosahexaenoic acid to omega-3 fatty acids ratio
DHA/PUFA	Docosahexaenoic acid to other polyunsaturated fatty acids than linoleic acid ratio
FALen	Description of average fatty acid chain length, not actual carbon number
FAw3/FAw67	Omega-3 fatty acids to omega-6 and -7 fatty acids ratio
FAw3/FAw9S	Omega-3 fatty acids to omega-9 and saturated fatty acids ratio
FAw3/FA	Ratio of omega-3 fatty acids to total fatty acids
FAw67/FAw9S	Omega-6 and -7 fatty acids ratio to omega-9 and saturated fatty acids ratio
FAw67/FA	Ratio of omega-6/7 fatty acids to total fatty acids
FAw9S/FA	Ratio of omega-9 and saturated fatty acids to total fatty acids
FR	Fischer's ratio; (Val+Leu+Ile)/(Phe+Tyr)
Free-C/Est-C	Free cholesterol to esterified cholesterol ratio
Glc/Cit	Glucose to citrate ratio
Glc/Pyr	Glucose to pyruvate ratio
Gln/Cit	Glutamine to citrate ratio
Gln/Glc	Glutamine to glucose ratio
Gln/His	Glutamine to histidine ratio
Gln/Ile	Glutamine to isoleucine ratio
Gln/Leu	Glutamine to leucine ratio
Gln/Phe	Glutamine to phenylalnine ratio
Gln/Pyr	Glutamine to pyruvate ratio
Gln/Tyr	Glutamine to tyrosine ratio
Gln/Val	Glutamine to valine ratio
Gp/Serum-TG	Glycoprotein acetyls to serum total triglycerides ratio
Gp/Tot-C	Glycoprotein acetyls to serum total cholesterol ratio
HDL2-C	Total cholesterol in HDL2 (Lipido)
HDL3-C	Total cholesterol in HDL3 (Lipido)
HDL-D	Mean diameter for HDL particles
His/Ile	Histidine to isoleucine ratio
His/Leu	Histidine to leucine ratio
His/Phe	Histidine to phenylalnine ratio
His/Tyr	Histidine to tyrosine ratio
His/Val	Histidine to valine ratio
IDL-C-eFR	Total cholesterol in IDL (Lipido)
Ile/Glc	Isoleucine to glucose ratio
Ile/Leu	Isoleucine to leucine ratio
Ile/Phe	Isoleucine to phenylalnine ratio
Ile/Serum-C	Isoleucine to serum total cholesterol ratio
Ile/Serum-TG	Isoleucine to serum total triglycerides ratio

Ile/Tyr	Isoleucine to tyrosine ratio
Ile/Val	Isoleucine to valine ratio
LA/DHA	Linoleic acid to docosahexaenoic acid ratio
LA/FAw67	Linoleic acid to omega-6 and -7 fatty acids ratio
LA/PUFA	Linoleic acid to other polyunsaturated fatty acids than linoleic acid ratio
Lac/Ala	Lactate to alanine ratio
Lac/Cit	Lactate to citrate ratio
Lac/Glc	Lactate to glucose ratio
Lac/Gln	Lactate to glutamine ratio
Lac/Pyr	Lactate to pyruvate ratio
LDL-C-eFR	Total cholesterol in LDL (Lipido)
LDL-D	Mean diameter for LDL particles
Leu/Glc	Leucine to glucose ratio
Leu/Phe	Leucine to phenylalanine ratio
Leu/Serum-TG	Leucine to serum total triglycerides ratio
Leu/Tyr	Leucine to tyrosine ratio
Leu/Val	Leucine to valine ratio
L-HDL-C/L-HDL-PL	Total cholesterol in large HDL to phospholipids in large HDL ratio
L-HDL-L/M-HDL-L	Total lipids in large HDL to total lipids in medium HDL ratio
L-HDL-L/S-HDL-L	Total lipids in large HDL to total lipids in small HDL ratio
M-HDL-C/M-HDL-PL	Total cholesterol in medium HDL to phospholipids in medium HDL ratio
M-HDL-L/S-HDL-L	Total lipids in medium HDL to total lipids in small HDL ratio
M-LDL-C/M-LDL-PL	Total cholesterol in medium LDL to phospholipids in medium LDL ratio
PC/Tot-CH	Phosphatidylcholine and other cholines to total cholines (and other N-trimethyl compounds) ratio
Phe/Tyr	Phenylalanine to tyrosine ratio
Phe/Val	Phenylalanine to valine ratio
Pyr/Cit	Pyruvate to citrate ratio
Serum-TG/Glc	Serum total triglycerides to glucose ratio
TG/PG	Ratio of triglycerides to phosphoglycerides
Tot-C/Est-C	Total cholesterol to esterified cholesterol ratio
Tyr/Val	Tyrosine to valine ratio
Val/Glc	Valine to glucose ratio
Val/Serum-TG	Valine to serum total triglycerides ratio
VLDL-D	Mean diameter for VLDL particles
VLDL-TG-eFR	Triglycerides in VLDL (Lipido)
XL-HDL-L/L-HDL-L	Total lipids in very large HDL to total lipids in large HDL ratio
L-HDL-L/M-HDL-L	Total lipids in very large HDL to total lipids in medium HDL ratio
XL-HDL-L/S-HDL-L	Total lipids in very large HDL to total lipids in small HDL ratio

Aalto-DD 16/2012

ISBN 978-952-60-4509-2
ISBN 978-952-60-4510-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Dept. of Biomedical Engineering and Computational Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**