Karsten Suhre *Editor*

Genetics Meets Metabolomics

from Experiment to Systems Biology

Springer

Genetics Meets Metabolomics

Karsten Suhre
Editor

Genetics Meets Metabolomics

from Experiment to Systems Biology

Springer

Editor
Karsten Suhre
Department of Physiology and Biophysics
Weill Cornell Medical College in Qatar
Education City – Qatar Foundation
Doha, State of Qatar

ISBN 978-1-4614-1688-3 e-ISBN 978-1-4614-1689-0 (eBook)
DOI 10.1007/978-1-4614-1689-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012939159

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Claire, Tijen and Amir, my loving family

Preface

Understanding the mechanisms that control human health and disease, in particular the role of genetic predispositions and their interaction with environmental factors is a prerequisite for the development of safe and efficient therapies for complex disorders, such as type 2 diabetes and cardiovascular disease. Over 100 years ago, Archibald Garrod already realized that inborn errors of metabolism are "*merely extreme examples of variations of chemical behavior which are probably everywhere present in minor degrees' and that this 'chemical individuality [confers] predisposition to and immunities from the various mishaps which are spoken of as diseases*" [1]. Recent advances in analytical technologies, in particular mass spectrometry, nuclear magnetic resonance spectroscopy, and liquid chromatography, have paved the way for the extensive characterization of a wide range of small molecules in many different types of biological samples. Aiming at the comprehensive and quantitative determination of ideally all key metabolites in a biological system, the emerging field of metabolomics has now joined ranks with other – *omics* technologies that benefitted from the development of high throughput measuring capabilities, such as genomics (next generation DNA sequencing) and transcriptomics (micro-arrays for mRNA expression analysis).

A limited number of books on metabolomics have been published in recent years. However, most focus on experimental questions and technical challenges of the field and are dedicated to a readership experienced in the field of bio-analytics. This book is complimentary to these specialist volumes as is centers on the application of metabolomics, with a special emphasis on the underlying genetics. Therefore the authors of this book took a more interdisciplinary approach. Their chapters address a wider readership of graduate students, postdoctoral researchers and experienced scientists from multiple domains. We hope that endocrinologists and biochemists with an interest in the genetics underlying metabolic phenotypes shall find new insights in the topics covered here, as shall geneticists who appreciate the fact that metabolic traits are more than just another set of quantitative variables to test for association. Most chapters shall also be accessible to clinical researchers who are neither specialists in genetics nor in biochemistry, but who wish to understand how

genetics and metabolomics come together in a systems-wide understanding of complex metabolic disorders. As promised in the title, all chapters of this book address exciting questions of where genetics meets metabolomics, all taking different viewpoints, ranging from experiment to systems biology. As this book is intended for an interdisciplinary readership, it contains aids for readers who are not deeply familiar with a particular domain, which are presented in the form of education boxes. These boxes describe central concepts that are known to researchers from one field, but that may be unfamiliar to others. We hope that this book shall inspire the new generation of researchers who address biological questions from a holistic point of view, combing genetics and metabolomics at all levels, from experiment to systems biology.

Doha, State of Qatar Karsten Suhre

Reference

1. Mootha VK, Hirschhorn JN (2010) Inborn variation in metabolism. Nat Genet 42:97–98

Contents

Contributors

Prof. Jerzy Adamski Helmholtz Zentrum München, Institute of Experimental Genetics, Genome Analysis Center, Neuherberg, Bavaria, Germany

Dr. Thure Adler German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Institute for Medical Microbiology, Immunology, and Hygiene, Technische Universität München, Munich, Germany

Dr. Luciana Afonso German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Juan Antonio Aguilar-Pimentel Department of Dermatology and Allergy, Biederstein, Clinical Research Division of Molecular and Clinical Allergotoxicology, TUM, Munich, Germany

Division of Environmental Dermatology and Allergy, Technische Universität München/Helmholtz Zentrum München, Neuherberg, Germany

Prof. Dr. Martin Hrabě de Angelis German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany

Anna Artati Helmholtz Zentrum München, Institute of Experimental Genetics, Genome Analysis Center, Neuherberg, Bavaria, Germany

Barbara M. Bakker Department of Pediatrics, Center for Liver, Digestive and Metabolic Diseases, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Michael P. Barrett Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

Dr. Lore Becker Friedrich-Baur-Institut, Department of Neurology, Ludwig-Maximilians-Universität München, Munich, Germany

German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Bavaria, Germany

PD. Dr. Johannes Beckers German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Bavaria, Germany

Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany

Prof. Dr. Raffi Bekeredjian Otto-Meyerhof-Zentrum, Department of Medicine III, Division of Cardiology, University of Heidelberg, Heidelberg, Germany

Dr. Alexander Bohla Comprehensive Pneumology Center, Institute of Lung Biology and Disease, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Rainer Breitling Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

Dr. rer. med Kathrin Budde Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany

Prof. Dr. Dirk H. Busch Institute for Medical Microbiology, Immunology, and Hygiene, Technische Universität München, Munich, Germany

Dr. Julia Calzada-Wack Institute of Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Christian Cohrs German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Hannelore Daniel Molecular Nutrition Unit, TUM, Freising, Germany

Saskia Decuypere Department of Parasitology, Unit of Molecular Parasitology, Institute of Tropical Medicine, Antwerp, Belgium

Jean-Claude Dujardin Department of Parasitology, Unit of Molecular Parasitology, Institute of Tropical Medicine, Antwerp, Belgium

Dr. Marc-Emmanuel Dumas Surgery and Cancer, Imperial College London, London, United Kingdom

Dr. Jack Favor Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. András Frankó German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Centerfor Environmental Health (GmbH), Neuherberg, Germany

Dr. Helmut Fuchs German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Valérie Gailus-Durner German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Lillian Garrett Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dominique Gauguier INSERM U872, Cordeliers Research Centre, Paris, France

Dr. Christian Gieger Helmholtz Center Munich – German Research Center for Environmental Health, Institute of Genetic Epidemiology, Neuherberg, Germany

Lisa Glasl Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Alexander Götz Comprehensive Pneumology Center, Institute of Lung Biology and Disease, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Jochen Graw Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Lining Guo Senior Director/Head of Project Management, Research and Development, Metabolon, Inc, Durham, NC, USA

Dr. Michael Hagn German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Wolfgang Hans German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. rer. med Robin Haring Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany

Dr. Sabine M. Hölter Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Marion Horsch German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Prof. Thomas Illig Research Unit of Molecular Epidemiology, Neuherberg, Bavaria, Germany

Hannover Unified Biobank, Hannover Medical School, Hannover

Melanie Kahle German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Gabi Kastenmüller Helmholtz Zentrum München, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany

Prof. Dr. Hugo Katus Otto-Meyerhof-Zentrum, Department of Medicine III, Division of Cardiology, University of Heidelberg, Heidelberg, Germany

Martin Kistler German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Tanja Klein-Rodewald Institute of Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Martin Klingenspor Molecular Nutritional Medicine, Else Kröner-Fresenius Center and ZIEL Research Center for Nutrition and Food Sciences, Technische Universität München, Freising, Weihenstephan, Germany

Prof. Dr. Thomas Klopstock Friedrich-Baur-Institut, Department of Neurology, Ludwig-Maximilians-Universität München, Munich, Germany

Dr. Thomas Kopf Department of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Bavaria, Germany

Florian Kronenberg Division of Genetic Epidemiology, Innsbruck Medical University, Innsbruck, Austria

Jan Krumsiek Helmholtz Zentrum München, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany

Dr. Christa Kühn Department of Molecular Biology, Leibniz Institutefor Farm Animal Biology (FBN), Dummerstorf, Germany

Dr. Kay A. Lawton Research and Development, Metabolon, Inc, Durham, NC, USA

Dr. Christoph Lengger German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. rer. nat. Marianna Lucio Helmholtz Zentrum München, Research Unit Analytical BioGeoChemistry, Neuherberg, Bavaria, Germany

Tonia Ludwig German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Holger Maier German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Susan Marschall German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Jonathan E. McDunn Oncology Research and Development, Metabolon, Inc, Durham, NC, USA

Kateryna Micklich German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Michael V. Milburn Research and Development, Metabolon, Inc, Durham, NC, USA

Gabriele Möller Helmholtz Zentrum München, Institute of Experimental Genetics, Genome Analysis Center, Neuherberg, Bavaria, Germany

Dr. Beatrix Naton German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Frauke Neff Institute of Pathology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Susanne Neschen German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Markus Ollert Department of Dermatology and Allergy, Biederstein, Clinical Research Division of Molecular and Clinical Allergotoxicology, TUM, Munich, Germany

Dr. Markus Peer Department of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Bavaria, Germany

Cornelia Prehn Helmholtz Zentrum München, Institute of Experimental Genetics, Genome Analysis Center, Neuherberg, Bavaria, Germany

Dr. Oliver Puk Institute of Developmental Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

PD. Dr. Ildikó Rácz Institute of Molecular Psychiatry, University of Bonn, Bonn, Germany

Dr. Michael Räß German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Birgit Rathkolb German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Chair for Molecular Animal Breeding and Biotechnology, Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany

Dr. Jan Rozman German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Molecular Nutritional Medicine, Else Kröner-Fresenius Center and ZIEL Research Center for Nutrition and Food Sciences, Technische Universität München, Freising, Weihenstephan, Germany

CEO John A. Ryals Metabolon, Inc, Durham, NC, USA

Manuela Sailer Molecular Nutrition Unit, TUM, Freising, Germany

Markus Scheerer German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Evelyn Schiller German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

PD. Dr. Philippe Schmitt-Kopplin Helmholtz Zentrum München, Research Unit Analytical BioGeoChemistry, Neuherberg, Bavaria, Germany

Dr. Prof. Gerd Schmitz Department of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Bavaria, Germany

Dr. Felix Schöfer German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Anja Schrewe German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Holger Schulz Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Ralph Steinkamp German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Claudia Stöger German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. Tobias Stöger Comprehensive Pneumology Center, Institute of Lung Biology and Disease, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Ferdinand Stückler Helmholtz Zentrum München, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany

Karsten Suhre Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City – Qatar Foundation, Doha, State of Qatar

Dr. Prof. Fabian J. Theis Helmholtz Zentrum München, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany

Irina Treise German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dimitrios Tziotis Helmholtz Zentrum München, Research Unit Analytical BioGeoChemistry, Neuherberg, Bavaria, Germany

Dr. Sibylle Wagner German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Dr. med. Prof. Henri Wallaschofski Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Mecklenburg-Vorpommern, Germany

Monja Willershäuser German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. David Wishart Department of Computing Science and Biological Sciences, University of Alberta, Edmonton, AB, Canada

Michael Witting Helmholtz Zentrum München, Research Unit Analytical BioGeoChemistry, Neuherberg, Bavaria, Germany

Prof. Dr. Eckhard Wolf Chair for Molecular Animal Breeding and Biotechnology, Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany

Annemarie Wolff-Muscate Institute of Developmental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Wolfgang Wurst Institute of Developmental Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Chair of Developmental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany

Dr. Ali Önder Yildirim Comprehensive Pneumology Center, Institute of Lung Biology and Disease, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Ramona Zeh German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany

Prof. Dr. Andreas Zimmer Institute of Molecular Psychiatry, University of Bonn, Bonn, Germany

Chapter 1
Introduction

Karsten Suhre

1 Chapters on Experiment Related Questions

A key element of any metabolomics study is the availability of high quality samples, ideally from a large number of biosamples. In *"Pre-conditions for high quality biobanking in large human epidemiological cohorts for metabolomics and other -omics studies"*, Thomas Illig highlights methods for high quality preservation of samples for later application of systematic molecular analyses like genomics, epigenomics or metabolomics (→ technology: biobanking).

The practical aspects of experimental metabolomics methods and questions that may determine a particular study design are provided by Jerzy Adamski and coworkers in their chapter *"Assay Tools for Metabolomics"*. They describe the different steps that are required for the successful collection of large quantitative metabolomics data sets (→ technology: high throughput mass spectrometry).

The selection of proper statistical tools for the analysis metabolomics data in combination with genetic variance is of utmost importance for the meaningful identification of associations between genotype and metabolic phenotype, as explained by Christian Gieger in his chapter on *"Statistical methods in genetic and molecular epidemiology and their application in studies with metabolic phenotypes"* (→ concept: genetic association).

K. Suhre (✉)
Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City – Qatar Foundation,
PO Box 24144, Doha, State of Qatar
e-mail: karsten@suhre.fr

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_1,

2 Chapters on Unicellular Organisms and Animal Models

The majority of cells in a human body are of microbial rather than human origin. Metabolic conversions performed by these bacterial communities closely interact with human metabolism. It is therefore not surprising that the human microbiome, such as of the intestines and on the skin, is influencing human health. The investigation of these processes in microbial communities is what Schmitt-Kopplin and co-workers term the *"meta-metabolome"* in their chapter on *"Ultrahigh resolution mass spectrometry based non-targeted microbial metabolomics"* (→ technology: ultrahigh resolution mass spectrometry; application: microbial metabolomics).

Using exact-mass mass spectrometry, Rainer Breitling and colleagues present in *"Metabolomic systems biology of protozoan parasites"* two case studies of metabolomic systems biology on two major protozoan pathogens, the African trypanosome Trypanosoma brucei, causative agent of sleeping sickness, and the Leishmania donovani parasites, responsible for visceral leishmaniasis (→ application: protozoan metabolomics).

Genetically modified mice are widely used as a model organism to study human diseases: mice are easy to handle and breed, there exist inbred strains, and the mouse genome sequence is available. However, not all genetically modified mice exhibit a clear disease phenotype when simply kept and fed in a cage. Therefore, defining appropriate challenges to induce disease phenotypes has become a major focus in current mouse studies. In *"Mouse genetics and metabolic mouse phenotyping"*, Helmut Fuchs, Martin Hrabě de Angelis and co-authors from the German Mouse Clinic argue the case of metabolomics to be used as a comprehensive phenotyping tool in such challenge experiments with genetically modified mice (→ concept: challenge experiments in mouse models).

Although at most times unknowingly, favourable genetic traits in livestock have been selected for since prehistoric times. Due to partial inbreeding, farm animals are also a great resource for genetic studies and complementary to fully in-bred animal models. Christa Kühn presents examples of application of *"Metabolomics in animal breeding"* and how metabolomics provides a new tool for optimisation of selection (→ concept: metabolomics assisted breeding).

3 Chapters on Human Health Related Topics

After highlighting the potential of metabolomics in bacterial and protozoan organisms, genetically inbred mice, and partially inbred livestock, the remaining chapters focus on human biology and its disorders. Differences in lifestyle are, to a large part, reflected in nutritional habits, and metabolomics appears to be a tool of choice in this area. However, Hannelore Daniel and Manuela Sailer argue that metabolomics in human nutrition research is still in its infancy. In their chapter *"Metabolomics applications in human nutrition"* they describe that two research tracks are presently emerging: assessment of food intake by identifying and quantifying marker

metabolites that originate from the intake of individual food components to provide better tools for assessing human food consumption, and characterizing the metabolic responses to dietary challenges to better define the health–disease relationship (→ concept: nutritional challenge experiments; application: human nutrition).

Syndromic diseases are characterized by their resistance to present a clear and unifying medical picture. A wider and mostly non-targeted phenotypic characterization of a large number of cases may eventually allow the identification of markers that help to obtain a better understanding of the underlying pathophysiology. Metabolomics provides access to measuring the "true" endpoints of biological processes and thereby promises to be a valuable tool in the study of syndromic diseases. Robin Haring, Kathrin Budde and Henri Wallaschofski present a concrete example of how metabolomics can help solving such problems in their chapter *"Metabolomics for the individualized therapy of androgen deficiency syndrome in male adults"* (→ concept: individualized therapy; application: syndromic diseases).

The full extent and complexity of the human metabolome is far from being understood, as explains David Wishart in *"Systems biology resources arising from the human metabolome project"*. This chapter provides a series of dedicated databases on endogenous and exogenous (foods, drugs) small molecules that provide valuable and often hand-curated information on most known metabolites, including their link to genetic variance in enzyme coding genes. He presents a hands-on example of how to use these online resources at the text-book example of the genetic disorder phenylketonuria (→ technology: databases; application: in-born errors of metabolism).

The identification of metabolic biomarkers for human disease is one of the central goals of clinical metabolic research. In *"Understanding cancer metabolism through global metabolomics"*, Mike Milburn and co-workers describe examples of how changes in metabolic profiles are used to identify cancer-related mutations in isocitrate dehydrogenase (IDH) genes. In their chapter the authors also describe a specific implementation of a multi-platform non-targeted metabolomics platform (→ technology: non-targeted multiplatform mass spectrometry; concept: metabolic biomarker; application: cancer).

Lipidomics is a major sub-field of metabolomics. Lipids are particularly intriguing due to the roles that play different levels of fatty acid chain length and desaturation in the aetiology of complex disorders. In *"Genetic and metabolic determinants of fatty acid chain length and desaturation, their incorporation into lipid classes and their effects on risk of vascular and metabolic disease"*, Thomas Kopf, Markus Peer and Gerd Schmitz provide an extensive and comprehensive overview of this exciting field (→ concept: lipidomics; application vascular and metabolic disease).

4 Chapters That Take a Systems Approach

Systems biology strategies to enhance the biological interpretation of haplotype – metabotype association networks derived from mQTL studies can provide a better understanding of pathophysiological mechanisms. Marc Dumas and Dominique

Gauguier introduce this concept in their chapter *"Mapping metabolic quantitative trait loci (mQTL) – a link between metabolome-wide association studies and systems biology"* (→ technology: nuclear magnetic resonance spectrometry; concept: metabolic quantitative trait loci).

Complex diseases such as coronary heart disease and type 2 diabetes mellitus are influenced by a large number of genes and environmental factors. In most cases the contribution of a single gene is small. Intermediate phenotypes, which are closer to the genetic cause of the disease deliver stronger biological read-outs and thereby allow drawing new conclusions on the functional background of the genetic variant. In his chapter *"Metabolic traits of intermediate phenotypes"*, Florian Kronenberg argues for the potential of metabolomics to provide a wide range of biologically relevant intermediate phenotypes (→ concept: intermediate phenotypes).

As shown in the previous chapters, disturbances in metabolism are at the root of a variety of human afflictions and complex diseases. In his chapter *"Genome-wide association studies with metabolomics"* Karsten Suhre shows how combining two highly sophisticated biochemical measurement methods – genetics and metabolomics – in genome-wide association studies can reveal deep insights into the genetic makeup of the human body's metabolic capacities (→ technology: genome-wide association studies; concept: the genetically determined metabotype).

In the preceding chapters many aspects of studies with metabolomics in relation to genetic and disease phenotypes have been described. Eventually, *"Systems biology meets metabolism"* in the final chapter with the same title by Jan Krumsiek and colleagues. The authors describe this new paradigm that is becoming increasingly popular, namely that of integrating data from multiple analyses into larger models. This paradigm is nowadays known as systems biology, and is expected to penetrate many classical molecular analyses.

The area of research where genetics and metabolomics meet is likely to represent a field where systems biology shall prosper highly in the years to come. We hope that the following chapters shall provide a thorough basis for the understanding of the underlying experimental techniques, concepts and potential biomedical applications of this exciting field.

Chapter 2
Pre-conditions for High Quality Biobanking in Large Human Epidemiological Cohorts for Metabolomics and Other – Omics Studies

Thomas Illig

1 New Generation, Large-Scale Cohort Studies Around the World

A first generation of large-scale prospective cohorts aiming to study cancer and other chronic diseases, that included biobanks with blood and/or urine samples, were initiated in the late 1980s and early 1990s, in the USA and Europe. In Europe, the largest study to date is the European Prospective Investigation into Cancer and Nutrition (EPIC) – a multi-centre cohort that includes a total of over 420,000 men and women in ten Western European countries, who all provided questionnaire data, anthropometric measurements and a blood sample at baseline. The recruitment of subjects into the EPIC study was completed mostly between 1992 and 1998, and this project is currently in a high-phase of scientific production, on a wide variety of issues and diseases (see website http://www.iarc.fr/epic). Currently, prospective biobank studies of a similar size, including 300,000–500,000 subjects, have started in the UK [1, 2], Sweden, Japan, Western Australia, the USA and Germany [3].

T. Illig, Ph.D. (✉)
Research Unit of Molecular Epidemiology,
Ingolstaedter Landstrasse 1, Neuherberg, Bavaria, Germany

Hannover Unified Biobank, Hannover Medical School, 30625, Hannover
e-mail: illig@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_2,

2 The German National Cohort (GNC): An Example for High Quality Sample Collection in a Large Epidemiological Setting

The German National Cohort will start in 2012 and will include a total of 200,000 women and men in the age range of 20–69 years (40,000 individuals under age 40 and 160,000 individuals over age 40 years), who will be recruited through a network of 18 local study centers spread all over North, East, South, West, and central Germany.

3 Collection and Storage of Biological Materials in the GNC to Reach High Sample Quality

A key element of the study protocol in the GNC is the collection of high quality samples from all study participants.

For blood, preservation of quality relies on minimizing pre-analytical artefacts that may be incurred during specimen collection, primary processing, transport and/or storage of the samples, including:

- Artefacts due to cell lysis leading to release of intracellular components that have concentrations several magnitudes higher in the intracellular as compared to the extracellular compartment, exemplified by release of potassium, LDH or catecholamines from red blood cells in haemolysis, or of proteolytic enzymes from leukocytes, which not only alters their serum or plasma concentration but may also degrade target analytes such as insulin;
- Artefacts due to cell metabolism, exemplified by the decrease in glucose concentration upon prolonged storage of blood or the continuing in-vitro production by cells of the amino acid homocysteine that has received attention as a marker of cardiovascular risk;
- Artefacts due to the enzymatic degradation of molecular species upon prolonged exposure to 4°C or higher;
- Molecular artefacts due to repeated freezing and thawing of stored biomaterials.

Given the huge number of potential analytes and taking into account that both analytes of interest and techniques may change over the long run of this study to an extent that cannot be foreseen today, avoidance of any artefacts is mandatory. This requires:

- The prompt and complete separation, ideally within 1 h of collection, of all particulate components of full blood to obviate the above detailed cell-derived artefacts.
- No delay in the aliquotation and freezing to obviate enzymatic degradation during prolonged transportation at 4°C or higher;
- Volumes small enough (190 μl) to guarantee single use only as opposed to repeat thaw-freeze cycles necessarily implicated in the storage of larger volumes.

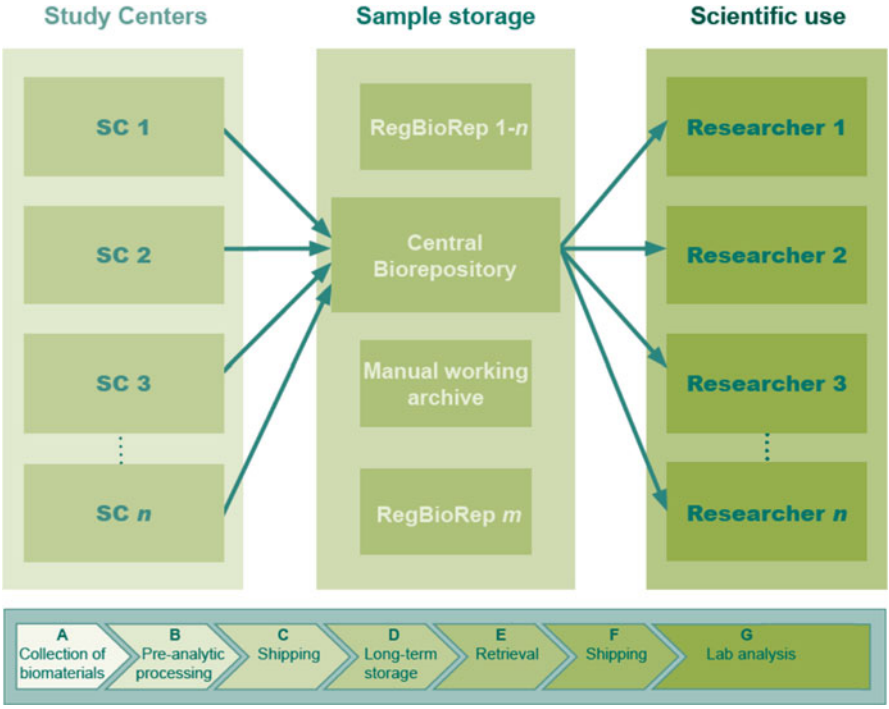


Fig. 2.1 Local processing and central storage of biomaterials including back-up storage in the German National Cohort (GNC) (SC = local study center; RegBioRep = Regional back-up storage facility; Central biorepository = Central automated −80°C biorepository; Manual working archive = Central manual storage facility in gas phase of liquid nitrogen tanks)

To reach the goal of high quality and wealth samples the following quality issues are applied in the GNC.

- Local processing of blood, urine, and other biomaterials, as opposed to a centralized one, complete enough to reach the level of ready-prepared small aliquots that can be transported to the central store on dry ice in the deep-frozen state (except for viable blood cells), which obviates the enzymatic disintegration incurred upon prolonged exposure to 4°C or higher (Fig. 2.1).
- Adherence to stringent standard operation procedures (SOPs) in all study centres (e.g. fast separation of cells from plasma/serum).
- Automation of all steps in preparation, storage, and retrieval of stored materials, promoting strict adherence to standard operation procedures (SOPs), maximizing reproducibility, and obviating artefacts that in manual processing inevitably occur on the long run due to individual failure. Thus, each of the 18 study centres will be equipped with a liquid handling platform.
- Storage of biomaterials from all participants throughout Germany in one central automated bio-repository and decentralized back-up storage (Fig. 2.2).
- Storage of many but rather small volume aliquots to avoid freeze thaw cycles and thereby to increase sample quality (Table 2.1).

Fig. 2.2 Large automated −80°C biorepository

Table 2.1 Biomaterials generated from blood in the German national cohort and stored volume sizes

Primary material	Volume	Processed material	Aliquots and storage at −180°C (unless stated otherwise)
Blood + clot activator	2 × 10 ml	Serum	30 × 0.19 ml
Blood + EDTA	2.0 ml	EDTA blood → hematology	Local lab., within < 6 h
	3 × 10 ml	EDTA plasma	48 × 0.19 ml
		EDTA-packed erythrocytes	6 × 0.19 ml
		Buffy coat + 90% of red cell layer for DNA extraction	1 × 9.0 ml (in manual −80°C freezers)
Blood + RNAse inhibitors	2.5 ml	RNA (Tempus/PAXgene)	1 × 2.5 ml (in manual −80°C freezers)
Blood in BD CPT	10 ml	Ficoll-isolated PBMC + DMSO	4 × 2 Mio PBMC (in liquid nitrogen vapor phase)
Total blood	**65 ml**		**92 aliquots stored**

Fig. 2.3 Sample storage in liquid nitrogen tanks

- Equipment with a professional laboratory information system (LIMS) for both the local and centralized pre-analytics and storage facilities to monitor the whole process from blood collection to sample storage and later sample retrieval.
- Storage of most blood and urine samples in gas phase of liquid nitrogen (Fig. 2.3).
- Diverse range of sample types (Table 2.2).

4 Sample Types Collected in the German National Cohort

It is planned to collect a broad range of high quality materials for future molecular research like metabolomics. It is planned to collect and store the following materials (a rationale for each sample material is given in Table 2.2):

- Blood and blood derivates (serum, EDTA plasma, erythrocytes, DNA, RNA, living cells from blood)
- Spot urine
- Faeces
- Nasal swabs
- Saliva

Current plans foresee storage facilities based on a combination of liquid nitrogen freezers for long-term storage and −80°C electric freezers for storage of samples

Table 2.2 Biological samples: rationale for inclusion

Sample type	Rationale	Collection
Blood	Different fractions available (plasma, serum, white, cells, red cells, and peripheral blood mononuclear cells)	Easy, low risk, well established, standardized procedures available
	Different types of components present (cells, proteins, DNA, RNA, hormones, nutrients, etc.)	Low costs of collection, except for RNA collection tubes and BD 'CPT' tubes for PBMC separation
	Suitable for a wide range of analytic procedures, including –omics technologies	
Urine	Different types of components present (cells, proteins, renal excretion products, etc.)	Easy, well-established, standardized procedures available
	Suitable for several analytic procedures, including –omics technologies	Low costs
	Provides additional (supplemental or new) information to blood samples	
Saliva	Provides specific information about oral microbiota	Issue of standardization of sample collection
	Different types of components present (cells, proteins, DNA, hormones, etc.)	Specific logistics for sample collection necessary
	Suitable for several analytic procedures, including –omics technologies	
Stool	Provides specific information about gut microbiota	Issue of standardization of sample collection
	Suitable for several analytic procedures, including –omics technologies	Specific logistics for sample collection necessary
Swabs	Provides specific information about nasopharyngeal microbiota	Standardized collection according to SOPs

that are foreseen to more imminently and frequently used for research projects. The biobanks will be equipped with a professional laboratory information system (LIMS) with combined database for both the local and centralized storage facilities. The use of freezer systems with sample automated retrieval mechanisms is under study.

5 Conclusion

The optimal collection of biomaterials for future molecular analysis like metabolomics represents a basic precondition for every large study population. Such molecular phenotyping indispensably requires a comprehensive array of biomaterials in optimal quality.

References

1. Elliott P, Peakman TC (2008) The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. J Epidemiol 37(2):234–244
2. Burton PR, Tobin MD, Hopper JL (2005) Key concepts in genetic epidemiology. Lancet 366(9489):941–951
3. Wichmann HE, Gieger C (2007) Biobanks. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 50(2):192–199

Chapter 3
Assay Tools for Metabolomics

Anna Artati, Cornelia Prehn, Gabriele Möller, and Jerzy Adamski

Abbreviations

AMU	Atomic mass unit
APCI	Atmospheric pressure chemical ionization
APPI	Atmospheric pressure photoionization
BMI	Body mass index
CE	Capillary electrophoresis
CE-MS	Capillary electrophoresis mass spectrometry
CI	Chemical ionization
CID	Collision-induced dissociation
EI	Electron impact ionization
ESI	Electrospray ionization
FIA	Flow injection analysis
FIA-MS	Flow injection analysis mass spectrometry
FT-ICR-MS	Fourier transform ion cyclotron resonance mass spectrometry
FT-IR	Fourier transform infrared spectrometry
GC	Gas chromatography
GC-MS	Gas chromatography mass spectrometry
GWAS	Genome-wide association studies
HMDB	Human metabolome database
HPLC	High performance liquid chromatography
IUPAC	International union of pure and applied chemistry
KEGG	Kyoto encyclopedia of genes and genomes

A. Artati, Ph.D. • C. Prehn, Ph.D. • G. Möller, Ph.D. • J. Adamski, Ph.D., M.Sc. (✉)
Helmholtz Zentrum München, Institute of Experimental Genetics, Genome Analysis Center, Ingolstädter Landstraße 1, Neuherberg, Bavaria, 85764, Germany
e-mail: anna.artati@helmholtz-muenchen.de; prehn@helmholtz-muenchen.de; gabriele.moeller@helmholtz-muenchen.de; adamski@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_3,

LC	Liquid chromatography
LC-MS	Liquid chromatography mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LIMS	Laboratory information and management system
LLOQ	Lower limit of quantification
LOD	Limit of detection
LOQ	Limit of quantification
m/z	Mass to charge ratio
MRI	Magnetic resonance imaging
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NMR	Nuclear magnetic resonance
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCI/NCI	Positive chemical ionization/negative chemical ionization
RF	Random forest
RP18	Reversed phase C18 alkyl chain modified silica
SIM	Single ion monitoring
SOP	Standard operating procedure
SPE	Solid phase extraction
SRM	Selected reaction monitoring
UHPLC	Ultra high performance liquid chromatography
UHPLC-MS	Ultra high performance liquid chromatography mass spectrometry
ULOQ	Upper limit of quantification

1 Metabolomics Is a Multidisciplinary Approach

1.1 Why Do Metabolomics?

Diversity of life takes place at different molecular levels, among others at the level of nucleic acids, proteins and metabolites. These molecules can be taken as readout of the status of a biological system.

Whereas the DNA-world (studied by *genomics*) can provide a lot of information on the potential function of genes and thereby supports the prediction of their functions, it is not very easy to derive real functional or dynamic data from it. The RNA-world (*transcriptomics*) and the protein-world (*proteomics*) both reflect much more the dynamics of a living organism. However, present technologies restrict the number of target molecules that are simultaneously discernible, so that a whole picture is not complete. Furthermore, many transcripts are translated into more than one protein, and many proteins are only functional in complexes with other proteins. As a result, genomics, transcriptomics, and proteomics merely indicate the cause of a

phenotypic response, but they are very limited in predicting what will actually happen in the whole organism. Their predictions are only clear in case of mutations inactivating genes.

The metabolite-world (*metabolomics*) reflects all functional activities, transient effects, as well as endpoints of biological processes determined by the sum of its genetic features, regulation of gene expression, protein abundance, and environmental influences. The original concept of metabolomics, i.e. measuring small metabolites in a body fluid, was pioneered by Linus Pauling in 1971 [1]. Although DNA-processing events like splicing, which result in different RNA products, can predict the number of consequently synthesized proteins [2, 3], the metabolome cannot be computed from the genome. In addition, changes in the metabolome happen even faster than those in the RNA [4].

The analytical methods of metabolomics reached a high level of sensitivity, dependable reproducibility, wide metabolite coverage, and high sample throughput. Hence, this research area is of interest to those studying the mechanisms of health and disease, nutrition effects, monitoring of biotechnological processes, or performing crop and food quality analyses. In this respect, metabolomics provides qualified large data sets required for systems biology approaches. As systems biology can generate hypotheses but not prove them, metabolomics is an excellent choice for the experimental verification.

1.2 Definitions

Metabolomics investigates metabolite homeostasis in health or analyzes dynamic metabolic responses of biological systems (cell, tissue, organism) to environmental challenges, toxic stimuli, genetic modifications or diseases [5, 6]. Beside the term *metabolomics* for the study of the metabolome (meaning all metabolites of a biological system) [7], the expression *metabonomics* [8] is used for the analysis of metabolites in challenged living systems, e.g. in an organism treated with a drug. Actually, both terms are used almost equivalently and employ similar analytical methods and data processing procedures. A more selective terminology is applied for metabolomics of specific chemical classes, i.e. *lipidomics* for the studies of lipids [5, 9] or *steromics* when steroids are analyzed [10]. Metabolites are often called *analytes* (for metabolites being analyzed) or *compounds* (especially if dealing with structures, formulas, and properties) Box 3.1.

1.3 Challenges of Metabolomics

Metabolomic research is performed to reach multiple aims such as to:

- Detect as many metabolites as possible
- Identify and annotate metabolites

Box 3.1 Metabolomics Terms

Metabolomics is the identification and quantification of ideally all metabolites in a biological system (cell or organism) to depict health homeostasis or a dynamic metabolic response to environmental challenges, toxic stimuli, genetic modifications, or diseases.
Metabonomics performs analyses of metabolites in biological systems challenged by a drug.
Profiling metabolomics looks for the catalogue of measurable metabolites.
Non-targeted metabolomics identifies differences between samples.
Targeted metabolomics quantifies selected sets of metabolites.
Analyte is a compound or metabolite being analyzed.
A ***biomarker*** is a molecule or a feature used to monitor a biologic process. It should be easily and reliably measurable so that a specific and robust quantification is provided. An example for a biomarker is the concentration of cholesterol used to monitor the effect of a lipid-lowering drug.
A biological ***matrix*** is a specific type of biological specimen, like a body fluid, tissue, breath air, or even a cell pellet, which contains the metabolites that shall be analyzed.
A ***phenotype*** is a specific characteristic feature. Examples are human weight, metabolite concentration or cell division rate in cell culture, quantified in measurable units like kg, mM or frequency, respectively.
Comorbidity describes the presence of a feature or condition existing simultaneously but independently to the observed phenotype. Comorbidity in medicine is either the presence or effect of one or more diseases in addition to a primary disease or disorder.
The limit of detection ***(LOD)*** describes the lowest analyte concentration that can be detected.
The lower limit of quantification (***LLOQ***) is defined as 10 times the standard deviation of the matrix (blank) and the upper limit of quantification (***ULOQ***) is experimentally defined by accuracy and linearity tests using spiked matrix samples.
The standard operating procedure ***(SOP***) is a written document depicting all requirements, details, steps and activities of a process to achieve uniformity of performance.

- Quantify metabolites
- Distinguish native metabolites from impurities or artifacts
- Determine indicative metabolites (biomarkers) for a given process
- Catalog metabolites (signatures) of organisms, tissues and organelles
- Resolve spatial or temporal metabolomes
- Provide metabolic phenotypes to genome-wide association studies (GWAS)

- Predict and/or analyze metabolic pathways
- Identify cross-talks between pathways
- Analyze mechanisms of disease or drug action
- Facilitate diagnostics.

Very often these aims cannot be realized simultaneously because of technological restrictions. The sizes of different metabolomes have been estimated to be from around 200,000 distinct metabolites in plants to a several fold smaller number in humans [11]. The process of annotation (cataloging) is still ongoing. The Human Metabolome Data Base (HMDB) reached around 8,000 different compounds [12] and the initiative LIPID MAPS refers to more than 9,000 different molecules for molecular lipids only [13]. However, not all metabolites are unequivocally annotated as yet. Identification of the molecular characters of compounds one by one is a tedious process and has to be parallelized in many laboratories worldwide. The needed approaches require extreme accuracy like that provided by ultra high performance liquid chromatography coupled with Fourier transform ion cyclotron resonance mass spectrometry (UHPLC-FT-ICR-MS) rather than fast analyses [14]. On the other hand, signature or biomarker search requires wide-scope profiling analtics as done by nuclear magnetic resonance (NMR) [15] or liquid chromatography mass spectrometry (LC-MS) [16].

Metabolomic analyses in biological samples detect endogenous metabolites, peptides, xenobiotics, dietary constituents and agents of environmental exposure [17]. The technical approaches for metabolomics are universally applicable in distinct species and have been successful in studies of yeasts [18], plants [19], mouse [20], human nutritional challenges [21], microbe-host interactions [22], natural products research [23], or even extraterrestrial organic matter [24]. Not only can metabolomic signature characteristics for a given biological process, like e.g. apoptosis [25] be depicted but the kinetics of biochemical pathways and metabolite conversions (called *flux*) can be monitored also [26]. Metabolomics has even been combined with genome-wide association studies (GWAS) in which new *genetically determined metabotypes* were discovered in humans [27]. Drug development and clinical trials profited from the contribution of metabolomics to the discovery of biomarkers of specific processes [28]. Metabolomics is a versatile tool in diagnostics [29–32].

1.4 Critical Elements of Metabolomics

Metabolomics is an integrative science. In order to deliver sound results it demands implementation of a set of non-separable and strict rules such as:

- Proper experimental design
- Standardized sample processing
- Versatile analytical methods
- Large scale bioinformatics

1.4.1 Experimental Design

Building up the experimental design one should consider which aim depicted in 1.3 is of major importance. This aim pre-defines requirements for the sample logistics and analytics. For animal models kept under controlled conditions (night/day cycle, diet, genetic background), reproducible experimental conditions (treatment, genetic modifications) are relatively easy to achieve. By the use of isogenic or inbred strains, the genetic variation can be reduced. Thereby, the experimenter can focus on treatment parameters, e.g. only one drug treatment over a time, or ageing over time. The experiment could even be repeated if required. On the contrary, in human population studies, it is more complicated to control the experimental setup. The major challenge is that human samples could be unique and no more recoverable, i.e. only a few experiments are possible. Ideally, several parameters like age, gender, race (or geographic origin), body mass index (BMI), circadian rhythm, nutrition and life style (e.g. nicotine or alcohol consumption, physical activity), medication and hormonal status (pregnancy, birth control) should be matched. In most cases, including clinical trials for drug efficacies, matching of these parameters is a challenge and often only a few hundred cases can be adequately matched from an initially large population of several thousands of participants. Without careful matching, additional variability is present as *comorbidity* disturbing the scientific outcome of experiments, i.e. it is not clear if the unmatched parameter(s) or the challenge cause the changes in the metabolome. Moreover, the experimental design should involve application of standard operating procedures (SOPs, as explained in Sect. 3.3) and well established analytical methods (referred to as *golden standard*) to ensure reproducibility and general applicability of the results.

1.4.2 Sample Preparation

An optimal metabolomics procedure would be to study the whole metabolism in an intact living organism using a non-invasive approach. Although such methods exists, e.g. magnetic resonance imaging (MRI), they can monitor only a small subset of metabolites [33]. Therefore, prior to analytics, the living system must be sampled.

Collection

The preparation includes sample collection and storage as well as metabolite extraction and preparation for analytics. In case of sample collection it has to be ensured that metabolites in the sample remain exactly the same as at the time of sampling. This is usually achieved by a “collect and freeze” procedure, e.g. by blood withdrawal, plasma preparation and subsequent storage at −80°C or by tissue fractionation for organelle isolation followed by immediate sample freezing. Under these conditions some metabolites are quite stable (e.g. amino acids) but some may decompose more quickly (e.g. diacylglycerols and phosphatidylethanolamines)

[34, 35]. For some short-living or reactive metabolites it might even be essential to add stabilizers prior to freezing (e.g. 2,6-di-tert-butyl-4-methylphenol (BHT) for plasma eicosanoids to prevent auto-oxidation [36]). Repetitive freeze-thaw cycles may influence metabolite concentrations and should be avoided by aliquoting the samples before freezing.

Extraction

Most samples (except for NMR-studies) undergo preparations to facilitate metabolite analytics. If proteins impair subsequent separation or analytics, they are removed by precipitation in a first step. The next step, the metabolite extraction (e.g. liquid-liquid or solid phase extraction) will greatly influence the variety of metabolites seen by the analytical methods. The use of isotonic phosphate buffered saline (PBS) will result in an extraction of hydrophilic metabolites, whereas the use of 90% methanol in water will foster the extraction of more hydrophobic compounds [37]. Consequently, in the first case more amino acids and in the latter case more lipids will be extracted. Very hydrophobic metabolites may even demand the addition of organic solvents, like chloroform, acetonitrile, methyl-*tert*-butylether, or as a more high-throughput alternative the use of RP18-solid phase extraction (SPE) cartridges. Please note that the choice of extraction procedure must be in accordance with the aims of the study, as there is no universal approach covering all metabolites. Sample collection and extraction might also be done in a single step. For example metabolites can be extracted from cultured cells by collecting and homogenizing the cells directly in cold methanol. This sampling procedure stops the metabolism for the sake of an optimal preservation of the metabolite. To be compatible with subsequent analytical procedures, extracts might be evaporated (or lyophilized) and reconstituted in different solvents suitable for respective analytics. In some cases it is not sufficient to only extract metabolites, but further processing like derivatization is needed. Derivatization of certain compounds is necessary to either make analysis possible at all to improve sensitivity of the analysis (see Sect. 2.1.1 for application example).

Matrix Effects

Because biological samples from different origin reveal high diversity in their chemical composition, sample processing must be adapted for different types of tissues, body fluids or cell cultures. The part of the biological sample, that only represents the medium (e.g. blood plasma) in which the metabolites of interest (e.g. amino acids) are dissolved in, is called a biological *matrix*. The change from one matrix to another may request an adaptation of protocols to ensure good quality and reproducibility of analytics. To this day, metabolomics has already been validated and applied in humans to a wide range of matrices including body fluids (plasma, serum, urine, cerebrospinal fluid, saliva) [32, 38], dried blood spots, [34], as well as tissues (and biopsies) [39], stool [40], lung lavage [41], and exhaled air [42].

1.4.3 Analytical Approaches

Presently, different analytical approaches include *profiling-*, *non-targeted-, and targeted- metabolomics* [43]. These approaches were developed to meet the distinct requirements for reaching study aims.

Different Types of Metabolomics

Profiling metabolomics performs survey or discovery analyses with a very high mass resolution but low sample throughput, and is interested in the identification rather than the quantification of metabolites. *Non-targeted metabolomics* provides information on the simultaneous presence of many chemical classes of metabolites (global view). It reaches a high sample throughput (e.g. 100 samples a week) with a large number of metabolites quantified and the possibility to identify differences in the abundance of metabolites. Non-targeted analytics is supported by in silico searches to annotate the signals. This approach not only allows the relative quantification of metabolites with mass spectra stored in databases, but furthermore it also detects unidentified metabolites not yet registered. These unidentified metabolites can be important in the context of a phenotype investigation and can be further identified with the help of a set of characteristic parameters (i.e. retention times and mass spectra). *Targeted metabolomics* allows the quantification of a pre-selected set of known metabolites and can reach a very high throughput (e.g. 1,000 samples per week). Targeted and non-targeted approaches have been shown to have some overlap in metabolites, and revealed a very good correlation of quantification [44].

Profiling and non-targeted metabolomics can be run *chemocentric* [45] or *ion-centric* [17]. The first approach is a method for global molecule detection with chemical identification, the second one uses ion detection without identification. The latter is faster and detects much more signals but has much greater chance of false positives. At present, metabolite coverage reached by the different metabolomics technologies ranges from 200 to about 1,000 compounds in a single run.

Major Features of Technologies for Metabolomics

The different approaches of metabolomics have been developed along the progress of technologies and scientific aims. Formerly, direct analyses (i.e. measurements without sample fractionation or metabolite pre-separation) were performed. They were based on NMR, FT-IR spectroscopy, Raman spectroscopy, and MS [29, 46]. These types of analyses performed well, but resolution was largely enhanced when separation steps were introduced prior to metabolite identification. The separation steps may include gas chromatography (GC), multidimensional gas chromatography (GC×GC), capillary electrophoresis (CE), liquid chromatography (LC), high performance liquid chromatography (HPLC) or ultra high performance liquid chromatography (UHPLC), which all can be coupled to an analytical instrument like NMR or MS [47].

Targeted metabolomics is based on GC-MS, LC-MS, UHPLC-MS or flow injection assay mass spectrometry (FIA-MS) [27, 29, 48, 49]. *Profiling* and *non-targeted metabolomics* are mostly performed by NMR, CE-MS, GC-MS, NMR, LC-FT-ICR-MS or UHPLC-MS [50, 51]. *Non-targeted metabolomics* often requires a high degree of parallel analyses (i.e. standardized simultaneous analyses on LC- and GC-MS) to cover as many metabolites as possible and to avoid a bias on specific chemical classes. It also needs special algorithms for metabolite identification with specific databases [17, 52, 53]. Table 3.1 shows an overview of significant features of the most common techniques. More information about the techniques will be given in Sect. 2.

1.4.4 Large Scale Bioinformatics

Metabolomics requires a substantial support from bioinformatics to handle large data sets generated by the analytics [54]. Such data sets have to follow specific requirements to be compatible worldwide [55] and standards for these requirements have been formulated [56]. Data generated by metabolomic approaches have been efficiently evaluated by bioinformatic methods such as principal component analysis (PCA), random forest (RF) or self-organizing maps [57–59]. Metabolomics, biostatistics, and bioinformatics converge in an approach called *chemometrics*, which is applied to identify metabolites and analyze their dependencies on phenotypes [60, 61]. Metabolomics perfectly integrates with other "omics" analyses and other phenotyping methods to verify hypotheses in systems biology approaches [62].

In order to achieve reliable and meaningful results, a good level of coordination between the four aspects design, sample processing, analytics, and bioinformatics is required, as they are usually performed in different laboratories and by distinct research teams.

2 Principles of Mass Spectrometry Assay Technologies

In mass spectrometric analyses the samples are ionized, and the resulting ions are accelerated in a defined electromagnetic field, successively selected according to their mass to charge ratio (m/z) and finally detected. A mass spectrum displays the plot of the m/z against its ion abundance. Please note that mass spectrometers do not determine atomic mass units (amu). In most cases separation techniques are applied in front of MS to achieve higher resolution.

2.1 Separation Technologies

To increase the number of molecules to be identified, samples may undergo one or more separation steps prior to mass spectrometric analyses of the compounds.

Table 3.1 Overview of features of common metabolomic technologies

Method	Sample preparation	Major application	Major advantages	Major disadvantages
FT-IR	Not required	Plant and bacterial samples	Whole metabolome analyses possible	Low resolution and sensitivity
		Profiling of metabolites	Suitable for high throughput	
		Discovery		
NMR	Not required	Soluble metabolites	Non-destructive to the sample	Low sensitivity
		Discovery	Quantitative	Large sample volume necessary
			Large spectra libraries available	Low throughput
GC-MS	Required and critical	Volatile metabolites (e.g. fatty acids, sugars)	Spectra libraries compatible among laboratories	Extraction necessary
		Diagnostics	Robust and reproducible method	Thermolabile molecules may decompose
		Discovery		Derivatization artefacts possible
LC-MS	Required	Soluble and lipophilic metabolites (e.g. prostanoids, amino acids)	Good resolution	Extraction necessary
		Diagnostics	Ion suppression effects minimized by LC	Ionization problems
		Discovery	High sensitivity	
			Robust and reproducible method	
FIA-MS	Required	Soluble and lipophilic metabolites	Very fast and robust method	Extraction necessary
		Diagnostics (e.g. newborn screen)	Simple handling	Ion suppression effects possible
				Ionization problems
				Only targeted metabolites visible
				Not the highest resolution

The most common separation procedures are shortly explained in the following chapter. However under specific conditions, the samples might be analyzed directly in MS without any preceding separations. This technique is called flow injection assay (FIA). More details and an example of that will be given in Sect. 2.2.3.

2.1.1 Gas Chromatography (GC)

GC is well suited to separate volatile compounds [63]. A broad variety of samples can be analyzed as long as the compounds are readily vaporized and thermally stable. To increase the volatility, some compounds have to be derivatized prior to GC analysis, as for example fatty acids that have to be transformed into their methyl esters. Like in other chromatographic techniques, a mobile and a stationary phase are required. The mobile phase is an inert carrier gas, e.g. helium, argon, or nitrogen. The most common stationary phase is a capillary column, typically 15–30 m long, coated on the inside with a thin (0.2 μm) film of a high boiling liquid (e.g. dimethyl polysiloxane). The carrier gas flows continuously through the injection port, column and detector. The sample is injected into the heated injection port, where it is vaporized and carried into the capillary column. In the capillary column, the sample analytes are separated according to their respective retention times on the stationary phase. Thereby, the retention time is dependent on the relative solubility of compounds in the liquid phase (and dependent on the relative vapor pressure).

The limitation of GC is that the substances must be volatile and thermally stable. For organic substances volatility is hardly achievable if the molecular weight of the compound exceeds 500 Da. High temperatures up to 300°C enhance volatility, but decomposition of the analyzed compounds can be the result. When derivatization is required, sample preparation is not trivial and can lead to a reduction of analyte concentrations or to an increase of background signal.

2.1.2 Liquid Chromatography (LC)

LC separates dissolved compounds in a liquid mobile phase along a solid stationary phase [64]. For separation of analytes with different chemical properties (e.g. a mixture of hydrophilic and hydrophobic compounds) different settings can be used. The most classical separation system for hydrophilic metabolites is called normal-phase HPLC. This system separates analytes based on their adsorption properties to the surface of a polar stationary phase (silica), as well as their polarity in a non-polar mobile phase (e.g. hexane). For more hydrophobic metabolites, a non-polar stationary phase together with a moderately polar mobile phase (e.g. water) is used, which is called *reversed phase* (RP) HPLC. The most common stationary phase in RP-HPLC is silica modified with C18 alkyl chains (RP18). Usually, the stationary phase is packed in metal tubes called *columns* (e.g. 4.6 mm internal diameter of 100 mm length). Typically, a HPLC is running with flow rates of about 0.5–2 mL/min and a back pressure of up to 200–400 bar. The performance of separation is strongly increased by using smaller particle sizes for the stationary phase material

(unfortunately associated with an increase in back pressure) [65], as applied in UHPLC. The usual time needed for one HPLC-run is about 5–60 min with a throughput of about 20–200 samples per day. Using UHPLC, analysis time can be reduced considerably with simultaneously improved resolution.

2.2 Mass Spectrometry

2.2.1 Ionization

Ionization of molecules is required for mass spectrometric analyses (Box 3.2). Only charged molecules (ions) enter the MS analyzer to be separated according to their m/z ratio by electromagnetic fields [66]. Depending on the separation technique and polarity of the analytes, different ionization techniques might be used (Fig. 3.1).

Box 3.2 Mass Spectrometry

The mass to charge ratio ***(m/z)*** is a dimensionless value used in mass spectrometric experiments, formed by dividing the mass number of an ion by its charge number.
Positive and ***negative ion modes*** are used to take advantage of ionization properties of the compounds if they are exposed to low or high pH of the solvent. Basic compounds are analyzed in positive and acidic compounds in negative ion mode.
The ***ion source*** is the part of a mass spectrometer that ionizes the compounds before they enter a mass analyzer.
A ***quadrupole*** is a setup of four poles, originating from four metal rods and creating a strong electromagnetic field oriented in parallel along the ion flight path. It serves as ion selector in mass spectrometer.
Ion-trap is a mass analyzer with high trapping capacity and efficiency, and with an ability to be emptied fast.
Ion suppression is observed when molecules of the matrix disturb the ionization efficiency of the molecules of interest and lead to a decrease of intensity of analyte signal.
Detector is a part of mass spectrometer recording the charge induced when an ion hits its surface.
SRM (selected reaction monitoring) is a data acquisition mode in tandem mass spectrometry where specific precursor (mother) and product (daughter) ions are selected in the first and second mass analyzer, respectively, in between undergoing a fragmentation in the collision cell. SRM is a tool to select a specific metabolite. SRMs are often used in targeted metabolomics.
Isobaric compounds have the same weight but different structures, e.g. leucine and iso-leucine. They might be difficult to differentiate in mass spectrometry if not separated previously by their chemical properties.

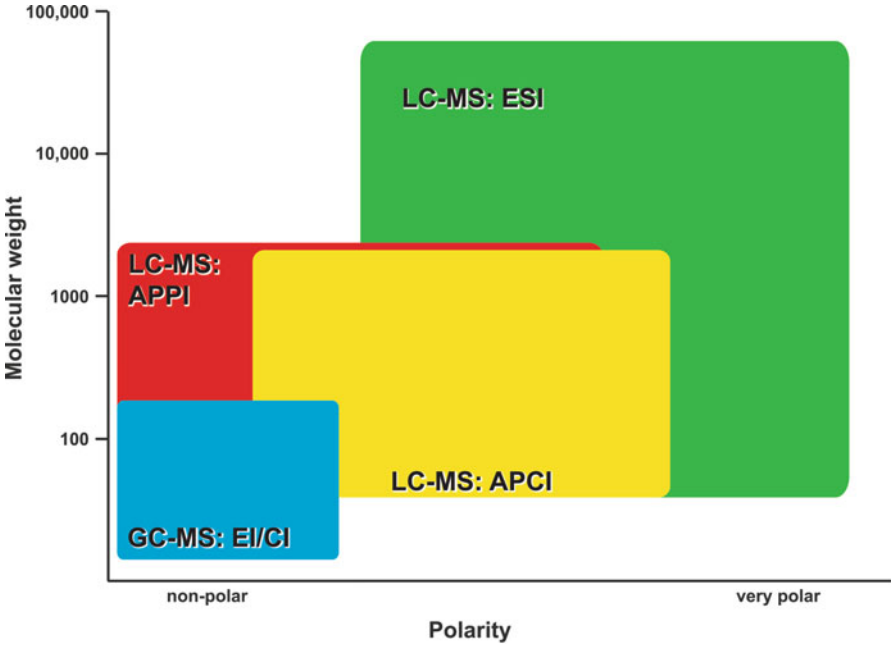


Fig. 3.1 Comparison of metabolite coverage by GC-MS and LC-MS ionization techniques. Small and non-polar molecules are best detected by the GC-MS ionization methods EI and CI. Larger and polar compounds are rather covered by the LC-MS method ESI, and medium-sized molecules can be best ionized by the LC-MS techniques APPI and APCI. Actual coverage may vary between different MS devices

For GC-MS there is a choice between electron impact ionization (EI) and chemical ionization (CI), both working in vacuum [67]. EI is a rather "hard ionization" where free high-energy electrons (70 eV) emitted from a filament (electrode) are used to bombard the analytes yielding many fragments of lower m/z. For CI, a reagent gas, mostly methane or ammonia, is let into the mass spectrometer. During this "soft ionization" the reagent gas interacts with emitted electrons forming reagent gas plasma. The reagent gas plasma transfers the charge to the analyte yielding a low amount of high m/z fragment ions often close to m/z of the original analyte ion (mother ion). CI can be performed in positive (PCI) or negative (NCI) mode.

For LC-MS there are three main ionization techniques, ESI (electrospray ionization), APCI (atmospheric pressure chemical ionization), and APPI (atmospheric pressure photoionization), which all are used at atmospheric pressure [68]. For ESI, the eluent (containing the analytes) is dispersed by help of a steel capillary. The capillary's wall is energized by high voltages and the charge is transferred to the solution as it passes the capillary's wall. In the next step, the solution is dispersed producing an aerosol. Using a continuous gas flow (nitrogen) and heating, the

solvent evaporates from the aerosol droplets resulting in a Coulomb explosion and the extant ions (sometimes multiple charged) enter the mass spectrometer. In APCI, the analytes are ionized after nebulization and heating of the solution by a so-called corona needle, which is extended into the spray cone. For APPI, photons emitted by UV-light and a volatile solvent, called *dopant* (e.g. toluol), are used to ionize the analytes. The latter technique works well with non-polar or low-polar compounds, which are not efficiently ionized by other ionization sources.

ESI is suitable for the analysis of molecules of very different size but not for very non-polar compounds. The ionization efficiency of APCI lies between the ESI and APPI techniques (Fig. 3.1). ESI and APCI can be used in *positive* and *negative mode*. Basic compounds in low pH solvents are readily protonated to produce positive molecular ions. These compounds are better analyzed in a positive ion mode. For efficient positive ionization a donor proton such as formic acid should be added to the mobile phase. The negative ion mode analysis is applied to acidic compounds. The addition of a proton acceptor, such as ammonium hydroxide, to the mobile phase facilitates negative ion formation (i.e. proton loss). The positive mode is more frequently used, because protons are often loosely associated with a molecule even when there are no obvious basic functional groups.

A typical obstacle in mass spectrometric analysis is the so-called *ion suppression*, where molecules of the matrix disturb the ionization efficiency of the molecules of interest and lead to a decrease of intensity for the respective analyte signal. The optimum conditions to keep the suppression as low as possible have to be found empirically (Box 3.2).

2.2.2 Principles of Mass Analyzers

After ionization, the ions enter the *mass analyzer* region of a mass spectrometer. This region consists of one or more single mass analyzers, separating ions according to their m/z ratio. Most common mass spectrometers use quadrupole, ion-trap, time-of-flight (TOF), or FT-ICR analyzers [46, 69, 70]. In the following we will describe quadrupole and ion-trap in more detail.

Quadrupole

For *targeted metabolomics*, quadrupoles are most suitable as they belong to the simplest and least expensive mass analyzers and give the most reproducible quantitative results. A quadrupole consists of four elongated and parallel ordered electrodes (rods), to which an alternating electromagnetic field is applied. Analyte ions move through the rods and reach the detector as long as they are in resonance with the frequency of the electromagnetic field. Otherwise they will be removed by being discharged at the electrodes. A quadrupole can operate either in scan or single ion monitoring (SIM) mode. In scan mode, the mass analyzer monitors a range of

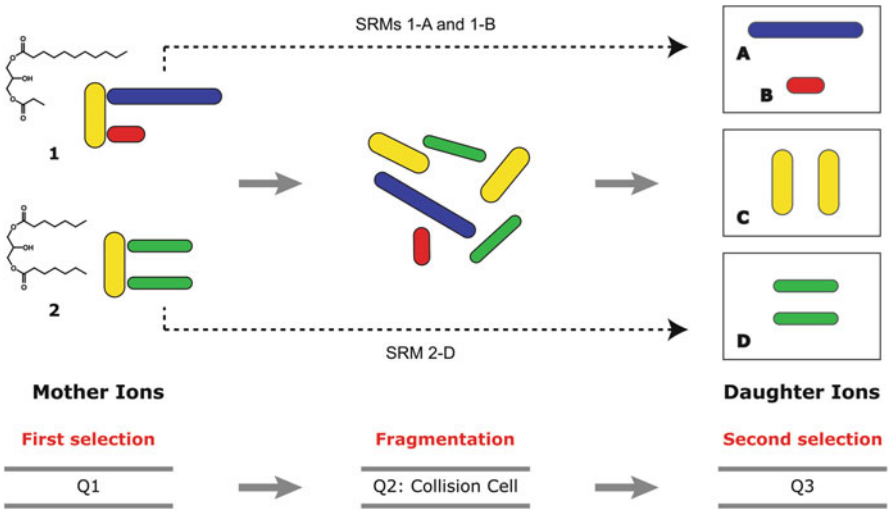


Fig. 3.2 Principle of SRM using tandem mass spectrometry. Application of single reaction monitoring (SRM) in tandem MS is exemplarily explained on two glycerol esters of identical molecular weight and m/z (316.44 amu, molecular formula $C_{17}H_{32}O_5$). The "mother ions" of the analytes are selected by quadrupole 1 (Q1). Subsequently they are broken into distinct fragments in Q2 due to their different chemical structure, although they show the same m/z. The resulting "daughter ions" can be separated in the second mass analyzer (Q3). Here, the daughter ions consist of different carboxylic acid residues (**a**: *blue*, **b**: *red*, **d**: *green*) and the glycerol moiety (**c**: *yellow*). The observation of a combination of a specific mother and daughter ion pair (e.g. ion 1 and ion A, or "1-A") is named SRM and can be used for the distinction of different molecules in a sample mixture. In this example, the SRMs 1-A and 1-B identify only molecule 1, while SRM 2-D is characteristic for molecule 2. On the other hand, the SRMs 1-C and 2-C would not discriminate between the molecules 1 and 2

mass-to-charge (m/z) ratios. In SIM mode, the mass analyzer monitors only one m/z ratio.

In some experiments, the separation of a distinct m/z by a single quadrupole is sufficient, but often the biological matrix is too complex in terms of having too many different compounds and thus requires another approach. Therefore, most metabolomic experiments are performed in *tandem mass spectrometer* (MS/MS) also named triple quadrupole mass spectrometer. In this setting, two sequentially assembled mass analyzers are separated by a collision cell (as shown schematically in Fig. 3.2). The first mass analyzer consists of a quadrupole mass filter (quadrupole 1 or Q1, also called mass analyzer 1 or MS_1), which allows to select ions according to their specific m/z. In MS_1 the originally ionized molecules, the so-called mother ions, are selected. In a next step, the ions can be fragmented in Q2, the collision cell, filled with an inert collision gas, e.g. nitrogen or helium. This process is called collision-induced dissociation (CID) and leads to the production of daughter ions. The third quadrupole (Q3, mass analyzer 2 or MS_2) acts in the same way as Q1 scanning the fragment ions emerging from Q2 according to their m/z. The resulting

detector output is described as MS/MS spectrum (or MS^2). Please note that the numbering of mass analyzers (e.g. MS_1) and mass spectra (e.g. MS^2) is different. By this approach, compounds of identical molecular masses but different molecular structures (e.g. different side chain length of the compounds) can be distinguished and individually analyzed. This would not be possible with a single quadrupole mass spectrometer. A higher resolution can be achieved when the Q3 is used as an ion-trap. In this case additional electrical lenses are mounted to Q3 in order to trap and fragment ions. With this setup further fragmentation of daughter ions is possible and MS^3 spectra are generated. These analyses are employed when, e.g. the position of double bonds in isomeric fatty acid side chains of phospholipids shall be determined.

Tandem MS additionally offers the possibility of selected reaction monitoring (SRM). In this approach, the instrument monitors only a selected molecule by its specific ion pair (mother and daughter ions). The electromagnetic field applied to MS_1 (Q1) is set to permit only a selected ion with a specific m/z (precursor ion or mother ion) to pass. After fragmentation in the collision cell, again only a specific product ion (daughter ion) will be selected by its specific m/z in MS_2 (Q3). Present instruments allow for nearly parallel observation of several SRMs. In that case the specific m/z ion pairs are monitored one after another in a very short time. The number of SRMs collected and the resolution are limited by the speed of the scanning mode (typically 2,000 m/z units per second). Multiple reaction monitoring (MRM) is widely used to describe the parallel acquisition of SRMs, but the IUPAC recommends not using this term anymore [71].

Ion-trap

For *non-targeted metabolomics*, an ion-trap mass analyzer is often preferred to a quadrupole as it exhibits higher sensitivity and is able to record more comprehensive mass spectra in low concentration ranges (0.01–0.1 mg/L). A 3-dimensional (3-D) ion-trap mass analyzer consists of a circular ring electrode and two end caps which together form a chamber. Ions entering the chamber are “trapped” there by electromagnetic fields. For detection, another field is applied to selectively eject ions from the trap (according to their mass). In 2003, 2-dimensional linear ion-traps have been introduced [72]. Their operation is very similar to that of conventional 3-D traps, but the linear design results in faster scan rates and enhanced sensitivity. This is due to better trapping efficiency, higher trap capacity, and the ability to be emptied faster. The MS^n capabilities of the ion trap mass spectrometer make it a powerful tool for the structural analysis of complex mixtures.

2.2.3 Example of Flow Injection Analysis (FIA)

A simplified approach in *targeted metabolomics* is the quantification of compounds using the flow injection analysis (FIA) approach (Fig. 3.3). FIA is a reasonable

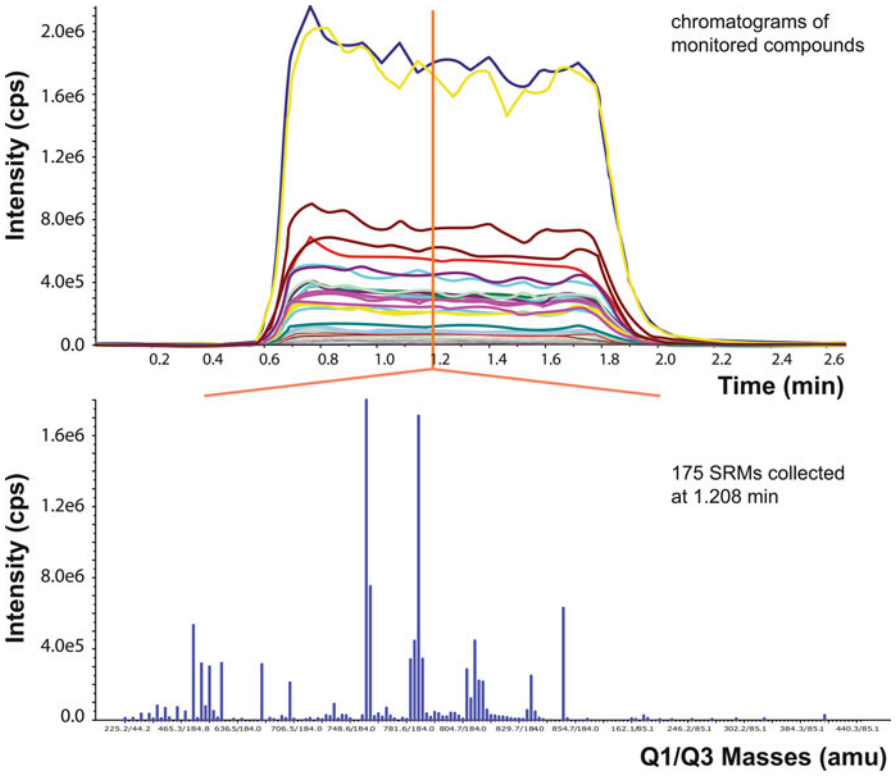


Fig. 3.3 Example of chromatogram and spectra resulting from FIA-MS. The sample mixed with mobile phase is directly infused into the MS without prior fractionation or separation by GC or LC. Identification and quantification of compounds is based on consecutive acquisition of several SRMs. *Upper panel*: chromatographic profiles of 175 individual compounds coded by different colors. *Lower panel*: spectrum of collected 175 SRMs at the 1.208 min

method when specific and distinct SRM spectra are known for different compounds to be analyzed. The compounds can be identified and quantified within a mixture based on their SRMs. For FIA a sample mixed with a suitable mobile phase is directly infused into an MS without prior fractionation or separation by GC or LC. Because this approach requires less analysis time in comparison to, e.g. LC-MS techniques, it is often considered for high throughput analysis. However, isomers differing in branch locations, like leucine and iso-leucine, cannot be discriminated by this method since they reveal the same SRMs.

2.2.4 Example of Targeted LC-MS/MS Analysis

In many cases molecules of interest reveal very similar chemical properties so that even LC cannot separate them easily. Nevertheless, by using the advance of tandem

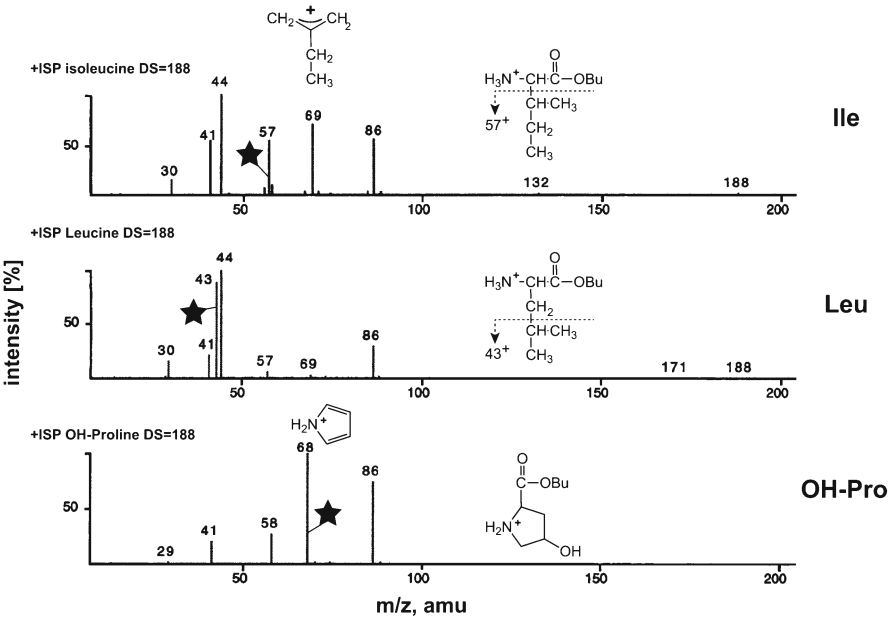


Fig. 3.4 Example of targeted MS analysis. Some substances might be not well separated by LC and furthermore, different compounds may produce ions of the same m/z. Shown are MS/MS spectra of three isobaric amino acids isoleucine (Ile), leucine (Leu) and hydroxy-proline (OH-Pro) which have been butylated. The amino acids reveal identical m/z 188^+ ions in Q1, which can be fragmented in Q2 at a collision energy of 45 eV, to create lower mass fragment ion species analyzed in Q3. The SRMs 188–69^+ (or 188–57^+), 188–43^+, and 188–68^+, can distinguish between Leu, Ile and OH-Pro. *Asterisks* label the mass of differentiating ions (Modified after [73])

mass spectrometry it is eventually possible to distinguish and quantify the compounds. This is exemplarily demonstrated for the isomeric amino acids leucine (Leu), isoleucine (Ile) and hydroxyproline (OH-Pro) (Fig. 3.4).

It is common to increase the sensitivity of amino acid detection by derivatization, for example with Edman's reagent (phenylisothiocyanate, PITC) or butyl ester. Unfortunately, some isobaric amino acids, like Leu, Ile or OH-Pro cannot even under this condition be distinguished due to poor chromatographic resolution, i.e. missing separation. However, by coupling the chromatography to mass spectrometry, e.g. LC-MS/MS, these amino acids can be differentiated [73] as follows: Presetting a specific collision energy, butylated acidic groups of Leu and Ile allow an initial loss of butyl formate creating m/z 86^+ product ions. A subsequent fragmentation creates even smaller ions (69^+, 43^+ and 68^+ for Ile, Leu and OH-Pro, respectively), which can finally be used to distinguish these amino acids. Although Leu and Ile both fragment to the characteristic carbonium ion structures at 43^+ and 69^+, Leu shows only a minor 69^+ product ion and Ile has no 43^+ but a relatively strong 57^+ instead (Fig. 3.4).

2.2.5 Challenges of Non-Targeted LC-MS Analyses

A drawback of *targeted metabolomics* is that only those metabolites can be detected and quantified, for which the apparatus has been tuned, although much more compounds of interest might be present in a biological sample. Additionally, the targeted metabolomics approach based on SRMs has its limitations because the number of SRM transitions that can be monitored in a single analysis is finite. It often requires multiple analyses of the same sample to cover all SRM transitions which prolongs analysis time [74].

For these reasons, *non-targeted metabolomics* is another valuable approach to be implemented along with targeted metabolomics. A common approach in non-targeted metabolomics is to analyze samples using a standardized (U)HPLC/GC separation and a particularly accurate and fast MS (featuring a good ion trap). Typically the same material is separated at different conditions (e.g. positive/negative mode, parallel LC and GC separation) to increase metabolome coverage and to minimize the bias. Resulting chromatograms reveal multiple peaks and for each peak mass spectra are collected. The latter are processed to prepare lists of potential molecular formulas (Fig. 3.5). Their identities are determined by searches in public databases (like KEGG, HMDB, MassBank, ChemACX, and ChemSpider) or in-house equipment-specific mass spectrum libraries [17]. An equipment-specific library has the advantage of higher resolution and a lower false-positive detection rate.

2.3 Detectors

After selection in mass analyzers the ions are quantitatively recorded by a detector. Common detectors used in MS-assay technology are electron multipliers and photon multipliers [75]. In an electron multiplier, positive or negative ions that hit an impact plate coated with copper or barium oxide (conversion dynode) cause the release of primary electrons. These electrons are accelerated by an electric potential towards a farther dynode and release further electrons. The electron cascade generated by dynodes is converted to an electrical potential by an amplifier to a measurable signal. In the case of a photon multiplier the electrons pass a plate containing phosphorus that emits photons and direct them to a photocathode.

3 Requirements for Quantification

3.1 Method Validation

Every analytical method has to be evaluated for reliability and quantification ranges [76]. This process is called validation and some of the tests required will be explained below.

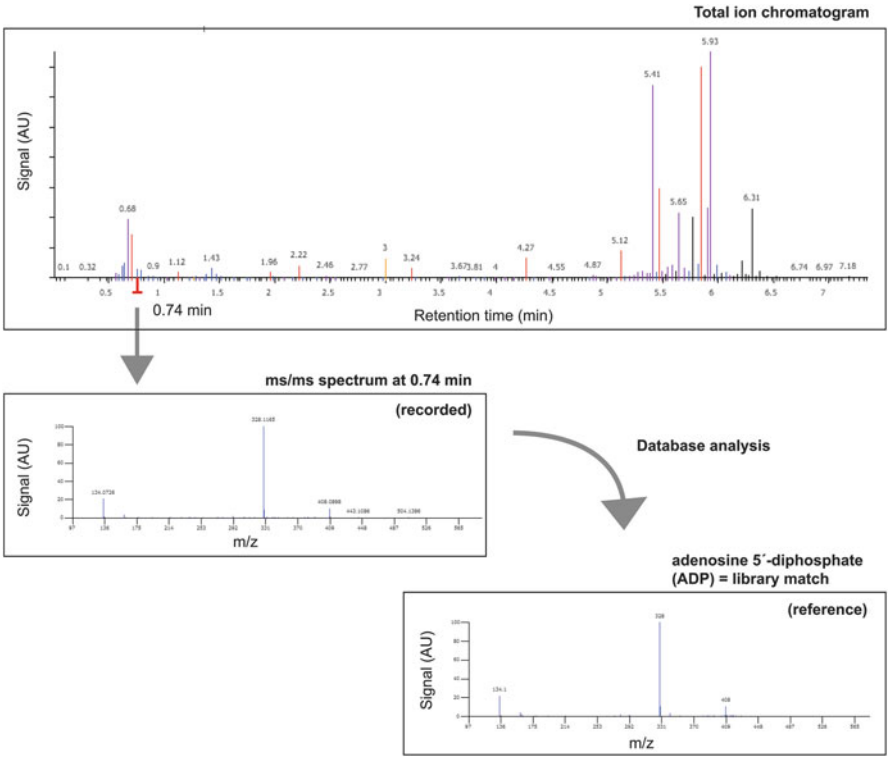


Fig. 3.5 Example of non-targeted LC-MS analysis. Total ion chromatogram (*upper panel*) of human plasma shows abundance and retention times of compounds. Ions at different time points are collected and recorded as mass spectra (*panel lower left*; a mass spectrum at 0.74 min) and compared with reference spectra (*panel lower right*) from databases to identify the compound

The *selectivity test* checks the ability of the method to specifically identify an analyte, e.g. the separation of an analyte peak from other matrix peaks in chromatography or the reliable identification of an analyte by mass spectrometry.

The *linearity test* is used to check the relation between concentrations and measured values (e.g. peak intensities) of an analyte. Ideally a linear correlation is intended, but this is not always reached. The final quantification is also limited by other features like LOD or LOQ (explained later in Sect. 3.2). Please note that quantification must not be performed outside the linear dynamic range of the calibration curve, as then the concentration is not proportional to its measured value (Fig. 3.6).

The *precision* describes the ability of a measurement to be consistently reproduced. It means that individual measured values for a compound should be the same when the analytical procedure is applied repeatedly to the same biological matrix. Precision is checked by (1) testing multiple aliquots of one homogeneous sample, (2) measurement of several sample preparations from the same biological matrix

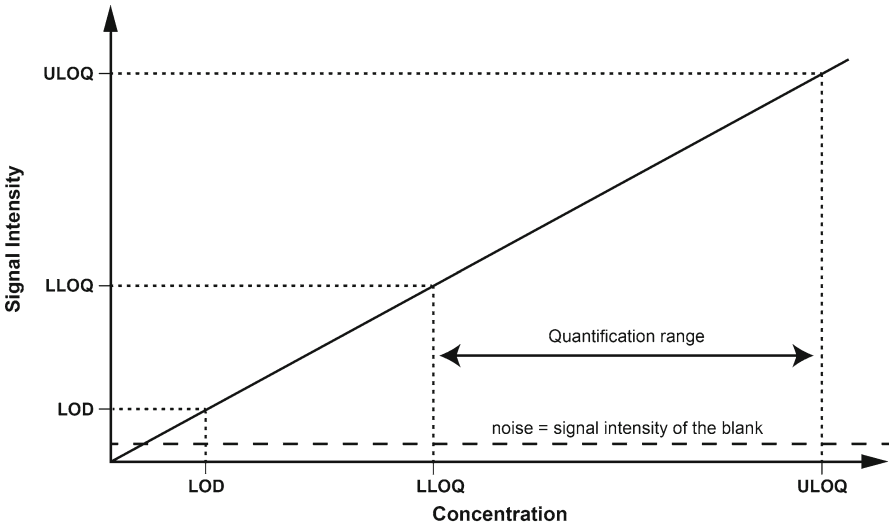


Fig. 3.6 Parameters describing quality of signal quantification. LOD (limit of detection) marks the concentration limit for detecting a metabolite. LLOQ and ULOQ (*lower* and *upper* limit of quantification, respectively) define the quantification range of the analytical method

and (3) repetition of measurements on different days. Please note that the parameter *precision* demonstrates the ability of the method to hit always the same value, but not the accuracy of it (see Fig. 3.7).

The *accuracy* describes the difference of the measured concentration of an analyte to the analyte's true concentration. Accuracy is tested by spiking real samples (preferably a biological matrix lacking the analyte of interest) with defined amounts of the analyte and comparing measured and spiked concentration. A metabolite-free matrix is rare, as biological samples are never free of metabolites, and therefore a surrogate matrix, e.g. PBS, has to be used instead.

The *robustness* of the method should be checked to get information about the influence of disturbances and variations that can happen in a laboratory deteriorating selectivity, linearity, precision, and accuracy. Examples of the parameters influencing the robustness are: shelf life of chemicals and solutions, changes in pressure or flow in HPLC, analyses on different apparatus with the same construction parameters, changes of brands or lots for chemicals, rotation of staff, sensitivity to impurities in the sample or changes of sample composition.

3.2 Sensitivity of Method

The *limit of detection (LOD)* is the lowest analyte concentration that can be detected (identified but not quantified) by the analytical method. Most frequently, the LOD

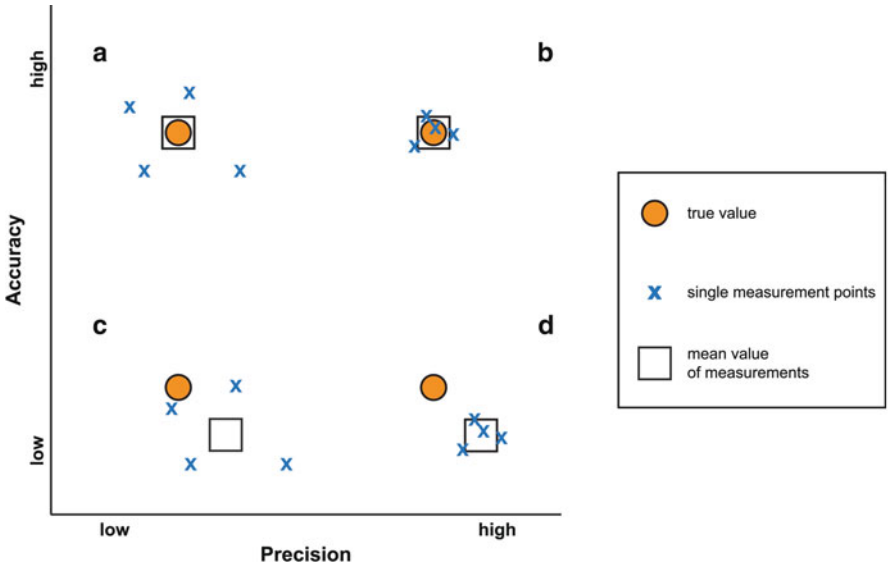


Fig. 3.7 Influence of accuracy and precision on the experimental readout. (**a**) High accuracy and low precision (workable). (**b**) High accuracy and high precision (optimum). (**c**) Low accuracy and low precision (method should be improved). (**d**) Low accuracy and high precision (method cannot be used for absolute quantification but for relative comparison of groups)

is defined as 3 times the mean matrix (blank) value (Fig. 3.6) [77]. All values below LOD must not be used for data analysis because they cannot be reliably differentiated from background noise.

For the *limit of quantification (LOQ)*, two outer limits are determined: the *lower limit of quantification (LLOQ)* and the *upper limit of quantification (ULOQ)*. Typically, the LLOQ is defined as 10 times the standard deviation of the blank (Fig. 3.6). ULOQ is experimentally defined by accuracy and linearity tests using spiked matrix samples. Values below LLOQ and above ULOQ should also not be used for data analysis because the accuracy of these values is not reliable.

3.3 Standard Operating Procedures (SOPs)

A lot of the technical variability can be minimized when sample processing (collection, storage, extraction, derivatization, application to analytical units, measurements, data collection, and documentation) follows standard operating procedures (SOPs) throughout the whole study [78]. Automation of as many steps as possible helps to keep variability even lower. For each biological matrix (e.g. tissue, cell culture or body fluids), the different steps of sample collection (e.g. heparin- or

EDTA-plasma), handling (e.g. homogenization or centrifugation), storage (tube type, temperature of storage) and extraction (e.g. solvent type) have to be standardized and validated with the subsequent analytical method. This is because the change from one matrix to another may request an adaptation of protocols to ensure good quality and reproducibility of the analytics. In case that a robot-mediated handling is used it must undergo process optimization as well, as the individual steps done by humans and a machine may have a different logistics. The performance of robots has to be monitored to avoid cross-contaminations (or carry-overs), which might be prevented by wash steps or use of disposables (e.g. liquid handling tips). Besides supporting higher data precision and reliability, automated steps can be easily incorporated into a laboratory information and management system (LIMS). LIMS is used to track and monitor all processes: sample storage, as well as extraction and processing prior to application on an analytical unit, and finally data collection and documentation. The procedures for operating the separation equipment, e.g. GC or LC, and analytical instruments like MS should involve periodic quality checks with standard samples. These could be synthetic (e.g. a mixture of several compounds in a defined matrix) or natural products (sample of human plasma pooled from many healthy individuals like "Standard Reference Material 1950 (SRM1950)" provided by NIST [79]). It is important to periodically check the performance of the instruments and the whole analytical pipeline not only with quality parameters as described by manufacturers but with samples similar to those in real experiments as well. The latter are more complex and more challenging. Every device and solution of the analytical process has its own life-time, e.g. columns for the HPLC will deteriorate with time and number of samples analyzed and also the vacuum pumps in MS-instruments will show lower efficiency with time. Therefore, performance of every sub-step should be monitored in order to perform appropriate replacements well in advance. A thorough documentation of maintenance and quality management is an important part of SOPs supporting dependable quantification.

References

1. Pauling L, Robinson AB, Teranishi R, Cary P (1971) Quantitative analysis of urine vapor and breath by gas–liquid partition chromatography. Proc Natl Acad Sci USA 68(10):2374–2376
2. Ota T, Suzuki Y, Nishikawa T et al (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36(1):40–45
3. Barash Y, Calarco JA, Gao W et al (2010) Deciphering the splicing code. Nature 465(7294): 53–59
4. Caldana C, Degenkolbe T, Cuadros-Inostroza A et al (2011) High-density kinetic analysis of the metabolomic and transcriptomic response of *Arabidopsis* to eight environmental conditions. Plant J. doi:10.1111/j.1365-313X.2011.04640.x
5. Griffiths WJ, Wang Y (2009) Mass spectrometry: from proteomics to metabolomics and lipidomics. Chem Soc Rev 38(7):1882–1896
6. Roessner U, Bowne J (2009) What is metabolomics all about? Biotechniques 46(5):363–365
7. Tweeddale H, Notley-McRobb L, Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. J Bacteriol 180(19):5109–5116

8. Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29(11):1181–1189
9. Oresic M (2009) Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. Nutr Metab Cardiovasc Dis 19(11):816–824
10. Ceglarek U, Shackleton C, Stanczyk FZ, Adamski J (2010) Steroid profiling and analytics: going towards sterome. J Steroid Biochem Mol Biol 121(3–5):479–480
11. Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48(1–2):155–171
12. Wishart DS, Knox C, Guo AC et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37(Database issue):D603–D610
13. Fahy E, Subramaniam S, Murphy RC et al (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50(Suppl):S9–14
14. Giavalisco P, Kohl K, Hummel J, Seiwert B, Willmitzer L (2009) 13°C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. Anal Chem 81(15):6546–6551
15. Waters NJ, Garrod S, Farrant RD et al (2000) High-resolution magic angle spinning (1)H NMR spectroscopy of intact liver and kidney: optimization of sample preparation procedures and biochemical stability of tissue during spectral acquisition. Anal Biochem 282(1):16–23
16. Griffin JL (2006) The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? Philos Trans R Soc Lond B Biol Sci 361(1465):147–161
17. Lawton KA, Berger A, Mitchell M et al (2008) Analysis of the adult human plasma metabolome. Pharmacogenomics 9(4):383–397
18. Young BP, Shin JJ, Orij R et al (2010) Phosphatidic acid is a pH biosensor that links membrane biogenesis to metabolism. Science 329(5995):1085–1088
19. Bais P, Moon SM, He K et al (2010) PlantMetabolomics.org: a web portal for plant metabolomics experiments. Plant Physiol 152(4):1807–1816
20. Altmaier E, Ramsay SL, Graber A, Mewes HW, Weinberger KM, Suhre K (2008) Bioinformatics analysis of targeted metabolomics–uncovering old and new tales of diabetic mice under medication. Endocrinology 149(7):3478–3489
21. Altmaier E, Kastenmuller G, Romisch-Margl W et al (2009) Variation in the human lipidome associated with coffee consumption as revealed by quantitative targeted metabolomics. Mol Nutr Food Res 53(11):1357–1365
22. Goodacre R (2007) Metabolomics of a superorganism. J Nutr 137(1 Suppl):259S–266S
23. Yuliana ND, Khatib A, Choi YH, Verpoorte R (2011) Metabolomics for bioactivity assessment of natural products. Phytother Res 25(2):157–169
24. Schmitt-Kopplin P, Gabelica Z, Gougeon RD et al (2010) High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. Proc Natl Acad Sci USA 107(7):2763–2768
25. Halama A, Moller G, Adamski J (2011) Metabolic signatures in apoptotic human cancer cell lines. OMICS 15(5):325–335
26. Rios-Estepa R, Turner GW, Lee JM, Croteau RB, Lange BM (2008) A systems biology approach identifies the biochemical mechanisms regulating monoterpenoid essential oil composition in peppermint. Proc Natl Acad Sci USA 105(8):2818–2823
27. Illig T, Gieger C, Zhai G et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42(2):137–141
28. Koulman A, Lane GA, Harrison SJ, Volmer DA (2009) From differentiating metabolites to biomarkers. Anal Bioanal Chem 394(3):663–670
29. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R (2007) Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics 8(9):1243–1266
30. Oresic M, Vidal-Puig A, Hanninen V (2006) Metabolomic approaches to phenotype characterization and applications to complex diseases. Expert Rev Mol Diagn 6(4):575–585
31. Griffiths WJ, Koal T, Wang Y, Kohl M, Enot DP, Deigner HP (2010) Targeted metabolomics for biomarker discovery. Angew Chem Int Ed Engl 49(32):5426–5445

32. Gowda GA, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D (2008) Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn 8(5):617–633
33. Hekmatyar SK, Wilson M, Jerome N et al (2010) (1)H nuclear magnetic resonance spectroscopy characterisation of metabolic phenotypes in the medulloblastoma of the SMO transgenic mice. Br J Cancer 103(8):1297–1304
34. Chace DH, Hillman SL, Van Hove JL, Naylor EW (1997) Rapid diagnosis of MCAD deficiency: quantitative analysis of octanoylcarnitine and other acylcarnitines in newborn blood spots by tandem mass spectrometry. Clin Chem 43(11):2106–2113
35. Zivkovic AM, Wiest MM, Nguyen UT, Davis R, Watkins SM, German JB (2009) Effects of sample handling and storage on quantitative lipid analysis in human serum. Metabolomics 5(4):507–516
36. Morrow JD, Roberts LJ 2nd (1994) Mass spectrometry of prostanoids: F2-isoprostanes produced by non-cyclooxygenase free radical-catalyzed mechanism. Methods Enzymol 233:163–174
37. Römisch-Margl W, Prehn C, Bogumil R, Röhring C, Suhre K, Adamski J (2011) Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics. doi:10.1007/s11306-011-0293-4
38. Whitfield PD, German AJ, Noble PJ (2004) Metabolomics: an emerging post-genomic tool for nutrition. Br J Nutr 92(4):549–555
39. Denkert C, Budczies J, Weichert W et al (2008) Metabolite profiling of human colon carcinoma–deregulation of TCA cycle and amino acid turnover. Mol Cancer 7:72
40. Monleon D, Morales JM, Barrasa A, Lopez JA, Vazquez C, Celda B (2009) Metabolite profiling of fecal water extracts from human colorectal cancer. NMR Biomed 22(3):342–348
41. Hu JZ, Rommereim DN, Minard KR et al (2008) Metabolomics in lung inflammation:a high-resolution (1)h NMR study of mice exposedto silica dust. Toxicol Mech Methods 18(5):385–398
42. Carraro S, Rezzi S, Reniero F et al (2007) Metabolomics applied to exhaled breath condensate in childhood asthma. Am J Respir Crit Care Med 175(10):986–990
43. Dunn WB, Ellis DI (2004) Metabolomics: current analytical platforms and methodologies. Trends Anal Chem 24(4):285–294
44. Suhre K, Meisinger C, Doring A et al (2010) Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS One 5(11):e13953
45. Kuchel PW (2010) Models of the human metabolic network: aiming to reconcile metabolomics and genomics. Genome Med 2(7):46
46. Brown SC, Kruppa G, Dasseux JL (2005) Metabolomics applications of FT-ICR mass spectrometry. Mass Spectrom Rev 24(2):223–231
47. Issaq HJ, Van QN, Waybright TJ, Muschik GM, Veenstra TD (2009) Analytical and statistical approaches to metabolomics research. J Sep Sci 32(13):2183–2199
48. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Targeted profiling: quantitative analysis of 1 H NMR metabolomics data. Anal Chem 78(13):4430–4442
49. Koal T, Deigner HP (2010) Challenges in mass spectrometry based targeted metabolomics. Curr Mol Med 10(2):216–226
50. Griffiths WJ, Karu K, Hornshaw M, Woffendin G, Wang Y (2007) Metabolomics and metabolite profiling: past heroes and future developments. Eur J Mass Spectrom 13(1):45–50
51. Zhao X, Fritsche J, Wang J et al (2010) Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. Metabolomics 6:362–374
52. Ohta T, Masutomi N, Tsutsui N et al (2009) Untargeted metabolomic profiling as an evaluative tool of fenofibrate-induced toxicology in Fischer 344 male rats. Toxicol Pathol 37(4): 521–535
53. Sreekumar A, Poisson LM, Rajendiran TM et al (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457(7231):910–914
54. Kastenmüller G, Römisch-Margl W, Wägele B, Altmeier E, Suhre K (2010) metaP-Server: a web-based metabolomics data analysis tool. J Biomed Biotechnol 2011:ID839862
55. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O (2010) The chemical translation service–a web-based tool to improve standardization of metabolomic reports. Bioinformatics 26(20):2647–2648

56. Sansone SA, Fan T, Goodacre R et al (2007) The metabolomics standards initiative. Nat Biotechnol 25(8):846–848
57. Haddad I, Hiller K, Frimmersdorf E, Benkert B, Schomburg D, Jahn D (2009) An emergent self-organizing map based analysis pipeline for comparative metabolome studies. Silico Biol 9(4):163–178
58. Enot DP, Beckmann M, Overy D, Draper J (2006) Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. Proc Natl Acad Sci USA 103(40):14865–14870
59. Patterson AD, Li H, Eichler GS et al (2008) UPLC-ESI-TOFMS-based metabolomics and gene expression dynamics inspector self-organizing metabolomic maps as tools for understanding the cellular response to ionizing radiation. Anal Chem 80(3):665–674
60. Trygg J, Holmes E, Lundstedt T (2007) Chemometrics in metabonomics. J Proteome Res 6(2):469–479
61. Holmes E, Nicholls AW, Lindon JC et al (2000) Chemometric models for toxicity classification based on NMR spectra of biofluids. Chem Res Toxicol 13(6):471–478
62. Kohl M (2011) Standards, databases, and modeling tools in systems biology. Methods Mol Biol 696:413–427
63. Honour JW (2006) Gas chromatography–mass spectrometry. Methods Mol Biol 324:53–74
64. Snyder LR, Kirkland JJ, Dolan JW (2009) Introduction to modern liquid chromatography. Wiley, New York
65. Kortz L, Helmschrodt C, Ceglarek U (2011) Fast liquid chromatography combined with mass spectrometry for the analysis of metabolites and proteins in human body fluids. Anal Bioanal Chem 399(8):2635–2644
66. Ardrey RE (2003) Liquid chromatography-mass spectrometry: an introduction. Wiley, London
67. Hübschmann HJ (2008) Handbook of GC/MS. Wiley VCH Verlag GmbH, Weinheim
68. Mcmaster mC (2005) LC/MS: a practical user's guide. Wiley, New York
69. Mims D, Hercules D (2004) Quantification of bile acids directly from plasma by MALDI-TOF-MS. Anal Bioanal Chem 378(5):1322–1326
70. Breitling R, Pitt AR, Barrett MP (2006) Precision mapping of the metabolome. Trends Biotechnol 24(12):543–548
71. Murray KK (2010) Glossary of terms for separations coupled to mass spectrometry. J Chromatogr A 1217(25):3922–3928
72. Schwartz JC, Senko MW, Syka JE (2002) A two-dimensional quadrupole ion trap mass spectrometer. J Am Soc Mass Spectrom 13(6):659–669
73. Casetta B, Tagliacozzi D, Shushan B, Federici G (2000) Development of a method for rapid quantitation of amino acids by liquid chromatography-tandem mass spectrometry (LC-MSMS) in plasma. Clin Chem Lab Med 38(5):391–401
74. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E (2009) Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal Chem 81(16):6656–6667
75. Dubois F, Knochenmuss R, Zenobi R (1999) Optimization of an ion-to-photon detector for large molecules in mass spectrometry. Rapid Commun Mass Spectrom 13(19):1958–1967
76. Kromidas S (1999) Validierung in der analytik. Wiley VCH Verlag GmbH, Weinheim
77. McNaughton AD, Wilkinson A (1997) Compendium of chemical terminology. Blackwell Science, Oxford
78. Holmes C, McDonald F, Jones M, Ozdemir V, Graham JE (2010) Standardization and omics science: technical and social dimensions are inseparable and demand symmetrical study. OMICS 14(3):327–332
79. McGaw EA, Phinney KW, Lowenthal MS (2010) Comparison of orthogonal liquid and gas chromatography–mass spectrometry platforms for the determination of amino acid concentrations in human plasma. J Chromatogr A 1217(37):5822–5831

Chapter 4
Statistical Methods in Genetic and Molecular Epidemiology and Their Application in Studies with Metabolic Phenotypes

Christian Gieger

1 Basic Genetics

The basic assumption in genetic epidemiology is that the human *genome* is the entity of a human's hereditary information and individual variations of this information are the major reasons for heritable disorders. The genome is made up of *DNA* (deoxyribonucleic acid) which consists of a long sequence of nucleotide base pairs of four types: adenine (A), cytosine (C), guanine (G), and thymine (T). Under native conditions, in the nucleus of a human cell, DNA is double stranded with complementary base pairing with nucleotide A binding only to nucleotide T, and nucleotide C binding only to nucleotide G. Double-stranded DNA is replicated by breakage of the two strands and construction of a new complementary strand for each, resulting two identical copies of the original. A single strand of DNA acts as a template for a complementary strand of RNA. In transcription, i.e. in the process of copying DNA into messenger RNA in gene expression, thymine (T) is replaced by uracil (U).

The human genome is distributed on 46 *chromosomes* consisting of 22 homologous pairs of autosomes and 1 pair of sex chromosomes. The complete set is the diploid complement. One chromosome in each of the 22 homologous pairs is derived from the mother and the other from the father. Gametes (sperm and ova) are haploid, i.e. they contain only one member of each homologous chromosomal pair (e.g. only one version of chromosome 14). All ova have chromosomal complement X and sperm are either X or Y. When sperm and ova fuse to form a zygote, the full set of chromosomes is formed. Thus the sperm determines the gender of the offspring with chromosomal complements (X, X) for females and (X, Y) for males.

In certain regions of the DNA, which we call *genes*, transcribed mRNA encodes instructions that tell the cell how to assemble amino acids to form proteins.

C. Gieger (✉)
Helmholtz Center Munich – German Research Center for Environmental Health,
Institute of Genetic Epidemiology, Ingolstaedter Landstr. 1, Neuherberg 85764, Germany
e-mail: christian.gieger@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_4,

We assume that mainly altered protein function caused by changes in the DNA sequence affects health and disease but also other molecular mechanisms as changes of methylation patterns may play an important role in disease manifestation. The two homologues will have the same sequence of genes in the same positions, but they will usually exhibit sequence variations at several loci and can therefore be distinguished. The haploid genome is about three billion base pairs long. 99.9% of the genome of any two unrelated individuals is identical. About 3% of the genome consists of coding sequences, and there are 20,000–25,000 protein-coding genes.

If the DNA sequence at a given genetic locus (often a gene) varies between different chromosomes in the population, each different version is an *allele*. If there are two alleles at a given locus, the allele that is less common in the population is called the minor allele. A cell is *homozygous* for a particular locus or gene when identical alleles of the locus or gene are present on both homologous chromosomes. A cell is *heterozygous* for a particular locus or gene when two different alleles occupy the gene's/locus's position on the homologous chromosomes. We define the terms *dominant* and *recessive* with respect to the way how alleles interact to form a phenotype. In cases with two different phenotypes, the phenotype which is developed for a heterozygote is called the dominant whereas the complementary is called recessive. In the case of dominance the heterozygote is phenotypically identical to the corresponding homozygote. In more complex dominance schemes the results of heterozygosity can be more complex. The *genotype* is the genetic constitution of a cell or an individual, i.e. the specific allelic makeup of the individual. *Genotyping* is the process of elucidating the genotype of an individual with a biological assay. Commonly used techniques include PCR, DNA sequencing, and nucleic acid hybridization to DNA microarrays.

The DNA sequence may vary between two versions of the same chromosome in several ways. Today, in particular in genome-wide association studies (GWAS) the most important structural class consists of single nucleotide polymorphisms. A single nucleotide polymorphism (*SNP)* represents a variation in a single nucleotide and different possible variants are the alleles of the SNP. SNPs may occur within protein coding sequences of genes, non-coding regions of genes, or in the intergenic regions. Usually the SNP is not causal for a disease but due to linkage disequilibrium (see below) in an association study the location of the causal disease variant can be inferred to an interval of highly correlated variants. Although individual SNPs might carry limited information, their ease of typing and large number mean that they are widely used in genetic association studies. However it is important to note that there are several more complex structural variations affecting more than a single position in the genome. For instance, copy number variations (CNV) are modifications of the DNA sequence resulting in alternated numbers of copies of the DNA sequences in the genome. CNVs can be either deletions of DNA sequences compared to the common state or duplicates having more than one copy of the commonly observed sequence. Recent large scale genome wide studies have shown that rare CNVs play a role in the development of complex diseases, in particular neurodevelopmental disorders [1–3] and in the variation of quantitative risk traits, like e.g. BMI [4], but the Wellcome Trust Case Control Consortium (WTCCC)

concluded from their study on common CNVs that these are unlikely to contribute greatly to the genetic basis of common human diseases [5].

A genotype represents the locus specific information for a homolog pair of chromosomes, e.g. for a SNP. A genotype has no natural order; in particular it does not contain information on assignment to mother or father. A *haplotype* is a composition of haploid genotypes and represents the allelic configuration at different loci along a single chromosome of a chromosome pair. Haplotypes are transmitted together and can be broken up by recombination events. In genetic associations studies a haplotype is a series of SNPs on a single chromosome. Array genotyping technologies allow for the determination of genotypes, but haplotypes can be only reconstructed by using statistical inference methods. Commonly used methods are EM algorithms and algorithms based Bayesian methods as implemented in the software Haploview [6], PHASE [7], or fastPHASE [8].

2 Basic Statistical Concepts

Statistical methods and models used in genetic association studies are basically identical to methods used in classical epidemiology. They rely on methods developed in descriptive and inferential statistics. There is only the need for some additional assumptions and definitions. We describe some basic concepts that are necessary for conducting a genetic study and assume for ease of description in the following that genetic variants are bi-allelic SNPs.

2.1 Allele Frequency

The *allele frequency* is the proportion of one allele relative to all alleles at a locus. It is calculated in a given study population and is often used to estimate the corresponding frequency of the allele in related populations. For instance data from the HapMap project (http://www.hapmap.org) and 1,000 Genomes project (http://www.1000genomes.org) is used as reference to predict allele frequencies in different ethnic groups as Europeans, Africans and Asians and can be used to depict the amount of genetic diversity. For a SNP, its *minor allele frequency* (*MAF*) is the frequency of the SNP's less frequent allele.

2.2 Linkage Disequilibrium

Linkage disequilibrium (LD) is the association or correlation of alleles of two or more SNPs that cannot be attributed to random inheritance, in other words LD is the presence of combinations of genotypes more often than expected in a pure random

process. Linkage equilibrium occurs when the genotype present for one SNP is independent of the presence of another genotype for a second SNP. LD can be quantified as the difference between observed and expected allele frequencies under the hypothesis of linkage equilibrium. Sets of SNPs can be merged together in *LD blocks* with high pairwise LD which are separated from each other by spots of high recombination breaking up LD structures. In association studies this block structure is used for testing whole regions for associations by selecting few 'tagging' SNPs from the LD block. These *tagging SNP* represent the whole region which includes often more than one gene but are usually in an association study not the causal variants. Subsequent functional analyses are necessary to further investigate these regions. These analyses include fine mapping of the region by sequencing to find potentially causal variants, genomic analysis of gene expression patterns, investigation of rare risk variants as causes of monogenic traits, epigenetic analyses and investigation of intermediate traits like metabolic profiles derived from metabolomics. The amount of LD between two markers can be measured as D' (D-prime, $0 \leq D' \leq 1$). D' is calculated as the difference between the observed number of co-occurrence of two alleles and their expected number under the assumption of linkage equilibrium divided by a normalising term that is the theoretical maximum for the observed allele frequencies. A value of 0 indicates that the two markers are in complete equilibrium, whereas a value of 1 represents the highest amount of disequilibrium [9]. D' is related to a second commonly used measure of LD the square of the allelic correlation coefficient R^2 ($0 \leq R^2 \leq 1$).

2.3 Hardy-Weinberg-Equilibrium

The *Hardy-Weinberg model* introduces a mathematical model that predicts under certain assumptions the frequency of offspring genotypes based on parental allele frequencies. It is formalized in a way that it describes equilibrium and not an evolutionary model. The Hardy-Weinberg model predicts that all allele frequencies will not change from one to the next generation. The *Hardy-Weinberg-Equilibrium (HWE)* is based on and requires the fulfillment of formal assumptions. These requirements are (1) random mating, i.e. individuals are pairing by chance, not preferably according to any of their genotypes or phenotypes, (2) no mutation, i.e. no new alleles occur at a locus, (3) no migration or emigration, (4) no genetic drift, i.e. the population is large enough that changes in allele frequencies due to random processes do not play an observable role, and (5) no selection, i.e. there exists no selective pressure for or against any trait. These assumptions are idealistic and can be used for genetic modeling but are rarely exactly met in practice, so that allele frequencies of a population slightly change from one generation to the next. But in large population based studies these conditions are often reasonably fulfilled so that HWE holds at least approximately for the vast majority of genetic loci.

Formally, for a SNP with two alleles, say A with allele frequency $P(A) = p$ and C with allele frequency $P(C) = q = (1\text{-}p)$, assuming HWE results in offspring with a

genotype frequency of $P(AA)=p^2$ for the homozygote AA, $P(AC)=2pq$ for the heterozygote AC and $P(CC)=q^2$ for the homozygote CC. This sums up to the same allele frequencies $P(A)=p$ for A and $P(C)=q$ for C also in the next generation so that the proportion is constant for the whole population. Statistically HWE means that in a population at a specific locus the alleles are independent and therefore in a sample the HWE can be checked by applying statistical tests for independence of distributions to the data, which in principal measure the amount of departure of the observed allele frequencies from the expected allele frequencies under the assumption of HWE. Violations of HWE can occur on a population level due to deviations from the basic assumptions (e.g. non-random mating and population stratification) but also on an experimental level due to any kind of non-random genotyping error. This is observed mainly due to insufficient genotyping quality resulting in missing genotype data (no calls) or calling errors (e.g. too many heterozygote genotypes by wrongly merging two genotype-call clusters). Genotyping errors can lead to spurious associations or false positives in cases where different patterns of genotyping errors occur for different levels of the phenotype, e.g. diseased and healthy [10]. Consequently, testing for HWE is also an additional instrument for data quality control.

2.4 Testing for Violation of Hardy-Weinberg-Equilibrium

In a sample a SNP can be tested for HWE or more precisely for whether there is a violation of the HWE by comparing the observed genotype count with the expected count under the assumption of true HWE. There are several statistical tests available to test the null hypothesis of HWE. In particular the χ^2-test for independence and Fisher's exact test are commonly in use. A detailed formal description and discussion of these tests is given e.g. by [11]. Briefly the χ^2-test statistic is a weighted sum of the squared distance between the observed and expected genotypes counts. The test statistic can be calculated from the genotype and allele frequencies and is compared to an appropriate quantile of the χ^2-distribution with one degree of freedom. If the observed genotype counts are unlikely, given the null hypothesis of HWE can be rejected. Alternatively Fisher's exact test can be used. A description of Fisher's exact test can be found in [12] and an example is given in [11]. Briefly the test calculates the probabilities of a particular number of heterozygotes given the allele frequencies, under the null hypothesis of HWE. The cumulative probability of obtaining at least the observed number of heterozygotes is calculated, and this is regarded as the *p*-value of the test.

Here, we want to give some general remarks on the practical use of these tests. The χ^2-test is easy to calculate and often used in practice but it should be noted that this test is not appropriate if any of the expected values are small because in this case the approximation by a χ^2-distribution becomes poor. This is in particular true for SNPs with a small MAF so that fields of the contingency table have small counts (<5 as a rule of thumb). For this reason Fisher's exact test became more frequently

used as this test allows for the exact calculation of the probability of a false rejection under the assumption of a true HWE. In a sample that is not ascertained based on a specific phenotype, the HWE test should be calculated for the full sample but in selected populations, in particular for case–control studies, HWE testing should be done with consideration. In most situations it is wise to perform the HWE test only in the subset of the controls because it is possible that a violation of HWE for associated SNP in the cases occurs due to a shift in allele frequency.

2.5 Design of Genetic Association Studies

In *genetic association studies* similar concepts for study design are used as in other well designed epidemiological studies. Commonly used designs include case–control studies comparing the frequency of SNP alleles in two different sets of individuals: The case group including individuals who suffer from the disease under investigation and the control group including individuals who are either disease free or are taken as a random sample of the general population assumed that the disease prevalence is not very high. An increased frequency of a specific allele or genotype in the case group compared with the control group indicates that the SNP allele may increase the disease risk. The decision whether a difference is significant is based on statistical inference such as χ^2-tests and logistic regressions. Often the aim of an association study is a detailed understanding of the disease pathogenesis. This can be investigated with a second type of study by appropriately selecting intermediate traits that are known to be risk factors for complex diseases. Such traits are most often measured on a quantitative scale allowing estimation of the amount of deviation from the normal state. Quantitative traits are best investigated in samples of a general, mostly healthy population and require other statistical methods such as t-tests and linear regression models. It should be kept in mind that in association studies significant associations are statistical correlations and do not necessarily indicate a causal relationship. Moreover, association can be mediated be other traits, as in the case of the *FTO* gene being associated with type 2 diabetes through its effect on obesity, or even confounding can occur, as in the case of apparent associations due to population admixture. A detailed overview of the variety of study designs and analysis methods is beyond the scope of this chapter but comprehensive descriptions are available in standard genetic epidemiological text books, e.g. [13].

2.6 Genetic Models and Single Locus Testing

For statistical testing of an association between a trait and genetic variants, in our case SNPs, a genetic model has to be specified to define the assumed underlying genetic mechanism. In general there are two model choices possible, namely allelic testing and genotypic testing. For disease traits, the first classifies cases and controls

according to their alleles; the second classifies cases and controls according to their genotypes. For genotypic testing the most general model allows for all possible genotypes an unrestricted effect. This leads for bi-allelic SNPs with three possible genotypes to a two-degrees of freedom test for the overall effect of the SNP. This model makes no structural assumption about differences in effect size between the genotypes but has an additional degree of freedom resulting in a test which is less powerful than models assuming the (true) genetic model, which is however in general unknown. In an *additive genetic model* each copy of the observed allele proportionally increases or decreases the effect on the outcome. A *dominate genetic model* for one of the alleles assumes that the heterozygote and the homozygote genotype of the allele are combined into a joint category. The same model is a *recessive genetic model* for the other allele. In genetic association studies the aim is to investigate whether a specific SNP is associated with the outcome under investigation. This is in general done by single locus testing without taking other SNPs into account, e.g. by conditioning on these potentially related loci in the regression model. But it might be nevertheless useful to include other important covariates into the model if they affect the phenotype independently of the genetic variant as they may increase the precision of the model. Moreover, adjustment for traits that are mediators between the SNP and outcome will remove seemingly direct associations with the outcome.

2.7 Multiple Testing

The type I error of a statistical test is defined as the probability to reject the null hypothesis "no effect of SNP on the phenotype" when there is no real association between the SNP and the phenotype; this means in other words we get a false positive result. The significance level of a test is the assumed type I error α for a single test. Genetic association studies, in particular genome-wide studies, require testing of large numbers of hypotheses for multiple SNPs and/or phenotypes. Very often these sets of SNPs and phenotypes, like in association studies with panels of metabolites, are considerably correlated. Analysis plans should appropriately consider adjustment for multiple testing. Reporting of significance in genetic association studies and genome-wide studies should be strictly based on study-wide or genome-wide significance.

The most straightforward but also most conservative correction for multiple testing is a Bonferroni correction. This correction method controls the family-wise error rate that is the probability of one or more type I errors among all tests. For instance, if we perform two tests each with a individual type I error α of 0.05 then the probability for at least one false positive decision is already $1-(1-0.05)^2=0.0975$. This inflation of the error probability means vice versa that controlling the family-wise error on a level of $\alpha=0.05$ requires to choose a test-wise type I error x accordingly. We have to assure that $1-(1-x)^N<0.05$ for N tests. Bonferroni correction uses an approximation $(1-x)^N\approx 1-N\cdot x$, so that we can choose $x=\alpha/N$,

i.e. we divide the test-wise type I error rate by the number of performed independent tests to get a corrected threshold for the aimed family wise error. The procedure assumes that all tests are independent and is therefore conservative in situations with correlated SNPs or phenotypes. Benjamini and Hochberg [14] proposed the false discovery rate (FDR) which is a less conservative adjustment. In contrast to controlling the family-wise error rate, the FDR controls the expected proportion of false discoveries among all rejected hypotheses. The q-value is defined to be the FDR analogue of the p-value [15, 16].

3 Genome-Wide Association Studies (GWAS)

The human genome contains several millions of SNPs. Some of these SNPs directly increase the susceptibility for a disease manifestation or cause individual modifications in phenotypes; others tag nearby causal variants making them useful as indirect markers of disease associations. *Genome-wide association studies (GWAS),* which are able to cover large amounts of the whole genomic variation, have become the standard analysis tool for risk loci discovery of complex diseases and related risk factors. This is possible due to the recent development of emerging chip technologies for DNA genotyping allowing for measuring of hundreds of thousands up to several millions of common SNPs. These chips are based on a dense coverage of common genetic variation provided by recent whole genome sequencing projects, like the HapMap project (http://www.hapmap.org) and the 1,000 Genomes project (http://www.1000genomes.org).

Genetic variants on genotyping arrays do not result necessarily in different protein sequences for both alleles but should cover such variants in high LD with the typed SNPs. In this sense GWAS based on genome-wide chips are an efficient and unbiased approach to find risk loci for common disorders. GWAS require sufficiently large sample sizes that allow for uncovering of risk loci with sufficient statistical power. It is, however, noteworthy that there are only few and far less than initially expected examples where single loci have a major impact with high explained variation on a phenotype. These so called 'low hanging fruits' have been detected in early GWAS with sample sizes of several thousands of individuals. In many of such cases these newly discovered high impact loci include a gene that is highly plausible as modulator of the phenotypes, e.g. a gene encoding C-reactive protein (CRP) [17] or the *SAA* gene family for acute-phase serum amyloid A [18]. In other cases the connection with a phenotype beyond the strong statistical association was less clear but has revealed a new pathway involved in the modulation of a phenotype and could be subsequently validated in functional studies, e.g. the *SLC2A9* gene as newly described high capacity urate transporter [19, 20].

It became obvious from early GWAS with several tens of thousands of individuals that most associated common variants taken for themselves have only moderate effects, often with odds ratios between 1.1 and 1.3 for disease traits, or with explained variances of <1% for quantitative traits. Consequently, for some traits there should

be hundreds of variants with such moderate and even smaller effect sizes. Recently, highly powered GWAS have used more than 100,000 individuals. This is realized in large multi-national consortia by putting together all available data from different studies in a single large meta-analysis. The studies have sufficient power to detect also association with small effect sizes. There is an ongoing discussion on the importance of such small effects. From an epidemiological point of view these loci may have a negligible relevance as they contribute only marginally to the overall risk of developing a disease. However it is important to note that the major aim of GWAS is not prediction of individual risk but rather the discovery of novel biological pathways involved in disease development [21]. This can in the future even provide information for an effective treatment. The ability to expose biology has already been shown in several large GWAS. For instance a large GWAS on serum lipid levels found many loci including genes encoding apolipoproteins, lipases and other key proteins known to be involved in lipid metabolism [22]. Moreover, the same study showed that 18 genes at the 95 lipid loci were previously linked to known Mendelian lipid disorders. This and other examples show that overall GWAS were very successful in discovery of relevant genes and can provide new hypotheses for the investigation of the genetics of complex diseases. A comprehensive database containing thousands of risk variants for several hundreds of traits and diseases is provided as Catalog of Published Genome-Wide Association Studies [23] by the National Human Genome Research Institute (NHGRI) (http://www.genome.gov/gwastudies/). However many newly identified loci do not implicate genes with already known functions, so that the underlying genetic etiology of complex disorders has to be explored with methods beyond statistical association analysis [24].

4 Candidate-Gene Association Studies Versus Genome-Wide Association Studies

Classical *candidate-gene association studies* investigate the effect of candidate markers in one or more genetic region on a phenotype of interest. Typically the genetic loci are characterized by sets of tagging SNPs. The SNPs tag the genetic region as they are selected to represent each distinct LD-block. Candidate gene studies limit the number of tests to small subsets of the genomic regions and focus on hypotheses for sets of genes which have prior evidence to be associated with the phenotype of interest. GWAS are in contrast to candidate-gene studies hypotheses-free and unbiased as they require no prior selection of candidates. Genetic association studies with sets of candidate genes have suffered from poor replicability of reported results. This has been among others impressively illustrated in a systematic review of replication studies by Hirschhorn and colleagues [25] who reviewed 600 association signals between common gene variants and disease previously reported in literature. They found that only six of 166 associations studied three or more times were consistently replicated, while for more than half of the loci replication attempts yielded inconsistent replication outcome. Successful replication of an

initial finding is the *sine qua non* in candidate gene association studies but one should also be aware that non-replication can reflect not only false positives calls in the original study but also problems in replication designs, like insufficient power or methodological inconsistency between study designs, which can lead to false-negative results in replications. The advent of GWAS is a notable change in paradigm of genetic association studies. Due to their unbiased approach in sufficiently powered studies the discovery of reliably validated association signals underwent a major improvement compared to previous candidate-gene association studies. However, it should not be ignored that a completely unbiased approach in GWAS has also a price, namely the relatively low power to detect signals with moderate and small effect sizes. This low power is mainly caused by the need to extensively correct for multiple testing. In GWAS several millions of SNPs are tested for association. Due to linkage disequilibrium many of these SNPs are highly correlated leading to an over-conservative correction when using the total number of SNPs in a Bonferroni correction. Pe'er et al [26] estimated based on data from HapMap Consortium a testing burden of one million independent tests in a dense genome-wide analysis in Europeans. Based on this estimate most GWAS have so far used $5*10^{-8}$ as threshold for genome-wide significance.

5 Meta-analysis of Genome-Wide Association Studies

Motivated by initial successes of GWAS in single studies, researchers started to extensively exchange and combine genetic data to uncover new genetic risk factors in large populations. The primary statistical tool for such a joint effort is a meta-analysis for pooling statistical evidence of genotype-phenotype associations. Prerequisite is the imputation of data collected with different genotyping platforms based on the same haplotype reference set to enable the exchange of data in a unified format. Imputation is a statistical method to predict known SNPs which are not genotyped with the used chip technology based on nearby genotyped SNPs and observed reference haplotypes from HapMap or another large scale sequencing effort. Software tools have been developed to solve this missing value problem efficiently [27, 28]. Not surprisingly, imputation methods are not perfect. Quality depends on the LD structure between genotyped and untyped SNPs. All imputation algorithms calculate for each imputed SNP for a given individual a posterior probability for the possible genotypes which can be transformed into the expected number of copies of an allele (allelic dosage). In subsequent association analysis this uncertainty of the imputations has to be taken into account. Standard linear and logistic regression allows incorporating imputation uncertainty by using allelic dosages instead of coded genotypes.

A meta-analysis of GWAS provides the possibility to combine study-specific statistics in single weighted statistics. In contrast to a full joint analysis of all available data, a meta-analysis does not need access to individual data but needs instead summary statistics from all included studies. This makes such large analyses

organizationally and computationally feasible and is the analytical basis of efforts in large consortia. Meta-analyses of GWAS use the same statistical methods that are commonly used for combining summary statistics in meta-analyses of published results, in particular it calculates joint effect estimates and p-values [29, 30]. By increasing the effective sample size and power, meta-analyses have shown to be extremely useful for the discovery of new genetic risk factors and gene functions. Divergent huge consortia like DIAGRAM for type 2 diabetes [31], CARDIoGRAM for coronary artery disease [32], GIANT for height [33], GLGC for serum lipids [22], or the International Consortium for Blood Pressure Genome-Wide Association Studies [34] have identified novel associations and have impressively illustrated the value of meta-analysis across GWAS.

Before starting a meta-analysis it should be carefully evaluated which studies to include. They should be comparable with respect to their study population and study design. Today most large meta-analyses are based on populations of European ancestry but often a validation in other ethnicities follows the main analysis. For example it may make sense to further explore loci discovered in Europeans also in populations of African ancestry as they tend to have shorter LD blocks allowing in some cases to further narrow down causal variants. In meta-analyses for quantitative traits an analysis in healthy samples from the general population is seen as gold standard for an unbiased analysis but inclusion of a minor fraction of individuals with disorders which might be physiologically related to the phenotype is accepted as exclusion of all such individuals is not always practically feasible. Sometimes even populations selected according to a phenotype are included as it is believed that the gain of power outperforms a possible dilution of effects. These considerations show that a perfect degree of homogeneity is mostly not possible, but a good study design can be approached by carefully selecting studies and defining a unified analysis plan.

An important requirement for any meta-analysis is the comparability of the phenotype as well as the genotype across all participating studies. In many studies phenotype comparability is relatively easy to assure, as for instance for human height or weight unified measurement devices are available. However, for other phenotypes such standards are not available and measurements depend on the device that is used in a particular study, e.g. for metabolite measures using different profiling platforms. Similar problems can occur for the definition of disease status as there might be heterogeneity due to different disease classifications or ascertainments, like self-reported or clinically validated status. It is essential that for all studies the phenotypes are transformed and coded in terms of the same scale or unit for quantitative traits and categories for categorical traits. For genotype data, it is of particular importance to define a unified way of data cleaning and quality control. Prior to analysis, typically several filtering steps are applied including call rates, imputation quality and HWE outliers. To ensure consistent comparability between studies, a unified annotation of the variant names and strand orientation is essential. Furthermore an explicit unique declaration of the version of the human genome assembly (e.g., NCBI build 35 or 36), the rs-identifier with version of NCBI dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), the chromosomal position relative to

the used genome assembly, the alleles and their strand orientation is required for an accurate analysis. Overall an elaborated concept for quality assurance is essential for a sound analysis step. Furthermore, the genotype coding has to be done with respect to the same reference allele. This assures that effect estimates or odds ratios are consistent across studies and can be combined to an overall estimate in the meta-analysis. Meta-analysis itself is usual performed using fixed effects methods (weighted z-scores or inverse variance method) or random effects methods [29]. Once the meta-analyses for all SNPs has been calculated it is useful to check the results for heterogeneity by computing Cochran's Q statistic as well as the I^2 statistic [35].

6 Population Stratification

Population stratification refers to any systematic difference in allele frequencies between different subpopulations of a larger population. A proper consideration is critical for GWAS analysis and meta-analysis of GWAS [36, 37]. Population stratification can be caused by different ancestries due to non-random mating, genetic drift or population admixture. A major concern in large scale association studies is that such studies tend in the presence of population stratification to generate more false positives than expected by chance. It plays an important role in genetic case–control association studies as they basically assume the differences in allele frequencies between cases and controls are due to the disease status and not due to differences based on mixtures of ancestrally distinct populations with different values of disease prevalence. However, population stratification can also occur in the analysis of quantitative traits whenever the membership to a genetically different subpopulation is correlated with the outcome. Several methods have been proposed to test and correct for population stratification in genetic association studies, in particular genome-wide studies. These approaches use either set of markers which are assumed to be uncorrelated to outcome loci or the complete set of available set of markers.

Population stratification has been discussed extensively in the literature and a number of methods have been proposed to address it. A commonly used method for correcting population stratification is genomic control (GC) [38–40]. An alternative approach is the usage of principle components in regressions to adjust in the model itself for population structure. The idea is to use the first few principal components of the correlation matrix of SNPs as covariates to capture underling structures not attributable to single SNPs but to structures [37]. After a genome-wide association analysis, it is essential to test the resulting test statistic for signatures of population stratification. It is standard to compare the genome-wide distribution of the test statistics with expected null distribution by calculation of the GC inflation factor λ and by inspecting Q–Q (quantile–quantile) plots of observed p-values. The GC inflation factor λ is defined as the ratio of the median of the empirically observed test statistics to the expected median and should be close to 1 for homogeneous

populations [40]. Q-Q-plots are useful to detect visually deviations of the observed test statistics or p-values from the expected null distribution. As the null distribution is calculated under null hypothesis of no association, an exclusion of known associations makes often sense. Inflated λ values or deviations in the Q–Q plots may indicate to uncorrected population stratification.

7 Case Study: Genetic Determinates of Uric Acid Levels and Their Associations with Gout

Uric acid is a metabolite that is created when the organism breaks down purine nucleotides (Fig. 4.1). Uric acid levels in blood serum are of clinical relevance as elevated concentrations can lead to gout and are associated with medical complications like kidney stones and hypertension. Physiological reasons for elevated uric acid levels are twofold: an increased production in liver and/or an insufficient excretion in kidneys. Although it is known that a strong genetic control influences the regulation of uric acid concentrations, until the first GWAS have been conducted there was no major gene regulating uric acid levels was known. This first wave of GWAS uncovered in up to 28,000 individuals a total of 9 loci with reproducible influence on uric acid [20, 41–43] (Fig. 4.1). The strongest effect on uric acid concentrations was detected for the gene *SLC2A9*, coding for the protein GLUT9, which has been later shown to serve as a high-capacity urate transporter in humans [19, 20]. The association is additive, with a large effect size of −0.35 to −0.40 mg/dL per copy of the minor allele (MAF = 23%). There is a pronounced gender difference

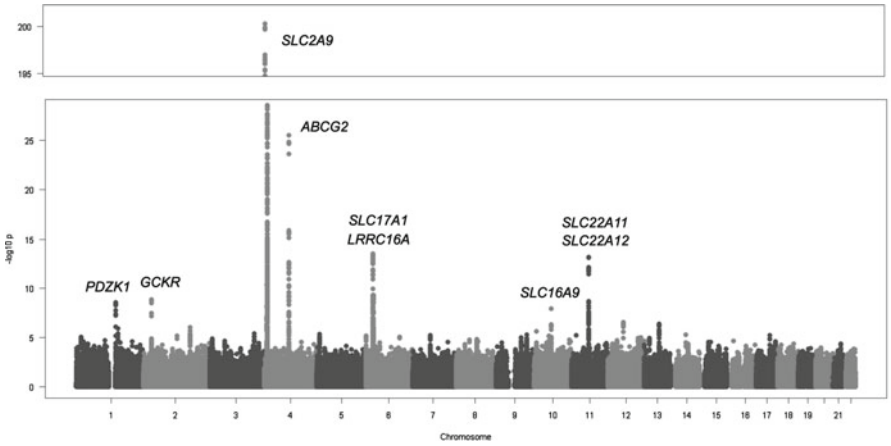


Fig. 4.1 Manhattan plot showing significance of association of all SNPs in the meta-analysis of uric acid from 14 studies totalling 28,141 participants of European descent. SNPs are plotted on the x-axis according to their position on each chromosome against association with uric acid concentrations on the y-axis (shown as − log10 p-value). Figure reproduced from [41]

with an explained variance of 1.2% in men and 6% in women. In addition there is a significant association with gout with an odds ratio of 1.68 per copy of the common risk allele. The second strongest association can be observed at *ABCG2* that is a member of the ATP-binding cassette (ABC) superfamily of membrane transporters. Woodward et al. [44] were the first to show that *ABCG2* is in fact a uric acid efflux transporter. Variants result in 53% reduced uric acid transport rates compared to wild-type *ABCG2* and show highly significant associations with gout (OR = 1.68 per risk allele). Remarkably the study shows that at least 10% (MAF = 11%) of all gout cases are attributable to a causal variant in the gene. These examples emphasize the power of GWAS on metabolite traits in expanding our understanding at the molecular level of disease mechanisms. In both cases the function of these genes were not known before the studies have been conducted. This shows the power of the analysis of intermediate traits as this can lead to valuable insights into the pathophysiology of complex diseases.

8 Missing Heritability

Genome-wide association studies have identified hundreds of genetic loci associated with complex diseases and related intermediate traits (Fig. 4.2). These studies have provided important insights into the genetic architecture of diseases and contributed to the understanding of underling disease mechanisms. However, only in few cases mostly for quantitative traits these studies could detect loci with large effect sizes which could explain a high fraction of the trait variation. The majority of variants identified so far are associated with small increases of disease risk or small to moderate differences in the level of a quantitative trait per copy of an allele. These loci explain mostly less than 1% of the variation of a quantitative trait and hence also only a small fraction of the heritability. This leads directly to the question where the missing heritability can be found and which approaches to generalize the commonly used procedures is most promising.

A major aim of GWAS is the generation of unbiased hypotheses by reporting a list of validated and novel loci which are associated with a trait solely based on large sample sizes and statistical modelling. In subsequent studies these loci can be further validated either based on biological knowledge databases, functional experiments and analyses in model organisms. To avoid wasting too much work and money for investigations of false positive calls, GWAS typically set rather stringent statistical thresholds for significance to strictly control false positive rates. Usually they use a Bonferroni correction threshold, e.g. $P < 5 \times 10^{-8}$, assuming testing in a total of one million independent genomic regions [26]. The consequence of such a conservative approach is a high false negative rate, i.e. the inability to detect loci that are truly associated with the trait having effect sizes that are too small to pass the ‘magic threshold’ of genome-wide significance. This leads even for GWAS with more the 100,000 individuals to estimates that up to several hundreds of truly associated SNPs with moderate effect sizes are still uncovered [33, 47, 48]. Another

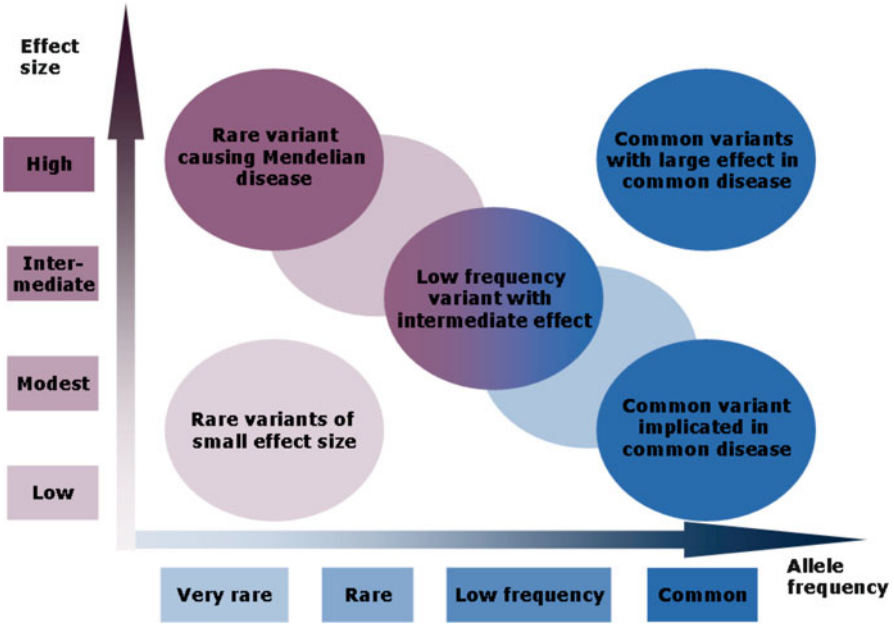


Fig. 4.2 Feasibility to identify genetic variants by risk allele frequency and effect size. GWAS are powered to identify common variants contributing to the inherited component of common diseases (Adapted from [45, 46])

reason for missing true positive loci is owed to the fact that large GWAS use predominantly one genetic model, namely additive models assuming a proportional increase in effect size per copy of the risk allele. While clear dominance should be covered well by an additive model assumption, a recessive model is not very likely to be detected by such a model. Even with very large sample sizes, we have still limited ability to detect rare to moderate frequent variants. This has several reasons: Firstly the majority of commonly used DNA chips are not designed to cover SNPs with a MAF < 1%. Even custom-designed chips have the problem to appropriately and reliably call the genotypes of rare variants as the commonly used calling algorithms work not very well for rare variants based on limited sample sizes. Moreover, due to very limited statistical power most genome-wide association studies do not even try to discover such low frequency variants and exclude them from the analysis plan from the beginning. There might be some progress due to upcoming sequencing projects but insufficient power is still an issue. It can be expected that only genes with highly penetrant rare variants show up. Rare variants with small to moderate effect size are extremely hard to detect by such studies (Fig. 4.2). Apart from these limitations due to study designs, chip designs and statistical power, there are some more fundamental considerations leading to reasons for missing heritability. Mangolio and colleagues [45] included in their seminal paper several more structural sources. They expect contributions to the elucidation of this puzzle by

performing systematical investigations of copy number variations (CNVs), epistasis, i.e. situations where several genes work together, and perhaps most promising epigenetics, i.e. changes in gene expression that are inherited but not caused by changes in the genetic sequence.

References

1. Sebat J, Lakshmi B, Malhotra D et al (2007) Strong association of de novo copy number mutations with autism. Science 316(5823):445–449
2. Stefansson H, Ophoff RA, Steinberg S et al (2009) Common variants conferring risk of schizophrenia. Nature 460(7256):744–747
3. The International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237–241
4. Jacquemont S, Reymond A, Zufferey F et al (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478:97–102
5. The Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464(7289):713–720
6. Barrett JC, Fry B, Maller J, Daly M (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265
7. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978–989
8. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644
9. Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49(1):49–67
10. Moskvina V, Craddock N, Holmans P et al (2006) Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered 61(1):55–64
11. Weir BS (1996) Genetic data analysis 2: methods for discrete population genetic data. Sinauer, Sunderland
12. Agresti A (2002) Categorical data analysis. (Wiley series in probability and statistics). Wiley-Interscience, Indianapolis
13. Ziegler A, König IR (2006) A statistical approach to genetic epidemiology: concepts and applications. Wiley-VCH, Indianapolis
14. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B Methodological 57:289–300
15. Storey J, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci 100:9440–9445
16. Storey JD (2002) A direct approach to false discovery rates. J Roy Stat Soc Ser B 64: 479–498
17. Dehghan A, Dupius J, Barbalic M et al (2011) Meta-analysis of genome-wide association studies in >80,000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123(7):731–738
18. Marzi C, Albrecht E, Hysi PG et al (2010) Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A. PLoS Genet 6(11):e1001213
19. Caulfield MJ, Munroe PB, O'Neill D et al (2008) SLC2A9 is a high-capacity urate transporter in humans. PLoS Med 5(10):e197

20. Vitart V, Rudan I, Hayward C et al (2008) SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat Genet 40(4):437–442
21. Hirschhorn JN (2009) Genomewide association studies–illuminating biologic pathways. N Engl J Med 360(17):1699–1701
22. Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466(7307):707–713
23. Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106(23):9362–9367
24. Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360(17):1759–1768
25. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4(2):45–61
26. Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32(4): 381–385
27. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4):413–417
28. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816–834
29. de Bakker PI, Ferreira MAR, Jia X et al (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17(R2):R122–R128
30. Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10(2):191–201
31. Voight BF, Scott LJ, Steinthorsdottir V et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42(7):579–589
32. Schunkert H, Köning IR, Kathiresan S et al (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43(4):333–338
33. Lango Allen H, Estrada K, Lettre G et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832–838
34. Ehret GB, Munroe PB, Rice KM et al (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478(7367):103–109
35. Ioannidis JP, Patsopoulos NA, Evangelou E (2007) Uncertainty in heterogeneity estimates in meta-analyses. BMJ 335(7626):914–916
36. Campbell CD, Ogburn EL, Lunetta KL et al (2005) Demonstrating stratification in a European American population. Nat Genet 237(8):868–872
37. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909
38. Bacanu SA, Devlin B, Roeder K (2002) Association studies for quantitative traits in structured populations. Genet Epidemiol 22(1):78–93
39. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66(6):1933–1944
40. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55(4):997–1004
41. Kolz M, Johnson T, Sanna S et al (2009) Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. PLoS Genet 5(6):e1000504
42. Döring A, Christian G, Mehta D et al (2008) SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet 40(4):430–436
43. Yang Q, Köttgen A, Dehghan A et al (2010) Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. Circ Cardiovasc Genet 3(6):523–530
44. Woodward OM, Köttgen A, Coresh J et al (2009) Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. Proc Natl Acad Sci USA 106(25):10338–10342

45. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747–753
46. McCarthy MI, Abecasis GR, Cardon LR et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356–369
47. Park JH, Wacholder S, Gail MH et al (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42(7):570–575
48. Yang J, Manolio TA, Pasquale LR et al (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43(6):519–525

Chapter 5
Ultrahigh Resolution Mass Spectrometry Based Non-targeted Microbial Metabolomics

Michael Witting, Marianna Lucio, Dimitrios Tziotis, and Philippe Schmitt-Kopplin

1 Microorganisms and Their Role in the Environment and Human Health

Microorganisms place the biggest part of the biomass in the world covering all branches of life and were the first form of life on earth [1, 2]. They were discovered 1675 by Anton van Leeuwenhoek using an self-designed light microscope [3–5]. Figure 5.1a shows a first drawing of microorganisms, at this time referred as *animalcules*, compared to a today used phylogenetic tree (Fig. 5.1b).

His contribution together with the works of Lazzaro Spallanzani, Louis Pasteur and Robert Koch, are regarded as the fundament of modern microbiology. It is estimated that the total amount of prokaryotic cells by itself on earth is $4–6 \times 10^{30}$, with $350–550 \times 10^{15}$ g carbon bound in their biomass [6]. Bacteria, as other microorganisms, are ubiquitous and can be found in various habitats and ecological niches, ranging from soil, acidic hot springs to radioactive waste [7]. Bacteria do not always occur as single individual cells but also as big communities, which are not only consisting of one isolated species but appear as multispecies and often associated with other higher organisms. Many bacteria fulfill an important role in the environment, such as nitrogen fixation from the atmosphere, whereas other are important to human health, e.g. the gut microbiome (= all microorganisms living in human gut). In the human flora, meaning the skin and gut flora, there are about ten times more bacterial cells than there are human cells in the body, with thousands of different species alone being in the gut [8, 9]. Nevertheless, several bacteria can serve as pathogens for plants, animals and humans. Diseases like tetanus, typhoid fever, diphtheria, syphilis, cholera or tuberculosis are pathogen born. Again, bacteria are

M. Witting • M. Lucio • D. Tziotis • P. Schmitt-Kopplin (✉)
Helmholtz Zentrum München, Research Unit Analytical BioGeoChemistry,
Ingolstädter Landstraße 1, Neuherberg, Bavaria 85764, Germany
e-mail: michael.witting@helmholtz-muenchen.de; marianna.lucio@helmholtz-muenchen.de;
Dimitrios.tziotis@helmholtz-muenchen.de; schmitt-kopplin@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_5,

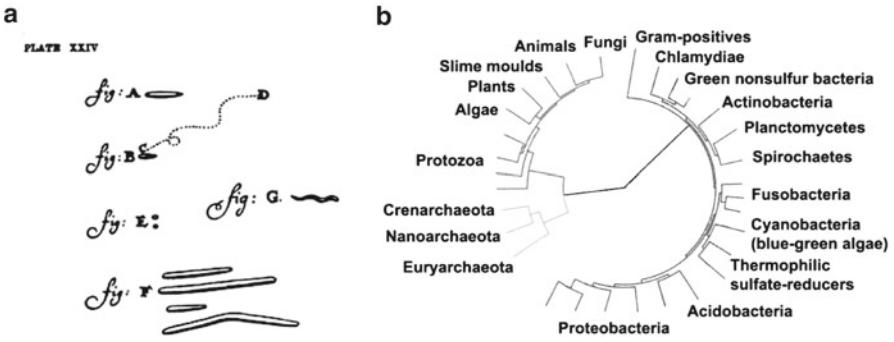


Fig. 5.1 (**a**) First drawings of bacteria, at this time called *animalcules*, obtained by Anton van Leeuwenhoek using a self-designed light microscope. (**b**) Modern phylogenetic tree of life today in use, presenting all branches of life. http://en.wikipedia.org/wiki/File:Collapsed_tree_labels_simplified.png

used to our benefit in food industry, e.g. in the production of cheese or yoghurt, or in biotechnology to produce fine chemicals. An example is Insulin, which is today produced exclusively in genetically modified *Escherichia coli* strains or yeast. Strain optimization is a key issue in biotechnology today. Beside their technical relevance and their role in health, microbes are an extensive part of fundamental research.

Each year more new species of microbes are described than in other phyla [10]. With these new species, novel pathways are going to be discovered. The best example is the reverse TCA cycle being long postulated, using the TCA cycle in reductive way for CO_2 fixation [11]. Different types of production of energy for survival are known from bacteria. It can be distinguished between phototrophs, lithotrophs, and organotrophs, using sunlight, inorganic compounds or organic compounds as energy source, respectively. This broad range of metabolic capabilities, mostly reflecting the ecological niche a bacterial species lives within, makes them an interesting field for metabolomics. In addition they feature a broad variety of secondary metabolic pathways, producing toxins or metabolites, specific for the environment they live in, e.g. chelators for sequestering iron or other essential metals. Moreover bacteria fulfill metabolic reactions not known from higher organisms. Research is going to identify more and more novel bacterial metabolites in future and will show us that the chemical diversity of microbial metabolomes is much bigger than what we can presently imagine.

2 Review of Current Advances in Microbial Metabolomics

Metabolomics, as the omics-technique being the closest to the observed phenotype and the real endpoint of the traditional view of biological information, flowing from DNA, over RNA and proteins to metabolites, is also one of the most challenging.

DNA, RNA and proteins are built out of defined chemical building blocks, whereas metabolites are ranging from very polar to non-polar compounds. Moreover they are present in a high dynamic range of concentrations ranging in orders of magnitude, making it (today) impossible to measure all of them in a single analysis. Despite this fact, metabolomics is currently growing fast in biotechnology, microbiology, systems biology and many other fields. Until today a lot of research in microbial metabolomics has been investigated on optimization of extraction protocols and evaluation of different analytical platforms. Quenching of metabolism, meaning the stopping of all biochemical reactions and inactivation of enzymes, is achieved by different techniques, whereas cold methanol or liquid nitrogen are the most commonly used. In the case of stability of extracted metabolites, the adenylate charge is preserved quite well and protects labile species from oxidation or degradation, making it well suited for microorganisms. A nice comparison of different quenching techniques and extraction methods from different sources can be found in Villas-Boas [12]. For the measurement of the metabolome, different analytical platforms are applied, whereas one of the oldest analytical platform is gas chromatography–mass spectrometry (GC-MS), which is still in use. GC-MS enables high-resolution separations, but it is only applicable to volatile compounds or such compounds that are volatile after a certain derivatization, restricting this method to metabolites smaller 500 Da. A recent study from E. Frimmersdorf et al. used GC-MS to elucidate how *Pseudomonas aeruginosa* adapts to different environments. Two different *P. aeruginosa* strains, PA01 and TBCF10839, were grown on either minimal medium with different carbon sources or a complex medium containing tryptone. During the exponential phase, metabolites directly available as carbon sources and metabolites belonging to the central carbon metabolism were found in higher concentrations, whereas in the stationary phase metabolites connected to production of exopolysaccharides, development of biofilms and rhamnolipids were found [13]. Such findings can be interesting and helpful for further investigations studying host-pathogen interactions of *P. aeruginosa*, which is a major threat to human health.

Beside GC-MS, liquid chromatography mass spectrometry (LC-MS) is widely used. It is applicable to even non-volatile compounds, while the use of ultrahigh-resolution liquid chromatography (UHPLC) or capillary liquid chromatography (capLC) provides separation efficiencies comparable to that of GC-MS, yielding several hundreds of thousands theoretical plates. The character of the used LC separation methods adds additional information about metabolite polarity for more precise metabolite identification. In metabolomics research mostly reversed phase (RP) and/or hydrophilic liquid interaction chromatography (HILIC) is performed to separate non-polar and polar compounds, respectively. Ballistic, meaning short time, gradients are often employed for a rough separation of metabolites, to overcome effects of ion-suppression, known from direct infusion experiments. This gradient times range from 1 or 2 to about 15 min, making high-throughput measurements possible. The amount of detected metabolites varies from several hundreds to thousands. After a first screening approach, longer gradients allowing more precise separation of metabolites are used for identity confirmation or purification of unknown substances. A recent study used RP-capLC-MS to investigate the effect of the gut microflora on the mouse blood system. Additionally different ionization techniques, electrospray ionization

(ESI) in positive and negative mode and atmospheric pressure chemical ionization (APCI) in positive mode, were used. Hertkorn et al. already showed on the basis of natural organic matter that using complementary approaches in ESI(+/−), APPI(+/−) and APCI(+/−) increases the available chemical space significantly [14]. Although the microbiome was not directly analyzed, the work of Wikoff et al. shows how big the influence of it on mammalian metabolism is. Analyzing the plasma of conventional and germ-free mice by RP-capLC-MS, they found several bacterial metabolites, being significantly different between the two groups. One example pointed out in this work is the degradation of food tryptophan by bacteria. Germ-free mice lacked bacterial degradation products like indole-3-propionic acid or indoxyl sulfate, build from indole by the bacteria, followed by an oxidation by cytrochrome P-450 and a phase II conjugation to sulfate in the liver [15].

A third way of metabolite analysis is capillary electrophoresis-MS (CE-MS), which separates charged species in an electrolyte filled fused silica capillary using a high electric field. It is applicable to charged compounds, like nucleotides or small organic acids from the TCA cycle. Exemplarily the work of T. Soga et al., who developed three different CE-MS methods for cationic metabolites, anionic metabolites and nucleotides and coenzyme A compounds, respectively, for metabolome analysis, should be mentioned here. Using this method setup, 352 metabolites, which are commercially available as standards, from different known metabolic pathways, could be determined and the whole platform was applied to 1,692 compounds from *Bacillus subtilis*. Different runs in 30 m/z steps were used to enhance sensitivity, leading to totally 33 repetitions per sample to cover the mass range from 70 to 1027 m/z. Unknown metabolites were searched against the LIGAND database and prediction of the compound identity using its charge, electrophoretic mobility and isotopic contributions was performed. Finally this approach was applied to metabolome changes in *B. subtilis* sporulation. The results, indicating accumulation of TCA cycle metabolites, are in good agreement with previous studies, demonstrating the power of this setup [16].

A major goal in metabolomics is to achieve whole metabolome coverage, which is so far only possible by combining different analytical methods. One of the most extensive works on microbial metabolomics was carried out by van der Werf et al. The group employed several methods to achieve comprehensive metabolome coverage. The methods included GC-MS with prior oximation and silylation to derivatize non-volatile compounds (OS-GC-MS), ion-pair liquid chromatography-mass spectrometry (IP-LC-MS), hydrophilic liquid interaction chromatography-MS (HILIC-MS), a LC-MS method for lipids, a non-polar GC-MS method and finally a GC-MS method for volatile compounds. As basis for their investigation, they used in silico predicted metabolomes for *Escherichia coli*, *Bacillus subtilis*, and *Sacharomyces cerevesiae*. The OS-GC-MS and IP-LC-MS methods showed the most detected metabolites [17].

The noted examples show nicely how the today used analytical setups are able to analyze the metabolome in a targeted way. Different analytical technologies needed to be combined to achieve good metabolome coverage. Also important, such work focusing on previous database knowledge is in most cases not able to find new metabolites or pathways. A promising approach, to fill this gap is non-targeted metabolomics using ultrahigh resolution analytics, which can help to gain further

results, that are beyond the scope of the previous presented. Beside this, techniques like GC-MS, LC-MS or CE-MS can be also used in a non-targeted manner, which is often used, but limited resolution avoids putative metabolite identification. For a further overview, read the reviews published elsewhere [18–20].

3 Non-targeted Metabolomics Using Ultrahigh-Resolution Analytical Platforms

As paradigm is changing from basic metabolomics setups to more holistic approaches and the goal to discover novel pathways and metabolites for more scientist ultrahigh resolution analytical platforms come to the fore. Systems biology is a key word metabolomics wants to deal with, assigning function to orphan genes. With ultrahigh resolution non-targeted metabolomics this is in the scope of today possible analytics. Non-targeted metabolomics, as an extension to metabolic fingerprinting or profiling, is using analytical tools like mass spectrometry (MS) or nuclear magnetic resonance (NMR) for hypothesis free elucidation of the metabolism. The goal remains still the same, differentiating metabolic alteration in different states, but the way to reach this goal is different. Most scientists are trained to focus on a particular target and not to think in a holistic manner. Still non-targeted metabolomics is in its infancy, not readily accepted, it gains more and more attention by the scientific community. For such a non-targeted approach, dealing with both known and unknown chemical entities, ultrahigh resolution analytics are the method of choice, discriminating between often thousands of different chemical entities. In addition, not only quality of measurements, but also quantity plays a role in metabolomic studies, such a ultrahigh resolution technique should be also high-throughput capable. Ion cyclotron resonance-Fourier transform mass spectrometry (ICR-FT/MS), can handle these needs as shown before [21, 22].

ICR-FT/MS, having a resolution up to 1,000,000 and more and a mass accuracy <100 ppb, is used as profiling technique in metabolomics. It allows annotation of potential metabolites and calculation of possible elemental formulas using exact mass information. ICR-FT/MS is mostly utilized in direct infusion mode, without prior chromatographic or electrophoretic separation of the metabolites, to take advantage of its high resolving power (Box 5.1). The only major drawback of this technique is that it cannot distinguish between isobaric substances, like hexoses or isomeric substances, for example. In spite of this, it is a very sensitive technique, which uses diluted samples, so that sample consumption is minimal, making it ideal for studies where resources are limited. The first example in which ICR-FT/MS was described for non-targeted metabolomics was published by Asaph Aharoni in 2002. His group investigated the changes in ripening strawberries. The high resolution made it possible to see different metabolites between green and red strawberries even in a window of 0.1 Da [23]. Since this time a whole bunch of paper describes the use of ICR-FT/MS for metabolomics studies, which are reviewed elsewhere [24, 25]. Automation of data acquisition is easily accessible using autosamplers for flow injection or automated robotic devices, like the Advion TriVersa Nanomate, for sample infusion [26–28].

Box 5.1 Theory of ICR-FT/MS, Mass Resolution, Mass Accuracy and Calculation of Possible Formulas Out of Exact Masses

Ion Cyclotron Resonance Fourier Transform Mass Spectrometer (ICR-FT/MS) is a mass analyzer based on cyclotron motion of ions in homogenous high magnetic fields. After ionization of the analytes by ESI, APCI, APPI or MALDI as either positively or negatively charged ions; they are focused by ion lenses transferred into the magnetic field of a superconducting magnet, where an oscillating electric field excites the ions to higher trajectories. The masses are then resolved by their different cyclotron (rotational) frequency of the ion rotation in the magnetic field If a moving molecule with a mass m and an electric charge q $(q = n \cdot e)$ is transferred into a magnetic field B which is orthogonal to the ion's velocity v, the Lorentz force F_L acts on the ion.

$$F_L = q \times v \times B$$

In the homogenous magnetic field, the moving charge has a constant velocity and moves on a stable circular trajectory with the radius r. Their by the centrifugal force equilibrates the magnetic force.

$$\frac{m \cdot v^2}{r} = q \cdot v \cdot B \ or \ \frac{m}{q} = \frac{B \cdot r}{v}$$

The angular velocity ω_c, also called as cyclotron frequency is only a function of m/z-ratio and the magnetic field strength B. Because of the right-hand rule positive and negative ions have contrary courses.

$$\omega_c = \frac{q \cdot B}{m} \ f = \frac{\omega}{2 \cdot \pi}$$

An additional alternating electric field orthogonal to the magnetic field causes the cyclotron resonance of a certain m/z-ratio. This field is applied between a pair of plates, called excitation plates. If the electromagnetic wave has the same frequency as a certain ion in the cyclotron cell, the resonance (absorption of energy) as consequence increases the ion's kinetic energy hence increase the radius of its trajectory. The excited oscillating ions are detected with a second pair of plates, rotated 90° to the excitation plates. The passing ions induce an alternating current between the detection plates (Fig. 5.3a). This so-called image current is a superimposition of several frequencies caused by several ions of different masses. Fast Fourier transformation is used to convert it to a mass spectrum.

ICR-FT/MS can reach high mass resolutions R, sometimes also referred as resolving power.

(continued)

Box 5.1 (continued)

$$R = \frac{M}{\Delta M}$$

M is the mass of an ion and ΔM is typically the peak width at 50% of the peak height, also called full width at half maximum (FWHM). The influence of different resolutions is illustrated in Fig. 5.3b, showing that a lower resolution cannot separate between different molecular species with the same nominal, but different exact masses. But not only resolving power is important, mass accuracy plays a crucial role for identification of metabolites. The mass error is often reported in ppm, meaning error in part per million.

$$\Delta m / z = \frac{measured\ mass - true\ mass}{true\ mass} \cdot 10^6\ ppm$$

Working with high-resolution instruments it is needed to take also the mass of electron (5.485799×10^{-4} u) into account. At 200, 400 and 800 Da, the mass of an electron would yield errors of 2.74, 1.37 and 0.69 ppm, respectively.

The determined exact masses can be used to calculate possible elemental formulas. Chemical formulas can be understood as linear combination of elements with distinct monoisotopic masses, following several chemical rules [32]. Using these rules, it is possible to calculate formulas out of exact masses obtained from ICR-FT/MS. Due to the high resolving power only a few different formulas fit the exact mass, which narrows down the list of potential candidates. Additionally isotopic information can be used to confirm the predicted formulas.

Data obtained from direct infusion ICR-FT/MS compared to LC-MS is small in case of file size (several MB) but rich in information. Further reduction is achieved by using mass lists from recalibrated mass spectra, picked by an appropriate peak detection algorithm. Such mass lists can be aligned to sample matrices for further statistical interpretation of the obtained data. For conversion of mass spectra from ICR-FT/MS to biological interpretable data we developed the MassTRIX server (Mass Translator into Pathways, www.masstrix.org). It is public accessible and corrects an uploaded mass list corresponding to ionization mode and possible adducts and compares the corrected masses against possible metabolites from KEGG, HMDB and LipidMaps within a certain erro range. These are mapped in second step to the respective pathway maps of a chosen organism, by calling the KEGG API. Additionally genes of interest can be highlighted [29]. Using this database annotation approach, only up to 15% of the experimental signals can be annotated; the remaining 85% may be given

some elementary composition, but no possible chemical structure. Thus new data evaluation approaches are being developed to consider the whole experimental dataset and unravel the yet unknown metabolites and their possible biological function.

One of these new approaches we use is the discipline of graph theory, which is widely used in bioinformatics and chemometrics due to its ability to provide efficient means of analysing as well as visualising real-world scenarios. Graph theory allows many pragmatic situations to be modelled in the form of a diagram consisting of a set of points (nodes) and a set of lines (edges) connecting parts of these points; a mathematical abstraction which yields the concept of a graph [30]. As its name implies, a graph can be represented graphically, and through this graphical representation we are able to study some of its properties and gain knowledge on the data it represents. A graph (also called a network) is in addition associated with a specialised matrix, which allows us to store it in silico and apply mathematical methods in order to analyse our data more thoroughly; a procedure known as 'network analysis'.

Network analysis can be applied on almost every scenario of ICR-FT/MS spectra in a number of ways. A first approach would be the mass-mass difference networks, in which each node represents an exact experimental mass and each edge represents a selected mass difference either taken from a predefined list of potential transformations, or detected on the fly through mass-difference clustering and correlation analysis [31]. In a purely biochemical context such a network model can be divided into *structural* and *functional* networks. In the case of structural networks, a list of selected theoretical mass differences is used in order to determine the adjacency relation between the nodes, i.e. detect transformations between the experimental masses. The resulting network can be described as a simulation of the real biochemical system which may extract the structural information expressed in an ICR-FT/MS dataset. An extension of such a model combined with a special visualisation technique gives rise to the concept of *functional networks*. A specific list of selected CHONS mass differences is used in a similar way in order to detect functional groups via the Kendrick mass defect approach.

In the context of metabolomics we have the possibility of modelling ICR-FT/MS datasets in the form of *correlation networks*. Such a task can be achieved by treating mass spectra either as row or column vectors out of which a correlation matrix is extracted (usually using Pearson correlation). By setting a threshold value on the correlation coefficient, the correlation matrix can be converted into a binary adjacency matrix which represents a network. In the case of row vector correlation the resulting network is a *metabolic correlation network* which through several methods of node quantification, hierarchisation, and clustering has the potential of contributing in biomarker identification. Combined with the structural network approach, this method has a great potential of non-targeted data reduction and comparison of the clusters to known KEGG pathways adds biological information. In the case of column vector correlation the mass spectra of the various samples can be modelled into a *sample-correlation network*, which may be used for clustering samples into biologically significant group. Such a network analysis approach is flexible enough to be used in both supervised and unsupervised ways. A typical hierarchy of these networks, their shape and the obtained information is illustrated in Fig. 5.2.

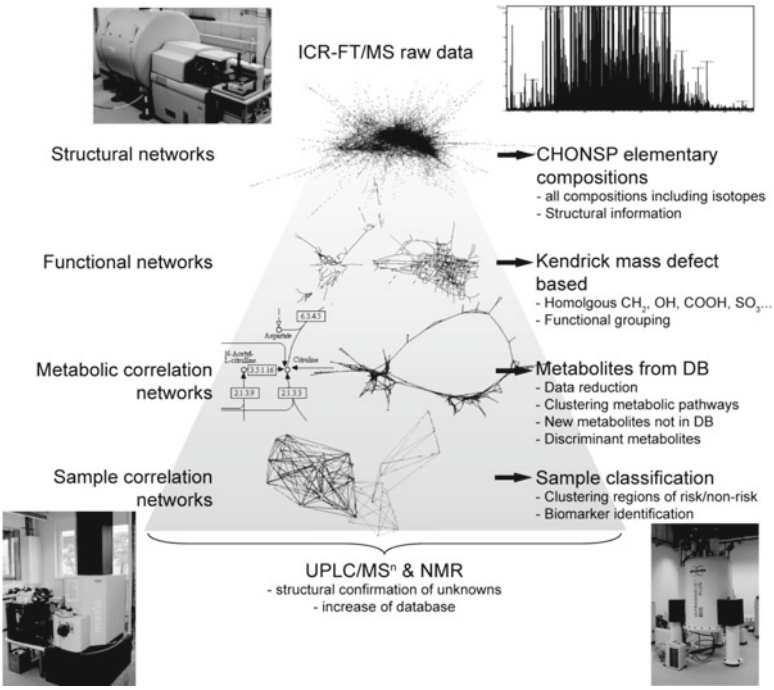


Fig. 5.2 Hierarchy of networks used for data analysis of ICR-FT/MS data. Structural networks calculate chemical formulas of experimental masses by using mass-mass difference information and applying chemical rules [32]. Functional networks are using the Kendrick mass defect approach in order to discover functional groups [33]. Metabolic correlation networks are reconstructed out of the correlation of mass spectra and mass-mass difference information which can be related to metabolic pathways of databases such as KEGG. They allow metabolite identification and clustering of potential biological significance. Sample correlation networks can be used to cluster the sampled patients into groups of biological importance, such as risk and non-risk. Combined with metabolic correlation networks they offer data reduction and can be used to detect metabolites that are discriminant in a specific experimental setup. UPLC/MS^n and NMR are used for confirmation of unknown metabolites

3.1 Metabologeography: Differentiating Genetically Close but Metabolomically Different Microorganisms

The comparison of the metabolome from the halophilic bacterium *Salinibacter ruber* using ICR-FT/MS was investigated by Rossello-Mora et al. to determine a geographical discrimination between different isolates. *S. ruber* can be found in different parts of the world. Totally 28 isolates, 10 Mediterranean, 13 Atlantic and 5 Peruvian, were cultivated under same conditions and both supernatant and cell pellet were analyzed on a Bruker APEX Qe ICR-FT/MS with 12 T superconducting

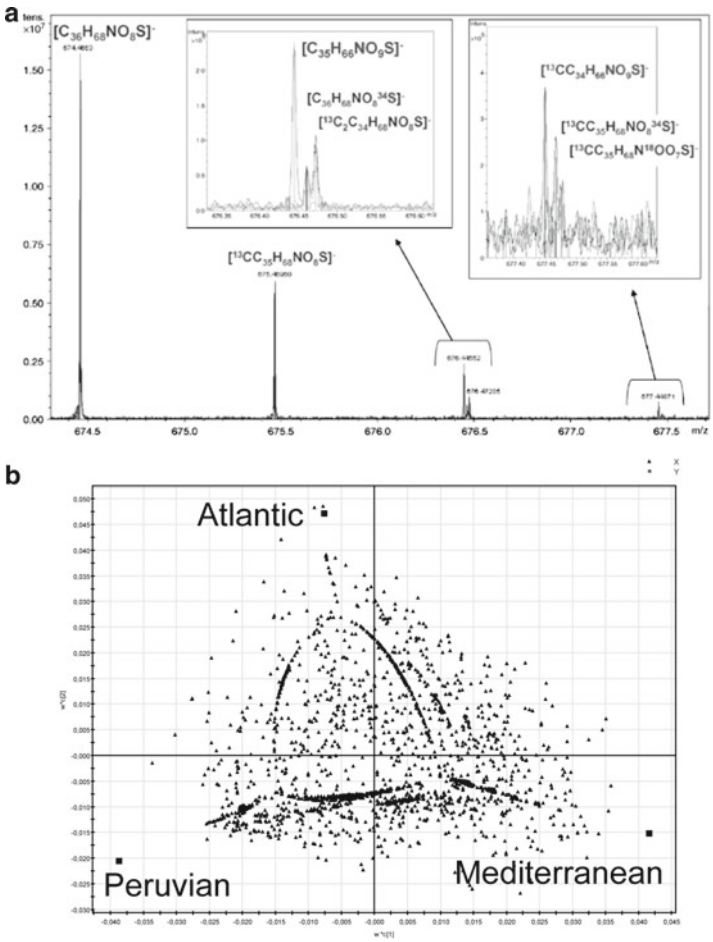


Fig. 5.3 (**a**) Detailed spectra on mass 674.4663 from a negative ionization ICR-FT/MS spectra and detailed isotopic information. The mass was identified as a sulfonolipid (**b**) Scatter loading plot of PLS model for differentiating the three origins of isolates

magnet and Apollo II ESI source. Spectra were acquired in positive and negative ionization mode. Possible elemental formulas were calculated on the resulting peak lists. Resolution of the measurement was high enough to distinguish between ^{34}S- and $^{13}C_2$-isotopes, as shown in Fig. 5.4a. Out of this data a matrix for further statistical analysis was created. Multivariate statistical analysis revealed a good separation of three different isolation regions (Fig. 5.4b). Furthermore, PLS-DA was able to separate the Mediterranean strains into their different origin locations.

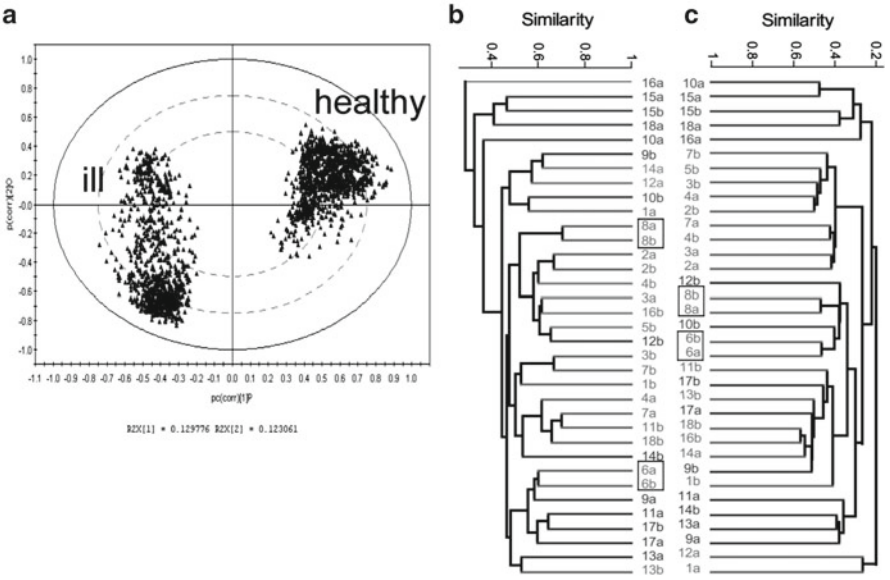


Fig. 5.4 (**a**) OPLS model to separate healthy individuals from Crohn's disease patients (**b**) similarity plot of microbial composition based on T-RFLP from fecal samples (**c**) Similarity plot of ICR-FT/MS data from fecal water extracts

Several sulfonolipid species were shown to be differentially present, according to the region of isolation [34].

A closer look at two Mediterranean strains, M8 and M31, showed that 10% of the genes encoded in M8 are absent in M31. Moreover metabolomic analysis, phage susceptibility and competition experiments revealed that these differences are not neutral [35]. The most recent work focused on the response of these two strains to environmental changes. ICR-FT/MS, together with multivariate statistics, separated different growth states and assigned significantly different metabolites. For the stationary phase, for example, metabolites belonging to the aminosugar, glycerolipid and glycerophospholipid metabolism showed decrease or increase [36].

3.2 Metabolic Biomarkers of Crohn's Disease

Crohn's disease (CD) is an inflammatory bowel disease, with unknown cause. One potential reason for the break-out can be a "dysbiosis" in the gut microbiome. Jansson et al. used ICR-FT/MS to profile the metabolome of 17 twin pairs, healthy and with CD. High resolution spectra from fecal water extracts were obtained

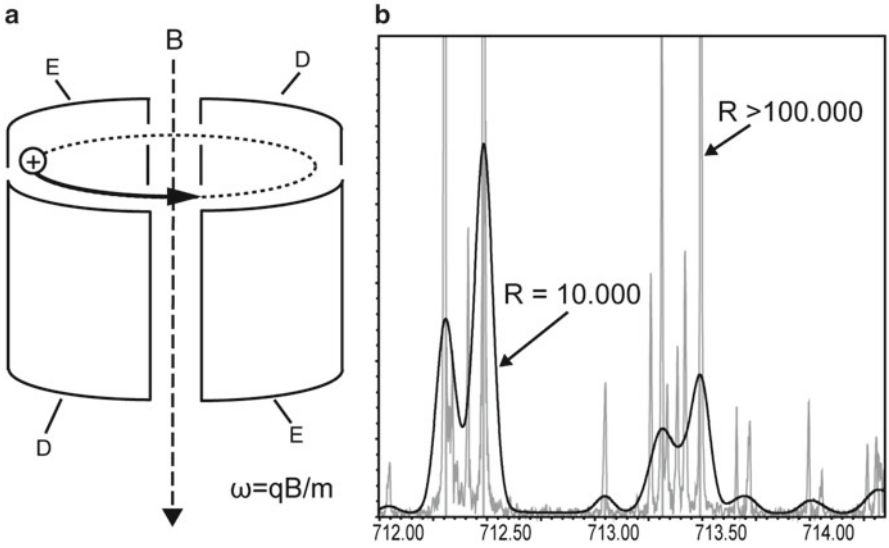


Fig. 5.5 (**a**) Principle of ICR-FT/MS as explained in Box 5.1. E = excitation plates, D = detection plate, B = magnetic field (**b**) Influence of resolution on measurement accuracy. In *grey*, a snapshot of a mass spectrum of a bacterial extract from *Pseudomonas aeruginosa* PA14 measured in positive ionization mode with a resolution of >100,000 is shown. In *black*, the same spectrum down scaled to a resolution of 10,000

using a Bruker APEX Qe ICR/FT-MS with 12 T superconducting magnet and Apollo II ESI source. Spectra were acquired in positive and negative ionization mode, internally recalibrated and exported to mass list files with a signal to noise threshold of 3 and aligned with in-house written software. Again multivariate statistics, including unsupervised and supervised techniques as principal component analysis (PCA), hierarchical cluster analysis (HCA) and partial least square discriminant analysis (PLS-DA) were used for data analysis. An OPLS model yielded good separation of healthy and ill individuals, illustrated in Fig. 5.5a. MassTRIX was used for annotation of possible metabolites to the masses. Several fecal metabolites where identified to contribute to the discrimination of the disease phenotype. This study shows how non-targeted metabolomics using ICR-FT/MS can be used for biomarker discovery out of non-invasive biosamples. In addition, good correlation could be found between the bacterial community profiles of fecal samples as analyzed based on polymerase chain reaction (PCR) amplification and terminal restriction fragment length polymorphism (T-RFLP) fingerprinting and the metabolite profiles (Fig. 5.5b, c) [37]. Currently ongoing work is focusing on the effect of the gut microbiome on type II diabetes, especially combining the deep metabotyping possibilities of ICR-FT/MS with the new deep sequencing technologies for DNA or RNA.

4 Outlook and Conclusion

ICR-FT/MS is able to supply potential candidates for novel metabolites and biomarkers. Its ultrahigh resolution and mass accuracy helps to narrow down the list of possible chemical formulas. Nevertheless ICR-FT/MS alone can provide a structure only in particular cases. Here "traditional" metabolomics techniques are needed, LC-MS allows separation from a complex mixture, the differentiation of possible isomers, the purification and structure elucidation by MS^n approaches and NMR. As example, a combination of UPLC, nano-LC-MS, ICR-FT/MS and bioassays is used in the research unit together with several cooperation partners to identify new quorum sensing bacterial signaling compounds (*N*-Acylated homoserine lactones, quinolones). Results from all analytical methods combined are combined and compared to attempt structural characterization without chemical synthesis of analytical standards and for example identified *N*-(3-hydroxydecanoyl homoserine lactone) as major AHL compound in the rhizosphere bacterium *Acidovorax sp.* N35 [38]. Finally, going more from isolated studies to systems biology should be the goal in microbial metabolomics, to produce a bigger and better understanding of the role of microorganisms in our environment and their interactions. As more and more high-throughput techniques in both targeted and non-targeted metabolomics are evolving, it is scaling up to the other members of the "omics"-family. This allows going from single gene deletions to whole genome libraries, to assign possible function to orphan genes. Bringing all "omics"-family members together will draw a picture that is even bigger and more defined than everyone separately can draw. More than this, living organisms and systems are more than the sum of genes, transcripts, proteins and metabolites and biology doesn't separate between them. The KEGG database contains today about 1,200 complete bacterial genomes compared to the biodiversity; this is a relatively small number. No one can expect today what else is out there to be discovered. At the moment we are about to scratch the tip of the metabolomics iceberg.

References

1. Schopf JW (2006) Fossil evidence of Archaean life. Philos Trans R Soc B Biol Sci 361(1470):869–885
2. Cavalier-Smith T (2006) Cell evolution and earth history: stasis and revolution. Philos Trans R Soc B Biol Sci 361(1470):969–1006
3. Leewenhoeck A (1684) An abstract of a letter from Mr. Anthony Leewenhoeck at Delft, sated Sep. 17. 1683. Containing some microscopical observations, about animals in the scurf of the teeth, the substance call'd worms in the nose, the cuticula consisting of scales. Philos Trans 14(155–166):568–574
4. van Leeuwenhoek A (1700) Part of a letter from Mr Antony van Leeuwenhoek, concerning the worms in sheeps livers, gnats, and animalcula in the excrements of frogs. Philos Trans 22(260–276):509–518
5. van Leeuwenhoek A (1702) Part of a letter from Mr Antony van Leeuwenhoek, F. R. S. concerning green weeds growing in water, and some animalcula found about them. Philos Trans 23(277–288):1304–1311

6. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci 95(12):6578–6583
7. Fredrickson JK, Zachara JM, Balkwill DL et al (2004) Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford site, Washington state. Appl Environ Microbiol 70(7):4230–4241
8. Steinhoff U (2005) Who controls the crowd? New findings and old questions about the intestinal microflora. Immunol Lett 99(1):12–16
9. O'Hara AM, Shanahan F (2006) The gut flora as a forgotten organ. EMBO Rep 7(7): 688–693
10. Rappá MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57(1):369–394
11. Buchanan B, Arnon D (1990) A reverse KREBS cycle in photosynthesis: consensus at last. Photosynth Res 24(1):47–53
12. Villas-Bôas SG, Roessner-Tunali U, Hansen MAE, Smedsgaard J, Nielsen J (2007) Metabolome analysis: an introduction, 1st edn. Wiley, Indianapolis
13. Frimmersdorf E, Horatzek S, Pelnikevich A, Wiehlmann L, Schomburg D (2010) How pseudomonas aeruginosa adapts to various environments: a metabolomic approach. Environ Microbiol 12(6):1734–1747
14. Hertkorn N, Frommberger M, Witt M et al (2008) Natural organic matter and the event horizon of mass spectrometry. Anal Chem 80(23):8908–8919
15. Wikoff WR, Anfora AT, Liu J et al (2009) Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. Proc Natl Acad Sci 106(10):3698–3703
16. Soga T, Ohashi Y, Ueno Y et al (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. J Proteome Res 2(5):488–494
17. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T (2007) Microbial metabolomics: toward a platform with full metabolome coverage. Anal Biochem 370(1): 17–25
18. Garcia DE, Baidoo EE, Benke PI et al (2008) Separation and mass spectrometry in microbial metabolomics. Curr Opin Microbiol 11(3):233–239
19. van der Werf MJ, Jellema RH, Hankemeier T (2005) Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. J Ind Microbiol Biotechnol 32(6):234–252
20. Mashego M, Rumbold K, De Mey M et al (2007) Microbial metabolomics: past, present and future methodologies. Biotechnol Lett 29(1):1–16
21. Gougeon RD, Lucio M, Frommberger M et al (2009) The chemodiversity of wines can reveal a metabologeography expression of cooperage oak wood. Proc Natl Acad Sci 106(23): 9174–9179
22. Liger-Belair G, Cilindre C, Gougeon RD et al (2009) Unraveling different chemical fingerprints between a champagne wine and its aerosols. Proc Natl Acad Sci 106(39):16545–16549
23. Aharoni A, Ric de Vos CH, Verhoeven HA et al (2002) Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. OMICS J Integr Biol 6(3):217–234
24. Brown SC, Kruppa G, Dasseux JL (2005) Metabolomics applications of FT-ICR mass spectrometry. Mass Spectrom Rev 24(2):223–231
25. Ohta D, Kanaya S, Suzuki H (2010) Application of Fourier-transform ion cyclotron resonance mass spectrometry to metabolic profiling and metabolite identification. Curr Opin Biotechnol 21(1):35–44
26. Huang N, Siegel MM, Kruppa GH, Laukien FH (1999) Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data. J Am Soc Mass Spectrom 10(11):1166–1173
27. Han J, Danell R, Patel J et al (2008) Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. Metabolomics 4(2): 128–140

28. Li X, Fekete A, Englmann M et al (2007) At-line coupling of UPLC to chip-electrospray-FTICR-MS. Anal Bioanal Chem 389(5):1439–1446
29. Suhre K, Schmitt-Kopplin P (2008) MassTRIX: mass translator into pathways. Nucleic Acids Res 36(suppl 2):W481–W484
30. Bondy JA MUSR (2007) Graduate text in mathematics: graph theory, 1st edn. Springer, New York
31. Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP (2006) Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. Metabolomics 2(3):155–164
32. Kind T, Fiehn O (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8(1):105
33. Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K (2001) Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. Anal Chem 73(19):4676–4681
34. Rossello-Mora R, Lucio M, Pena A et al (2008) Metabolic evidence for biogeographic isolation of the extremophilic bacterium Salinibacter ruber. ISME J 2(3):242–253
35. Pena A, Teeling H, Huerta-Cepas J et al (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. ISME J 4(7):882–895
36. Brito-Echeverría J, Lucio M, López-López A et al (2011) Response to adverse conditions in two strains of the extremely halophilic species Salinibacter ruber. Extremophiles 15(3): 379–389
37. Jansson J, Willing B, Lucio M et al (2009) Metabolomics reveals metabolic biomarkers of Crohn's disease. PLoS One 4(7):e6386
38. Fekete A, Frommberger M, Rothballer M et al (2007) Identification of bacterial N-acylhomoserine lactones (AHLs) with a combination of ultra-performance liquid chromatography (UPLC), ultra-high-resolution mass spectrometry, and in-situ biosensors. Anal Bioanal Chem 387(2):455–467

Chapter 6
Metabolomic Systems Biology of Protozoan Parasites

Rainer Breitling, Barbara M. Bakker, Michael P. Barrett, Saskia Decuypere, and Jean-Claude Dujardin

1 The Special Role of Metabolomics in Systems Biology

Metabolomics has a special position in the hierarchy of systems biological research: although the technological developments that enable the comprehensive profiling of metabolite levels are very recent, a large number of the classical success stories in systems biology have involved metabolic networks. The detailed, quantitative data collected in enzymological studies were the foundation for the development of metabolic control analysis (MCA), a mathematical analysis of metabolic systems that

R. Breitling (✉)
Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Joseph Black Building B3.10, Glasgow G12 8QQ, United Kingdom

Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, Groningen, 9747 AG, The Netherlands
e-mail: rainer.breitling@glasgow.ac.uk

B.M. Bakker
Department of Pediatrics, Center for Liver, Digestive and Metabolic Diseases, University Medical Center Groningen, University of Groningen, Hanzeplein 1, Groningen, 9713 GZ, The Netherlands
e-mail: b.m.bakker@med.umcg.nl

M.P. Barrett
Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Sir Graeme Davies Building, Glasgow, United Kingdom
e-mail: Michael.Barrett@glasgow.ac.uk

S. Decuypere • J.-C. Dujardin
Department of Parasitology, Unit of Molecular Parasitology, Institute of Tropical Medicine, Antwerp, B-2000, Belgium
e-mail: sdecuypere@itg.be; JCDujardin@itg.be

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_6,

clearly showed that many popular intuitive concepts, such as the idea of rate-limiting reactions, need to be replaced by refined notions, such as distributed control [1, 2]. This had important implications for drug development, as only enzymes with a substantial control coefficient for important metabolic pathways will be promising drug targets – and these enzymes were not always the same that were predicted by the study of enzymes in isolation or qualitative analysis of pathway behavior [3]. By showing that a quantitative and integrative approach can yield biologically relevant and testable insights, especially for large complex systems, MCA had a major stimulating effect on the emergence of systems biology.

In the post-genomic era, another metabolome-centered approach has been particularly popular and powerful: constraint-based modeling, as exemplified by flux balance analysis (FBA), can be used to predict the metabolic potential of entire organisms using homology-based genome annotations [4, 5]. Such predictions can be used to identify essential enzymatic reactions, including synthetic lethal mutations, and predict the capacity to grow in various environmental conditions, as well as the metabolic rearrangements expected as a result of gene knockouts or enzyme inhibition. Despite our incomplete knowledge of metabolic enzymes and the neglect of any quantitative kinetic information in FBA, the general predictions arrived at by this method can be quite accurate. Recent improvements allow for the incorporation of additional thermodynamic and regulatory constraints [6–12], and some methods even aim at generating dynamic models from constraint-based descriptions using a variety of parameter prediction methods [13–15].

Metabolomics is not only a very successful application area of systems biology, but it has a privileged position also in a more fundamental sense: it is positioned at the extreme end of the "dogma" of molecular biology, with information flowing from the genome, via the transcriptome and the proteome, to the metabolome [16]. This places the metabolome closest to the actual phenotype. Recent experimental work indicates that this closeness of the metabolome to the phenotype is not just conceptual, but has a biological basis as well. Genetic analysis of the phenotypic, metabolomics, proteomic and transcriptomic variation in a large population of recombinant inbred individuals showed that genetic polymorphisms that cause phenotypic variation are most clearly reflected at the metabolite level, while transcriptome variation, for instance, seems to be largely buffered by the cellular network and does not regularly result in phenotypic diversity [17]. The importance of checking metabolic endpoints as indicators of phenotypic status has long been recognized for biomarker development: many of the most common disease biomarkers are indeed metabolites. In the field of parasitic diseases, only few studies have attempted so far to develop biomarkers of protozoan infections. However, the results of the first studies monitoring metabolic changes in blood and urine in animals infected with African trypanosome yielded promising results [18]. Several unique metabolite markers of trypanosome infection could be identified in the course of an infection. Identification of blood or urine biomarkers to characterize the type or stage of a parasitic disease could mark a major advance beyond the current diagnostic standards of microscopic inspection of cerebral spinal fluid (trypanosomiasis) or bone marrow and spleen aspirates (leishmaniasis).

2 Trypanosome Metabolism in Systems Biology

Protozoan parasites, and in particular *Trypanosoma brucei*, have been important model organisms for metabolomic systems biology from an early stage [19]. These single-celled organisms live as extracellular parasites in the human blood stream, later invading other organs, including the brain, where they cause African sleeping sickness, a debilitating and widespread illness that is difficult to treat effectively. Trypanosomes are transmitted by insect vectors (tsetse flies of the genus *Glossina*) and undergo a complex life cycle with various morphological forms both in the transmitting insect and in the infected human. Due to the molecular similarities between protozoan parasites and their human hosts, existing drugs tend to have severe side effects, similar to cancer chemotherapeutics. The resulting poor compliance aggravates the already serious problem of emerging drug resistance.

Quantitative mathematical models of metabolism were developed to assist with the identification of new treatment strategies [20]. They made use of the greatly reduced metabolic complexity of the parasites, which rely on host supply for a large fraction of their metabolites and exploit the constant supply of blood glucose in the human body for ATP generation almost exclusively based on glycolysis. Comprehensive kinetic data measured under controlled conditions were available for most of the enzymes of the glycolytic pathway, and these were crucial for the generation of the first computational model of trypanosome glycolysis. This was used for a detailed metabolic control analysis, revealing how inhibition of each of the enzymes in the system would affect cellular viability [21].

The first computational model has gone through several rounds of revision, updating parameters based on new enzymatic assay conditions, and is now one of the key resources for metabolomics systems biology of protozoan parasites [22]. In addition, sequencing of the complete *T. brucei* genome has lead to the initiation of a well-curated metabolic pathway database (TrypanoCyc, [23]).

3 Metabolomic Profiling of Trypanosomes

The initial metabolic model building effort in trypanosomes was based on highly specific quantitative measurements of enzyme kinetics and metabolite levels. Recently, this approach has been complemented by post-genomic untargeted studies of the cellular metabolome. Breitling et al. have shown that it is possible to reconstruct hypothetical metabolic maps for trypanosomes *ab initio*, using high-accuracy mass spectrometry analysis of cell extracts [24]. In their proof-of-principle study, they analyzed metabolite extracts from bloodstream form trypanosomes collected from rat blood using Fourier transform mass spectrometry. Exploiting the high mass accuracy of the instrument (better than 1 ppm), they were not only able to assign putative chemical formulas to many observed metabolites, but could also infer putative chemical relationship between those compounds. Each of the commonly occurring biochemical transformations (except isomerizations) corresponds

to a characteristic mass difference. If the masses, and consequently the mass differences, are observed at sufficient accuracy, this information can be used to infer the possible biochemical connectivity between the observed metabolites, assuming that many of them are connected in a global metabolic network [25].

The major advantage of *ab initio* metabolomics studies is their global scope, limited only by the technological biases of the analytical machinery (although these can sometimes be substantial). This allows an extension of our metabolic understanding beyond the well-characterized pathways of central metabolism [26]. While protozoan parasites are not characterized by an extraordinary diversity of secondary metabolites, they do show some unusual features, such as the use of trypanothione (a bis-glutathione-polyamine conjugate) instead of glutathione as the main redox active metabolite. The metabolism of polyamines is the only proven point of action for a currently used trypanocidal drug, the ornithine analogue difluoromethylornithine (DFMO or eflornithine), acting as an irreversible inhibitor of the enzyme ornithine decarboxylase and thus blocking polyamine biosynthesis [27]. Other aspects of trypanothione metabolism are also considered as promising targets for therapeutic intervention [28]. Other uncommon metabolic pathways may still await discovery and would be valuable targets for parasite-specific drugs.

The *ab initio* network reconstruction based on high-accuracy metabolomics profiles can be performed using the MetaNetter plugin for the Cytoscape network visualization tool [29]. On the basis of a list of observed exact masses and list of expected metabolic transformations, a putative metabolic network is reconstructed. The plugin also enables a number of basic topological analyses and the combination of mass difference analysis and correlation-based network reconstruction, which can provide additional orthogonal evidence for or against certain connections in the network.

Even in those cases where the inferred connection between two metabolites does not correspond to an enzymatically catalyzed reaction, the network context provides important information about the chemical identity of observed metabolites [30]. For instance, if a metabolite cannot be assigned a single molecular formula because of limited mass accuracy, the presence of confidently assigned connected molecules can be used for disambiguation. This concept has been implemented in a Bayesian statistical framework and can be employed globally, to find the best overall assignment of molecular identities that is maximally consistent with the predicted metabolic relationships [30].

4 Metabolomics of *Leishmania* and the Genetic Challenge

Metabolomic systems biology for some of the other protozoan parasite species beyond trypanosomes has been more difficult, mostly due to a lack of baseline knowledge of the metabolic capacities of these organisms. An example is provided by parasites of the genus *Leishmania*, which have successfully colonized a large variety of vertebrate and invertebrate hosts and cause a wide range of

manifestations, ranging from asymptomatic infections to mild skin lesions and lethal visceral forms of leishmaniasis [31]. Detailed kinetic models of metabolism are currently out of reach for these understudied pathogens, given that the necessary systematic measurement of enzyme parameters has not been performed.

Nonetheless, metabolomics can play an important role for understanding phenotypic diversity [31]. Indeed, microsatellite analyses revealed that the genetic variability in *L. donovani* strains on the Indian subcontinent is extremely limited at the sequence level, most probably due to a recent population bottleneck of the parasites. An ongoing full-genome sequencing project identified only about 3,500 SNPs segregating in the 17 strains sequenced so far, in a genome of about 32.4 Mb (GeMInI consortium, manuscript in preparation); this is less than 10 times the diversity of the human host population, which is itself genetically quite homogeneous. Of these SNPs, only about 400 are predicted to cause non-synonymous changes at the protein level. Nonetheless, the infection caused by these genetically homogeneous parasites is clinically heterogeneous, as is the responsiveness to drugs. The genetic diversity of the host underlying differential susceptibility for infection obviously plays a role in this clinical heterogeneity.

However, the parasite genome itself, which seems so homogeneous on the sequence level, also holds some surprises when looking at the structural aspects of the genome [32]. Firstly, the karyotype of clinical strains was shown to be extremely plastic, which results in an unusually high degree of aneuploidy, with no two strains having identical karyotypes. Experimental induction of drug resistance clearly showed how *Leishmania* parasites use this mechanism to up-regulate a series of genes present in individual chromosomes [33]. Secondly, a number of genes are present in large tandem repeats and are prone to expansion/contraction. The *L. donovani* sequencing project showed how genes encoding for major effectors of splicing and translation (rDNA transcription units and mini-exon genes) were significantly rearranged among drug-resistant strains [34]. Last but not least, in some chromosomal regions, sets of single-copy genes are flanked by direct repeats making these loci prone to the generation of extra-chromosomal circular DNA (by homologous recombination). Such extrachromosomal elements are frequently encountered under experimental drug pressure [33]. Standard genetic approaches to characterize parasite physiology are hampered by this volatility, but it can be hypothesized that the ultimate effect of the large variety of structural genetic variants will be a small number of discrete "metabotypes". Metabolomic profiles could therefore become important biomarkers for discriminating clinical subtypes and monitoring drug resistance [31].

In a pilot study, t'Kindt et al. have recently shown that untargeted metabolic profiles obtained by high-accuracy mass spectrometry can be useful to distinguish drug-sensitive and -resistant strains [35]. The samples used in that study were derived from parasite isolates from visceral leishmaniasis patients in Nepal that responded differently to standard drug treatment. The small sample number (2 drug-sensitive and 3 drug-responsive strains) precluded any firm conclusions about the metabolomic mechanism of drug resistance, but highlighted the potential of metabolomics to differentiate drug-sensitive and -resistant phenotypes. Approximately

one third of the 340 detected metabolites showed distinct abundance patterns in the two groups of samples; validation studies with a larger number of well-characterized clinical isolates are currently underway. These metabolomic studies of parasite diversity are complemented with whole-genome analysis of the same set of parasite strains. The latter will allow determining how the genomic structural variation described earlier is translated into changes at the metabolic and phenotypic level.

Another important aspect of metabolomics applied to *Leishmania* infection is the more intimate relationship between host and parasite [36]: similar to the malaria parasite, *Leishmania* is an intracellular pathogen. It hides from the host immune response by proliferating in the macrophages of the host. To do so efficiently, the parasite must be able to exploit, and probably also manipulate, the metabolic productivity of the host cell to its own advantage. Metabolomics promises to be an essential tool to study this interaction between intracellular parasites and their host cells at the metabolic level. A first generation of parasitological studies comparing infected vs. non-infected host cells have already been performed for the malaria parasite [36] and have demonstrated the usefulness of metabolomics to highlight the host metabolomic pathways subject to modulation by the malaria parasite. Ideally, host and pathogen metabolomes from infected cells should be fractionated to study the metabolite cross-talk between both systems, but additional technical advances are required to achieve a reliable separation and sampling.

5 Establishing Metabolomics Platforms for Protozoan Parasites

The successful application of high-accuracy mass spectrometry-based metabolomics has required a substantial amount of novel method development. The first step, and probably the most challenging, is the development of a suitable sampling protocol. For metabolomic studies of protozoan parasites, the production of reproducible in vitro parasite cultures and reliable, quantitative metabolite extraction is critical and needs to be optimized specifically for each type of parasite. The close causal relationship between the metabolome and the phenotype makes the metabolome susceptible to any changes in in vitro growth conditions (e.g., medium, temperature, and growth rate). When aiming to compare different parasite strains on the metabolomic level, it is absolutely essential to harvest the parasite cultures at the same stage of growth or development; ideally, time-course based comparisons should be done. As the parasites are typically grown in very complex media, containing a large variety of metabolites at sometimes very high concentrations, washing of the cell pellets is critical. For *Leishmania* samples t'Kindt et al. could show that removal of phospholipids from the medium is most difficult, and at least three successive washing steps are recommended [37]. They also explored the efficiency of a large number of different cell disruption protocols to achieve reliable and comprehensive metabolite extraction. The tested methods included heating on a heating block, mixing in a Thermomixer, Ultra Turrax or Dispomix, mechanical shearing in

a vortexer, and milling in a Retsch mill. The latter approach showed the lowest total metabolite yield, and all cold approaches were superior to the hot extraction in their ability to reproducibly detect a number of target compounds (such as NAD and reduced trypanothione). The authors recommended the cold thermomixer approach for its reproducibility, good yield, and ease and rapidity of handling. They also showed that chloroform:methanol:water extraction (20:60:20 v/v/v) was superior to all other extraction solvents tested (aqueous methanol, aqueous ethanol, aqueous isopropanol, aqueous acetonitrile, and methanol:chloroform). In particular, none of the chloroform-free solvents resulted in visible cell disruption. With the optimized protocol and a single analytical condition, 118 metabolites from the LeishCyc database were putatively identified in the sample, corresponding to a coverage of about 20% of the predicted metabolome [37].

Another area of essential method development focused on the computational processing of the resulting mass spectra. Metabolomics poses unique challenges in this respect. For example, due to the high sensitivity of the instrument, each compound (real metabolite) in a sample generates on the order of 10 (and often many more) signals in the spectrum at various masses, in addition to the one peak of interest at the correct mass [38]. These additional peaks are due to natural isotopes, fragmentation in the electrospray ion source, and a wide range of poorly understood chemical modifications, such as the formation of multiple adducts. Consequently, only a very small fraction of the signals are of interest for the biologist. This is not a problem for targeted analyses, which immediately zoom in on expected metabolites; but for untargeted studies, which are the unique strength of metabolomics, the resulting data complexity can be a serious nuisance. We have developed a comprehensive set of computational tools, mzMatch, to clean up the data using correlation patterns across samples and within the chromatogram, which is combined with a large collection of other useful processing and visualization tools (available for download at http://mzmatch.sourcefourge.net/). This software is also integrated with the statistical software R, which can be employed for downstream processing and data interpretation.

A particularly critical component of the computational toolbox is a method for mass calibration of the spectra using ubiquitous contaminant ions [39]. Initial metabolomics mass spectra were burdened by very intense signals from extraneous molecules, which appeared throughout the chromatogram. Standard peak picking methods tended to report these background ions as the most relevant masses in the spectra, leading to almost uninterpretable results. However, as the contaminants are not overlapping real metabolite signals, due to the high mass resolution of the instruments, they can be identified based on their characteristic chromatograms. On the other hand, as they are ubiquitous, i.e. shared between all or most spectra in a dataset, they can be used to align spectra after the unavoidable mass drift during long-term studies. Finally, a considerable fraction of the contaminants has been described in earlier mass spectrometric studies of chemical compounds. Once they can be matched to their molecular formula, their exact mass is known and can be used for internal mass calibration. In one case study on trypanosome metabolite profiles acquired on an Orbitrap mass spectrometer, this internal

calibration using background ions resulted in an improvement of mass accuracy by a factor of almost 10, to a median accuracy of 0.21 ppm [39]. Given the critical importance of mass accuracy for metabolite identification and *ab initio* network reconstruction (Box 6.1), this processing step greatly improved the usability of the data for downstream systems biology applications.

Box 6.1 The Advantages of High Mass Accuracy in Liquid-Chromatography Mass Spectrometry for Metabolomics

Comprehensive profiling of cellular metabolomes requires a combination of analytical technologies [40]. Nuclear magnetic resonance is the preferred method for quantitation of abundant metabolites, while mass spectrometry coupled to various separation techniques (gas chromatography, liquid chromatography or capillary electrophoresis) is most sensitive for broad coverage of the metabolome [31]. The studies described here use liquid chromatography coupled to mass spectrometry instruments with a particular high mass accuracy [24, 35, 37, 39].

Two general types of mass analyzer were used. The first one uses ion cyclotron resonance (ICR) to trap the analyte ions in a strong magnetic field and analyses their mass-dependent resonance frequency spectra by Fourier transformation to separate the signals of different metabolites in the sample (ICR-FTMS; [41]). The second one traps the analytes in the electrical field between two specially shaped electrodes, around which the ions orbit (Orbitrap; [42]). The mass-dependent signal is contained in the frequency of oscillation along the main axis of the electrode and is again extracted by Fourier transformation. This type of instrument has the advantage that no strong magnetic fields and associated supercooled magnets are needed; this has brought high-accuracy mass spectrometry within the reach of many metabolomics laboratories.

With these instruments, the mass accuracy can be better than 1 ppm [39], which means that the mass uncertainty corresponds approximately to the mass of an electron relative to the mass of three glucose molecules. This extreme accuracy has several major advantages for the interpretation of the complex mass spectra:

1. As metabolites are made up of a small number of elements, each with a characteristic non-integer mass, their accurate mass can serve as a unique molecular identifier. Only a limited number of potential chemical formulas can explain a particular accurate mass: once the accuracy is well below 1 ppm, there is typically only a single potential match in the biochemical databases [43]. Additional information, e.g. from chromatographic retention times and tandem mass spectrometry fragmentation patterns, is required to distinguish isomers with the same formula but different structure [44].

(continued)

Box 6.1 (continued)

2. Accurate masses imply accurate mass differences. Mass differences can be used to infer the chemical relationships between the observed molecules. This can be employed to predict *ab initio* metabolic networks [25], but also to identify chemical derivatives of the true metabolites, which cause artificial signals in the spectra [38].
3. Even in targeted experiments, where standard compounds with known chromatographic properties and tandem mass spectrometry are used to identify molecules, accurate mass can help to distinguish molecule classes of very similar behavior. For example, in a lipidomic analysis of *Leishmania donovani* lipids were separated into well-resolved classes on a silica gel column run in hydrophilic interaction chromatography mode, and their fatty acid composition was characterized by tandem mass spectrometry, but high accuracy mass spectrometry was necessary to discriminate acyl- and acyl-alkyl-lipids [45].

6 Future Perspectives

To fulfill its potential as a major systems biology tool in parasite studies, metabolome analysis has to be combined in novel ways with a wide range of other technologies and concepts. The SilicoTryp project (for trypanosomes) and the GeMInI initiative (for *Leishmania*) are currently exploring some of these integrative approaches [22, 31].

The combination of metabolome profiling, which provides a static snapshot of cellular physiology, with more dynamic approaches will be essential for the way forward. Dynamic information can be obtained by fluxomic studies, using stable isotope labels to trace the fate of metabolic precursors in the metabolic network [46–48]. Metabolic fluxes in this type of experiment can either be quantified by steady-state labeling patterns in proteinogenic amino acids or followed at high temporal resolution after a pulse of labeling. Various experimental perturbations can be considered to generate informative dynamic behavior: changing carbon sources to mimic the available substrates in human host and insect vector; applying drugs at sublethal doses; inducing parasite differentiation to follow the development process.

To interpret the dynamic data, and metabolome data in general, integration with dynamic computational models will be crucial. The dynamic model of trypanosome central metabolism is an excellent starting point for this kind of effort, but it needs to be expanded towards the larger metabolic network, using targeted enzymatic studies, genome-scale modeling and the incorporation of new hypothetical pathways predicted by *ab initio* network inference [22, 26]. Another dynamic extension will involve the kinetic description of transcription, translation, and protein and

mRNA turnover [22]. This has already been attempted for a single enzyme in the glycolysis model [49], and deep-sequencing approaches are now able to generate the necessary data on a genome-wide scale. The diversity of confidence in the kinetic parameters for each of these model extensions will need to be addressed using newly developed statistical tools that allow dynamic modeling under uncertainty [50, 51].

The metabolome of a parasite will not only adapt dynamically to different life stages or drug regimes, but it will also differ between genetically divergent pathogens [35]. As discussed above, the genomic architecture of parasites can be highly volatile even within a relatively homogeneous population. Large-scale whole-genome sequencing of natural (clinical) isolates will provide a unique background against which to place the interpretation of metabolome profiles obtained for the same isolates.

Acknowledgements We thank the SilicoTryp and GeMInI consortia for their contribution to the research that forms the basis of this review. The SysMO-funded SilicoTryp project aims at the reconstruction of a comprehensive computational model of trypanosome biology. It includes partners from the University of Glasgow, UK, University College London, UK, University of Groningen, NL, University of Heidelberg, Germany, and University of Edinburgh, UK. The GeMInI consortium initiated by the Institute of Tropical Medicine, Antwerp, Belgium, combines large-scale genomic sequencing and metabolomic analysis for a better understanding of the natural diversity of leishmaniasis. It involves partners from the B.P. Koirala Institute of Health Sciences, Nepal, Strathclyde University, UK, Sanger Institute, UK, University of Glasgow, UK, University of Groningen, NL, and the Institute of Tropical Medicine, Antwerp, Belgium.

References

1. Fell DA (1992) Metabolic control analysis: a survey of its theoretical and experimental development. Biochem J 286(Pt 2):313–330
2. Fell DA (1998) Increasing the flux in metabolic pathways: a metabolic control analysis perspective. Biotechnol Bioeng 58(2–3):121–124
3. Bakker BM, Westerhoff HV, Opperdoes FR, Michels PA (2000) Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs. Mol Biochem Parasitol 106(1):1–10
4. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7(2):130–141
5. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–897
6. Kummel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics 7:512
7. Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol 2(2006):0034
8. Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics 24(18):2044–2050
9. Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. Metab Eng 6(4):285–293
10. Shlomi T, Eisenberg Y, Sharan R, Ruppin E (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol 3:101

11. Covert MW, Palsson BO (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. J Biol Chem 277(31):28058–28064
12. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys J 95(3):1487–1499
13. Ko CL, Voit EO, Wang FS (2009) Estimating parameters for generalized mass action models with connectivity information. BMC Bioinformatics 10:140
14. Smallbone K, Simeonidis E, Swainston N, Mendes P (2010) Towards a genome-scale kinetic model of cellular metabolism. BMC Syst Biol 4:6
15. Kotte O, Heinemann M (2009) A divide-and-conquer approach to analyze underdetermined biochemical models. Bioinformatics 25(4):519–525
16. Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: current technologies and future trends. Proteomics 6(17):4716–4723
17. Fu J, Keurentjes JJ, Bouwmeester H et al (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genet 41(2):166–167
18. Wang Y, Utzinger J, Saric J et al (2008) Global metabolic responses of mice to *Trypanosoma brucei* brucei infection. Proc Natl Acad Sci USA 105(16):6127–6132
19. Barrett MP, Bakker BM, Breitling R (2010) Metabolomic systems biology of trypanosomes. Parasitology 137(9):1285–1290
20. Bakker BM, Michels PA, Opperdoes FR, Westerhoff HV (1997) Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. J Biol Chem 272(6):3207–3215
21. Bakker BM, Michels PA, Opperdoes FR, Westerhoff HV (1999) What controls glycolysis in bloodstream form *Trypanosoma brucei*? J Biol Chem 274(21):14551–14559
22. Bakker BM, Krauth-Siegel RL, Clayton C et al (2010) The silicon trypanosome. Parasitology 137(9):1333–1341
23. Chukualim B, Peters N, Hertz-Fowler C, Berriman M (2008) TrypanoCyc – a metabolic pathway database for *Trypanosoma brucei*. BMC Bioinformatics 9(suppl 10):P5
24. Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP (2006) Ab inition prediction of metabolic networks using Fourier transform mass spectrometry data. Metabolomics 2(3):155–164
25. Breitling R, Pitt AR, Barrett MP (2006) Precision mapping of the metabolome. Trends Biotechnol 24(12):543–548
26. Breitling R, Vitkup D, Barrett MP (2008) New surveyor tools for charting microbial metabolic maps. Nat Rev Microbiol 6(2):156–161
27. Bacchi CJ, Nathan HC, Hutner SH, McCann PP, Sjoerdsma A (1980) Polyamine metabolism: a potential therapeutic target in trypanosomes. Science 210(4467):332–334
28. Heby O, Persson L, Rentala M (2007) Targeting the polyamine biosynthetic enzymes: a promising approach to therapy of African sleeping sickness, Chagas' disease, and leishmaniasis. Amino Acids 33(2):359–366
29. Jourdan F, Breitling R, Barrett MP, Gilbert D (2008) MetaNetter: inference and visualization of high-resolution metabolomic networks. Bioinformatics 24(1):143–145
30. Rogers S, Scheltema RA, Girolami M, Breitling R (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics 25(4):512–518
31. Scheltema RA, Decuypere S, T'Kindt R et al (2010) The potential of metabolomics for Leishmania research in the post-genomics era. Parasitology 137(9):1291–1302
32. Dujardin JC (2009) Structure, dynamics and function of *Leishmania* genome: resolving the puzzle of infection, genetics and evolution? Infect Genet Evol 9(2):290–297
33. Leprohon P, Legare D, Raymond F et al (2009) Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. Nucleic Acids Res 37(5):1387–1399
34. Imamura H, Decuypere S, Tim Downing T et al (2010) Whole comparative genome sequencing reveals several levels of genomic diversity among *Leishmania donovani* strains in Indian subcontinent. In: 12th international conference on parasitology, Melbourne
35. T'Kindt R, Scheltema RA, Jankevics A et al (2010) Metabolomics to unveil and understand phenotypic diversity between pathogen populations. PLoS Negl Trop Dis 4(11):e904

36. Kafsack BF, Llinas M (2010) Eating at the table of another: metabolomics of host-parasite interactions. Cell Host Microbe 7(2):90–99
37. t'Kindt R, Jankevics A, Scheltema RA et al (2010) Towards an unbiased metabolic profiling of protozoan parasites: optimisation of a *Leishmania* sampling protocol for HILIC-orbitrap analysis. Anal Bioanal Chem 398(5):2059–2069
38. Scheltema R, Decuypere S, Dujardin J et al (2009) Simple data-reduction method for high-resolution LC-MS data in metabolomics. Bioanalysis 1(9):1551–1557
39. Scheltema RA, Kamleh A, Wildridge D et al (2008) Increasing the mass accuracy of high-resolution LC-MS data using background ions: a case study on the LTQ-Orbitrap. Proteomics 8(22):4647–4656
40. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T (2007) Microbial metabolomics: toward a platform with full metabolome coverage. Anal Biochem 370(1): 17–25
41. Brown SC, Kruppa G, Dasseux JL (2005) Metabolomics applications of FT-ICR mass spectrometry. Mass Spectrom Rev 24(2):223–231
42. Hu Q, Noll RJ, Li H et al (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40(4):430–443
43. Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics 7:234
44. Kind T, Fiehn O (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8:105
45. Zheng L, T'Kind R, Decuypere S et al (2010) Profiling of lipids in *Leishmania donovani* using hydrophilic interaction chromatography in combination with Fourier transform mass spectrometry. Rapid Commun Mass Spectrom 24(14):2074–2082
46. Sauer U (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol 15(1):58–63
47. Zamboni N, Sauer U (2009) Novel biological insights through metabolomics and 13 C-flux analysis. Curr Opin Microbiol 12(5):553–558
48. Sauer U (2006) Metabolic networks in motion: 13 C-based flux analysis. Mol Syst Biol 2:62
49. Haanstra JR, Stewart M, Luu VD et al (2008) Control and regulation of gene expression: quantitative analysis of the expression of phosphoglycerate kinase in bloodstream form *Trypanosoma brucei*. J Biol Chem 283(5):2495–2507
50. Miskovic L, Hatzimanikatis V (2011) Modeling of uncertainties in biochemical reactions. Biotechnol Bioeng 108(2):413–423. doi:10.1002/bit.22932
51. Xu TR, Vyshemirsky V, Gormand A et al (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. Sci Signal 3(134):ra20

Chapter 7
Mouse Genetics and Metabolic Mouse Phenotyping

Helmut Fuchs, Susanne Neschen, Jan Rozman, Birgit Rathkolb, Sibylle Wagner, Thure Adler, Luciana Afonso, Juan Antonio Aguilar-Pimentel, Lore Becker, Alexander Bohla, Julia Calzada-Wack, Christian Cohrs, András Frankó, Lillian Garrett, Lisa Glasl, Alexander Götz, Michael Hagn, Wolfgang Hans, Sabine M. Hölter, Marion Horsch, Melanie Kahle, Martin Kistler, Tanja Klein-Rodewald, Christoph Lengger, Tonia Ludwig, Holger Maier, Susan Marschall, Kateryna Micklich, Gabriele Möller, Beatrix Naton, Frauke Neff, Cornelia Prehn, Oliver Puk, Ildikó Rácz, Michael Räß, Markus Scheerer, Evelyn Schiller, Felix Schöfer, Anja Schrewe, Ralph Steinkamp, Claudia Stöger, Irina Treise, Monja Willershäuser, Annemarie Wolff-Muscate, Ramona Zeh, Jerzy Adamski, Johannes Beckers, Raffi Bekeredjian, Dirk H. Busch, Jack Favor, Jochen Graw, Hugo Katus, Thomas Klopstock, Markus Ollert, Holger Schulz, Tobias Stöger, Wolfgang Wurst, Ali Önder Yildirim, Andreas Zimmer, Eckhard Wolf, Martin Klingenspor, Valérie Gailus-Durner, and Martin Hrabě de Angelis

H. Fuchs, Ph.D. • S. Neschen, Ph.D. • S. Wagner, VMD • L. Afonso, Ph.D. • C. Cohrs, Ph.D.
A. Frankó, Ph.D. • M. Hagn, Ph.D. • W. Hans, Ph.D. • M. Horsch, Ph.D. • M. Kahle
M. Kistler • C. Lengger, Ph.D. • T. Ludwig • H. Maier, Ph.D. • S. Marschall, Ph.D.
K. Micklich • B. Naton, Ph.D. • M. Räß, Ph.D. • M. Scheerer • E. Schiller • F. Schöfer, Ph.D.
A. Schrewe, Ph.D. • R. Steinkamp, Ph.D. • C. Stöger, Ph.D. • I. Treise
M. Willershäuser • R. Zeh • V. Gailus-Durner, Ph.D.
German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: hfuchs@helmholtz-muenchen.de; susanne.neschen@helmholtz-muenchen.de;
sibylle.wagner@helmholtz-muenchen.de; luciana.afonso@helmholtz-muenchen.de;
christian.cohrs@helmholtz-muenchen.de; andras.franko@helmholtz-muenchen.de;
michael.hagn@helmholtz-muenchen.de; wolfgang.hans@helmholtz-muenchen.de;
horsch@helmholtz-muenchen.de; melanie.kahle@helmholtz-muenchen.de;
martin.kistler@helmholtz-muenchen.de; lengger@helmholtz-muenchen.de;
tonia.ludwig@helmholtz-muenchen.de; holger.maier@helmholtz-muenchen.de;
s.marschall@helmholtz-muenchen.de; kateryna.micklich@helmholtz-muenchen.de;
beatrix.naton@helmholtz-muenchen.de; michael.raess@helmholtz-muenchen.de;
markusscheerer@helmholtz-muenchen.de; evelyn.schiller@helmholtz-muenchen.de;
felix.schoefer@helmholtz-muenchen.de; anja.schrewe@helmholtz-muenchen.de;
steinkamp@helmholtz-muenchen.de; claudia.stoeger@helmholtz-muenchen.de;
irina.treise@helmholtz-muenchen.de; monja.willershaeuser@helmholtz-muenchen.de;
ramona.zeh@helmholtz-muenchen.de; gailus@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_7,

J. Rozman, Ph.D.
German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Molecular Nutritional Medicine, Else Kröner-Fresenius Center and ZIEL Research Center
for Nutrition and Food Sciences, Technische Universität München,
Gregor-Mendel-Straße 2, Freising, Weihenstephan 85764, Germany
e-mail: jan.rozman@helmholtz-muenchen.de

B. Rathkolb, VMD
German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Chair for Molecular Animal Breeding and Biotechnology, Gene Center,
Ludwig-Maximilians-Universität München, Feodor Lynen-Straße 25, Munich 81377, Germany
e-mail: birgit.rathkolb@helmholtz-muenchen.de

T. Adler, VMD
German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Institute for Medical Microbiology, Immunology, and Hygiene,
Technische Universität München, Trogerstraße 30, Munich 81675, Germany
e-mail: thure.adler@helmholtz-muenchen.de

J.A. Aguilar-Pimentel, Ph.D.
Department of Dermatology and Allergy, Biederstein,
Clinical Research Division of Molecular and Clinical Allergotoxicology, TUM,
Biedersteiner Straße 29, Munich 80802, Germany

Division of Environmental Dermatology and Allergy,
Technische Universität München/Helmholtz Zentrum München,
Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: aguilar.pimentel@lrz.tum.de

L. Becker, Ph.D.
German Mouse Clinic, Institute of Experimental Genetics,
Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Friedrich-Baur-Institut, Department of Neurology,
Ludwig-Maximilians-Universität München, Ziemssenstraße 1a, Munich 80336, Germany
e-mail: lore.becker@helmholtz-muenchen.de

A. Bohla, Ph.D. • A. Götz, Ph.D. • T. Stöger, Ph.D. • A.Ö. Yildirim, VMD
Comprehensive Pneumology Center, Institute of Lung Biology and Disease,
Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH),
Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: alexander.bohla@helmholtz-muenchen.de; alexander.goetz@helmholtz-muenchen.de;
tobias.stoeger@helmholtz-muenchen.de; oender.yildirim@helmholtz-muenchen.de

J. Calzada-Wack, M.D. • T. Klein-Rodewald, Ph.D. • F. Neff, M.D.
Institute of Pathology, Helmholtz Zentrum München, German Research Center for
Environmental Health (GmbH), Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: calzada@helmholtz-muenchen.de; tanja.klein@helmholtz-muenchen.de;
frauke.neff@helmholtz-muenchen.de

L. Garrett, Ph.D. • L. Glasl • S.M. Hölter, Ph.D. • A. Wolff-Muscate
J. Graw, Ph.D. • O. Puk, Ph.D.
Institute of Developmental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1,
Neuherberg 85764, Germany
e-mail: lillian.garrett@helmholtz-muenchen.de; lisa.glasl@helmholtz-muenchen.de;
hoelter@helmholtz-muenchen.de; wolf-muscate@helmholtz-muenchen.de;
graw@helmholtz-muenchen.de; oliver.puk@helmholtz-muenchen.de

G. Möller, Ph.D. • C. Prehn, Ph.D. • J. Adamski, Ph.D., M.Sc.
Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München,
Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: gabriele.moeller@helmholtz-muenchen.de; prehn@helmholtz-muenchen.de;
adamski@helmholtz-muenchen.de

I. Rácz, Ph.D. • A. Zimmer, Ph.D.
Institute of Molecular Psychiatry, University of Bonn,
Sigmund-Freud-Straße 25, Bonn 53105, Germany
e-mail: iracz@uni-bonn.de; a.zimmer@uni-bonn.de

J. Beckers, Ph.D. • M. Hrabě de Angelis, Ph.D. (✉)
German Mouse Clinic, Institute of Experimental Genetics, Helmholtz Zentrum München,
German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1,
Neuherberg 85764, Germany

Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan,
Technische Universität München, Freising 85350, Germany
e-mail: beckers@helmholtz-muenchen.de; hrabe@helmholtz-muenchen.de

R. Bekeredjian, M.D. • H. Katus, M.D.
Otto-Meyerhof-Zentrum, Department of Medicine III, Division of Cardiology,
University of Heidelberg, Im Neuenheimer Feld 410, Heidelberg 69120, Germany
e-mail: raffi.bekeredjian@med.uni-heidelberg.de; hugo_katus@med.uni-heidelberg.de

D.H. Busch, M.D.
Institute for Medical Microbiology, Immunology, and Hygiene,
Technische Universität München, Trogerstraße 30, Munich 81675, Germany
e-mail: dirk.busch@microbio.med.tum.de

J. Favor, Ph.D.
Institute of Human Genetics, Helmholtz Zentrum München, German Research Center
for Environmental Health (GmbH), Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: favor@helmholtz-muenchen.de

T. Klopstock, M.D.
Friedrich-Baur-Institut, Department of Neurology,
Ludwig-Maximilians-Universität München, Ziemssenstraße 1a, Munich 80336, Germany
e-mail: thomas.klopstock@med.uni-muenchen.de

M. Ollert, M.D.
Department of Dermatology and Allergy, Biederstein, Clinical Research Division of Molecular
and Clinical Allergotoxicology, TUM, Biedersteiner Straße 29, Munich 80802, Germany
e-mail: ollert@lrz.tum.de

1 Mouse Genetics

Historically, the coexistence of humans and mice can be traced back many centuries, with both species following similar paths in their movement to different parts of the world. Often maligned for their pathogenic contamination of food stores, mice gained some appreciation during the Roman era when so-called “fancy mice” were bred for their unusual coat colours. It was not until more recent times that in the twentieth century a retired American teacher, Abby Lathrop, began breeding these “fancy mice” in a more systematic fashion. This work, combined with the development of the first inbred mouse strains by Ernest Castle and Clarence Little, laid the foundations for the mouse becoming one of the most important organisms to model human diseases [1].

Among the many reasons why the mouse is an essential animal model for medical research is the availability of several different inbred mouse strains. All mice within the same inbred strain are as genetically homogenous as a pair of monozygotic twins. Thus, when these mice are used to determine the effect of specific mutations or treatments, the inter-individual differences attributed to genetic polymorphisms can be over-looked. From a practical standpoint, employing mice also has the advantage that, as small animals, they require minimal space and energy and can be generated relatively fast for experimentation. Furthermore, as a mammal, mice have anatomical and physiological similarities to humans including certain stages of their development and physiological pathways.

The mouse genome is – depending on the level to be considered – 90% to 99% identical with the human genome as well as the next to have been sequenced [2–6]. As a result, the accessibility of the mouse genomic sequence made this small animal

H. Schulz, M.D.
Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: schulz@helmholtz-muenchen.de

W. Wurst, Ph.D.
Institute of Developmental Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Chair of Developmental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany
e-mail: wurst@helmholtz-muenchen.de

E. Wolf, VMD
Chair for Molecular Animal Breeding and Biotechnology, Gene Center, Ludwig-Maximilians-Universität München, Feodor Lynen-Straße 25, Munich 81377, Germany
e-mail: ewolf@lmb.uni-muenchen.de

M. Klingenspor, Ph.D.
Molecular Nutritional Medicine, Else Kröner-Fresenius Center and ZIEL Research Center for Nutrition and Food Sciences, Technische Universität München, Gregor-Mendel-Straße 2, Freising, Weihenstephan 85764, Germany
e-mail: mk@tum.de

an even more important tool for the modeling of human diseases. Biotechnology enabled us to generate mutant mouse lines with different technologies and for various purposes to answer a series of scientific questions. Initially, spontaneous mutants were identified, bred and analyzed followed in the 1960s with the use of irradiation to generate mouse mutants that carried, in most cases, chromosomal aberrations [7]. Later on, chemical mutagenesis became a very powerful tool to produce mouse mutants. In its initial phase, these experiments were important to study the toxicological impact of substances on human health, and to develop rules for maximum permissible concentrations for hazardous substances and radiation. The mouse mutants were used to calculate the mutation frequency by comparing the obtained mutant mice with a so-called "specific locus test" [8]. But very soon it turned out that mutants from mutagenesis experiments harbored the potential as an excellent tool to study mammalian genetics and to research gene functions. In particular one substance, the synthetic alkylating compound *N*-ethyl-*N*-nitroso-urea (ENU) came into special focus of geneticists, since the mutations caused by this mutagen are in most cases point mutations [9]. This reflects many situations in human diseases. Thus ENU was used as the prime substance to generate mouse mutants in large scale mutagenesis screens world-wide (e.g. [9–12]) (Box 7.1)

Box 7.1 ENU Mutagenesis

In ENU mutagenesis projects, male mice are injected with a dose of the mutagen *N*-ethyl-*N*-nitroso-urea (ENU). After injection, the ENU starts to distribute in the whole body and in particular targets early stem cell spermatogonia. ENU alkylates genomic DNA during cell division, by transferring its ethyl group to nucleophilic sites of nucleic acids, ultimately leading mainly to point mutations [23, 24]. As ENU acts as a poison, the ENU-treated males get sterile for a certain period, but recover dose- and strain-dependent after an individual sterility period. The inbred strain C3HeB/FeJ turned out to both tolerate high doses of injected ENU and to regain fertility to about 50% of injected mice following an administered ENU dosage from 80 to 90 mg/kg body weight in 3 weekly intervals. After a recovery period of 3 weeks, injected mice are mated with untreated wild-type female mice. To assure that offspring is produced from the mutagenized sperm, only litters born after a timeframe accordant to two spermatogenesis cycles of 49 days each [25] are weaned for phenotyping. The resulting offspring will be heterozygous for the mutations that occurred in the spermatogonia of the sperm cells that succeeded in fertilization. It was calculated that every F1 mouse is carrying an average of 20 independent mutations [26].

There are variations in ENU-mutagenesis projects: in a dominant screen, the F1 offspring of an ENU-treated male is screened for the occurrence of interesting and medically relevant phenotypes. The detected individuals with

(continued)

Box 7.1 (continued)

phenotypic variations have to be proven as inherited mutations by a confirmation cross, where the F1 animal is mated with a wild-type animal. In the resulting N2 generation a fraction of the offspring should carry the characteristic phenotype. According to Mendelian rules it should be 50%, but in many cases incomplete penetrance or other effects reduce the outcome to a lower number. In the case that some offspring with the characteristic phenotype are detected the mutant is confirmed. As a next step, the mutant line has to be maintained and phenotypically characterized. Until recently a major challenge in ENU mutagenesis projects was that the causative mutation is a priori unknown, and has to be identified via further breeding and analytical efforts. With today's sequencing capabilities the detection of causative mutations in ENU mutagenized mice has become very efficient and no longer represents a major bottleneck.

A variation of ENU mutagenesis is to screen for recessive phenotypes. In this case, the F1 offspring is further bred to produce offspring that are homozygous for the mutated allele. The G2 generation is then subjected to phenotypic analysis. A further variation of ENU-technology involves breeding ENU-treated males to females with large chromosomal aberrations in order to detect recessive ENU mutations in a genomic region of interest [9]. Here, the effect of the ENU-mutations can be analyzed in a hemizygous status that has to be specifically located within the defined region of the chromosomal aberration. In addition, in so-called sensitized screens, ENU-technology is used to screen for modifier genes that interact with a specific allele. The rationale is to cross ENU-treated males with females that carry already a mutation in a known gene. The F1 animals can be screened for new phenotypes that are caused by the interaction of the known locus and the new ENU-based mutation on a different locus. This approach may be used to identify mutations in genes that directly or indirectly interact with the known mutation and that modify (enhance or reduce) a mutant phenotype [27].

Some research institutes that carried out ENU-screening projects for longer periods have frozen DNA and sperm samples from F1 animals of treated ENU males. There is the possibility to screen the DNA-archives of F1 animals for mutations in a specific gene, and to generate live animals from their frozen sperm.

Whereas mutagenesis projects are phenotype-driven "reverse genetic" approaches, gene-driven "forward genetic" technologies directly focus on the desired gene for the creation of mutations within a specified locus. The most widely used technology is the classical knock-out. In a knock-out mouse, essential parts of the gene of interest are eliminated or exchanged in the genome by homologous recombination in

embryonic stem cells, which are then integrated in early mouse embryos and have to undergo germ line transmission to produce mutant animals. The result is a mouse where the specific gene product is manipulated in a way that it cannot fulfill its function any more. The most important purpose of knock-out mice is to study the function of a gene in combination with a resulting phenotype.

The application of knock-out technology is not successful in every species of interest. For example, the development of knock-out applications in rats – which is one of the most important "competitors" of the mouse as a small mammalian species to model human diseases – is still in an early phase. Even if for behavioral studies and some aspects of metabolic phenotyping the rat might be preferred, the possibility to generate defined mutants with a standard technology made the mouse a valuable tool. Furthermore, in the meantime the generation of knock-out mice is commercially available.

With the availability of conditional mutagenesis approaches the knock-out technology has become increasingly refined: it is now possible to design tissue specific knock outs by breeding mice that carry a defined construct that can be activated by crossing with lines expressing a Cre-recombinase under any tailored promoter. Another special tool is the possibility of inducible expression of knock-out constructs at a certain time point, e.g. by application of substances like tamoxifen. These refinements enable researchers to study the effects of a gene only in a specific organ or at a specific age, which is of special interest for genes that have different functions in different organs, or have indispensible functions during embryogenesis.

The availability of a large array of versatile technologies made the scientists to consider the establishment of central resources for mouse mutants and their data. There are three levels of challenges:

- **Production** of mutant mouse lines,
- **Phenotyping** of the mutant mouse lines, and finally
- **Archiving and distribution** of mice and data.

The generation of mutant mouse lines has been targeted by systematic projects for the production of mutants for every single gene. Starting with some national projects like the German Gene Trap Consortium (GGTC), the European, U.S. and North American efforts were combined within EUCOMM [13], KOMP, and Norkomm [14] initiatives. The international knock out mouse consortium [15] now unifies these projects to a worldwide network of research institutes that systematically produce mutant mouse lines that are accessible for the scientific community.

Phenotyping of mutant mouse lines until recently was exclusively done by specialized laboratories that have a focus on a certain disease, organ or molecular pathway. In the last decade, a growing need developed to perform systematic and comprehensive phenotyping of each mutant mouse line. It became evident that it is of importance to broadly characterize mutant mouse lines, as most genes have pleiotropic functions. For this purpose, mouse clinics were founded to address this challenge, where mutant lines are analyzed not only for the occurrence of a specific phenotype, but to obtain a complete check up [16] for a large number of medically relevant areas, like behavior, neurology, nociception, dysmorphology, bone and

cartilage, cardio-vascular function, metabolism, clinical chemistry, immunology, allergy, steroids, eye and vision, lung function, gene expression and pathology.

The first mouse clinic with open access for the scientific community was the German Mouse Clinic at the Helmholtz Zentrum in Munich [17–18] (www.mouseclinc.de). In 2002 the European program EUMORPHIA [19] (www.eumorphia.org) was started to harmonize phenotyping protocols across the continent. The EUMORPHIA program was the basis for the foundation of the European mouse disease clinic (EUMODIC, www.eumodic.org). Within the EUMODIC program four mouse clinics, the Institut Clinique de la Souris, the Wellcome Trust Sanger Institute, the MRC in Harwell and the German Mouse Clinic work together to reach the goal of analyzing 500 mutant mouse lines that originate from the EUCOMM resource. The data are freely accessible for the scientific community and the public and can be accessed via the Europhenome database (www.europhenome.org). In addition, this program combines the efforts of, on the one hand, these four mouse clinics in Europe and, on the other hand, specialized laboratories for each disease area that will focus their phenotyping efforts on interesting mutant lines for a detailed characterization in their field of research.

The EUMODIC program is a proof of principle approach that scientists are able to organize and coordinate a large scale phenotyping effort across countries and institutions. The next step will be to establish an international consortium, the International Mouse Phenotyping Consortium (IMPC, www.mousephenotype.org) where scientists from Europe, the U.S. and Canada, Asia and Australia network together to reach the highly ambitious goal to phenotype a mutant mouse line for each single gene within a time period of 10 years [20].

Mutant mouse lines are a valuable resource and that therefore need to be conserved and distributed to interested researchers upon request. After generation and phenotyping of a mutant mouse line, further analysis might not be possible with existing tools, but the line should be kept available for future work. It is not possible to maintain all mutant lines as live animal stocks even if all resources worldwide would be used. Thus preservation of frozen embryos or frozen sperm is currently the method of choice. The Jackson Laboratory (www.jax.org) provides one of the world-wide leading resources where many mutant lines as well as the most demanded mouse strains are archived and can be retrieved. The European Mouse Mutant Archive [21, 22] (EMMA, www.emma-net.org) is the leading European repository for archiving and distribution of mouse lines. It collaborates closely with The Jackson Laboratory and other international mouse repositories in the Federation of International Mouse Resources (FIMRe).

In order to speed up the distribution of the mouse lines to interested scientist, the know-how to re-derive a mouse line (e.g. via in-vitro fertilization, IVF) is made available to animal facilities and veterinarians by offering courses for cryo-preservation techniques. The possibility to send sperm or embryos instead of live mice is also of advantage regarding the sanitary status of research animals. Most pathogens can be cleared via washing steps that are included in the cryo-preservation protocols.

In addition to the preservation and distribution of the mouse lines themselves, it becomes increasingly important to disseminate the available large-scale datasets from mouse mutants. Computer scientists are involved in the development of databases that store data collected from each mouse line and information about the availability and accessibility of a certain mutant. The leading resource to collect this information is the mouse genome informatics database that is offered via the website of The Jackson Laboratory (www.informatics.jax.org).

The capacities for phenotyping, archiving and distribution of mouse models are currently not matching the demand by the biomedical research community. Moreover, sustainable funding of the underlying infrastructure is usually lacking. In Europe, these problems are being addressed by the Infrafrontier Project (www.infrafrontier.eu), an initiative of research institutions, research organizations and funding institutions that aims at establishing the pan-European Infrafrontier Research Infrastructure for open-access to scientific platforms and services for phenotyping, archiving and distribution of mouse disease models. This will provide also the basis for large-scale international efforts such as the International Mouse Phenotyping Consortium (IMPC).

2 Phenotyping

2.1 Standardized Large-Scale Phenotyping Approaches

The phenotyping of mutant mouse lines is still the bottleneck in the pipeline from mutant mouse generation via phenotyping and archiving. However, this step is crucial for the decision, which gene is associated with a certain disease, and which mutant mouse can serve as a model for the disease. Thus, efforts concentrated on increasing the power and efficiency of standardized comprehensive phenotyping of mouse models need to be performed. Within this section, we describe the systemic phenotyping approaches by using the German Mouse Clinic (for more details and protocols see [17, 28–29]) as an example, with a major focus on metabolic techniques.

The phenotyping of mouse lines in the German Mouse Clinic (GMC) is divided into a primary screen, for an almost complete comprehensive characterization in the fields allergy, behavior, clinical chemistry, diabetes, dysmorphology, bone and cartilage, energy metabolism, steroids, eye and vision, immunology, lung function, molecular phenotyping, neurology, nociception, and pathology, as well as secondary, more deep-drilling and hypothesis-driven screens. Secondary and tertiary screens are designed to obtain a detailed phenotypic analysis in selected areas where alterations within the primary screen were detected. In order to be able to cover all these important research areas, the GMC is a consortium in which experts from various fields of mouse physiology and pathology in close cooperation with clinicians contribute to the phenotyping of a mutant mouse line.

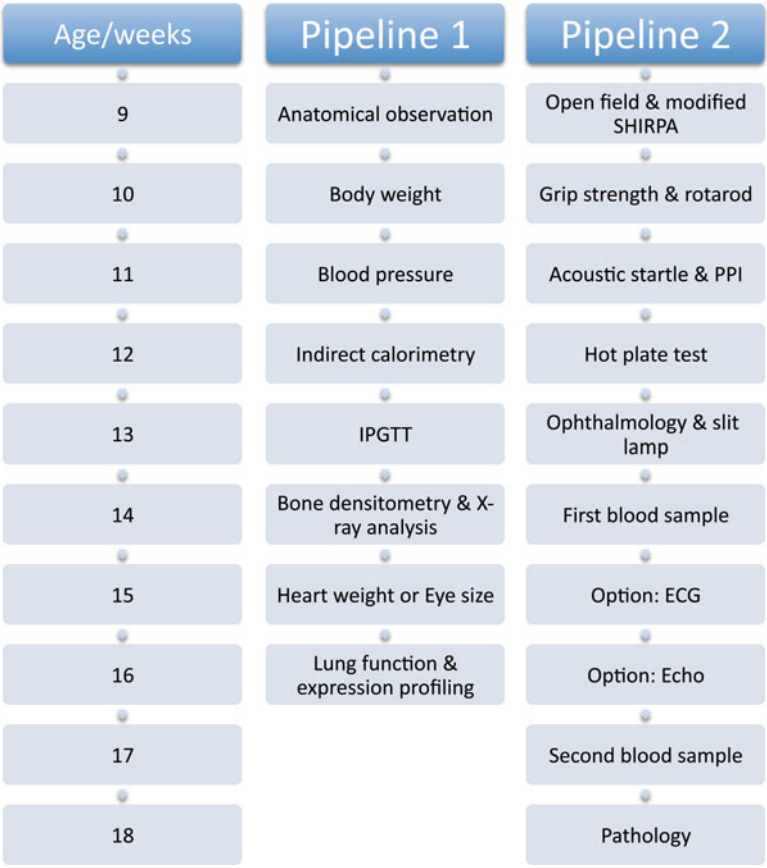


Fig. 7.1 The pipeline structure for the systematic and standardized primary phenotyping of mutant mouse lines in the German Mouse Clinic

In the primary screen, a cohort of 80 animals (20 male mutants, 20 female mutants, 20 male controls and 20 female controls) is analyzed in two pipelines (each of them with 10 animals per sex and genotype). In total, more than 500 parameters are determined from each individual mouse. The complete cohorts for phenotyping are shipped to the GMC-facility at the age of 7 weeks. The age range of animals within the same cohort should not exceed 1 week. After import into the German Mouse Clinic, the mice have 2 weeks to adapt to the new environment, and then the phenotyping starts by analyzing them for 1 week in each specific test.

As shown in Fig. 7.1, in the first pipeline phenotyping starts with an anatomical observation. At the age of 11 weeks, blood pressure is analyzed, followed by indirect calorimetry and intraperitoneal glucose tolerance test in the two subsequent weeks. Then, an X-ray image is taken and analyzed for bone structural abnormalities, and bone density is measured. One week later, a blood sample is collected from overnight fasted mice to determine blood lipid and glucose values. Laser

interference biometry is used to measure the eye size. From a sub-cohort of the mice the heart weight is determined, while in the remaining animals lung function is analysed. From the latter animals organs are collected that can be used for expression profiling experiments.

The second pipeline starts with behavior and neurological investigations by open field test, SHIRPA-test (this is a battery of behavioral and neurological observations to assess basic functions like general appearance, movement, reflexes [30, 31]), rotarod and grip strength analysis as well as acoustic startle and pre-pulse inhibition test. Nociception is assessed using the hot-plate test. Further information about eye function is obtained using funduscopy and slit-lamp analysis. Afterwards a blood sample is collected, which is prepared and distributed for analysis of clinical chemistry, haematology, immunology, allergy and steroid level parameters. Detected alterations in blood parameters can be confirmed in a second sample that is taken 3 weeks later. In-between the blood sampling procedures, there is the option to investigate cardio-vascular parameters via echo- or electrocardiography. The mice of the second pipeline end up in the pathology screen for macroscopic and microscopic analysis.

After finalizing both primary screening pipelines the complete dataset of the mutant mouse line is analyzed and discussed with all scientists of the different modules: In many cases, the synopsis presented by the scientists from specialized areas uncovers new information that might have been considered as irrelevant, if the data had been raised without any connection between the different partners. By taking into account parameters from other screens that are biologically interconnected with each other, even borderline significant findings take on a new light. Thus, the primary phenotypic analysis helps to create new ideas and phenotyping hypotheses for more detailed characterization in secondary analysis. Secondary experiments are offered by every screen of the GMC. There, a confirmation of the findings from the primary screen can be performed with an independent cohort, and the phenotype will be characterized in more detail with more sophisticated technologies that are too laborious, time consuming, expensive and of a too high resolution of detail to be implemented in the primary phenotyping. As a few examples computer tomography for metabolic and bone characterization, electro-myopathy or -encephalopathy for neurological analysis, analysis of olfactory function and recognition memory for behaviour phenotyping can be mentioned. Some further analysis can be done in specialized labs like, for example, the neuropathic pain model for secondary nociceptive analysis.

Imaging technologies are becoming more and more important in mouse phenotyping. Still, for primary phenotyping the range of applications is currently limited to X-ray or echo analysis due to the need for high investment and the labor- and time-intensive way to run the experiments. But in the near future, imaging techniques will even replace traditional well-established methods. There are computer tomography based measures like micro-CT or pQCT (peripheral quantitative computed tomography) available as well as magnetic resonance tomography (MRT) machines for small animals. The next step is to integrate PET (positron emission tomography) devices into mouse CT or MRT machines. The progress in the improvement of existing, and further development of new imaging technologies is immense.

In the last decade technical progress made digital X-ray imaging faster and cheaper, while at the same time the obtainable resolution was improved to reach an even better quality than X-ray films. These perspectives can really make us confident for the future.

Another aim in implementing imaging procedures is the improvement of animal protection: Reduction, refinement and replacement are the three goals focused on minimizing the load on experimental animals. Imaging technologies will make major contributions with respect to refinement of experiments as does the generation of as many data from single animals for reduction.

2.2 Metabolic Phenotyping

Phenotyping of mouse mutants for metabolic parameters has become more and more important since due to the increasing incidence human diseases like diabetes, obesity or the metabolic syndrome are in focus of scientific research. In this respect, there are many possibilities to collect data from mice. The easiest obtainable parameter is the body weight at different age levels of the mouse. More information will be gained, if the body composition is analyzed. This can be achieved either by DEXA (dual energy X-ray absorptiometry) technology, which is able to discriminate between fat mass, lean mass and bone tissue by the application of two X-ray beams of different energy levels. The technology is limited in applications, as only mice of a weight over 18 g can be assessed with a reliable accuracy. Nuclear magnetic resonance (NMR) based methods provide a more modern way for the determination of fat mass and lean mass. In addition, body fluids can be analyzed by this method, which is much quicker than DEXA. If body composition needs to be determined in the most accurate way, Soxhlet extraction will be the method of choice. Using computer tomography, the fractions of subcutaneous and visceral fat in the body can be quantified and compared.

From a small amount of blood, clinical chemical parameters can be determined like glucose, cholesterol and triglyceride levels. Further parameters of interest might be high and low density lipoprotein (HDL/LDL)-cholesterol, non-esterified fatty acids (NEFA) and glycerol that can be measured via automated analyzers.

The standard version might be to determine these parameters under non-fasted conditions, but fasting will result in additional information. Thus, both a fasted and a non-fasted version of the sample will yield separate information of high value.

Indirect calorimetry is an important part of the puzzle to put together a complete picture of the metabolic situation in a mouse line. This is a method of estimating energy expenditure by measuring respiratory gases – oxygen consumption and carbon dioxide release. The analysis of mice in special cage systems for indirect calorimetry will, depending in the system used, yield information about:

- Oxygen consumption
- Carbon dioxide production
- Respiratory exchange ratio

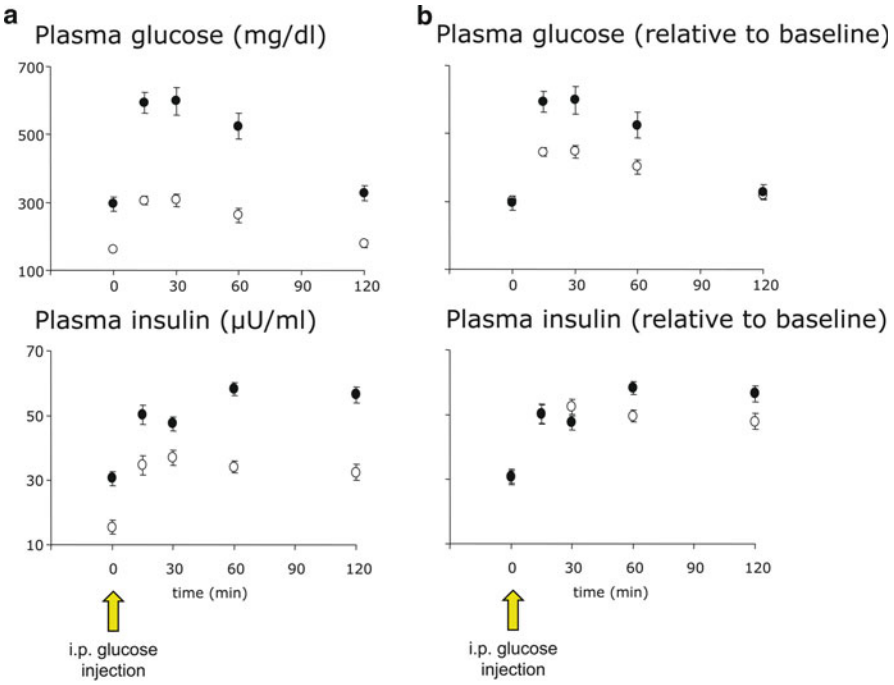


Fig. 7.2 Only 1 week of high-fat diet feeding induces glucose intolerance in mice. Data are shown from the intraperitoneal glucose tolerance tests (i.p.GTT) performed in age-matched males of a lean "general purpose" mouse strain fed with either a standard laboratory diet (*open circles*) or a 27 gm% high-fat diet for 1 week (*closed circles*). (**a**) Depicts time-dependent changes in plasma glucose and insulin concentration excursions at baseline (time 0 min) and following an intraperitoneal glucose injection. (**b**) Illustrates relative plasma glucose and insulin excursions from the respective baseline of each group. Data represent means ± SEM of 7–9 mice/group

- Heat production
- Food and water consumption
- Locomotor activity

In addition, body weight and body temperature should be assessed. Thus, a complete picture about energy input and energy utilization can be obtained.

The glucose tolerance test (GTT) collects information about how the body handles glucose. The same test is also applied in human diagnostics for a first indication of diabetes. In mice, the glucose can be administered into the body either by intraperitoneal (i.p.) injection (which is the easier and most frequently used mode of application) or by the oral route using a gavage. After a fasting period the basal glucose level is determined, and the glucose is administered. Dependent on the protocol used the blood glucose level is measured at several time points, e.g. 15, 30, 60 and 120 min after administration. The normal reaction is a maximum blood glucose value between 15 and 30 min and then the body regulates the level back to the basal level at the last measurement at 120 min (see Fig. 7.2). A higher peak value as well

as a missing or delayed back regulation to the basal glucose level, both represented by the area under the curve, indicates problems in blood glucose regulation.

Impaired glucose tolerance and elevated fasting plasma glucose concentrations might result from an impaired capacity of pancreatic β-cells to adequately secrete insulin in response to increases in blood glucose concentrations. Alternatively (or in addition) insulin resistance, defined as the decreased sensitivity or responsiveness of tissues (e.g. liver, skeletal muscle, adipose tissue, brain, heart) to insulin action, could account for a reduced glucose clearance from plasma during GTTs. Both pathophysiological conditions have been linked to type 2 diabetes, the metabolic syndrome, atherosclerosis, and cardiovascular disease. In order to quantify and localize defects in insulin action or β-cell function the glucose-clamp technique serves as a valuable diagnostic tool. In a euglycemic-hyperinsulinemic clamp, insulin action on the whole body as well as at the organ level is assessed. In a hyperglycemic clamp, primarily pancreatic β-cell sensitivity in response to elevations in plasma glucose (β-cell function) is determined [32].

Steroid screen provides information on the overall concentrations of signal molecules quantified by LC-MS [33] at the Genome Analysis Center (www.gac-munich.de). This screen allows quantification of different steroids in mouse plasma and in tissue including stress- and glucose-balance-relevant glucocorticoids. There is the possibility of a secondary screen of targeted metabolomics assay quantifying hexoses, amino acids, biogenic amines and lipids [34].

2.3 Challenge Experiments

For the analysis of complex human diseases, the genetic predisposition has to be taken into consideration but also environmental factors like life style and aging. In this respect one might talk about a triangle of genotype, envirotype and phenotype, where all three factors interact with each other [35]. For modeling human diseases with the mouse, challenge experiments are used to mimic influences of environmental factors. The challenge experiments might provoke phenotypic reactions in mutant animals that would remain hidden without the challenge.

The term “challenge experiment” covers a wide area that spans from short-term, acute reactions (like a glucose tolerance test) to long-term experiments that might last several months where the influences reflect a mild but chronic situation. For the analysis of conditions that correspond to the situations of modern human life, challenges that mimic the environmental impact on human health are of special interest. In Fig. 7.3 an example for an environmental challenge platform is shown. The challenge platform comprises five areas that target the surfaces where the human body gets into interaction with its environment: diet, lung, stress, infection and activity. For each of the five parts of the platform, challenge tests ranging from acute to chronic impact were developed: Applying different diets is a powerful tool to influence physiologic parameters in mice. To stimulate reactions of the lung,

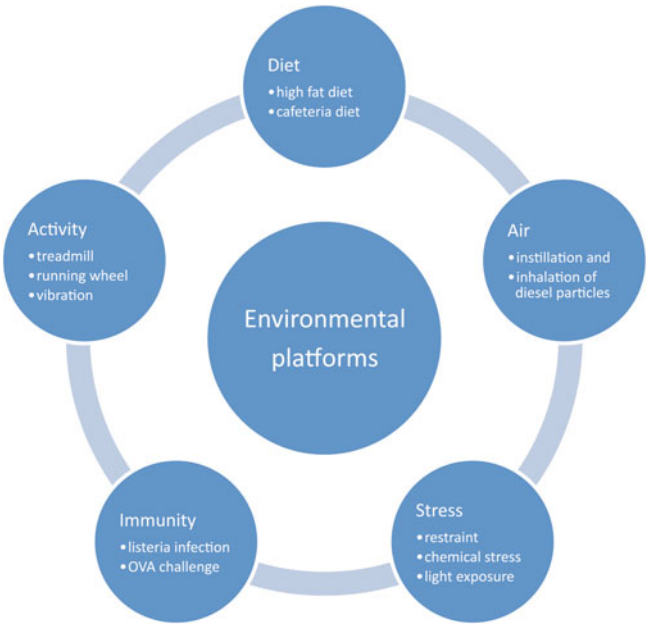


Fig. 7.3 Example of an environmental challenges platform that is applied in the German Mouse Clinic at the Helmholtz Zentrum München in Munich to analyze genotype-environmental interactions. Five areas with different challenge experiments are available that are designed to mimic environmental influences on human health

instillation (acute reaction) or inhalation (chronic reaction) of diesel particles is used. For immune reactions, ovalbumin (OVA) challenge and infection with Listeria monocytogenes are standard experiments. Stress conditions can be simulated in mice by restraint, by the application of chemical substances that produce oxidative stress, or by light exposure. In an activity platform, mice can run either on a treadmill system or are allowed to run voluntarily on running wheels that are placed into the home cages. Another possibility for activity is to put mice on a vibration platform to provoke muscle contraction. The challenge tests can be even combined with each other, and a specialized read-out from the primary phenotyping pipelines can be applied.

2.3.1 High Fat Diet Challenge

Design and application of "customized" environmental challenges, which take the type of genetic modification and its putative disease consequences into account, is an important tool for studying gene-environment-phenotype interactions. Hepatic insulin resistance appears to be a major – although not well understood – core defect

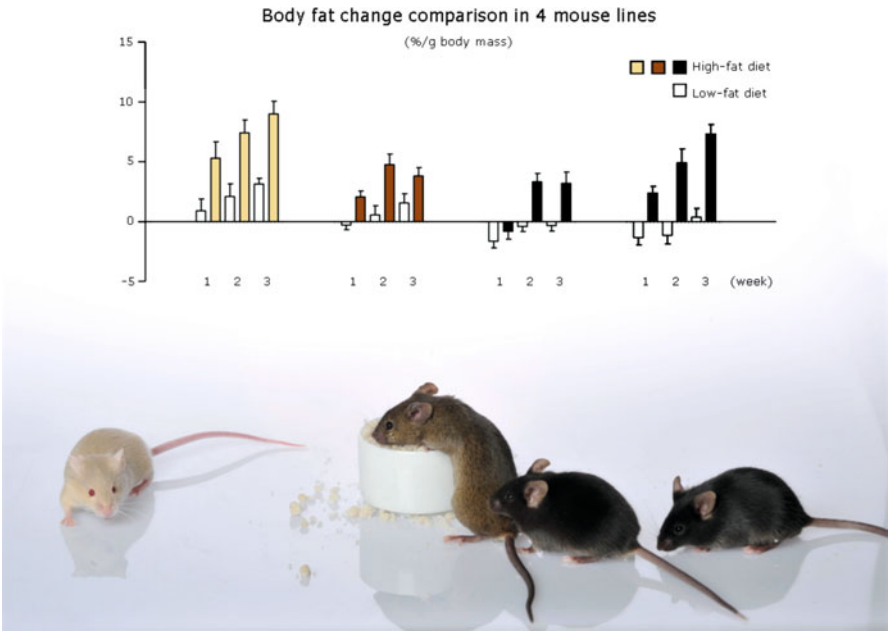


Fig. 7.4 Mice bred on four different genetic backgrounds show marked differences in their response to gain body fat when exposed to the same high fat diet challenge for 1, 2, or 3 weeks (*colored bars*). The *open bars* represent litter- and age-matched controls treated similar to the respective high-fat diet challenged groups but fed with a low fat (standard laboratory mouse diet) diet. Data represent means ± SEM of 9–14 mice/group

in type 2 diabetes pathogenesis and is often associated with non-alcoholic fatty liver disease. A short (1–3 weeks) high-fat diet challenge, with safflower oil being the major fat source, is associated with increases in hepatic lipid deposits and decreases in hepatic insulin sensitivity. Thus, applying such a diet challenge is a powerful approach to dissect pathomechanism implicated in lipid-mediated hepatic insulin resistance. Nevertheless, one has to keep in mind that physiological effects caused by such a high-fat diet challenge differ markedly with regards to magnitude and temporal patterns depending on the genetic background of a mouse model (Fig. 7.4).

Another animal model, where a diet challenge is applied to trigger a particular pathophenotype is the New Zealand Obese (NZO) mouse. Exclusively males develop polygenic obesity-associated type 2 diabetes (diabesity) with a phenotype penetrance of at maximum 50%. Dietary carbohydrate restriction markedly aggravates obesity but completely prevents hyperglycemia and β-cell destruction in NZO males. Exposing carbohydrate-restricted male NZO mice after an age of 18 weeks to dietary carbohydrates very rapidly induces overt diabetes in all individuals. Therefore, such a diet challenge allows for precise investigation of time-dependent mechanisms underlying progressive β-cell degeneration.

2.3.2 Drug Challenge

Mouse models are of great value for investigating mechanisms of drug action on both the whole body as well as at the organ level. In contrast to humans, where multiple tissue biopsies are limited, mouse plasma, urine, bile, tissue, etc. samples are not subject to these limitations. Thus, changes in phenotypic parameters in combination with metabolic (e.g. via targeted or non-targeted metabolomics) and transcriptomic signatures, histological features, etc. indirectly or directly related to drug action can be evaluated in detail. The generation of comprehensive datasets ("drug-types") therefore enables a more precise modeling of highly dynamic, multidimensional processes and promises a substantial gain of knowledge by integrating a systems biology approach. Insights from drug challenge experiments in mouse models thus contribute to a better understanding of drug target organs, their side effects, underlying mechanisms separating drug responders from non-responders, and novel options in the therapy of human diseases.

2.4 Data Analysis

In order to handle the huge amount of data that is generated during the phenotyping process, the help of database systems is needed. Laboratory information management systems (LIMS) based on relational databases can store demographic data for each single mouse (e.g. date of birth, genetic background, genotype, pedigree information) as well as the complete phenotyping data and accompanying meta-data records (information on protocols how the data was taken) collected from each individual.

As databases may be designed for the specific needs of a facility, most phenotyping centers have programmed their own customized LIMS systems. MausDB [36], the mouse and phenotyping data management system of the German Mouse Clinic covers a broad spectrum of functionalities, and has been provided to the scientific community under an open source license. It can be downloaded via: http://www.helmholtzmuenchen.de/ieg, and has been in use by at least a dozen facilities worldwide. For public access, data from individual mouse clinics and phenotyping institutions is uploaded to public databases such as the Mouse Phenome Database (http://phenome.jax.org/), or the Europhenome database [37] (www.europhenome.org).

The analysis of the enormous data set that is collected within the phenotyping pipelines is a major challenge. The primary screen aims to generate working hypotheses for follow-up studies. Based on the observed data, researchers have to come to a final decision whether and in which areas to invest further efforts. An essential task is the development of appropriate data analysis techniques to support scientists in the decision as to whether the mutant group differs from the control group. Inferential statistics may serve as a helpful tool. Parametric tests such as the Student's *t*-test or, in the case of variables that do not conform to the normal distribution, non-parametric

tests, can be a first approach. For some variables, sex has to be treated as a confounder and, therefore, data from male and female mice has to be considered separately or two-way analysis of variance (ANOVA) is applied. However, for some variables, more complex analysis techniques have to be applied, especially in cases where additional factors confound the data.

Collecting data from control mice of the same genetic background over a long time period provides the possibility of comparing a current mutant line with this pooled control sample. In this case, reference ranges can be calculated for each variable and may be used for the decision as to whether the observation of a mutant line is altered compared to the pooled set of data from control animals. This way of analysis takes into account periodical influences and reduces the number of false positive decisions.

Generally speaking, statistics serves as a tool for finding evidence in support of the hypothesis of differing groups based on the observed data. However, it will not be able to substitute an experienced scientist in interpreting the findings.

The standardized generation of data sets from large-scale phenotyping projects offers the possibility to run meta-analysis approaches and data mining exercises on the tremendous amount of data. The analysis of the complete phenotyping database as a resource (e.g. the Europhenome database) bears the potential to uncover novel correlations between parameters and patterns associated with some disease areas that might not be detectable through the analysis of single mutant lines alone. While activities in this field are still in the starting phase, and the development of software tools for this purpose is in progress, first results are already available. For example, syn-expression groups of genes in different organs were discovered by the analysis of data sets from molecular phenotyping activities, where transcript profiles using a microarray containing 21,000 cDNA probes in a series of disease models were assessed. Using microarray experiments, expression patterns of in total 90 organs from 46 mutant mouse lines were analyzed, and identified up to 232 differentially expressed genes in 45 organs [38]. The approach helped in identifying the recurring regulation of particular genes and groups of co-expressed genes.

2.5 Application of Large Scale Mouse Phenotyping

There is a variety of publications available that used a comprehensive standardized mouse phenotyping approach to address a specific scientific question. The papers might be grouped into four main areas:

- Discovery of unknown gene functions and pleiotropic effects
- Models for human diseases (e.g. models for diabetes or metabolic diseases)
- Gene-environment interactions
- Systemic phenotyping for target validation

In the following section a few examples will be mentioned for each of these applications.

2.5.1 Discovery of Unknown Gene Functions and Pleiotropic Effects

Loss-of-function mutations in the ubiquitously expressed transcription factor FoxP2 impairs the ability to speak in humans. A comprehensive phenotyping approach in mice carrying a human Foxp2 version revealed that while the physiological functions of the lung and all other organs are normal, this gene specifically affects the morphology and function of brain circuits involved in speech and language abilities [39].

CIN85 is involved in receptor trafficking and cytoskeletal dynamics, and plays a vital role specifically in D2 dopamine receptor endocytosis [40]. Interestingly, loss-of-function mutations in this gene do not only lead to hyperactivity, but also to deviations in several metabolic parameters, and D2 receptor function has been shown to be involved in the regulation of appetite, energy intake and obesity.

Missing-in-metastasis (MIM/MTSS1) is a tissue-specific regulator of actin and plasma membrane dynamics, whose altered expression levels have been linked to metastatic behavior of various cancers using in vitro assays. The in vivo analysis of MIM null mice displayed a severe urinary concentration defect. These functional alterations correlated with the compromised integrity of kidney epithelia intercellular junctions. These data demonstrated a new function of MIM that modulates the actin cytoskeleton/plasma membrane interactions to promote the maintenance of cell-cell contacts in kidney epithelia [41].

2.5.2 Models for Human Diseases

Pitx3 has recently been shown in a GWAS study to be associated with sporadic forms of Parkinson's disease (PD), which account for approximately 90% of PD cases. Analyzing the mutant mouse line Eyeless which carries a mutation in the *Pitx3* gene [42] revealed that the Eyeless mutants do not only recapitulate the motor impairments and the dopaminergic dysfunctions typical for PD, but that they also show alterations in nociception, which opened up new avenues for further investigations of the underlying mechanisms.

Two ENU induced mouse models of human renal diseases have been characterized: The *Umod* (A227T*)* mouse line as model of uromodulin storage disease, as well as mouse line *Slc12a1*(I299F) as a model of type I Bartter syndrome [43, 44]. Uromodulin storage disease is a dominantly inherited condition associated with progressive renal failure and hyperuricemia in humans, ultimately resulting in end stage renal disease. The mouse models characterized share most of the clinical symptoms concerning kidney function with affected human patients. The systemic phenotypic characterization of the mutant mouse revealed additional effects on energy and bone metabolism.

The mouse model of human type I Bartter syndrome *Slc12a1* (I299F) is the first viable and fertile mouse model described, displaying most symptoms seen in human patients suffering from antenatal Bartter syndrome. In contrast to most published human cases, which are homozygous carriers of *SLC12A1* mutations suffering from polyuria already during gestation leading to prenatal polyhydramnios, the mouse

model did not show pathological changes during gestation or suckling period. Nevertheless, this mouse line shares the typical pathophysiology and might be a valuable model to test new therapeutic strategies for salt loss tubulopathies.

2.5.3 Gene-Environment Interactions

Mice deficient for the *Eps8* gene display reduced body weight, partial resistance to age- or diet-induced obesity, overall improved metabolic status and live longer than wild-type mice. It was possible to identify the mechanisms behind this phenotype as lower body weight was not caused by reduced food intake but it was correlated with decreased intestinal nutrient absorption due to reduced intestinal microvilli length. An analysis of the subcellular localization of Eps8 in intestinal cells suggested that Eps8 is localized in intestinal microvilli. Since microvilli serve to augment the absorptive surface of the intestine, their reduction in *Eps8*-KO mice explained the absorption defect and the calorie restriction phenotype observed in these animals [45].

2.5.4 Systemic Phenotyping Used for Target Validation and Modeling Therapeutic Intervention

Oxidative stress is a candidate mechanism in ischemic stroke and NADPH oxidase type 4 (NOX4) was identified as a major source of oxidative stress and as a putative therapeutic target. *Nox4* knockout mice were analyzed in a comprehensive phenotype screen under standard conditions without detecting abnormalities that would suggest potential side effects of a drug decreasing NOX4 function. However, after both transient and permanent cerebral ischemia these mice were largely protected from oxidative stress, blood–brain-barrier leakage, and neuronal apoptosis [46].

Acknowledgements We would like to thank Reinhard Seeliger, Nicole Boche, Sabrina Bothur, Anna Dewert, Jan Einicke, Ralf Fischer, Birgit Frankenberger, Sandra Geißler, Michaela Grandl, Brigitte Herrmann, Christine Hollauer, Elfi Holupirek, Maria Kugler, Jacqueline Müller, Elenore Samson, Florian Schleicher, Daniela Schmidt, Waldemar Schneider, Ann-Elisabeth Schwarz, Bettina Sperling, Waldtraud Stettinger, Lucie Thurmann, Susanne Wittich, Anja Wohlbier, and Claudia Zeller as well as the GMC animal caretaker team Manuela Huber, Boris Schön, Heidi Marr, Annica Miedl, Tina Reichelt, Michael Gerstlauer, Renate Huber, and Horst Wenig for expert technical help. This work has been funded by the German Federal Ministry of Education and Research to the German Center for Diabetes Research (DZD e.V.) and to the GMC (NGFNplus grant No. 01GS0850, 01GS0851, 01GS0852, 01GS0853, 01GS0854, GS0868, 01GS0869) as well as by an EU grant (EUMODIC, LSHG-2006-037188, German Mouse Clinic).

References

1. Silver LM (1995) Mouse genetics. Oxford University Press, New York
2. Okazaki Y, Furuno M, Kasukawa T et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420(6915):563–573, 5

3. Dermitzakis ET, Reymond A, Lyle R et al (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature 420:578–582
4. Wade CM, Kulbokas EJ 3rd, Kirby AW et al (2002) The mosaic structure of variation in the laboratory mouse genome. Nature 420(6915):574–578
5. Reymond A, Marigo V, Yaylaoglu MB et al (2002) Human chromosome 21 gene expression atlas in the mouse. Nature 420:582–586
6. Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562, 5
7. Ehling UH (1966) Dominant mutations affecting the skeleton in offspring of x-irradiated male mice. Genetics 54:1381–1389
8. Favor J, Sund M, Neuhäuser-Klaus A, Ehling UH (1990) A dose-response analysis of ethylnitrosourea-induced recessive specific-locus mutations in treated spermatogonia of the mouse. Mutat Res 231:47–54
9. Wilson L, Ching YH, Farias M et al (2005) Random mutagenesis of proximal mouse chromosome 5 uncovers predominantly embryonic lethal mutations. Genome Res 15:1095–1105
10. Kile BT, Hentges KE, Clark AT et al (2003) Functional genetic analysis of mouse chromosome 11. Nature 425:81–86
11. Srivastava AK, Mohan S, Wergedal JE, Baylink DJ (2003) A genomewide screening of N-ethyl-N-nitrosourea-mutagenized mice for musculoskeletal phenotypes. Bone 33:179–191
12. Takahasi KR, Sakuraba Y, Gondo Y (2007) Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis. BMC Mol Biol 8:52
13. Friedel RH, Seisenberger C, Kaloff C, Wurst W (2007) EUCOMM–the European conditional mouse mutagenesis program. Brief Funct Genomic Proteomic 6:180–185
14. Guan C, Ye C, Yang X, Gao J (2010) A review of current large-scale mouse knockout efforts. Genesis 48:73–85
15. Ringwald M, Iyer V, Mason JC et al (2011) The IKMC web portal: a central point of entry to data and resources from the international knockout mouse consortium. Nucleic Acids Res 39:D849–D855, Database issue
16. Abbott A (2009) The check-up. Nature 460:947–948
17. Gailus-Durner V, Fuchs H, Becker L et al (2005) Introducing the German mouse clinic: open access platform for standardized phenotyping. Nat Methods 2:403–404
18. Fuchs H, Gailus-Durner V, Adler T et al (2009) The German mouse clinic: a platform for systemic phenotype analysis of mouse models. Curr Pharm Biotechnol 10:236–243
19. Brown SD, Chambon P, de Angelis MH, Consortium E (2005) EMPReSS: standardized phenotype screens for func-tional annotation of the mouse genome. Nat Genet 37:1155
20. Abbott A (2010) Mouse project to find each gene's role. Nature 465:410
21. Hagn M, Marschall S, Hrabě de Angelis M (2007) EMMA – the European mouse mutant archive. Brief Funct Genomic Proteomic 6:186–192
22. Wilkinson P, Sengerova J, Matteoni R et al (2010) EMMA – mouse mutant resources for the international scientific community. Nucleic Acids Res 38:D570–D576, Database issue
23. Justice MJ, Noveroske JK, Weber JS, Zheng B, Bradley A (1999) Mouse ENU mutagenesis. Hum Mol Genet 8:1955–1963
24. Noveroske JK, Weber JS, Justice MJ (2000) The mutagenic action of N-ethyl-N-nitrosourea in the mouse. Mamm Genome 11:478–483
25. Hecht NB (1986) Regulation of gene expression during mammalian spermatogenesis. In: Rossant J, Pedersen RA (eds) Experimental approaches to mammalian embryonic development. Cambridge University Press, New York
26. Augustin M, Sedlmeier R, Peters T et al (2005) Efficient and fast targeted production of murine models based on ENU mutagenesis. Mamm Genome 16:405–413
27. Rubio-Aliaga I, Soewarto D, Wagner S et al (2007) A genetic screen for modifiers of the delta1-dependent notch signaling function in the mouse. Genetics 175:1451–1463
28. Gailus-Durner V, Fuchs H, Adler T et al (2009) Systemic first-line phenotyping. Methods Mol Biol 530:463–509
29. Fuchs H, Gailus-Durner V, Adler T et al (2011) Mouse phenotyping. Methods 53:120–135

30. Rogers DC, Fisher EM, Brown SD et al (1997) Behavioral and functional analysis of mouse phenotype: SHIRPA, a proposed protocol for comprehensive phenotype assessment. Mamm Genome 8:711–713
31. Schneider I, Tirsch WS, Faus-Kessler T et al (2006) Systematic, standardized and comprehensive neurological phenotyping of inbred mice strains in the German mouse clinic. J Neurosci Methods 157:82–90
32. Neschen S, Morino K, Hammond LE et al (2005) Prevention of hepatic steatosis and hepatic insulin resistance in mitochondrial acyl-CoA: glycerol-sn-3-phosphate acyltransferase 1 knockout mice. Cell Metab 2:55–65
33. Haller F, Prehn C, Adamski J (2010) Quantification of steroids in human and mouse plasma using online solid phase extraction coupled to liquid chromatography tandem mass spectrometry. Nat Protoc. doi:10.1038/nprot.2010.22
34. Illig T, Gieger C, Zhai G et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42:137–141
35. Beckers J, Wurst W, Hrabě de Angelis M (2009) Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling. Nat Rev Genet 10:371–380
36. Maier H, Lengger C, Simic B et al (2008) MausDB: an open source application for phenotype data and mouse colony management in large-scale mouse phenotyping projects. BMC Bioinformatics 26:169
37. Morgan H, Beck T, Blake A et al (2010) EuroPhenome: a repository for high-throughput mouse phenotyping data. Nucleic Acids Res 38:D577–D585, Database issue
38. Horsch M, Schädler S, Gailus-Durner V et al (2008) Systematic gene expression profiling of mouse model series reveals coexpressed genes. Proteomics 8:1248–1256
39. Enard W, Gehre S, Hammerschmidt K et al (2009) A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell 137:961–971
40. Shimokawa N, Haglund K, Hölter SM et al (2010) CIN85 regulates dopamine receptor endocytosis and governs behaviour in mice. EMBO J 29:2421–2432
41. Saarinkangas J, Mattila PK, Varjosalo M et al (2011) Missing-in metastasis MIM/MTSS1 promotes actin assembly at intercellular junctions and is required for integrity of kidney epithelia J. Cell Science 124:1245–1255
42. Rosemann M, Ivashkevich A, Favor J et al (2010) Microphthalmia, parkinsonism, and enhanced nociception in Pitx3 (416insG) mice. Mamm Genome 21:13–27
43. Kemter E, Rathkolb B, Rozman J et al (2009) Novel missense mutation of uromodulin in mice causes renal dysfunction with alterations in urea handling, energy, and bone metabolism. Am J Physiol Renal Physiol 297:F1391–F1398
44. Kemter E, Rathkolb B, Bankir L et al (2010) Mutation of the Na+-K+-2Cl- cotransporter NKCC2 in mice is associated with severe polyuria and a urea-selective concentrating defect without hyperreninemia. Am J Physiol Renal Physiol 298:F1405–F1415
45. Tocchetti A, Soppo CB, Zani F et al (2010) Loss of the actin remodeler Eps8 causes intestinal defects and improved meta-bolic status in mice. PLoS One 5:e9468
46. Kleinschnitz C, Grund H, Wingler K et al (2010) Post-stroke inhibition of induced NADPH oxidase type 4 prevents oxidative stress and neurodegeneration. PLoS Biol 8:e1000479

Chapter 8
Metabolomics in Animal Breeding

Christa Kühn

1 Application of Metabolomics in Livestock Species

Metabolomics in its close definition is a rather young field in farm animal production. Initially, metabolomic analyses in farm animals had been initiated for many non-genetic applications e.g., control of drug abuse, control of embryo and oocyte quality in reproductive processes or for detection of product origin of food, whereas genetic variability essentially has been ignored in these fields. Only recently, the fields "Physiological Genomics/Genetics" and "Refined phenotypic description of animal models" have emerged, that fit into the current concept entitled "Genetics meets Metabolomics: from Experiment to Systems Biology". Up to now, the non-genetic application fields however, still comprise the majority of attempts applying metabolomic technologies in animal breeding comprising:

- Detection of drug abuse/toxicology:
- Product control/product processing
- Reproduction physiology – assessment of oocyte, sperm or embryo quality
- Nutritional physiology:
- Biomarker for early/easy disease detection
- Physiological genomics/Genetics
- Refined phenotypic description of animal models

C. Kühn (✉)
Department of Molecular Biology, Leibniz Institute for Farm
Animal Biology (FBN), Wilhelm-Stahl-Allee 2, Dummerstorf 18196, Germany
e-mail: kuehn@fbn-dummerstorf.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_8,

2 Non-genetic Applications of Metabolomics in Animal Production

2.1 Detection of Drug Abuse/Toxicology

Major areas of drug abuse in animals frequently addressed by metabolomic analytical methods are sports competitions (e.g., horse racing, show jumping, dressage) and meat production in livestock. The purpose of these attempts is focused on two aspects. On the one hand, any procedure to blur the true physical ability of a potential selection candidate by drug abuse will impede progress in animal breeding, because truly superior individuals with the best genetic potential cannot be identified. On the other hand, residuals of many banned drugs pose a severe threat to the consumers' health when contained in animal products. Thus, detection of drug abuse has always been an important issue in livestock production. Untargeted metabolomic approaches have been performed [1] in order to identify metabolic pattern associated with abuse of recombinant growth hormone. These untargeted metabolic profiles serve as surrogate biomarkers replacing specific metabolic responses that may be animal dependent regarding threshold or immunological response. The investigation of anabolic steroid administration in cattle by means of metabolomic fingerprints had been one of the pioneering fields of metabolomics in livestock [2]. Since then, many other studies on the detection of steroid intake followed extending even to fish [3]. The problem that metabolites of natural steroids are sometimes unknown is insignificant for the untargeted metabolomics approaches. Instead, detailed, accurate analyses of untargeted metabolomic fingerprints even provided indication on the chemical nature of the respective metabolites [4].

2.2 Product Control

The origin of livestock products (e.g., milk, meat) is becoming increasingly important for animal production. Especially for products of protected origin (e.g. Roquefort cheese or Parma ham) and products from organic farming, the origin of the product is a major determinant how the product is valued by the consumer. Additionally, for reasons of consumer's food safety and for livestock protection against epidemic diseases, importation of livestock products from foreign countries is often strictly regulated even under the premise of the global Free Trade Agreement (FTA). The implementation of these regulations and the control of the origin of livestock products is a severe challenge. For meat, metabolomic studies using NMR spectroscopy or GC/MS and LC/MS/MS techniques with multivariate analyses proved efficient for discriminating the origin of the product [5, 6]. The advantage of an untargeted NMR metabolomics approach is that it does not require prior input of hypotheses on the nature of the molecules discriminating the origin of the products. Frequently, this information is not available in advance and instead, is part of the output of respective

untargeted metabolomic analyses extended by a targeted metabolite profiling [5]. Other attempts have been made to discriminate between milk samples from organic and conventional farming by means of GC/MS and LC/MS/MS techniques with consecutive, secondary metabolite identification using established reference libraries [6]. In addition to their geographical or production chain origin, livestock products have to be tested for their identity to avoid illegal replacement of high-value raw material by non-approved by-products of similar origin and chemical composition. Surowiec et al. [7] exemplarily demonstrated that mechanically recovered meat, which is no consumable meat according to EC regulations, can be discriminated from true meat of the respective species by GC-MS and multivariate data analysis.

The safety of products from animals fed genetically modified plants or being genetically modified themselves is heavily discussed in Europe. Assessment of potential hazards to consumers' health requires an unbiased, comprehensive analysis of all product components. Alterations in the composition of respective products can be monitored by non-targeted metabolomic approaches, which do not require a priori knowledge about the metabolites affected, e.g. respective studies investigated whether consumption of milk from transgenic goats exerted beneficial or deleterious effects on serum metabolites in piglets [8].

Another aspect of animal product control is the monitoring of production processes, e.g. the cheese maturation or the post-slaughter processes encompassing meat recovery from the carcass or postmortem ageing. Post-mortem aging is pivotal for beef flavor and tenderness and related enzymatic activities like proteolysis or glycogenolysis. These processes manifest themselves in substantial changes of the intramuscular metabolites, which can be used for monitoring meat maturation [9]. A prolonged post-mortem aging of beef is known to result superior taste and tenderness compared to a shortened, however less costly meat maturation. Differences in the metabolomic profiles of muscle samples depending on the time of post-mortem aging have been revealed by NMR spectroscopy with consecutive metabolite identification [9] offering perspectives for an efficient, impartial product control. While single molecules relevant for this purpose had been determined previously, the respective analyses had required a substantial number of different methods for quantification, which can now be replaced by comprehensive metabolomic methods.

Critical points for use of metabolomic data in product control will be the detection of metabolic fingerprints or distinct metabolites that truly and exclusively result from the origin under consideration and do not reflect ambiguous effects of other environmental or genetic factors modulating the metabolome [10].

2.3 Reproduction Physiology: Assessment of Oocyte, Sperm or Embryo Quality

The reduced complexity of ovarian follicular fluids and of the culture medium for oocytes and embryos has enabled already early successful studies on the metabolome of the ovarian follicle and the metabolic capacity of the developing oocyte/embryo by

metabolomics means. These studies can be discriminated regarding two different objectives. In vivo approaches on follicular fluids focused on the identification of biomarkers for fertility, which is a key trait in any farm animal production scheme [11]. The other objective was to replace the mostly subjective morphological criteria of assessing oocyte and embryo quality in artificial reproduction techniques. In addition to serving as biomarkers for quality assessment, metabolomic data could also provide valuable information about the background of potentially impaired physiological processes in the in-vitro culture protocols [12]. Most metabolomic studies in reproduction physiology took a targeted approach by investigating specific metabolite subpopulations, e.g. fatty acids.

2.4 Nutritional Physiology

Discussion about greenhouse gas production has initiated a revival of research on microbial processes during ruminant digestion. The power of metabolomic approaches enabled the description of the multi-facetted rumen metabolism with an unprecedented comprehensiveness [13]. Consequently, differences between different ruminant diets regarding production of greenhouse gas precursors could be described applying NMR and GC-MS metabolome technology. Exemplarily, methylamine, a long-recognized precursor of methane, had been described elevated in high grain diets. Together with next-generation sequencing metagenomics technology, metabolomics in nutritional physiology will offer new prospects to elucidate structure and function of the rumen metabolism, which is discussed as a major contributor to greenhouse gas production

Nutritional physiology in farm animals also applies metabolomic approaches for comprehensively investigating the effects of nutritional programming, i.e. for monitoring epigenetic effects of the diet during early ontogenesis of the individual [14, 15].

2.5 Biomarker for Early/Easy Disease Detection

There are a substantial number of infectious and non-infectious diseases in farm animals, for which a sensitive and specific detection is difficult to obtain. Especially in cases of an unknown causal agent, a non-targeted metabolomics approach is attractive to establish profiles discriminating the healthy from the pathological condition. Respective approaches were taken for the detection of osteochondrosis in companion animals and horses [16] in agreement with similar attempts in humans and for the diagnosis of Bovine Spongiform Encephalopathy (BSE, mad cow disease) in cattle. [17] BSE was a prime example, because at the time of the outbreak of the disease, there was no conclusive evidence regarding the nature of the disease, but a very strong demand to reliably identify cattle affected by the disease prior to a clinical outbreak.

3 General Concepts of Selection in Animal Breeding: Rationale for Requirement of Reliable Genotype/Phenotype Predictors

Historically, animal breeders thrive to obtain reliable, cheap predictors of the genetic make-up of an individual already early in ontogenetic life to maximize selection response regarding target traits like performance, disease resistance or product quality. Especially, this is inevitable for phenotypes that are not expressed or measurable in the selection candidates themselves (e.g. milk production in bulls, egg production in roosters, all carcass traits). For those traits, alternative routes of assessing the genetic potential of a selection candidate have to be taken: either by looking at the phenotype of close relatives or by making use of biomarkers that serve as predictors for phenotypic traits. These markers may be causal DNA mutations or DNA variants closely associated in linkage disequilibrium with the causal mutation. Alternatively, also enzyme activities or metabolite concentrations can serve as biomarkers.

3.1 Major Determinants of Selection Response

The selection gain, i.e. the superiority of the offspring generation compared to the parental generation, per time interval is the key determinator of success in animal breeding. According to Rendel and Robertson [18], the selection gain in a population per year is determined by the selection differential, the accuracy of selection, the genetic variability in the test population and the generation interval.

A high selection intensity (i.e. a large selection differential: the relative superiority of individuals selected for producing the next generation compared to the entire generation in which selection is performed) requires a large number of potential parents to be tested for their genetic merit for the target trait and the subsequent frequent reproduction of these selected parents. If either only a minor proportion of the population can be screened for the respective phenotype (e.g. due to sex-limited trait expression like milk or egg production or due to high costs like for feed efficiency) or if the selected individuals show a poor reproduction, selection gain is severely shortened. A high accuracy of selection implies that the selection process is able to precisely identify the best individuals of a population regarding the genetic potential for the target trait. Most economically important traits in animal breeding are complex traits substantially influenced by environmental effects. Thus, detection of the most superior individuals for a specific trait is more or less impaired by non-genetic factors. The generation interval (the age of the parents when producing the next generation) is a major issue for selection gain per time interval especially in species with late maturity and for those traits that require a long period of observation (e.g. longevity).

According to these key determinators of selection gain, the precise knowledge of the genetic potential of a large number of potential parental individuals unbiased by

non-genetic effects already early in life is optimal for successful breeding programs. Starting in the 1950s, many concepts were developed and implemented for breeding schemes that included testing of relatives and progeny and application of quantitative genetic methods to merge information on phenotypes and genetic relationships within a population to calculate breeding values for members of the respective population. These breeding values that ideally should be unbiased by environmental effects can be calculated for each individual within a population. However, only for those members of the population with a large number of relatives/progeny, the reliability of those breeding values enables an authentic picture of the genetic potential of an individual. In spite of these limitations, at the population level, the calculation of breeding values based on performance records of the target individual (if available) and its relatives proved to be very successful resulting in the impressive increase of production in farm animal species. However, the idea still flourished to have additional, specific, reliable genetic predictors of the genetic potential of an individual. In addition to an increase in the reliability of the prediction of the genetic merit of an individual, such genetic predictors/markers would also enable to substantially reduce the time and cost consuming performance tests of progeny or other relatives of the target individuals.

3.2 Concepts for Detection of Genetic Predictors in Livestock Populations

Already in the early days of domestication, farmers unknowingly used genetic markers for improvement of their breeding stock: they selected for coat color and body shape pattern (e.g., coat color spotting, ear forms, head combs) of their lines, because they assumed that individuals with a specific shape would be more likely to share the superior properties of the respective line than individuals with an entirely different shape.

These coat color and body shape pattern were, however, of limited usefulness, especially for selection within breeds or lines with all individuals sharing the identical phenotype. Nevertheless, the idea to use some kind of marker as a predictor was very attractive, because it might be able to address all major determinants of selection gain: it should be cheap enabling screening of large numbers of the population and it should be expressed reliably and independent of sex, age or environment. In 1983, Beckman and Soller [19] published a pioneering paper about marker assisted selection implementing Restriction Fragment Length Polymorphisms (RFLPs), the first class of DNA markers available. Since then, the number of DNA markers available has increased substantially moving on from Variable Number of Tandem Repeat (VNTR), microsatellites, single nucleotide polymorphisms (SNPs) to copy number variation (CNVs). Currently, several million genetic polymorphisms are available for major farm animal species like horse, pig, cattle or chicken and even extending to aquaculture species like salmon [20, 21]. With these tools at hand resulting from

the major attempts of farm animal genome sequencing, entirely new concepts to identify useful genetic markers and to reveal their physiological background have become feasible.

In this respect, farm animal species offer substantial advantages compared to human or laboratory animal populations: long-term outcross selection lines with large phenotypic differences, multi-generation pedigrees, routine population-wide phenotyping, opportunities to standardize environmental conditions, targeted mating and a genome organization more similar to human than most laboratory animals.

Taking advantage of these conditions, a number of specific resource populations had been set up in order to identify markers closely linked to quantitative trait loci (QTL). QTL are those loci in the genome, which exhibit an effect on the genetic variability of a quantitative (complex) trait [22]. QTL are of particular interest in farm animal production, because the vast majority of economically important traits are quantitative and complex by nature.

4 Application of Metabolomics in Animal Breeding

4.1 History

Already decades ago, animal breeders endeavored to use available biochemistry kits to obtain early predictors of animal performance that might serve as biomarkers. The idea behind this was that the activity of key metabolic enzymes and the resulting metabolites being the direct readout of biochemical pathways should be more closely correlated to the underlying genetic make-up of an individual than its complex conventional phenotype, e.g., milk performance or growth [23, 24]. At that time, in animals only very few DNA markers were available. Thus, the main aim of these studies had been to obtain biomarkers for early recognition of the genetic potential of farm animals for the respective production performance. Many of these studies focused on enzymes, hormones and metabolites, which were known to be major contributors, regulators and/or end products of the energy metabolism. These attempts can be looked upon as the ancestors of a targeted metabolomic approach in animal breeding and currently, the respective detection methods still serve as a kind of gold standard for validating results from modern metabolomic analyses. However, the success of these first attempts was and is limited due to the restricted number of metabolites under investigation. Nowadays, there are an increasing number of appropriate polymorphic genetic loci within the target species that become available for genetic linkage or association analyses. As a result the focus of metabolomic studies in animal breeding shifted into two major directions: (1) Metabolomic profiles as refined metabotypes for a better phenotypic description of the animals and (2) merging metabolomics and genomics to reveal major genetic determinants of key physiological processes providing tools for improved animal selection and husbandry/nutrition.

4.2 Metabolomic Profiles as Refined Metabotypes for a Better Phenotypic Description

In animal breeding, metabolomic profiles might either serve as criterion useful for current selection or represent a refined indicator of historical selection. The metabolic profiles can objectively reveal a desired or undesired phenotype within a group of individuals homogeneous by cursory inspection. This concept includes targeted approaches to reveal specific genetically determined pattern of distinct, known metabolites as well as non-targeted approaches for identification of new biomarkers for specific phenotypes, e.g. disease resistance or metabolic status. This approach is an extension of the initial attempts for a refined phenotyping of farm animals beyond simple measurement of basic performance traits. Examples for metabolomics studies screening for better phenotypic predictors of complex physiological phenotypes were recently published for dairy cattle targeting at an improved description of the metabolic status post-partum [25, 26]. By nature, metabolic profiling for this purpose has a substantial focus on methodological aspects. Metabolomic approaches, which initially did not take genetically determined variation into consideration, also might be recruited for selection purposes extending the idea of refined performance testing of farm animals. An example of this idea is the monitoring of meat maturation (see above, Sect. 2.2), which initially was intended to reveal suboptimal meat aging conditions. However, keeping the environment constant the same technology might now as well be applied to select for a genetically determined, superior capacity of muscle processing and, consequently, an increased meat quality in the target population.

4.3 Merging Metabolomics and Genomics to Reveal Major Genetic Determinants of Key Physiological Processes

As indicated above, farm animal populations provide a very appropriate setting for the understanding the molecular background of genetically determined variation in complex polygenic traits. Long-term selection in outbred populations resulted in lines of individuals with remarkable phenotypic differences regarding many complex traits of general biological and medical interest like growth, lipid deposition or nutrient conversion. However, the physiological background of these differences, the shifts in the dynamics of the biological systems is only poorly understood. A better understanding of the metabolic pathways affected by differences in the genetic architecture of individuals could provide distinct biomarkers for deleterious alterations already in ontogenesis. Additionally, it improves the identification of causal genetic variants underlying the phenotypic variation representing the optimal markers for selection. Finally, a profound knowledge of the physiological background of genetically determined phenotypic variability provides a starting point for the future development of specific diets or treatments for an improved, sustainable

animal production and animal welfare (see Sect. 4.6). Although still at its infancy in livestock, examples for respective metabolomic approaches on selected farm animal populations exist, e.g. in chicken and cattle.

4.3.1 Chicken

Selection in chicken resulted in layer and broiler lines with remarkable performance of the metabolic systems: broiler chicken achieve a remarkable feed conversion rate (food: body mass gain) of less than 2:1 [27]. However, the underlying changes of the metabolic system driven by the genetic variation are only partly understood, which is also true for the accompanied deleterious collateral effects. Correspondingly, targeted metabolomics profiling as a means to provide a system-level understanding of the physiological background of the metabolic shifts was a straightforward approach in this situation [28]. Four broiler lines starting from two different base populations were selected for high or low growth rate, or for high and low body fat content, respectively. Selection resulted in extreme differences in growth rate and body composition, specifically fat deposition. The four different selection lines were compared in a targeted metabolomics approach distinctly focusing on lipidomics. At a key time point, at 5 weeks of age, which had been determined by previous transcriptomic analyses, plasma samples were analyzed by GC-MS after separating lipid classes by preparative HPLC. In addition to reporting the molar concentrations of the target molecules, subsequent attempts were undertaken to calculate the activity of key enzymes of lipid metabolism from the molar concentrations of precursor and product metabolite. Metabolomic data showed that the differences in adiposity between the selection lines had developed by different metabolic shifts during selection in the two different base populations. Whereas in one base population, the enhanced hepatic conversion of feed to very low density lipoproteins (VLDL) triglycerides and its peripheral uptake and storage seems to be increasing adiposity in the selection line for high body fat content, in the other base population the decreased adiposity of one selection line seemed to be result of the inability to utilize and store VLDL triglycerides. Thus, in spite of a concordant divergent adiposity generated by both selection experiments, metabolomic analysis suggested that different biological pathways are affected and that different genetic mechanisms should be causal to the differences.

4.3.2 Cattle

Another example merging animal metabolomics and genomics at a more fundamental stage is the analysis of a major locus modulating pre- and postnatal growth in cattle. Linkage and association studies in two divergent cattle populations had mapped this locus to cattle chromosome 6, which had been confirmed by several other powerful studies in this species. Comparative data from human and mice indicated that this locus presumably might be a genetic modulator of growth of mammalian

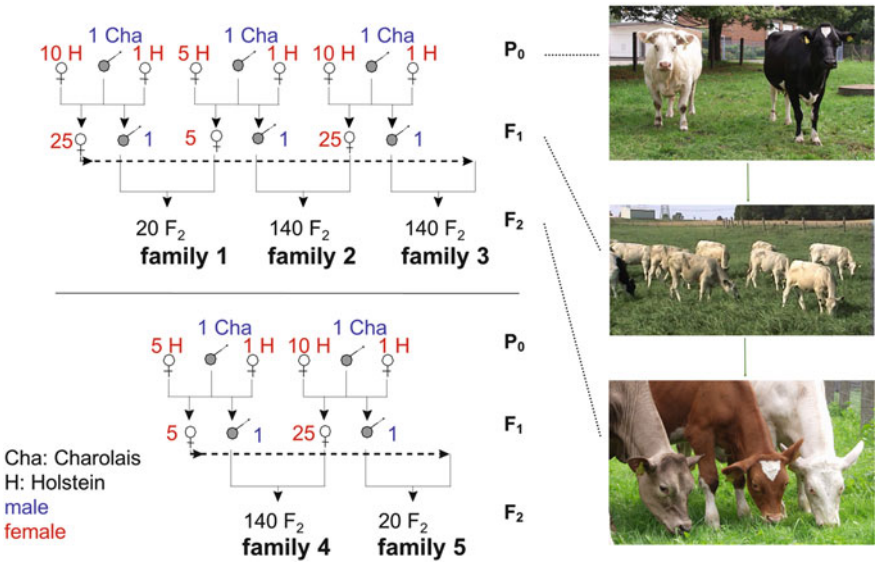


Fig. 8.1 Mating scheme SEGFAM Design of the SEGFAM resource population. The SEGFAM population was created in a F2 design by mating each of five sires from a major European beef breed, the Charolais, to dams from the most important dairy cattle breed worldwide, the Holsteins. Thus, the population comprises distinctly divergent metabolic types prone to nutrient accretion (beef) or nutrient secretion (dairy). Subsequently, the generated female F1 offspring of one Charolais sire were mated to a single F1 son from another Charolais sire in order to generate a F2 generation of half and fullsibs. All individuals in the F1 as well as in the F2 generation were born from multiple ovulation and embryo transfer to virgin Holstein heifers

growth in many species. However, the nature of this locus and potential physiological pathways affected remained unknown. Genetic analyses in an outcross F_2 cattle design (SEGFAM, Fig. 8.1) from two major European breeds identified a mutation in the non-structural maintenance of chromosome (SMC) condensing complex subunit G (NCAPG) gene to be the most likely background for the QTL [29]. A respective confirmation could be obtained in a historically and geographically distant Japanese cattle population [30–33]. The mutation affecting amino acid position 442 of the NCAPG protein encoded either the mutated allele methionine (442 M) or the ancient allele isoleucine (442I). The NCAPG protein belongs to the family of condensins. Some fundamental functions of the protein had been described in Drosophila and HeLa cell models: during mitosis, NCAPG is important for chromosome condensation and interacts with DNA methyltransferase DNMT3B. However, the distinct role of the NCAPG gene in mammalian physiology and specifically in regulation of body growth was largely unknown.

At this point, a metabolomic analysis of the F_2 resource population proved to be invaluable for the functional annotation of the NCAPG gene and for an initial understanding of the physiological pathways behind divergent growth. In addition to the very standardized environmental conditions regarding housing and feeding and

sample collection, the F_2 resource population also provided large half- and full-sib families generated by multiple ovulation and embryonic transfer. Use of embryonic transfer excluded any systematic maternal genetic effects on prenatal development of the F_2 individuals due to divergent intrauterine milieu. A close network of phenotypic recording from birth to a detailed carcass dissection after slaughter enabled additional links between metabolomics data and phenotype. In addition to the general advantages of a farm animal resource population compared to studies in human population, the SEGFAM population had a further valuable characteristic for the investigation of the physiological background of the NCAPG I442M mutation: the population segregated for a mutation (Q204X) in a second gene very well established as a major modulator of mammalian growth: the GDF8 gene encoding myostatin. Thus, the effects of both mutations, NCAPG I442M and GDF8 Q204X, could be comparatively investigated on an identical genetic background.

Close monitoring the body weight gain during ontogenesis had indicated that congruently both loci primarily acted during the onset of puberty, at about 8 months of age, which is known to be a key period of mammalian growth. Consequently, this represented the most interesting interval for a metabolome analysis using plasma samples that were obtained after a 12 h fasting period. Other previous data had suggested that the GDF8 Q204X and the NCAPG I442M mutation both affected body weight gain as well as lipid deposition. Thus, the focus of the subsequent targeted metabolomic analysis using electrospray ionization tandem mass spectrometry (ESI/MS/MS) with the Biocrates AbsoluteIDQ targeted metabolomics technology analogously to a method previously applied to human serum lipidomics [34] was directed to indicators of lipid metabolism (acylcarnitines, glycerophosphatidylcholines, and sphingomyelins). These groups comprised 63% of all identified metabolites. Other metabolites measured were amino acids, sugars and biogenic amines.

Although phenotypically, the effects of the NCAGP I442M and the GDF8 Q204X mutation on dimension and key time points of divergent body weight gain were very similar at a sketchy view, close inspection revealed substantial differences in detailed phenotype. Whereas the mutated 442 M NCAPG allele exerted effects directed on a proportional increase in growth of all body compartments, the effects of the mutated 204X GDF8 allele resulted in a disproportional growth, primarily of the muscle tissue. These phenotypic differences were reflected by completely different pattern of effects on metabolic profiles for both mutations (Figs. 8.2 and 8.3). For the NCAPG I442M mutation, specific associations with metabolites from the arginine metabolism were detected, whereas not even a respective tendency was observed for the GDF8 Q204X mutation. This proves the locus specificity of the association between NCAPG I442M mutation and metabolites. Upon supplementation, arginine has well-known effects on alleviation of the intrauterine growth restriction in mammals. Arginine is also precursor of NO, which plays a multifaceted, extremely important role in energy metabolism and vascularization. Interestingly, arginine is also frequently proposed as a food additive to increase muscle growth and to decrease lipid deposition, e.g. in human athletes. These effects exactly match the effects of the NCAPG I442M mutation in cattle. Subsequent metabolic analyses at time points earlier and later than 8 months of age did not show

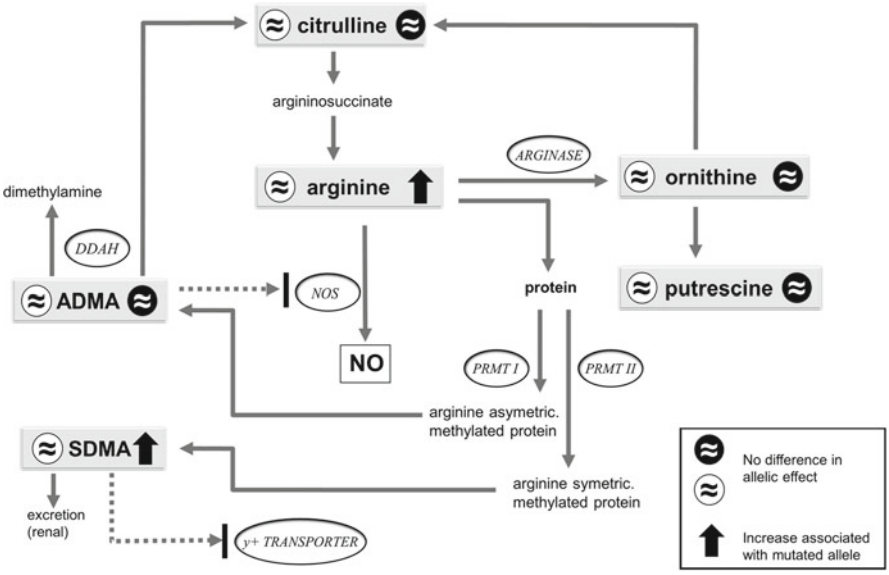


Fig. 8.2 Association arginine metabolism. Comparison of the associations of the bovine NCAPG I442M and the GDF8 Q204X mutations with plasma metabolites from the arginine metabolism. Plasma samples were obtained from male individuals of a F_2 resource population originating from Charolais and the Holstein breed. All metabolites in *grey boxes* were quantified. Signals to the *left* of the quantified metabolites indicate the effect of the mutated GDF8 204X allele compared to the wild type 204Q allele, signals to the *right* of the quantified metabolites indicate the effect of the mutated NCAPG 442 M allele compared to the wild type 442I allele. *Solid arrows* indicate metabolic pathways; *dashed arrows* indicate a regulatory function of the metabolite on the indicated enzyme. *ADMA* asymmetrically methylated dimethyl arginine, *SDMA* symmetrically methylated dimethyl arginine, *NO* nitrogen oxide, *NOS* NO synthase, *PRMT I* protein arginine methyltransferase type I, *PRMT II* protein arginine methyltransferase type II, *DDAH* dimethylarginine dimethylaminohydrolase

any correlation of the respective metabolites at 3, 8, and 14 months of age in the SEGFAM population. This is another indication on the relevance sampling time for the interpretation of metabolomic profiles.

The association of the NCAPG I442M mutation with pre- and post-natal growth and the comprehensive metabolomics characterization of the allelic effects provide an example of a suitable genetic predictor of performance serving as marker for improved of cattle breeding.

For the GDF8 Q204X mutation, the most prominent association was observed with the free plasma carnitine. Similarly to the situation for NCAPG/arginine, this association pattern was strictly locus-specific to GDF8 Q204X and not due a general growth-related phenomenon as is demonstrated by a complete lack of the respective association for the NCAPG I442M mutation. Carnitine is important for transportation of fatty acids into the mitochondrion for beta-oxidation and is controversially discussed as a feed additive for growth promotion and decrease of lipid deposition. The fact that genetic variability substantially affects plasma carnitine levels might explain the controversy in the literature regarding to outcome of carnitine

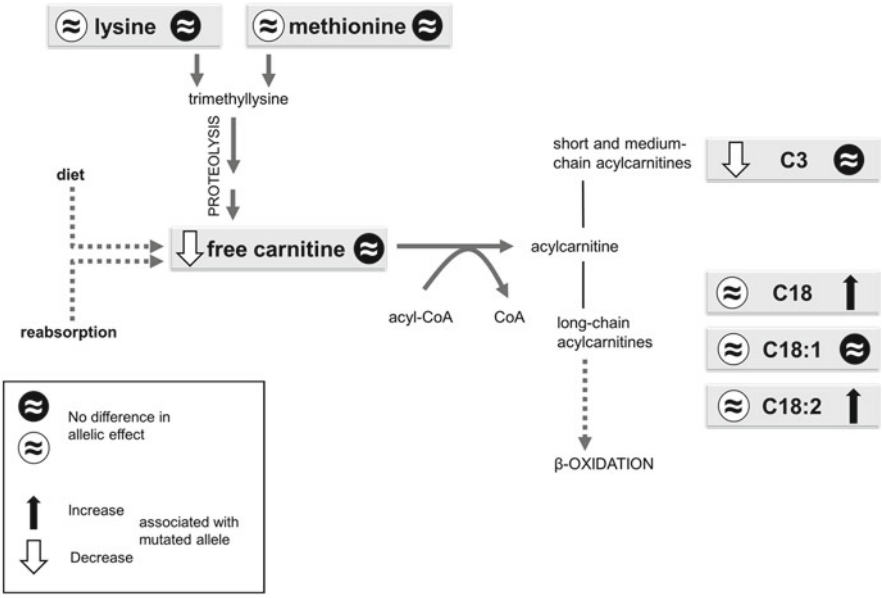


Fig. 8.3 Association carnitine metabolism. Comparison of the associations of the bovine NCAPG I442M and the GDF8 Q204X mutations with plasma metabolites from the carnitine metabolism. Plasma samples were obtained from male individuals of a F_2 resource population originating from Charolais and the Holstein breed. All metabolites in *grey boxes* were quantified. Signals to the *left* of the quantified metabolites indicate the effect of the mutated GDF8 204X allele compared to the wild type 204Q allele, signals to the *right* of the quantified metabolites indicate the effect of the mutated NCAPG 442 M allele compared to the wild type 442I allele. *Solid arrows* indicate metabolic pathways, *dashed arrows* indicate a flow of metabolites from or to specific physiological functions. *C3* Propionylcarnitine, *C18* Stearoylcarnitine, *C18:1* Oleoylcarnitine, *C18:2* Linoleylcarnitine

supplementation into diets. Respective studies investigating the potential dietary effects of feed additives usually ignore a genetic variability among probands (see also Sect. 4.6).

These results in cattle merging a targeted metabolomic profiling and genetic polymorphisms provided the first link ever regarding the role of NCAPG in mammalian physiology and demonstrate the huge potential of the respective approach not only for farm animal breeding but also for using farm animal data for elucidating general phenomena in mammals.

4.4 Metabolomics for Improved Phenotypic Description of Animal Models for Human Diseases

Increasingly, companion and farm animals serve as important animal models for human diseases. In this context, a refined analysis of the phenotype is required to accurately compare human disease and animal models and also to better characterize the pathobiology of the disease. Metabolomic analyses represent one tool to

achieve this refined phenotyping. An example for the use of metabolomics for improved phenotypic description of animal genetic models for human diseases is the neuronal ceroid lipofuscinosis (Batten disease), a group of inherited neurodegenerative diseases in human and animals comprising several genetically distinct members. The specific inherited defect in the South Hampshire sheep breed is caused by a mutation in the ovine CLN6 gene and shows a very similar progression of the pathological symptoms to the effects of a respective mutation in the human CLN6 gene. NMR and GC-MS analyses and subsequent multivariate pattern recognition tools were applied on cerebrospinal fluid and brain tissue samples of CLN6 mutant sheep at different ages to identify biochemical abnormalities during the time course of disease development in sheep [35]. These strategies are directed towards developing biomarkers for early diagnostics and towards a better understanding of the pathophysiological consequences of the specific disease. The consideration of different time points highlights an important feature of metabolomics profiles: its temporal dependency. Besides sheep, especially pig and dog are frequently monitored with metabolomics tools to gain insight into the pathobiology of human diseases either by inducing pathological situations (pigs frequently used in cardiovascular studies [36]) or by taking advantage of existing genetic models (e.g. metabolic defects in dogs [37]).

4.5 Genomic Predictions Based to Metabolic Profiles (Representing a Direct Read-Out of Biological Processes)

Currently, the calculation of genomic breeding values and its application in genomic selection schemes revolutionizes animal breeding. Genomic breeding values are calculated from genotype information of a large number of SNP (>> 10,000) and a previously established estimation algorithm [38]. These algorithms had been derived and tested in large training and validation populations [39]. In this context, metabolomic data established in a well-characterized part of the population can represent a very refined phenotype with a close link to tits physiological background. The combination of specific phenotyping with genome-wide marker data could then enable an improved picture of the genetic architecture of complex traits. Increasingly, there is a growing interest in the livestock-derived concept of genomic evaluation also for improved prediction of genetic predisposition for complex traits like obesity/diabetes in humans [40].

4.6 Metabotype-Guided Animal Nutrition and Husbandry

The revelation of substantial endogenous differences in important dietary metabolites due to genetic variation (see Sect. 4.3.2) indicated that the previous concepts evaluating the effects of different food additives will benefit from accounting for individual animal effects. It had been recently demonstrated that individuals with specific genotypes of a mutation of the leptin gene in cattle responded differently to

a Zilpaterol hydrochloride supplementation in the diet regarding intramuscular fat deposition and carcass weight [41]. Cattle with an endogenously elevated carnitine or arginine levels (Sect. 4.3.2) might respond substantially different to a carnitine or arginine supplementation in the diet compared to animals with low endogenous levels. Knowing about the key time points and the respective metabolic pattern associated with increased growth enables a dosed, but appropriate supply of dietary resources in order to optimally navigate farm animals through the respective production scheme to the benefit of both, feed efficiency and animal welfare.

Pioneering experimental studies directed on nutrigenomics have been performed in dogs. The canine species is outstanding regarding the intra-species variability of phenotypes across all mammalian species. However, apart from inherited metabolic defects the more subtle variation of the canine metabolome is just now being considered. Breed specific predisposition to chronic ailments like kidney/bladder stones or sensitive bowl problems are well-known. But the questions what physiological mechanisms are behind these predispositions or if/how a genotype x diet interaction contributes to the ailments are still unanswered. One of the driving forces for respective metabolomic studies with a classical nutrigenomic objective was to obtain an increased knowledge of the dietary demands of the divergent dog breeds. The respective data are a prerequisite for the tailored design of appropriate diets for the different canine metabolic phenotypes.

Targeted and non-targeted metabolomics studies using NMR and FIE-MS plus GC-MS techniques on urine samples were undertaken in order to screen for subclinical, potentially breed-specific refined metabolic phenotypes [33, 42]. The studies monitored groups of dogs from two distinct breeds on a standardized diet as well as dogs from a number of different breeds fed unspecified diets in divergent private housing conditions. Although there is a well-known large variation of many metabolites regarding their urine concentration, the study uncovered a number of breed-specific metabolic fingerprint characteristics. An example was the strong discriminating signals from a large number of phenolic molecules potentially originating from metabolism of cinnamates and flavonoids and their gut flora derivate in the diet. Dietary phenols as flavonoids are frequently discussed as food additives with potential beneficial effects in human diets. A differential utilization of these molecules or their gut flora derivate might modulate the effect of respective feed additives. Interestingly, Beagle dogs, which are a very popular animal model in many clinical studies, displayed a metabolomics pattern distinctively different from other dog breeds in these analyses.

References

1. Kieken F, Pinel G, Antignac JP et al (2011) Generation and processing of urinary and plasmatic metabolomic fingerprints to reveal an illegal administration of recombinant equine growth hormone from LC-HRMS measurements. Metabolomics 7:84–93
2. Dumas ME, Debrauwer L, Beyet L, Lesage D, André F, Paris A, Tabet JC (2002) Analyzing the physiological signature of anabolic steroids in cattle urine using pyrolysis/metastable atom bombardment mass spectrometry and pattern recognition. Anal Chem 74:5393–5404

3. Samuelsson LM, Forlin L, Karlsson G, Adolfsson-Erici M, Larsson DGJ (2006) Using NMR metabolomics to identify responses of an environmental estrogen in blood plasma of fish. Aquat Toxicol 78:341–349
4. Rijk JCW, Lommen A, Essers ML et al (2009) Metabolomics approach to anabolic steroid urine profiling of bovines treated with prohormones. Anal Chem 81:6879–6888
5. Jung Y, Lee J, Kwon J, Lee KS, Ryu DH, Hwang GS (2010) Discrimination of the geographical origin of beef by H-1 NMR-based metabolomics. J Agric Food Chem 58:10458–10466
6. Boudonck KJ, Mitchell MW, Wulff J, Ryals JA (2009) Characterization of the biochemical variability of bovine milk using metabolomics. Metabolomics 5:375–386
7. Surowiec I, Fraser PD, Patel R, Halket J, Bramley PM (2011) Metabolomic approach for the detection of mechanically recovered meat in food products. Food Chem 125:1468–1475
8. Brundige DR, Maga EA, Klasing KC, Murray JD (2010) Consumption of pasteurized human lysozyme transgenic goats' milk alters serum metabolite profile in young pigs. Transgenic Res 19:563–574
9. Graham SF, Kennedy T, Chevallier O et al (2010) The application of NMR to study changes in polar metabolite concentrations in beef longissimus dorsi stored for different periods post mortem. Metabolomics 6:395–404
10. Moorby JM, Fraser MD, Parveen I, Lee MRF, Wold JP (2010) Comparison of 2 high-throughput spectral techniques to predict differences in diet composition of grazing sheep and cattle. J Anim Sci 88:1905–1913
11. Bender K, Walsh S, Evans ACO, Fair T, Brennan L (2010) Metabolite concentrations in follicular fluid may explain differences in fertility between heifers and lactating cows. Reproduction 139:1047–1055
12. Singh R, Sinclair KD (2007) Metabolomics: approaches to assessing oocyte and embryo quality. Theriogenology 68:S56–S62
13. Ametaj BN, Zebeli Q, Saleem F et al (2010) Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows. Metabolomics 6:583–594
14. Nissen PM, Nebel C, Oksbjerg N, Hanne C (2011) Metabolomics reveals relationship between plasma inositols and birth weight: possible markers for fetal programming of Type 2 diabetes. J Biomed Biotechnol. doi:10.1155/2011/378268
15. Nyberg NT, Nielsen MO, Jaroszewski JW (2010) Metabolic trajectories based on H-1 NMR spectra of urines from sheep exposed to nutritional challenges during prenatal and early postnatal life. Metabolomics 6:489–496
16. Lacitignola L, Fanizzi FP, Francioso E, Crovace A (2008) 1 H NMR investigation of normal and osteo-arthritic synovial fluid in the horse. Vet Comp Orthopaed 21:85–88
17. Lasch P, Schmitt J, Beekes M et al (2003) Antemortem identification of bovine spongiform encephalopathy from serum using infrared spectroscopy. Anal Chem 75:6673–6678
18. Rendel JM, Robertson A (1950) Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. J Genet 50:1–8
19. Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic-improvement – methodologies, mapping and costs. Theor Appl Genet 67:35–43
20. Matukumalli LK, Lawley CT, Schnabel RD et al (2009) Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4:4
21. Dominik S, Henshall JM, Kube PD et al (2010) Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (Salmo salar) breeding population. Aquaculture 308:S56–S61
22. Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. Theor Appl Genet 46:319–330
23. Flach D, Dzapo V, Wassmuth R (1984) Stoffwechselparameter als Indikatoren für die Leistungsveranlagung von Rindern. I. Beziehungen von Schilddrüsenhormonen, Insulin, Kreatin-Kinase, Glutamat-Dehydrogenase und Glutathion-Reduktase zu Kriterien der Milchleistung. Z Tierzüchtg Züchtungsbiol 101:188–197

24. Fuhrmann H, Eulitzmeder C, Geldermann H, Sallmann HP (1989) On the evaluation of hormone and metabolic profiles after infusion of glucose, propionate and butyrate in cattle. Berl Munch Tierarztl 102:188–193
25. Klein MS, Almstetter MF, Schlamberger G et al (2010) Nuclear magnetic resonance and mass spectrometry-based milk metabolomics in dairy cows during early and late lactation. J Dairy Sci 93:1539–1550
26. Dettmer K, Almstetter MF, Appel IJ et al (2010) Comparison of serum versus plasma collection in gas chromatography – mass spectrometry-based metabolomics. Electrophoresis 31: 2365–2373
27. Arthur JA, Albers AA (2003) Industrial perspective on problems and issues associated with poultry breeding. In: Muir WM, Aggrey SE (eds) Poultry genetics, breeding and biotechnology. CABI, Oxford
28. Walzem RL, Baillie RA, Wiest N et al (2007) Functional annotation of genomic data with metabolic inference. Poult Sci 86:1510–1522
29. Eberlein A, Takasuga A, Setoguchi K et al (2009) Dissection of genetic factors modulating fetal growth in cattle indicates a substantial role of the non-SMC condensin I complex, subunit G (NCAPG) gene. Genetics 183:951–964
30. Setoguchi K, Watanabe T, Weikard R et al (2011) The SNP c1326T>G in the non-SMC condensin I complex subunit G (NCAPG) gene encoding a p.Ile442Met variant is associated with an increase in body frame size at puberty in cattle. Anim Genet. doi:10.1111/j1365-2052.2011.02196.x
31. Weikard R, Altmaier E, Suhre K et al (2010) Metabolomic profiles indicate distinct physiological pathways affected by two loci with major divergent effect on Bos taurus growth and lipid deposition. Physiol Genomics 42A:79–88
32. Setoguchi K, Furuta M, Hirano T et al (2009) Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (CW-2) on chromosome 6 and the Ile-442-Met substitution in NCAPG as a positional candidate. BMC Genet 10:43
33. Beckmann M, Enot DP, Overy DP et al (2010) Metabolite fingerprinting of urine suggests breed-specific dietary metabolism differences in domestic dogs. Brit J Nutr 103:1127–1138
34. Gieger C, Geistlinger L, Altmaier E et al (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4:e1000282
35. Pears MR, Salek RM, Palmer DN et al (2007) Metabolomic investigation of CLN6 neuronal ceroid lipofuscinosis in affected South Hampshire sheep. J Neurosci Res 85:3494–3504
36. Solberg R, Loberg EM, Andresen JH et al (2010) Resuscitation of newborn piglets. Short-term influence of FiO(2) on matrix metalloproteinases, caspase-3 and BDNF. PLoS One 5(12): e14261
37. Whitfield PD, Noble PJM, Major H et al (2005) Metabolomics as a diagnostic tool for hepatology: validation in a naturally occurring canine model. Metabolomics 1:215–225
38. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
39. Vanraden PM, Van Tassell CP, Wiggans GR et al (2010) Invited review: reliability of genomic predictions for North American holstein bulls. J Dairy Sci 92:16–24
40. de Los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880–886
41. Engler M, Defoor P, Marquess L (2010) Impact of a leptin SNP and zilpaterol hydrochloride on growth and carcass characteristics of finishing steers. In: 32nd conference of the international society for animal genetics, Edinburgh
42. Viant MR, Ludwig C, Rhodes S, Günther UL, Allaway D (2007) Validation of a urine metabolome fingerprint in dog for phenotypic classification. Metabolomics 3:453–463

Chapter 9
Metabolomics Applications in Human Nutrition

Hannelore Daniel and Manuela Sailer

Human metabolism is a continuum. It shifts constantly between anabolic conditions after food intake and catabolic states between meals or during extended starvation periods. At all times there is need of a constant supply of nutrients and metabolites for ATP production and of building blocks for the continuous remodeling of cellular structures. However, the sources of fuels used to maintain metabolic functions are variable (carbohydrates versus lipids versus proteins) depending on frequency of eating and fasting and the quantity and quality of food intake. The profile of the metabolites in any biological sample obtained, taken at any time is a snapshot of an ever-changing "integrated metabolome". Human plasma, urine or breath metabolomes contain not only endogenously produced metabolites but also the nutrients and metabolites provided by the diet and, in addition, metabolites derived from the microbiota hosted in the human large intestine which are partly absorbed and appear later in blood and urine (Fig. 9.1).

1 Metabolomics for Food Intake Assessment and Microbiome Effects

All food items consumed contain literally millions of metabolites with a huge concentration range. Whereas foods of animal origin match crossly in chemical composition the human metabolome, foods of plant origin contain in addition to the nutrients a large spectrum of compounds of the plant secondary metabolism. These compounds are produced by plants mainly for attraction or defense of other creatures. Substances to give color to leaves, flowers and fruits or to provide flavor and taste are

H. Daniel (✉) • M. Sailer, M.Sc.
Molecular Nutrition Unit, TUM, Gregor-Mendel Str. 2,
Freising 85354, Germany
e-mail: daniel@wzw.tum.de; manuela.sailer@wzw.tum.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_9,

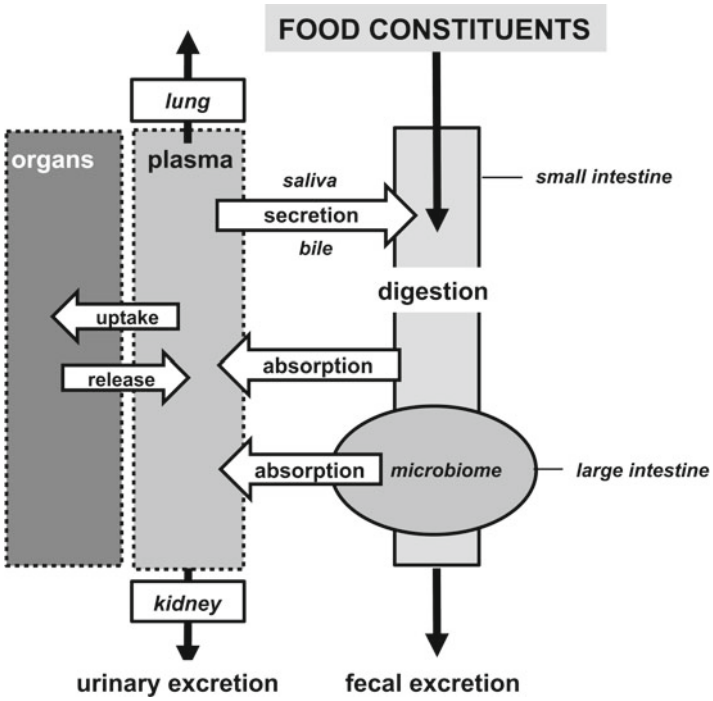


Fig. 9.1 Origin of metabolites in human biofluids. Schema to illustrate that human plasma or urine samples contain a surrogate metabolome of food derived nutrients and metabolites as well as metabolites produced in endogenous metabolism or by the microbiota in human intestine

made for attracting animals that help to disseminate pollen or seeds. Other chemicals are produced for defense against attacking invaders such as bacteria, fungi, insects or higher animals or even to cope with UV-stress.

The quantitatively most important components in the "plant food metabolome" are represented by complex carbohydrates, proteins and lipids as plant energy stores and in addition by vitamins, minerals and trace elements. Most of the constituents of the secondary plant metabolism are in the core structure polyphenols with multiple substituents and oxidation states and may comprise >10,000 different entities. The concentration and the pattern of these chemicals depend on plant genetics but also on environmental conditions and are considered to provide a typical fingerprint of the plant species that is served as a food. Most of these phenolic structures, however, cannot be degraded or cleaved in mammalian metabolism. They might be considered in the first place as xenobiotics, although a huge body of literature suggests that they may contribute to human health with a multitude of proposed functions. Like all other xenobiotics, almost all plant compounds undergo substantial modification in phase I and phase II metabolism. Already in the intestinal epithelial cell after absorption and during passage through liver, conjugation with glucuronic acid, sulfate or other groups takes place and numerous metabolites are formed that are secreted via bile back into the intestinal lumen and are found in circulation for excretion into urine.

For some of the plant secondary compounds only their metabolic conversion by bacteria in the large intestine enables absorption into the host's circulation. Depending on the composition of the microbiota, the so-called "microbiome", huge differences in the production of a given plant compound and in its appearance in blood can be found. Humans can easily be qualified as responders or non-responders to certain compounds by the ability of their individual microbiota for production of these metabolites that are later found in plasma and urine [1]. For example, the microbiota converts daidzein found in soy into equol or matairesinol and secoisolariciresinol found in linseed and other plants into enterolactone. When metabolomics applications in human samples cover these metabolites, a huge variability in the study group may arise since the capability of the individual microbiomes for production of these metabolites is different and the magnitude by which they are formed varies considerably amongst individuals.

Food metabolomics has approached the question of whether metabolite profiling in human plasma and urine can help to solve a key problem in human nutrition research and this is the assessment and quantification of food intake. So far, studies mostly rely on a food frequency questionnaire with reporting of food items eaten in quantity over a given time. These data are then used to calculate from food composition databases the mean dietary intake of distinct nutrients. Although these approaches are valuable in identifying food patterns that characterize a human based on its dietary habits, the approaches fail to deliver concise information on food consumption and the resulting nutrient intake. One reason for the lack of precision is the underreporting of food intake as a commonly observed problem when using food questionnaires in human cohorts. The second more important problem is the huge compositional variability of foods based on the large variety of natural and processed products with second line variability in composition depending on season and conditions during harvesting and processing. Metabolomics applied to plant and animal tissues for assessing composition and variability based on genetics or environmental cues is important to fill this gap as for edible parts, food composition data bases can be extended on basis of comprehensive metabolomic analysis.

Recent studies have demonstrated that self-reported dietary habits or patterns can indeed be linked to plasma or urinary metabolome data as derived via targeted LC-MS/MS or NMR-based analysis [2, 3]. Moreover, a variety of studies have succeeded in identifying individual compounds in plasma or urine that are derived from ingestion of individual food items such as onions, cacao or green tea in well controlled human supplementation studies [4–6]. A recent study for example identified urinary proline betaine as a marker of citrus fruit consumption in studies with volunteers. The findings were also validated in a larger cohort with around 2,000 participants that could be classified as citrus-consumers or noncitrus-consumers based on this novel compound [7]. Although proline betaine is not metabolized in humans and therefore has per se a food marker quality, its concentrations in fruits as well as processed fruit-juices is highly variable and therefore quantification in human urine will not necessarily allow a quantitative assessment of intake of the corresponding fruits or drinks, yet, it could allow to define consumers dietary patterns and habits. Other markers of dietary intake relate to meat consumption with creatine, carnitine and trimethylamine-N-oxide excretion in urine associated with a high voluntary

meat and fish intake [8]. Although it can be expected that food metabolomics will provide more such surrogate markers of food intake, it remains to be seen whether those will allow a quantitative intake assessment of individual food items.

2 Compartmentation of the Human Metabolome

Metabolomics, when applied in human studies faces the problem to sort out immediate effects of the diet and the contribution of the microbiota on the metabolite profiles to resolve the "endogenous host metabolome". Metabolism is determined by the dynamics of biosynthesis and degradation of proteins (turn-over) that function either as enzymes, receptors, transporters, channels, hormones and other signaling molecules or that provide structural elements for cells, organs or the skeleton. Between the proteins, there is a variable flow of metabolic intermediates, which serve as building blocks in synthesis of homo- or heteromeric macromolecules or which act as precursors in the synthesis of other low molecular intermediates and as substrates for ATP production. However, the metabolome in composition and concentrations is variable in space and time. Every organ, different cells within an organ and different intracellular compartments display different metabolite compositions and those are different from the metabolite profiles in human blood or other body fluids. Cell membranes with integral transporter proteins and a membrane potential difference separate and compartmentalize the intracellular and extracellular (i.e. plasma) metabolomes and concentration differences for the same metabolite in the cell and in the extracellular space can be as large as 2–3 orders of magnitude.

There are only a few studies that have simultaneously determined metabolite concentrations in tissues and blood in experimental animals and even less studies are available for the human condition. One of the few examples from studies in humans is based on the analysis of intracellular free amino acid levels from muscle biopsies in comparison to plasma levels [9]. These studies revealed that for example glutamine and taurine concentrations in tissues reach almost 20 mM (based on intracellular water) while plasma concentrations were 570 μM in case of glutamine and around 70 μM in case of taurine. For taurine, therefore a concentration ratio of around 220-fold was obtained followed by a 70-fold higher intracellular than extracellular concentration for glutamic acid and a 33-fold difference for glutamine while those of the essential amino acids differed only by factors of 2- to 6-fold. When urinary levels of the same amino acids are analyzed and standardized to creatinine excretion plasma to urine ratios can vary by up to 100-fold. These data collected from different studies are presented in Fig. 9.2. Numerous transporters in the plasma cell membrane are responsible for those apparently huge concentration ratios – with some of the transporters acting as exchangers others working as uniporters. In addition, these transporters can undergo rapid changes in activity state in response to changes in plasma hormone levels and thereby alter the plasma metabolome. When plasma metabolites are taken to reconstruct metabolic perturbations or disease states by using pathway or network analysis tools, it should be considered that plasma — is not an

	plasma[a]	cellular[b]	urine[c]
Taurine	70	15440	72
Glutamine	570	19450	36
Glycine	210	1330	107
Alanine	330	2340	30
Methionine	20	110	6
Valine	220	260	5

[a] mean, µmol per liter: Bergström et al. 1974, Journal of Applied Physiology
[b] mean, µmol per liter intracellular water: Bergström et al. 1974, Journal of Applied Physiology
[c] mean of first morning urine, mmol per mole of creatinine: Parvy et al. 1988, Clinical Chemistry

Fig. 9.2 Compartmentation of metabolite pools. Demonstration of quite impressive concentration differences (µmol/L) between extracellular (plasma) and intracellular compartments as well as urine based on selected free amino acids. Data are compiled from different studies in human volunteers

open system. Cell membranes separate the compartments and transport phenomena (absorption and secretion) could play a pivotal role. However, pathway maps and analysis tools frequently used contain no relevant information on cellular transport processes which – with a few exceptions – are also largely uncharacterized in specificity and regulation within physiological settings.

Obtaining human tissue samples for metabolite profiling is a more difficult task than just collecting blood or urine. For tissue metabolomes, (minimal) invasive techniques such as biopsies or microdialysis may be applied. Because of these difficulties almost all reported metabolomics studies in humans so far have used plasma, serum or urine. Those samples however represent just flow-through compartments in which thousands of nutrients and metabolites provided by the diet after digestion, absorption and clearance by the liver are mixed with those released by the various organs and cell types, including those derived from microbial metabolism. Figure 9.3 summarizes schematically the multidirectional metabolite fluxes and metabolite origins that constitute the human blood and urine metabolomes with volatile compounds exhaled in breath and excretion of water soluble products by the kidneys.

3 The Human Fasting Metabolome

In addition to food- and microbiota-derived components of the metabolome, there are many factors effecting the metabolite patterns and concentrations with intrinsic factors such as genotype, gender, hormonal status or age, and extrinsic factors such

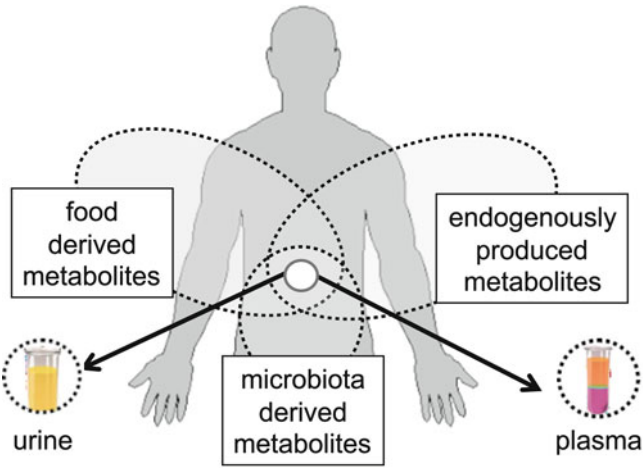


Fig. 9.3 The plasma metabolome as an interface. As a central compartment, plasma changes in composition continuously based on the multidirectional exchange of metabolites. Those are derived either from digested and absorbed food constituents, are produced in the organs and released from there into blood or are synthesized by the gut microbiota and absorbed from the large intestine. Blood delivers also the volatiles to lung followed by exhalation in breath and provides metabolites to kidney with selective, yet not complete re-absorption into circulation and excretion into urine. Secondary metabolite fluxes occur via secretion for example of saliva and bile into the intestine with complete or incomplete re-absorption into circulation

as medication, smoking, stress, pathologies, physical activity, socio-economic status or cultural habits. It is generally considered that a representative metabolome in a human volunteer or patient is obtained when samples are collected in fasting state.

Food intake followed by digestion, absorption and appearance of dietary constituents in plasma as well as the concomitant hormone responses can be followed in the plasma metabolome in a time-dependent fashion [10, 11]. Meal composition will define the metabolome patterns as well as their time-dependent changes. Yet, after a minimum of around 3 h and a maximum of 6–8 h of post-prandial state, food-derived changes in the plasma metabolome are essentially cleared. For practical but not necessarily scientific reasons sample collection in most human studies is done after overnight fasting. Occasionally on the day before sampling, or at least for the last meal served food intake is standardized to minimize carry-over effects of the diet/meal and to reduce "biological noise". However, this human fasting metabolome (8–14 h after the last meal) is also only one time-point in a continuum of metabolic adaptations.

The overnight fasting state is a catabolic state and is characterized by a progressive decline of liver glycogen stores with a growing demand of substrates for hepatic gluconeogenesis for meeting the needs of cells that cannot utilize other fuels except glucose. Simultaneously, lipolysis rate is increased with a rise in levels of free fatty acids in plasma that now serve as energy substrates mainly for muscle and liver metabolism. Glycerol released from adipose tissue as well as glucogenic amino

acids derived from increased protein breakdown serve as precursors for glucose production in liver. Beta-oxidation of fatty acids in muscle and liver can be monitored by time-dependent increases in short and medium-chain acylcarnitines in plasma. A measure of protein breakdown (or diminished protein synthesis rate) is provided by substantial increases in plasma concentrations for example of the branched chain amino acids and their ketoacid-homologues as well as increased levels of some aromatic amino acids [12]. As fasting proceeds, these changes are amplified and the overflow of acetyl-CoA from fatty acid ß-oxidation is in liver translated into an enhanced production of ketone bodies such as acetoacetate and ß-hydroxybutyrate appearing in plasma and urine whereas acetone as a volatile decarboxylation product of acetoacetate is easily now detected in exhaled breath.

The hormone profile of the fasting state is characterized by low plasma insulin but high glucagon and catecholamine levels that determine the catabolic changes in fuel selection and metabolite fluxes. After eating within 30 min, this situation is usually quickly reversed to an anabolic situation, associated with an increase in insulin levels. Figure 9.4 shows both situations for a few marker metabolites (free fatty acids and branched chain amino acids) and insulin levels. It demonstrates the major changes in uptake and release of nutrients/metabolites in a catabolic state (extended fasting) and an anabolic condition such as during food intake or after an oral glucose tolerance test with intake of 75 g of glucose.

4 Diet-Induced Metabolome Changes

Several studies, using different methods for plasma metabolite profiling have recently demonstrated that solely the intake of glucose via an oral glucose tolerance test causes major changes in the plasma profiles of numerous metabolites [13–15]. Glucose and insulin levels rise as expected but the most pronounced changes in the metabolome are observed for bile acids, followed by a large spectrum of amino acids and some keto-acids. Bile acid concentrations rise as fast as the blood glucose levels and increase several-fold but the origin of this phenomenon remains to be understood. Plasma levels of about 10 amino acids (mainly the branched chain and aromatic amino acids but also ornithine and citrulline) decline substantially and remain below fasting levels for up to 3 h after glucose intake. This is best explained by the activation of amino acid transporters in the plasma membrane of insulin-sensitive cells that now increase the uptake of these amino acids from plasma [16, 17]. ß-hydroxybutyrate and other ketoacids decrease in concentration as a measure of the anabolic action of insulin with suppression of lipolysis and fatty acid oxidation. It needs to be stressed, that these changes observed after glucose intake are all underlying also any meal-induced changes in the plasma metabolome with nutrients and metabolites from the food appearing in plasma while simultaneously the increased insulin levels cause most of the described changes in the pool of endogenous metabolites.

Insulin action in muscle, adipose tissue and liver reverses all fasting induced changes to a catabolic situation with an increase in hepatic and muscle glycogen

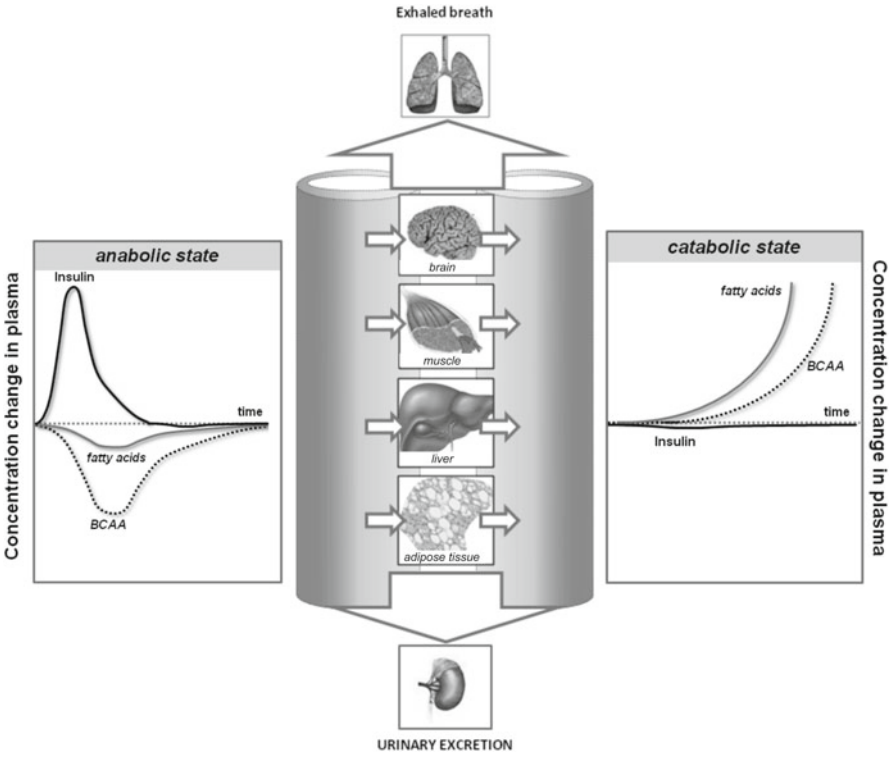


Fig. 9.4 Dynamics of plasma metabolite changes. Anabolic (after food intake) and catabolic states (between meals and fasting) cause major changes in plasma metabolite levels. The catabolic state is characterized by low insulin levels but increased levels of glucagon and catecholamines. The need for ATP production under these conditions is mainly met by enhanced rates of lipolysis, proteolysis and gluconeogenesis. Fatty acids are released from the adipose tissue as main energy substrates for ß-oxidation and acetyl-CoA production whereas increasing branched chain amino acid levels (BCAA) are indicators for a high rate of protein breakdown with gluconeogenic amino acids serving as building blocks (next to glycerol from triglyceride break-down) for hepatic production of glucose to supply obligatory glucose-utilizing cells such as brain or red blood cells. After food intake insulin secretion promotes energy storage by increasing uptake of fatty acids and lipid storage in adipose tissue whereas increased uptake of amino acids into muscle and liver increases protein synthesis. Utilization of glucose as prime energy substrate by muscle, adipose tissue and liver is also enhanced by insulin and hepatic synthesis of glycogen is increased while gluconeogenesis and glycogenolysis simultaneously are inhibited by insulin

synthesis, increased glucose oxidation and increased protein synthesis when sufficient amino acids are provided (Fig. 9.4). Insulin resistance (IR) is frequently associated with severe obesity in humans and is centrally involved in the metabolic syndrome as a condition in the development of type 2 diabetes (NIDDM: non-insulin dependent diabetes mellitus). Despite elevated plasma insulin and glucose levels, tissues such as muscle and liver show reduced or blunted insulin actions leading to an imbalance in the metabolic control with a more pronounced activity of anti-anabolic hormones such as glucagon and catecholamines. This suggests that IR

may cause similar metabolite profiles as in healthy volunteers found during prolonged fasting. Recent metabolite profiling approaches in humans with impaired glucose tolerance or with established NIDDM have indeed identified a subset of metabolites with changes that mimic crossly a prolonged fasting state in healthy volunteers. Amongst the most discriminating plasma metabolites that currently best define an IR state or NIDDM are some metabolites derived from carbohydrate metabolism but also branched chain amino acids, aromatic (Phe, Tyr) amino acids and ketoacids derived from fatty acid and amino acid degradation [18]. Free fatty acids and some glycerophospho- and sphingolipids also show characteristic changes in plasma associated with IR [19, 20]. Urinary analysis of samples obtained from humans with IR or NIDDM also identified predominantly amino acids and their derivatives as the most discriminating metabolites [20].

5 Genetic Determinants of Nutrition-Related Metabolites

Although human genotyping and genome-wide association studies (GWAS) including those related to nutrition are conducted on large scale, comprehensive metabolite profiling approaches are rarely embedded into GWAS. There are however numerous studies that have used targeted approaches with determination of only a few metabolites in plasma or urine in combination with genotyping and those cover almost every part of human metabolism and every metabolite group. Genotype-nutrition associations with analysis of changes in individual plasma metabolites have been described for a large number of vitamins including ascorbic acid [21], vitamin E [22], vitamin A [23] or plasma lipids with a strong focus on polymorphisms in genes encoding apolipoproteins [24, 25]. In relation to lipid metabolism, LDL- or HDL-receptors [26–28] various lipases as well as nuclear receptors, mainly the PPAR-peroxisome proliferator-activated receptor family [29, 30] as well as fatty acid desaturases [31] have been studied for genetic effects on plasma profiles of metabolites. Amongst those gene/protein, classes fatty acid desaturases FADS1 and FADS2 revealed strong associations with altered plasma ratios in selected glycerophospholipids and both contribute to the conversion of n3 or n6 fatty acids into higher polyunsaturated fatty acids (PUFA) such as arachidonic acid and eicosapentanoic acids from which series 2 or series 3 eicosanoids are derived. Various genetic studies have identified SNP´s in these enzymes and those seem to associate in general with altered PUFA-patterns in plasma cell membrane and plasma phospholipids and appear to associate with a variety of diseases or at least disease dispositions [32–34].

Amongst the vitamins, a well-studied example is related to a prominent SNP (C677T Ala-Val) in the methylentetrahydrofolate-reductase (MTHFR) that shows strong associations with the folate status but also with the incidence of various cancer entities. This SNP has also been proposed to associate with cardiovascular diseases (CVD), depression or neural tube birth defects. MTHFR encodes a key enzyme of the remethylation of homocysteine to methionine and requires 5′-methyltetrahydrofolate as a cofactor. The C677T variant of the enzyme shows increased thermolability, reduced activity and altered regulation by cofactors. In particular

when diets low in folate and methyl groups are supplied, this SNP shows particular strong associations with altered metabolite levels and disease endpoints. Folate levels in serum are reduced in the TT versus CC variants whereas homocysteine levels increase inversely [35]. In patients with cardiovascular diseases including stroke patients plasma homocysteine levels in TT genotypes are even more increased and this has led to the conclusion that plasma-homocysteine levels are an independent risk factor in the development and progression of cardiovascular diseases. Yet, interventions have not provided decisive evidence that folate supplementation can reduce the incidence of CVD. For public health promotion, the US, Canada and Chile introduced some 15 years ago a mandatory folate supplementation of flour. Follow up studies assessing plasma homocysteine levels in the population failed to provide the expected effects on the incidence of CVD despite decreased plasma homocysteine levels in the population [36, 37]. This may serve as an example on the difficulties to proof that putative metabolite biomarkers indeed have a causal link to a disease state and possess predictive quality. However, with the extension of metabolomics approaches to GWAS and other cohort studies hopefully more robust markers can be identified that withstand verification in proper designed intervention studies.

Finally, folate metabolism including the gene variants of MTHFR provides currently the only example of an available mathematical model or systems biology approach related to human nutrition [38]. Based on kinetic constants of 11 enzymes in the cytosolic folate cycle, concentrations of 14 relevant metabolites and known regulatory mechanisms a systems model was established that reproduced quite well most experimental findings on folate metabolism. Since a key enzyme in the cycle is the dihydrofolate-reductase which is also a target in tumor therapy with "antifolates" such as methotrexate, the model appears also useful for applications in pharmacology. In addition, taking the genetic variants of MTHFR into the model, it predicted decreased enzymatic activity for the risk allele, with increased concentrations of homocysteine as seen in vivo [39]. This seminal work should guide the development of other systems biology approaches to describe nutritional processes by predictive models.

6 Summary

Metabolomics in human nutrition research is in its infancy. Two tracks however are emerging. One relates to the assessment of food intake by identifying and quantifying marker metabolites that originate from the intake of individual food components which hopefully provides in future better tools in assessing human food consumption. The second track is dedicated to better characterizing metabolic responses to dietary components and dietary habits with the goal to better define the health-disease relationship. This will also include more combinations of comprehensive genotyping and phenotyping of volunteers or patients. Metabolite profiling for deriving markers or metabolite patterns that characterize a disease or pre-disease state need also defined challenge studies in humans to overcome the intrinsic enormous plasticity

Box 9.1 Challenge Tests in Human Metabolomics Studies

Challenge test in human metabolomics studies appear particularly useful for identifying interesting phenotypes and for linking phenotypic differences in response to a challenge to genetic heterogeneity. Numerous environmental pressures have shaped over millennia our genome providing the basis of the inherent plasticity and robustness of our metabolic control. Pushing the metabolic control circuits to their limits by challenging the organism for example with an oral glucose tolerance test, a lipid loading test, an exercise tests or extended fasting periods provides conditions better suited for separating normal/healthy from impaired or diseased. The paradigm for this approach is the oral glucose tolerance test that is applied in clinical routine for diagnosis of insulin resistance and diabetes type 2. Lipid loading tests similarly have demonstrated slow and fast responders for plasma triglyceride level changes that associate with certain disease risks. Although challenge tests allow better phenotypic differentiation between individuals/genotypes, they are demanding for both volunteers and researchers as well as for the logistics of clinical or epidemiological studies. The most simple challenge and easily standardized is a prolonged fasting period of 24 h with a comparison of metabolite responses to those obtained after normal overnight fasting. In addition, oral glucose tolerance tests have been defined with best practice advice by expert panels and those may be included as well for example with time-dependent finger prick blood sampling when sensitivity of metabolomics platforms allow small plasma volumes to be analyzed. To be able to compare metabolomics data and the obtained phenotypic responses across study centers and cohorts an international effort is needed to define the challenges with minimum standard operating procedures.

in human metabolic adaptation. Such challenges (see Box 9.1) with a time dependent analysis of changes in the metabolome require an international effort in standardization to obtain data that can be compared across studies and cohorts.

References

1. Watanabe S, Yamaguchi M, Sobue T et al (1998) Pharmacokinetic of soybean isoflavones in plasma, urine and feces of men after ingestion of 60 g baked soybean powder. J Nutr 128:1710–1715
2. O'Sullivan A, Gibney MJ, Brennan L (2011) Dietary intake patterns are reflected in metabolomic profiles: potential role in dietary assessment studies. Am J Clin Nutr 93:314–321
3. Altmaier E, Kastenmüller G, Römisch-Margl W et al (2011) Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. Eur J Epidemiol 26:145–156

4. Llorach R, Urpi-Sarda M, Jauregui O, Monagas M, Andres-Lacueva C (2009) An LC-MS-based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption. J Proteome Res 8:5060–5068
5. Stalmach A, Troufflard S, Serafini M, Crozier A (2009) Absorption, metabolism and excretion of Choladi green tea flavan-3-ols by humans. Mol Nutr Food Res 53(Suppl 1):S44–S53
6. Mullen W, Edwards CA, Crozier A (2006) Absorption, excretion and metabolite profiling of methyl-, glucuronyl-, glucosyl- and sulpho-conjugates of quercetin in human plasma and urine after ingestion of onions. Br J Nutr 96:107–116
7. Heinzmann SS, Brown IJ, Chan Q et al (2010) Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. Am J Clin Nutr 92:436–443
8. Stella C, Beckwith-Hall B, Cloarec O et al (2006) Susceptibility of human metabolic phenotypes to dietary modulation. J Proteome Res 5:2780–2788
9. Bergstrom J, Furst P, Noree LO, Vinnars E (1974) Intracellular free amino acid concentration in human muscle tissue. J Appl Physiol 36:693–697
10. Haimoto H, Sasakabe T, Umegaki H, Wakai K (2009) Acute metabolic responses to a high-carbohydrate meal in outpatients with type 2 diabetes treated with low-carbohydrate diet: a crossover meal tolerance study. Nutr Metab 29:52
11. Skurk T, Rubio-Aliaga I, Stamfort A, Hauner H, Daniel H (2010) New metabolic interdendencies revealed by plasma metabolite profiling after two dietary challenges. Metabolomics 7:388–399
12. Rubio-Aliaga I, de Roos B, Duthie SJ et al (2010) Metabolomics of prolonged fasting in humans reveals new catabolic markers. Metabolomics 7:375–387
13. Zhao X, Peter A, Fritsche J et al (2009) Changes of the plasma metabolome during an oral glucose tolerance test: is there more than glucose to look at? Am J Physiol Endocrinol Metab 296:E384–E393
14. Shaham O, Wei R, Wang TJ et al (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Mol Syst Biol 4:214
15. Wopereis S, Rubingh C, van Erk MF et al (2009) Metabolic profiling of the response to an oral glucose tolerance test detects subtle metabolic changes. PLoS One 4:e4525
16. Hyde R, Peyrollier K, Hundal HS (2002) Insulin promotes the cell surface recruitment of the SAT2/ATA2 system A amino acid transporter from an endosomal compartment in skeletal muscle cells. J Biol Chem 277:13628–13634
17. Deo RC, Hunter L, Lewis GD et al (2010) Interpreting metabolomic profiles using unbiased pathway models. PLoS Comput Biol 6:e1000692
18. Gall WE, Beebe K, Lawton KA et al (2010) Alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. PLoS One 5:e10883
19. Suhre K, Meisinger C, Döing A et al (2010) Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS One 5:e13953
20. Zhao X, Fritsche J, Wang J et al (2010) Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. Metabolomics 6:362–374
21. Cahill LE, El-Sohemy A (2009) Vitamin c transporter gene polymorphisms, dietary vitamin c and serum ascorbic acid. J Nutrigenet Nutrigenomics 2:292–301
22. Wright ME, Peters U, Gunter MJ et al (2009) Association of variants in two vitamin e transport genes with circulating vitamin e concentrations and prostate cancer risk. Cancer Res 69:1429–1438
23. Manolescu DC, El-Kares R, Lakhal-Chaieb L et al (2010) Newborn serum retinoic acid level is associated with variants of genes in the retinol metabolism pathway. Pediatr Res 67: 598–602
24. Talmud PJ, Hawe E, Martin S et al (2002) Relative contribution of variation within APOC3/A4/A5 gene cluster in determining plasma triglycerides. Hum Mol Genet 11:3039–3046
25. Moreno JA, López-Miranda J, Marín C et al (2003) The influence of the apolipoprotein E gene promoter (−219 G/T) polymorphism on postprandial lipoprotein metabolism in young normolipemic males. J Lipid Res 44:2059–2064

26. Kulseth MA, Berge KE, Bogsrud MP, Leren TP (2010) Analysis of LDLR mRNA in patients with familial hypercholesterolemia revealed a novel mutation in intron 14, which activates a cryptic splice site. J Hum Genet 55:676–680
27. Torres AL, Moorjani S, Vohl MC et al (1996) Heterozygous familial hypercholesterolemia in children: low-density lipoprotein receptor mutational analysis and variation in the expression of plasma lipoprotein-lipid concentrations. Atherosclerosis 126:163–171
28. McCarthy JJ, Lewitzky S, Reeves C et al (2003) Polymorphisms of the HDL receptor gene associated with HDL cholesterol levels in diabetic kindred from three populations. Hum Hered 55:163–170
29. Volcik KA, Nettleton JA, Ballantyne CM, Boerwinkle E (2008) Peroxisome proliferator-activated receptor [alpha] genetic variation interacts with n-6 and long-chain n-3 fatty acid intake to affect total cholesterol and LDL-cholesterol concentrations in the atherosclerosis risk in communities study. Am J Clin Nutr 87:1926–1931
30. Robitaille J, Brouilette C, Houde A et al (2004) Association between the PPARalpha-L162V polymorphism and components of the metabolic syndrome. J Hum Genet 49:482–489
31. Bokor S, Dumont J, Spinneker A et al (2010) Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fat (2010) Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fat aid ratios. J Lipid Res 51:2325–2533
32. Rzehak P, Heinrich J, Klopp N et al (2009) Evidence for an association between genetic variants of the fatty acid desaturase 1 fatty acid desaturase 2 (FADS1 FADS2) gene cluster in the fatty acid composition of erythrocyte membranes. Br J Nutr 101:20–26
33. Kröger J, Zietemann V, Enzenbach C et al (2011) Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European prospective investigation into caner and nutrition (EPIC)-potsdam study. Am J Clin Nutr 93:127–142
34. Schaeffer L, Gohlke H, Müller M et al (2006) Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition phospholipids. Hum Mol Genet 15:1745–1756
35. Ashfield-Watt PA, Pullin CH, Whiting JM et al (2002) Methlenetetrahydrofolate reductase 677 C–>T genotype modulates homocysteine responses to a folate-rich diet or a low-dose folic acid supplement: a randomized controlled trial. Am J Clin Nutr 76:180–186
36. Bazzano LA, Reynolds K, Holder KN, He J (2006) Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials. JAMA 296:2720–2726
37. Albert CM, Cook NR, Gaziano JM et al (2008) Effect of folic acid and B vitamins on risk of cardiovascular events and total mortality among women at high risk for cardiovascular disease: a randomized trial. JAMA 299:2027–2036
38. Nijhout HF, Reed MC, Budu P, Ulrich CM (2004) A mathematical model of the folate cycle: new insights into folate homeostasis. J Biol Chem 279:55008–55016
39. Reed MC, Nijhout HF, Neuhouser M et al (2006) A mathematical model gives insights into nutritional and genetic aspects of folate-mediated one-carbon metabolism. J Nutr 136:2653–2661

Chapter 10
Metabolomics for the Individualized Therapy of Androgen Deficiency Syndrome in Male Adults

Robin Haring, Kathrin Budde, and Henri Wallaschofski

1 Serum Testosterone Concentrations as a Biomarker of Men's General Health Status

1.1 Biology, Physiological Mechanisms, and Epidemiological Evidence

Testosterone is the major circulating androgen in men and essential for the development and maintenance of specific reproductive tissues as the testis and other characteristic male properties including increased muscle strength, bone mass, and hair growth [1]. In serum, most of the circulating testosterone (50–60%) is bound to sex hormone-binding globulin (SHBG), while a smaller fraction (40–50%) is loosely bound to albumin, leaving only 1–3% to circulate as "free" testosterone not bound to protein [2]. In order to maintain testosterone concentrations at appropriate levels, a dynamic network of different interacting factors involved in the excretion and metabolic clearance must be in balance. Briefly, the testosterone action in target cells depends on the amount of steroid which can penetrate into the cells, the extent of metabolic conversions within the cells, the interactions with the receptor proteins, and finally upon the action of the androgen receptors at the genomic level [3]. Given the overall halftime of testosterone in men of about 11 min [4], the flux through this network must be great in order to balance the breakdown of testosterone by a continuous supply.

R. Haring (✉) • K. Budde • H. Wallaschofski, M.D.
Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Ferdinand-Sauerbruch-Strasse, Greifswald, Mecklenburg-Vorpommern 17475, Germany
e-mail: robin.haring@uni-greifswald.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_10,

Based on findings from various prospective epidemiological studies [5–9], it is now well established that total testosterone (TT) concentrations show an age-related decline with mean serum TT concentrations at age of 75 years being about two thirds of those at age of 25 years [10]. But over and above this proven age-related decline on a population-level, a considerably larger inter-individual variability of TT concentrations could be observed at any age. Dependent on the genetic background, accompanying comorbidity, medications, or adverse lifestyle behaviors, individual TT concentrations could be either well preserved against this physiological decline or decrease even more progressively [6]. Although the physiological basis and the extent of the suggested co-factors underlying the large inter-individual variability in TT concentrations are not yet fully elucidated, disturbances in the biosynthesis and actions of testosterone caused by acute illness or chronic diseases are well known [2].

Conversely, findings from prospective epidemiological cohort studies accumulated evidence suggesting low serum TT concentrations as an independent predictor of various cardiovascular risk factors including obesity [11], dyslipidemia [12], hypertension [13], metabolic syndrome [14–16], and type 2 diabetes [17, 18]. It has also been repeatedly observed in different prospective population studies, that low TT concentrations are independently associated with an increased mortality risk [19–22]. But although a multitude of prospective studies showed that low TT concentration precede the onset of various cardiovascular risk factors, others found reduced TT concentrations in men with type 2 diabetes [23], history of myocardial infarction [24], metabolic syndrome [25], obesity [26], and comorbidity [26]. Additionally, a prospective population-based study among 1,490 men aged 20–79 years clearly demonstrated that aging men with obesity, metabolic syndrome, type 2 diabetes, or dyslipidemia had a significantly higher risk to develop low TT concentrations over the five-year follow-up period compared to metabolically healthy aging men [6].

Given the bidirectional nature of the revealed low TT – cardiovascular risk factor associations, reverse causality remains a possibility, since it is still unclear whether low TT concentrations contribute to or are a very early consequence of mechanisms finally leading to a higher cardiovascular risk factor burden (Fig. 10.1). Therefore, low TT concentrations should be considered rather as risk marker instead of risk factor. A risk factor is defined by an aetiologic or causal role in a certain disease process, whereas a risk marker is mainly useful to improve predictive ability [27]. In addition, neither case–control studies, nor prospective cohort studies, observed an independent association between low TT concentrations and incident fatal or nonfatal cardiovascular disease (CVD) events [28], further limiting TT to a risk marker of subclinical, intermediate CVD. In conclusion, if recent epidemiological data tells us anything, it is to perceive serum TT concentrations as a biomarker of good health and overall well-being in men [29]. Circulating TT concentrations show a physiologic decline in conjunction with aging, cardiovascular comorbidity, obesity, medications, and depression. Therefore, testosterone assessment may play a role as personalized risk marker, rather than as an independent causal cardiovascular risk factor [30].

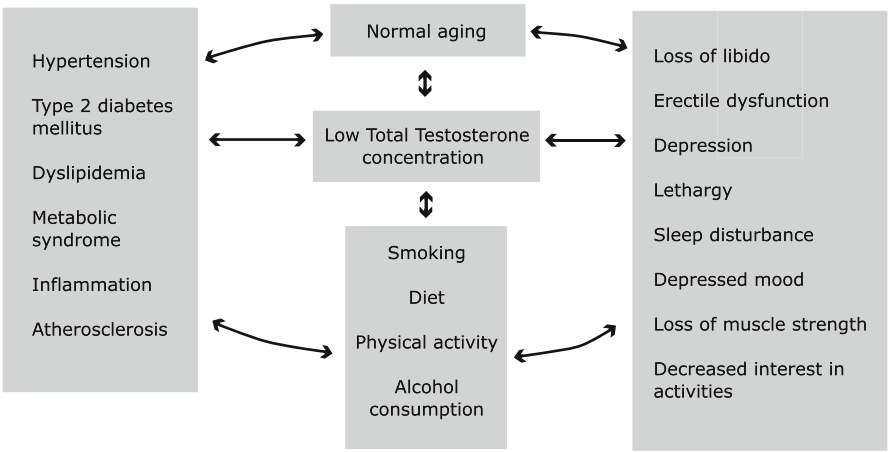


Fig. 10.1 Low total testosterone concentrations in the context of cardiovascular comorbidity, normal aging, health-related lifestyle, and symptoms suggestive of androgen deficiency in men

2 Current Practice for the Therapy of Androgen Deficiency Syndrome in Male Adults

The primary clinical use of testosterone replacement therapy is the diagnosis of primary or secondary hypogonadism caused by "classical" disorders such as Klinefelter syndrome, Kallmann syndrome, or pituitary insufficiency [31]. There is no uncertainty that these patients should receive testosterone replacement therapy [32]. However, in the majority of patients, the diagnosis of low TT concentrations parallels with advanced age, accompanying acute or chronic diseases, medication use, and adverse health-related lifestyle behaviors [33, 34]. Thus, androgen deficiency is a frequent diagnosis in aging men with a prevalence of about 10–20%, depending on the applied cut-off and studied population [6, 35]. Androgen Deficiency Syndrome (ADS) is a "syndromic" diagnosis including both, clinical symptoms together with persistent low TT concentrations (Fig. 10.2). Thus, the TT measurement is a crucial diagnostic criterion requiring proper evaluation and interpretation.

2.1 Laboratory Diagnosis of Androgen Deficiency

Current guidelines unequivocally highlight the measurement of morning serum TT concentrations by a reliable assay as the initial diagnostic test to assess the male androgen status [29, 36, 37]. The presently used, extensively automated procedures

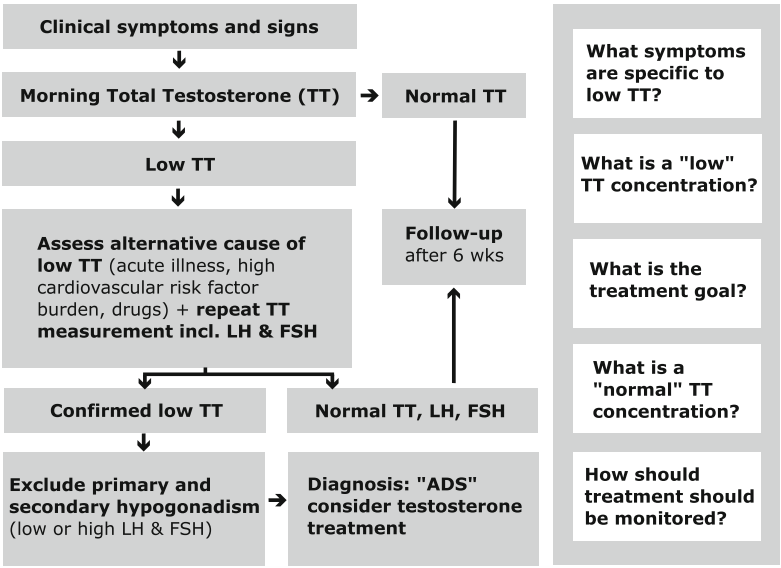


Fig. 10.2 The diagnostic concept of Androgen Deficiency Syndrome (ADS) and its inherent uncertainties and limitations. *FSH* follicle stimulating hormone, *LH* luteinizing hormone

for the analyses of TT concentrations in routine diagnostics are commercial platform-based immunoassays. But even the modern immunologic methods for TT measurement show a significant lack of specificity [38]. Furthermore, the substantial inter-assay and inter-laboratory differences in measured absolute TT concentrations [38–40] proved immunologic procedures insufficient for the diagnostically interesting low concentration range in elderly and comorbid men [41]. Hence, the more precise mass spectroscopic procedures, demonstrating considerably lower intra- and inter-laboratory variability [42], are increasingly considered to be the gold standard for clinical serum TT measurement [36–38, 43, 44]. Beyond these analytical factors, several pre-analytical and physiological co-factors also need to be considered to properly interpret measured TT concentrations. Beside the above mentioned physiological decline and the substantial inter-individual variability of male TT concentrations, a distinct circadian rhythm, the individual's fasting status, and comorbidities (especially severe illness) must also be taken into account. Furthermore, the blood collection technique, sample management and storage, as well as the used reagents have been shown to influence measured TT concentrations [45, 46]. Taken together, measured TT concentrations must be evaluated cautiously not only for reasons of insufficient diagnostic and analytic quality of immunologic measurements in the low concentration range, but also for reasons of pre-analytical and physiological influence factors.

2.2 Clinical Diagnosis

Even more uncertainty exists with regard to the clinical symptoms and signs presumably related to low serum TT concentrations. Conditions like erectile dysfunction, low libido, decreased muscle mass and strength, increased body fat, decreased bone density, decreased vitality, and depressed mood were suggested to be testosterone-associated (Fig. 10.1) [37]. But none of these symptoms are specific to low TT concentrations in men. Therefore, one or more of these symptoms must be corroborated with repeatedly measured low morning TT concentrations, to constitute the diagnosis of ADS and consider the initiation of testosterone replacement therapy (Fig. 10.2). Similarly, an observational study among 3,369 men aged 40–79 years reported that only sexual symptoms had a syndromic association with low TT concentrations [47]. In contrast, a meta-analysis of 17 randomized placebo-controlled trials showed that testosterone replacement therapy only moderately improved the number of sexual symptoms and had no effect on erectile function [48]. This finding is in line with the notion that TT concentrations required for maintaining normal sexual activity are low (which explains the reason why some contracted men still have an erection due to the androgens produced by the adrenal gland) and the factors commonly involved in sexual dysfunction in elderly men are not hormonal [49]. In published reports, the lack of discrimination between androgenic effects on the different domains of sexual function – erectile function, sexual desire, orgasmic function, intercourse satisfaction, and overall satisfaction – along with inadequate sample sizes and statistical power further contribute to misconceptions and misuse of testosterone replacement therapy [48]. In conclusion, the clinical conditions related to low TT concentrations are of non-specific nature and only suggestive, but not diagnostic, of ADS (Box 10.1). This absence of definite clinical correlates of ADS is reflected by the difficulties of the guideline expert panelists to issue firm recommendations and criteria for the initiation of testosterone replacement therapy [36].

Box 10.1 Testosterone and the Androgen Deficiency Syndrome

Testosterone is the major circulating androgen in men, showing an age-related physiological decline starting at 55 years. Furthermore, a considerable inter-individual variability of total testosterone (TT) concentrations could be observed at any age, hampering the definition of a “normal” TT concentration. Late-onset hypogonadism or Androgen Deficiency Syndrome (ADS) is a “syndromic” diagnosis including both, persistent low TT concentrations together with clinical symptoms including erectile dysfunction, low libido, decreased muscle mass and strength, increased body fat, decreased vitality, and depressed mood. But due to its unspecific symptoms, treatment goals, and monitoring parameters, there are many uncertainties concerning the diagnosis, therapy, and monitoring of ADS to date.

2.3 Current Practice of Testosterone Therapy

Once a treatment decision has been made, improvements in signs and symptoms of ADS together with serum TT concentrations in the mid to lower range of young adult males should be sought [36, 37]. To achieve these therapeutic goals, injectable, oral, and transdermal preparations of natural testosterone are currently available. Due to inadequate data, much discussion exists about the critical threshold to determine a definite cut-off for the optimal serum TT concentration in terms of efficacy and safety, as well as a risk-benefit equation for intervention. A review and meta-analysis of 51 treatment studies concluded that the safety of testosterone replacement therapy and its adverse cardiovascular effects are still unknown [50]. The finding of an increase in hemoglobin and hematocrit and a small decrease in high-density lipoprotein cholesterol were of unknown clinical significance. Results from a randomized, double-blinded, placebo-controlled trial in 274 frail elderly men aged 65–90 years, showed that the effects of 6-month testosterone replacement therapy on muscle strength, lean mass, and quality of life were not maintained at six months post treatment [51]. This finding suggests a limited long-term benefit of testosterone replacement therapy in elderly men. The potential risks of oversupplementation during testosterone replacement therapy were exemplified by a discontinued trial among elderly men (mean age 74 years) with mobility limitations, after testosterone replacement was associated with an increased risk of adverse cardiovascular events in the intervention group [52]. The starting doses applied in this trial were higher than those recommended by the manufacturer, and the treatment goal in these patients (34.7 nmol/L) was considerably higher than that recommended by the Endocrine Society (13.9–17.4 nmol/L) [36].

Taken together, the current evidence about the safety of testosterone replacement therapy in men with regard to patient-important outcomes is of low quality and limited by short follow-up periods [50]. Thus, testosterone replacement therapy should only be initiated in the presence of TT concentrations clearly below the lower normal limit for younger men, together with unequivocal signs and symptoms of TT deficiency, in the absence of other reversible causes of decreased TT concentrations, and after screening for contraindications [31]. Once initiated, testosterone replacement therapy should induce and maintain secondary sex characteristics and improve sexual function, sense of well-being, muscle mass and strength, and bone mineral density [36]. Accordingly, the response to testosterone replacement therapy should be assessed and monitored by patients' well-being, sexual activity, occasional measurement of serum TT concentrations, hemoglobin and hematocrit, bone density, and prostate parameters [31, 37].

But as stated above, the presumed syndromic nature of androgen deficiency is often difficult to disentangle, since symptoms and signs suggestive of ADS are readily accounted for by comorbidities and borderline TT deficiency is a frequent biochemical accompaniment of systemic disease – the reason why "pure" ADS is quite uncommon. Furthermore, we do not yet have an operational definition for "normal" TT concentrations at different ages, nor have we identified specific signs and symptoms

to accurately discriminate between those who need treatment and those who do not [32]. Thus, the conjunction between low TT concentrations and several non-specific symptoms, constituting the diagnosis of ADS or late-onset hypogonadism, remains a controversial concept. What is needed most to clarify the outlined uncertainties (Fig. 10.2) and to formally investigate the efficacy of testosterone replacement therapy are improved laboratory methods including standardized and harmonized sex hormone measurements independent of method, time, and place [53], as well as large-scale, long-term, randomized controlled trials of symptomatic middle-aged and elderly men with well-documented low TT concentrations [29].

In conclusion, the key problems surrounding the diagnosis, therapy, and monitoring of ADS include the unanswered question of (1) Which symptoms are specific to low serum TT concentrations? (2) Which treatment goal meets an individual's metabolic needs or what is a "normal" testosterone concentration at different ages? and (3) Which parameters should be monitored during testosterone treatment? Against the background of these limitations and uncertainties, the application of metabolomics offers a variety of scientific opportunities to improve the diagnosis, therapy, and monitoring of ADS in male adults.

3 Metabolomics for the Improved Diagnosis, Therapy, and Monitoring of ADS

3.1 The Principle Techniques of Metabolomics

Small molecules are the end result of all regulatory and metabolic processes at the cellular level within tissues in all organisms. Metabolomics is the global analysis of all, or nearly all of these cellular metabolites [54]. There are two major, complementary approaches in metabolomics: *Targeted analysis* is the most developed analytical approach in metabolomics and is used to measure the concentration of a limited number of precisely known metabolites [54]. With the advantage of a truly quantitative high-throughput approach, targeted analyses are limited by the number of detectable metabolites and the necessary *a priori* knowledge of the targeted metabolites. Therefore, this method is only limited applicable to discover novel metabolic biomarkers or to survey global metabolome-wide changes. The complementary approach of *metabolic profiling* allows measurement of a large set of metabolites in a global (non-targeted), semi-quantitative manner. Since this approach uniquely identifies and simultaneously quantifies a wide range of metabolites and their profile, it is most commonly used for biomarker discovery and the monitoring of global metabolic changes in response to toxic insults, disease processes, or drug therapy.

Although a wide range of techniques can be applied for metabolomics, the two principal methods used to analyze metabolites in body fluids such as urine and plasma are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) [55]. A routine single NMR measurement, generated under full automation in

an acquisition time of 5–10 min, provides semi-quantitative and structural information on a plethora of metabolites in an untargeted multi-marker approach. NMR has also the advantages of being non-destructive to biosamples and requiring minimal sample preparation and disturbance to spare time-consuming and labor-intensive extraction or derivatization steps [55]. MS is currently the most developed technique to quantitatively analyze specific metabolites or a defined set of metabolites with a high sensitivity and throughput. In addition, the rapid technical progress increasingly enables an analytical flexibility that makes MS-based metabolomics amenable to targeted as well as untargeted metabolomics approaches. But in order to apply metabolomics as a novel tool for clinical routine diagnostics (targeted metabolomics) and for biomarker discovery (metabolic profiling) a combination of both techniques is currently needed.

3.2 The Application of Metabolomics to Biomarker Discovery

Through the characterization of metabolic phenotypes and the individual readout of metabolic states, metabolomics promises to discover biomarkers and elucidate biological pathways of human disease [56]. Thus, metabolomics has the potential to contribute significantly to the advent of personalized medicine [57–59]. Indeed, recent metabolomics studies in human urine and plasma samples identified numerous metabolites associated with cardiovascular risk factors [60] including blood pressure [61], type 2 diabetes [62], and atrial fibrillation [63]. Furthermore, metabolomics studies demonstrated the phenotypic heterogeneity of CVD and the limitations of single diagnostic biomarkers [64]. The discovery of new metabolic biomarkers, using untargeted metabolic profiling as well as targeted metabolite quantification, holds the promise to be translated into clinical tools for the application to personalized medicine [65]. The following passage aims to evaluate the potential application of metabolomics to improve the diagnosis, therapy, and monitoring of ADS in the aging male.

3.3 Are There Any Specific Symptoms Related to Low Testosterone Concentrations?

Although a plethora of clinical symptoms and signs have been suggested to be related to low TT concentrations (Fig. 10.1), their non-specificity contributes to the existing doubts about the postulated syndromic nature of low TT concentrations [66–69]. Thus, the discovery of symptoms specific to low TT concentrations would considerably advance the diagnostic reliability of ADS. Metabolic profiling has been shown to help and advance the diagnosis of diseases such as type 2 diabetes [62], Alzheimer, osteoarthritis, or kidney disease [56]. Among 2,422 normoglycemic individuals followed for 12 years, targeted metabolite quantification by liquid

chromatography-tandem mass spectrometry (LC-MS) revealed several amino acids highly predictive of incident type 2 diabetes [62]. These findings were replicated in an independent, prospective cohort and help to understand the role of amino acid metabolism in the pathogenesis of type 2 diabetes. It has also been shown that metabolic profiles of human serum can correctly diagnose not only the presence, but also the severity of coronary heart disease [70]. Classifying patients into three groups according to stenosis of one, two, or three of the coronary arteries, the employed metabolic profiles correctly distinguished all of the groups, whereas none of the traditional clinical risk factors differentiated correctly [70]. Another promising use of metabolomics is indicated by a number of publications on the diagnosis of infant inborn errors of metabolism [71]. Metabolomics also discovered the existence of substantial phenotypic differences with regard to gut microflora from presumably identically bred laboratory rats [72].

These surprising insights suggest that metabolomics may also play an important role in the advanced understanding of the phenotypic heterogeneity of ADS in aging males. The diagnostic information obtained from hypothesis-free metabolic profiling complemented with a targeted quantitative approach is likely to yield a biomarker set of multiple metabolites that could provide comprehensive insights into pathophysiological metabolic processes specific to the onset and progression of ADS that were previously not assessable with traditional single biomarkers such as TT, cholesterol, or fasting glucose [73]. Thus, metabolic profiling may help to overcome the single biomarker conservatism by involving several biomarkers or biomarker combinations to account for the suggested multifactorial causation of ADS. Considering metabolomics as the most proximal reporters of biochemical alterations in response to disease processes or drug therapy [58], metabolic signatures have the unique potential to disclose linkages between physiological, behavioral, and environmental characteristics and thereby to account for the multifactorial causation of ADS.

3.4 Which Treatment Goal Meets an Individual's Metabolic Needs or What is a "Normal" Testosterone Concentration?

In order to pin down the therapeutic effects of an intervention or to elucidate the biochemical alterations caused by a disease, the definition of normality is crucial. As stated above, there are still considerable uncertainties with regard to the definition of normal TT concentrations at different ages. Facing current guidelines [36, 37] which provide only little orientation and arbitrarily defined, fixed TT cut-offs, new approaches are strongly needed. Some investigators suggested the use of age-specific percentile cut-offs (e.g. <10th percentile in each 10-year age group) to account for the age-related physiological decline in TT concentrations [6, 74, 75]. But still, our understanding of individual set points for circulating TT concentrations (below which one, but not another, individual may develop metabolic changes indicative of TT deficiency) or the concept of reserve capacity (the possibility that men with TT

concentrations below the fixed cut-off still may have adequate concentrations to meet their metabolic needs) is very limited. Thus, there is currently no consensus about the definition of normal TT concentrations at different ages, further diluting any efforts to transparently define treatment goals for individuals under testosterone replacement therapy.

Here, metabolomics offers the potential to measure testosterone metabolites and relatives in serum, plasma, and urine to provide a broader metabolic picture for the evaluation of the single absolute serum TT measurement. In particular the urinary steroid profile, which is mainly based on testosterone, reflect the metabolic pathways of androgenic compounds and is essential for the diagnosis of diseases related to steroid secretion [76]. Very recently, LC-MS-based metabolomics was applied for the direct analysis and quantification of major urinary metabolites as markers of exogenous steroid administration in routine doping controls [77]. This method was subsequently validated and applied to a clinical testosterone replacement study comparing a group of healthy male volunteers with a placebo group [77]. Similar studies replicated this method, providing a fast and sensitive analytical procedure for the simultaneous separation, determination, and quantification of testosterone derivatives in human urine [78, 79]. Thus, monitoring steroid conjugates in human urine via metabolomics-based urinary testosterone measurements could yield alternative markers of endogenous testosterone production and testosterone administration.

Furthermore, incorporating detailed information on an individual's metabolic status is likely to advance our understanding of a normal TT concentration in accordance to an individual's metabolic needs. Although the definition of "normal" is in general a tricky endeavor, metabolomics readily generates a multivariate data set that allows the statistical description of biochemical normality, as evidenced by the statistically defined normal model derived from data collected by the COMET consortium [80]. But metabolomics could not only provide a broader picture of the metabolic state of the ADS patient, but also a pre-intervention metabolic signature to evaluate the adequacy, performance, and varying individual beneficial or adverse responses to testosterone replacement therapy.

3.5 Which Parameters Should be Monitored During Testosterone Treatment?

Metabolomics also offers the potential to obtain a multi-metabolite characterization of the metabolic state of the patient under testosterone replacement and to monitor simultaneously the changes in concentrations of a wide range of molecules. Hence, pharmaco-metabonomics has been developed to better understand the inter-individual variability in drug response, predict the individual treatment response, and to ultimately personalize drug treatment [81]. Pharmaco-metabonomics is sensitive to both, the genetic and the modifying environmental influences to determine the individual baseline metabolic profile and to assess the outcome of a drug intervention. Since perturbations of the metabolic state of an individual generally

manifests as particular patterns of metabolites, these metabolic signatures could be used to monitor individuals under testosterone replacement therapy. To control for the large inter-individual variability in drug response and potential confounders, serial sampling can be performed so that each patient serves as his own control. As a proof of principle, studies in patients undergoing controlled interventions such as exercise stress test [82] or oral glucose challenge [73, 83] showed that most metabolites displayed concordant changes in cases and controls, while the metabolites with significant discordant changes in cases remained unchanged in controls. In prospective biomarker studies, the metabolomics approach was proven to identify, categorize, and profile kinetic patterns of early metabolic biomarkers of planned and spontaneous myocardial infarction [84, 85]. Based on plasma samples from 36 patients undergoing a planned myocardial infarction, LC-MS-based metabolite profiling identified different metabolic profiles in the early phase of myocardial injury [85]. More recently, a case–control study of 140 coronary artery disease patients applied targeted LC-MS-based metabolomics to identify two metabolites that were associated with coronary artery disease and subsequent cardiovascular events [86]. Taken together, metabolomics not only successfully differentiated between individuals with and without impaired metabolic regulation, but also revealed novel insights into the investigated disease mechanisms and pathways.

Given the fact that human metabolomics studies are subject to a large inter-individual variability and potential confounders such as age, gender, diet, and comorbidities, the importance of a standardized protocol for the application of metabolomics in a clinical setting with regard to socio-demographic, environmental, nutritional, and behavioral factors could not be underestimated. Also the impact of diurnal and seasonal variations on human metabolic profiles warrants further systematic evaluation. However, metabolomics has been used in a number of studies to define a normal biochemical profile and proved to significantly differentiate effects related to sex, age, diet, various diseases, and drugs. Thus, pharmaco-metabonomics is envisioned to provide real-time metabolic profiles as dynamic markers of a biological status reflecting the individual treatment response and to reveal indicators of treatment efficiency in patients under testosterone replacement therapy.

Metabolic profiling may help to identify a set of metabolites that predict differences in the response to testosterone treatment and to provide biomarker candidates for testosterone replacement therapy monitoring. It has been shown that the application of metabolic profiling of the response to an oral glucose tolerance test detected subtle metabolic changes [73]. Also in animal studies, the urine metabolic profile was used to predict how an individual metabolized a certain drug and their susceptibility to the side effects of that drug [59]. Applying these principles to the treatment and monitoring of ADS would have enormous implications for personalized medicine and optimized risk-benefit ratio for individuals under testosterone replacement therapy. In a possible two-step procedure, single or combined biomarkers of treatment response initially identified by a hypothesis-free metabolic profiling approach, could then be quantified by high-throughput targeted metabolite analyses [80]. The identified particular metabolite pattern or ratio could be subsequently applied to metabolomics-based prediction models to assess the efficacy and safety of

individualized ADS therapies. Given the time-dependent fluctuations of metabolites that occur in response to testosterone treatment, the fairly easy collection of biosamples such as urine or blood is another advantage of metabolomics-based profiling. Furthermore, metabolomics provide an excellent analytical and biological reproducibility and the cost per sample and analyte is low. Taken together, metabolomics offer a huge potential to individually initiate and monitor drug therapies, as to achieve maximal efficacy and avoid adverse drug reactions, to ultimately help to make the vision of personalized medicine become reality.

4 Metabolic Phenotyping in the Field of Andrology

Advanced analytical platforms featured the growth of a wide range of "omics" sciences including genomics, transcriptomics, proteomics, and metabolomics, enabling the measurement of various levels of biomolecular organization in complex mammalian systems. But compared to other molecular profiling techniques, metabolomics provides a more direct and dynamic snapshot of the current physiological status of an individual, representing the omics-level closest to the phenotype and therefore most capable of reflecting the non-linear impact of environmental and lifestyle factors on disease risk. Studies published to date have illustrated the potential for applying metabolomics to the field of andrology. Despite their limited number and pilot-scale study populations, it becomes increasingly clear that metabolomics could improve our capability to diagnose ADS and to be of great value for the clinical management of ADS. Facing the outlined uncertainties in the current diagnosis and therapy of ADS, and the advantageous non-invasive low-cost character of metabolomics, according exploratory and hypothesis generating studies may likely prove beneficial.

In summary, metabolomics is starting to have an impact not only on disease diagnosis and prognosis, but also on drug treatment efficacy and safety monitoring. Through the rapid advances of high-throughput molecular profiling techniques, the future growth of metabolomics research is anticipated to accelerate the implementation and adoption of the metabolomics technology on a much broader basis. In parallel, there is a huge body of knowledge supporting the belief that age changes are characterized by increasing entropy, which results in the random loss of molecular fidelity, and accumulates to slowly overwhelm maintenance systems [87]. When the escalating loss of molecular fidelity ultimately exceeds repair and turnover capacity, vulnerability to pathology or age-associated diseases increases [88]. Thus, the aging process and its associated diseases are more intrinsically malleable than had been previously thought [88]. This holds also true for ADS as a companion of age-associated comorbidity and decreasing physical function and performance. Therefore, the introduction of metabolomics into the field of endocrinology in general, and andrology in particular, provides a novel tool to advance our understanding of the underlying pathophysiological mechanisms of ADS and to further enhance the quality of its diagnosis, therapy, and monitoring.

References

1. Mooradian AD, Morley JE, Korenman SG (1987) Biological actions of androgens. Endocr Rev 8:1–28
2. Kaufman JM, Vermeulen A (2005) The decline of androgen levels in elderly men and its clinical and therapeutic implications. Endocr Rev 26:833–876
3. Rommerts FFG (2004) Testosterone: an overview of biosynthesis, transport, metabolism and non-genomic actions. In: Nieschlag E, Behre H (eds) Testosterone: action, deficiency, substitution. Cambridge University Press, Cambridge
4. Sandberg AA, Slaunwhite WR Jr (1956) Metabolism of 4-C14-testosterone in human subjects I. Distribution in bile, blood, feces and urine. J Clin Invest 35:1331–1339
5. Feldman HA, Longcope C, Derby CA, Johannes CB, Araujo AB, Coviello AD, Bremner WJ, McKinlay JB (2002) Age trends in the level of serum testosterone and other hormones in middle-aged men: longitudinal results from the Massachusetts male aging study. J Clin Endocrinol Metab 87:589–598
6. Haring R, Ittermann T, Volzke H, Krebs A, Zygmunt M, Felix SB, Grabe HJ, Nauck M, Wallaschofski H (2010) Prevalence, incidence and risk factors of testosterone deficiency in a population-based cohort of men: results from the study of health in Pomerania. Aging Male 13: 247–257
7. Harman SM, Metter EJ, Tobin JD, Pearson J, Blackman MR (2001) Longitudinal effects of aging on serum total and free testosterone levels in healthy men. Baltimore longitudinal study of Aging. J Clin Endocrinol Metab 86:724–731
8. Morley JE, Kaiser FE, Perry HM 3, Patrick P, Morley PM, Stauber PM, Vellas B, Baumgartner RN, Garry PJ (1997) Longitudinal changes in testosterone, luteinizing hormone, and follicle-stimulating hormone in healthy older men. Metabolism 46:410–413
9. Zmuda JM, Cauley JA, Kriska A, Glynn NW, Gutai JP, Kuller LH (1997) Longitudinal relation between endogenous testosterone and cardiovascular disease risk factors in middle-aged men. A 13-year follow-up of former multiple risk factor intervention trial participants. Am J Epidemiol 146:609–617
10. Vermeulen A, Kaufman JM, Giagulli VA (1996) Influence of some biological indexes on sex hormone-binding globulin and androgen levels in aging or obese males. J Clin Endocrinol Metab 81:1821–1826
11. Khaw KT, Barrett-Connor E (1992) Lower endogenous androgens predict central adiposity in men. Ann Epidemiol 2:675–682
12. Haring R, Baumeister SE, Völzke H, Dorr M, Felix SB, Kroemer HK, Nauck M, Wallaschofski H (2011) Prospective association of low total testosterone concentrations with an adverse lipid profile and increased incident dyslipidemia. Eur J Cardiovasc Prev Rehabil 18:86–96
13. Torkler S, Wallaschofski H, Baumeister SE, Völzke H, Dörr M, Felix SB, Rettig R, Nauck MA, Haring R (2010) Inverse association between total testosterone concentrations, incident hypertension, and blood pressure. Aging Male 14:176–182
14. Haring R, Volzke H, Felix SB, Schipf S, Dorr M, Rosskopf D, Nauck M, Schofl C, Wallaschofski H (2009) Prediction of metabolic syndrome by low serum testosterone levels in men: results from the study of health in Pomerania. Diabetes 58:2027–2031
15. Kupelian V, Page ST, Araujo AB, Travison TG, Bremner WJ, McKinlay JB (2006) Low sex hormone-binding globulin, total testosterone, and symptomatic androgen deficiency are associated with development of the metabolic syndrome in nonobese men. J Clin Endocrinol Metab 91:843–850
16. Laaksonen DE, Niskanen L, Punnonen K, Nyyssonen K, Tuomainen TP, Valkonen VP, Salonen R, Salonen JT (2004) Testosterone and sex hormone-binding globulin predict the metabolic syndrome and diabetes in middle-aged men. Diabetes Care 27:1036–1041
17. Schipf S, Haring R, Friedrich N, Nauck MA, Lau K, Alte D, Stang A, Völzke H, Wallaschofski H (2011) Low total testosterone is associated with increased risk of incident type 2 diabetes mellitus in men: results from the study of health in Pomerania (SHIP). Aging Male 14:168–175

18. Vikan T, Schirmer H, Njolstad I, Svartberg J (2010) Low testosterone and sex hormone-binding globulin levels and high estradiol levels are independent predictors of type 2 diabetes in men. Eur J Endocrinol 162:747–754
19. Araujo AB, Dixon JM, Suarez EA, Murad MH, Guey LT & Wittert GA (2011) Endogenous testosterone and mortality in men: a systematic review and meta-analysis. J Clin Endocrinol Metab 96:3007–3019
20. Haring R, Volzke H, Steveling A, Krebs A, Felix SB, Schofl C, Dorr M, Nauck M, Wallaschofski H (2010) Low serum testosterone levels are associated with increased risk of mortality in a population-based cohort of men aged 20–79. Eur Heart J 31:1494–1501
21. Khaw KT, Dowsett M, Folkerd E, Bingham S, Wareham N, Luben R, Welch A, Day N (2007) Endogenous testosterone and mortality due to all causes, cardiovascular disease, and cancer in men: European prospective investigation into cancer in Norfolk (EPIC-Norfolk) prospective population study. Circulation 116:2694–2701
22. Laughlin GA, Barrett-Connor E, Bergstrom J (2008) Low serum testosterone and mortality in older men. J Clin Endocrinol Metab 93:68–75
23. Barrett-Connor E, Khaw KT, Yen SS (1990) Endogenous sex hormone levels in older adult men with diabetes mellitus. Am J Epidemiol 132:895–901
24. Swartz CM, Young MA (1987) Low serum testosterone and myocardial infarction in geriatric male inpatients. J Am Geriatr Soc 35:39–44
25. Laaksonen DE, Niskanen L, Punnonen K, Nyyssonen K, Tuomainen TP, Salonen R, Rauramaa R, Salonen JT (2003) Sex hormones, inflammation and the metabolic syndrome: a population-based study. Eur J Endocrinol 149:601–608
26. Wu FC, Tajar A, Pye SR, Silman AJ, Finn JD, O'Neill TW, Bartfai G, Casanueva F, Forti G, Giwercman A, Huhtaniemi IT, Kula K, Punab M, Boonen S, Vanderschueren D (2008) Hypothalamic-pituitary-testicular axis disruptions in older men are differentially linked to age and modifiable risk factors: the European male aging study. J Clin Endocrinol Metab 93:2737–2745
27. Wang TJ (2008) New cardiovascular risk factors exist, but are they clinically useful? Eur Heart J 29:441–444
28. Ruige JB, Mahmoud AM, De Bacquer D, Kaufman JM (2011) Endogenous testosterone and cardiovascular disease in healthy men: a meta-analysis. Heart 97:870–875
29. McLachlan RI (2010) Certainly more guidelines than rules. J Clin Endocrinol Metab 95: 2610–2613
30. Maggio M, Basaria S (2009) Welcoming low testosterone as a cardiovascular risk factor. Int J Impot Res 21:261–264
31. Nieschlag E, Behre H (2004) Clinical uses of testosterone in hypogonadism and other conditions. In: Nieschlag E, Behre H (eds) Testosterone: action, deficiency, substitution. Cambridge University Press, Cambridge
32. Isidori AM, Lenzi A (2007) Testosterone replacement therapy: what we know is not yet enough. Mayo Clin Proc 82:11–13
33. Snyder PJ (2008) Decreasing testosterone with increasing age: more factors, more questions. J Clin Endocrinol Metab 93:2477–2478
34. Travison TG, Araujo AB, Kupelian V, O'Donnell AB, McKinlay JB (2007) The relative contributions of aging, health, and lifestyle factors to serum testosterone decline in men. J Clin Endocrinol Metab 92:549–555
35. Araujo AB, O'Donnell AB, Brambilla DJ, Simpson WB, Longcope C, Matsumoto AM, McKinlay JB (2004) Prevalence and incidence of androgen deficiency in middle-aged and older men: estimates from the Massachusetts male aging study. J Clin Endocrinol Metab 89: 5920–5926
36. Bhasin S, Cunningham GR, Hayes FJ, Matsumoto AM, Snyder PJ, Swerdloff RS, Montori VM (2010) Testosterone therapy in men with androgen deficiency syndromes: an endocrine society clinical practice guideline. J Clin Endocrinol Metab 95:2536–2559
37. Wang C, Nieschlag E, Swerdloff R, Behre HM, Hellstrom WJ, Gooren LJ, Kaufman JM, Legros JJ, Lunenfeld B, Morales A, Morley JE, Schulman C, Thompson IM, Weidner W,

Wu FC (2009) Investigation, treatment, and monitoring of late-onset hypogonadism in males: ISA, ISSAM, EAU, EAA, and ASA recommendations. J Androl 30:1–9
38. Wang C, Catlin DH, Demers LM, Starcevic B, Swerdloff RS (2004) Measurement of total serum testosterone in adult men: comparison of current laboratory methods versus liquid chromatography-tandem mass spectrometry. J Clin Endocrinol Metab 89:534–543
39. Dorgan JF, Fears TR, McMahon RP, Aronson Friedman L, Patterson BH, Greenhut SF (2002) Measurement of steroid sex hormones in serum: a comparison of radioimmunoassay and mass spectrometry. Steroids 67:151–158
40. Hsing AW, Stanczyk FZ, Belanger A, Schroeder P, Chang L, Falk RT, Fears TR (2007) Reproducibility of serum sex steroid assays in men by RIA and mass spectrometry. Cancer Epidemiol Biomarkers Prev 16:1004–1008
41. Taieb J, Mathian B, Millot F, Patricot MC, Mathieu E, Queyrel N, Lacroix I, Somma-Delpero C, Boudou P (2003) Testosterone measured by 10 immunoassays and by isotope-dilution gas chromatography-mass spectrometry in sera from 116 men, women, and children. Clin Chem 49:1381–1395
42. Vesper HW, Bhasin S, Wang C, Tai SS, Dodge LA, Singh RJ, Nelson J, Ohorodnik S, Clarke NJ, Salameh WA, Parker CR Jr, Razdan R, Monsell EA, Myers GL (2009) Interlaboratory comparison study of serum total testosterone [corrected] measurements performed by mass spectrometry methods. Steroids 74:498–503
43. Stanczyk FZ, Clarke NJ (2010) Advantages and challenges of mass spectrometry assays for steroid hormones. J Steroid Biochem Mol Biol 121:491–495
44. Thienpont LM, Van Uytfanghe K, Blincko S, Ramsay CS, Xie H, Doss RC, Keevil BG, Owen LJ, Rockwood AL, Kushnir MM, Chun KY, Chandler DW, Field HP, Sluss PM (2008) State-of-the-art of serum testosterone measurement by isotope dilution-liquid chromatography-tandem mass spectrometry. Clin Chem 54:1290–1297
45. Haring R, Spielhagen C, Nauck M (2011) Challenges in the measurement of serum testosterone concentrations as a biomarker of mens health. J Lab Med 35:1–5
46. Wheeler MJ, Barnes SC (2008) Measurement of testosterone in the diagnosis of hypogonadism in the ageing male. Clin Endocrinol (Oxf) 69:515–525
47. Wu FC, Tajar A, Beynon JM, Pye SR, Silman AJ, Finn JD, O'Neill TW, Bartfai G, Casanueva FF, Forti G, Giwercman A, Han TS, Kula K, Lean ME, Pendleton N, Punab M, Boonen S, Vanderschueren D, Labrie F, Huhtaniemi IT (2010) Identification of late-onset hypogonadism in middle-aged and elderly men. N Engl J Med 363:123–135
48. Isidori AM, Giannetta E, Gianfrilli D, Greco EA, Bonifacio V, Aversa A, Isidori A, Fabbri A, Lenzi A (2005) Effects of testosterone on sexual function in men: results of a meta-analysis. Clin Endocrinol (Oxf) 63:381–394
49. Wespes E (2010) Current approaches to erectile dysfunction and testosterone deficiency. Minerva Urol Nefrol 62:431–435
50. Fernandez-Balsells MM, Murad MH, Lane M, Lampropulos JF, Albuquerque F, Mullan RJ, Agrwal N, Elamin MB, Gallegos-Orozco JF, Wang AT, Erwin PJ, Bhasin S, Montori VM (2010) Clinical review I: adverse effects of testosterone therapy in adult men: a systematic review and meta-analysis. J Clin Endocrinol Metab 95:2560–2575
51. O'Connell MD, Roberts SA, Srinivas-Shankar U, Tajar A, Connolly MJ, Adams JE, Oldham JA, Wu FC (2011) Do the effects of testosterone on muscle strength, physical function, body composition, and quality of life persist six months after treatment in intermediate-frail and frail elderly men? J Clin Endocrinol Metab 96:454–458
52. Basaria S, Coviello AD, Travison TG, Storer TW, Farwell WR, Jette AM, Eder R, Tennstedt S, Ulloor J, Zhang A, Choong K, Lakshman KM, Mazer NA, Miciek R, Krasnoff J, Elmi A, Knapp PE, Brooks B, Appleman E, Aggarwal S, Bhasin G, Hede-Brierley L, Bhatia A, Collins L, LeBrasseur N, Fiore LD, Bhasin S (2010) Adverse events associated with testosterone administration. New Engl J Med 363:109–122
53. Vesper HW, Botelho JC (2010) Standardization of testosterone measurements in humans. J Steroid Biochem Mol Biol 121:513–519
54. Shulaev V (2006) Metabolomics technology and bioinformatics. Br Bioinformatics 7:128–139

55. Lindon JC, Nicholson JK (2008) Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. Annu Rev Anal Chem 1:45–69
56. Lindon JC, Holmes E (2007) A survey of metabonomics approaches for disease characterisation. The handbook of metabonomics and metabolomics. Elsevier, London
57. Holmes E, Wilson ID, Nicholson JK (2008) Metabolic phenotyping in health and disease. Cell 134:714–717
58. Lewis GD, Asnani A, Gerszten RE (2008) Application of metabolomics to cardiovascular biomarker and pathway discovery. J Am Coll Cardiol 52:117–123
59. Nicholson JK, Lindon JC (2008) Systems biology: metabonomics. Nature 455:1054–1056
60. Barderas MG, Laborde CM, Posada M, de la Cuesta F, Zubiri I, Vivanco F, Alvarez-Llamas G (2011) Metabolomic profiling for identification of novel potential biomarkers in cardiovascular diseases. J Biomed Biotechnol 2011:1–9
61. Holmes E, Loo RL, Stamler J, Bictash M, Yap IK, Chan Q, Ebbels T, De Iorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453:396–400
62. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE (2011) Metabolite profiles and the risk of developing diabetes. Nature Med 17:448–453
63. Mayr M, Yusuf S, Weir G, Chung YL, Mayr U, Yin X, Ladroue C, Madhu B, Roberts N, De Souza A, Fredericks S, Stubbs M, Griffiths JR, Jahangiri M, Xu Q, Camm AJ (2008) Combined metabolomic and proteomic analysis of human a trial fibrillation. J Am Coll Cardiol 51:585–594
64. Makinen VP, Soininen P, Forsblom C, Parkkonen M, Ingman P, Kaski K, Groop PH, Ala-Korpela M (2008) 1 H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. Mol Syst Biol 4:167
65. Nicholson JK (2006) Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol 2:52
66. Gould DC, Petty R, Jacobs HS (2000) For and against: the male menopause–does it exist? BMJ 320:858–861
67. Morley JE, Perry HM 3rd (2003) Androgen treatment of male hypogonadism in older males. J Steroid Biochem Mol Biol 85:367–373
68. Perheentupa A, Huhtaniemi I (2007) Does the andropause exist? Nature Clin Pract 3:670–671
69. Seidman SN (2006) Normative hypogonadism and depression: does 'andropause' exist? Int J Impot Res 18:415–422
70. Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK, Bethell HW, Clarke S, Schofield PM, McKilligin E, Mosedale DE, Grainger DJ (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1 H-NMR-based metabonomics. Nat Med 8:1439–1444
71. Moolenaar SH, Engelke UF, Wevers RA (2003) Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. Ann Clin Biochem 40:16–24
72. Robosky LC, Wells DF, Egnash LA, Manning ML, Reily MD, Robertson DG (2005) Metabonomic identification of two distinct phenotypes in Sprague-Dawley (Crl:CD(SD)) rats. Toxicol Sci 87:277–284
73. Wopereis S, Rubingh CM, van Erk MJ, Verheij ER, van Vliet T, Cnubben NH, Smilde AK, van der Greef J, van Ommen B, Hendriks HF (2009) Metabolic profiling of the response to an oral glucose tolerance test detects subtle metabolic changes. PLoS One 4:e4525
74. Jankowska EA, Biel B, Majda J, Szklarska A, Lopuszanska M, Medras M, Anker SD, Banasiak W, Poole-Wilson PA, Ponikowski P (2006) Anabolic deficiency in men with chronic heart failure: prevalence and detrimental impact on survival. Circulation 114:1829–1837
75. Schatzl G, Madersbacher S, Temml C, Krenn-Schinkel K, Nader A, Sregi G, Lapin A, Hermann M, Berger P, Marberger M (2003) Serum androgen levels in men: impact of health status and age. Urology 61:629–633

76. Mareck U, Geyer H, Opfermann G, Thevis M, Schanzer W (2008) Factors influencing the steroid profile in doping control analysis. J Mass Spectrom 43:877–891
77. Badoud F, Grata E, Boccard J, Guillarme D, Veuthey JL, Rudaz S, Saugy M (2011) Quantification of glucuronidated and sulfated steroids in human urine by ultra-high pressure liquid chromatography quadrupole time-of-flight mass spectrometry. Anal Bioanal Chem 400: 503–516
78. Bean KA, Henion JD (1997) Direct determination of anabolic steroid conjugates in human urine by combined high-performance liquid chromatography and tandem mass spectrometry. J Chromatogr B Biomed Sci Appl 690:65–75
79. Strahm E, Kohler I, Rudaz S, Martel S, Carrupt PA, Veuthey JL, Saugy M, Saudan C (2008) Isolation and quantification by high-performance liquid chromatography-ion-trap mass spectrometry of androgen sulfoconjugates in human urine. J Chromatogr 1196–1197:153–160
80. Robertson DG, Reily MD, Cantor GH (2007) Metabonomics in preclinical pharmaceutical discovery and development. In: Lindon JC, Nicholson JK, Holmes E (eds) The handbook of metabonomics and metabolomics. Elsevier, London
81. Clayton TA, Lindon JC, Cloarec O, Antti H, Charuel C, Hanton G, Provost JP, Le Net JL, Baker D, Walley RJ, Everett JR, Nicholson JK (2006) Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature 440:1073–1077
82. Sabatine MS, Liu E, Morrow DA, Heller E, McCarroll R, Wiegand R, Berriz GF, Roth FP, Gerszten RE (2005) Metabolomic identification of novel biomarkers of myocardial ischemia. Circulation 112:3868–3875
83. Shaham O, Wei R, Wang TJ, Ricciardi C, Lewis GD, Vasan RS, Carr SA, Thadhani R, Gerszten RE, Mootha VK (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Mol Syst Biol 4:214
84. Baumgartner C, Lewis GD, Netzer M, Pfeifer B, Gerszten RE (2010) A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury. Bioinformatics 26:1745–1751
85. Lewis GD, Wei R, Liu E, Yang E, Shi X, Martinovic M, Farrell L, Asnani A, Cyrille M, Ramanathan A, Shaham O, Berriz G, Lowry PA, Palacios IF, Tasan M, Roth FP, Min J, Baumgartner C, Keshishian H, Addona T, Mootha VK, Rosenzweig A, Carr SA, Fifer MA, Sabatine MS, Gerszten RE (2008) Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. J Clin Invest 118: 3503–3512
86. Shah SH, Bain JR, Muehlbauer MJ, Stevens RD, Crosslin DR, Haynes C, Dungan J, Newby LK, Hauser ER, Ginsburg GS, Newgard CB, Kraus WE (2010) Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ Cardiovasc Genet 3:207–214
87. Hayflick L (2007) Entropy explains aging, genetic determinism explains longevity, and undefined terminology explains misunderstanding both. PLoS Genet 3:e220
88. Kirkwood TB (2005) Understanding the odd science of aging. Cell 120:437–447

Chapter 11
Systems Biology Resources Arising from the Human Metabolome Project

David Wishart

1 Introduction

Small molecules are intimately connected to every part of the genome and the proteome. In fact, it might be argued that metabolites (or small molecules) actually sit atop the "omic" pyramid of life (Fig. 11.1). Effectively they serve as the "canaries of the genome". Indeed, just a single base change in a gene, can lead to a 10,000 fold change in the concentration of a metabolite or drug [1, 2]. The importance of small molecules for understanding basic biology cannot be overemphasized. Indeed, for more than 80 years the entire field of biochemistry has been dedicated to understanding how certain small molecules control, or are controlled by, larger biological molecules or systems [3, 4]. More recently, fields such as metabolomics [5], systems biology (see Box 11.1) and chemical genomics (see Box 11.2) have emerged with a more holistic mandate for understanding and exploring chemical-biological interactions. All three disciplines are concerned with either measuring small molecules or using small molecules to assess their effects on the genome, the transcriptome or the proteome of a given cell, tissue or organism.

Understanding the interaction between small molecules and larger biological systems is also vitally important to the field of medicine. Consider the following facts: >95% of all diagnostic clinical assays test for small molecules [6], 89% of known drugs are small molecules [7], 50% of all drugs are derived from pre-existing metabolites [8] and 30% of identified genetic disorders involve diseases of small molecule metabolism[9]. Furthermore, almost all of the leading causes of chronic disease arise from adverse interactions of small molecules with our genome or proteome [10–12]. These include obesity (dietary sugars, fats), diabetes (dietary sugars), heart disease (dietary fats, cholesterol), cancer (pollutants and mutagens), and

D. Wishart, Ph.D. (✉)
Department of Computing Science and Biological Sciences, University of Alberta,
2-21 Athabasca Hall, Edmonton, AB T6G 2E8, Canada
e-mail: david.wishart@ualberta.ca

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_11,

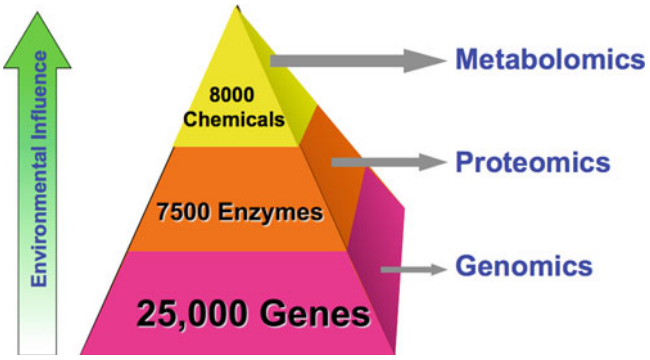


Fig. 11.1 A schematic illustration showing how metabolomics is positioned relative to proteomics and genomics. The genome codes for the proteome which, in turn, acts on the metabolome. In humans there are approximately 25,000 genes (the genome), 7,500 enzymes and transporters (the "active" proteome) and ~8,000 endogenous metabolites (the metabolome)

Box 11.1 What is Systems Biology?

Systems biology is an integrated discipline that combines high throughput experimental techniques such as genomics, proteomics and metabolomics with computational techniques such as bioinformatics and computer simulation in an attempt to fully understand or mechanistically model a biological system, such as a cell, organ or organism. One of the first applications of system biology involved the construction of a mathematical model, in 1952, that explained the action potential propagating along a neuronal cell axon [18]. The formal study of systems biology was launched by systems theorist Mihajlo Mesarovic in 1966 with an international symposium at the Case Institute of Technology in Cleveland, Ohio entitled "Systems Theory and Biology." Leroy Hood, who founded the Systems Biology Institute in Seattle in 2000, is widely credited for linking systems biology to modern "omics" technologies.

Box 11.2 What is Chemical Genomics?

Chemical genomics is primarily concerned with the characterization of genomic (and other "omic") responses to chemical compounds. Chemical genomics uses techniques such as microarrays or high throughput DNA sequencing in conjunction with natural or synthetic chemical library screens. The goal of chemical genomics is to help with the rapid identification of novel drugs and drug targets or to improve the understanding of biochemical signaling mechanisms.

adverse drug reactions (drugs or drug metabolites). It is easy to forget that maternal dietary imbalances and deficits still affect tens of thousands of children in the industrialized world (neural tube defects), childhood cancer – [13, 14] but also account for millions of cases of fetal alcohol syndrome, blindness and mental retardation in the developing world [15]. Clearly metabolites are important, not only for diagnostic purposes, but also for treating and understanding diseases. Yet despite the clear importance that small molecules have in medicine and biology, their place at the scientific table has largely been usurped by the "big" molecules (i.e. DNA, RNA and proteins). Indeed, for nearly two decades most of our biomedical resources have been targeted towards sequencing the human genome [16] or unraveling the human proteome [17]. Consequently, relatively little has been directed towards characterizing the human metabolome.

2 The Human Metabolome Project

In an effort to better characterize the human metabolome, a group of Canadian researchers launched the Human Metabolome Project (HMP) in 2005 [19]. The primary goal of the HMP is to use advanced experimental techniques and literature/text mining to compile as much information about the "detectable" human metabolome – or more appropriately, the human metabolomes as possible. Over the past 7 years this information has been periodically released and updated through a variety of public, web-accessible databases. These databases include the Human Metabolome Database or HMDB [20], which covers endogenous human metabolites (including some very common food, drug and microbial metabolites), DrugBank [21] which contains data on exogenous drugs and drug metabolites, the Toxin and Toxin-Target Database or T3DB [22] which covers pollutants, poisons and environmental toxins and FooDB [23] which contains data on foods (i.e. phytochemicals) and food additives. The approximate size of these databases (in terms of compound numbers) is shown in Fig. 11.2. This figure also illustrates the concentration ranges typically reported for these compounds.

For the most part, the data in these databases has been compiled through extensive literature reviews and careful manual curation by the HMP's team of biocurators, bioinformaticians and chemists. However, in addition to these literature-derived data sets, the HMP has also been performing comprehensive, quantitative metabolic profiling of three medically important biofluids: (1) blood (or serum); (2) urine and (3) cerebrospinal fluid (CSF). Experimental data collected by the HMP on the human cerebrospinal fluid metabolome has been described in two separate papers [24, 25] and the results are maintained in a database at: http://www.csfmetabolome.ca. Currently this database contains nearly 1,000 CSF-specific metabolites along with their corresponding concentrations and disease associations. More recently the HMP completed a comprehensive characterization of human serum metabolome. This effort yielded more than 4,200 serum/blood-specific metabolites and employed at least five different metabolomics platforms [26]. The results are

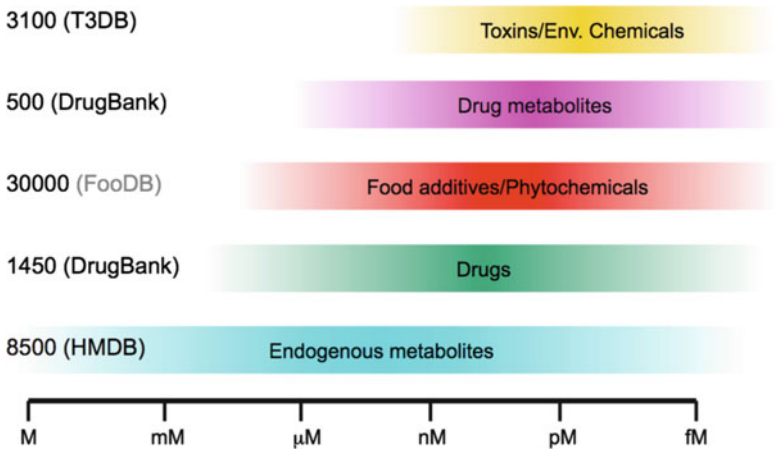


Fig. 11.2 An illustration showing the approximate size (# of compounds) and concentration range (molar, millimolar, micromolar, nanomolar, picomolar) of the compounds found in different human metabolomes. HMDB covers the endogenous metabolites, DrugBank covers approved drugs and nutraceuticals, FooDB covers food components and food additives, T3DB covers toxins, pollutants and poisons

housed in on online human serum metabolome database located at: http:///www.serummetabolome.ca. It is expected that the human urine metabolome will be finished by mid 2012, although the HMP has already published one paper describing the partial characterization of this very important biofluid [27].

3 Defining Different Human Metabolomes

While all humans share essentially the same endogenous metabolome, each human has a different exogenous metabolome. The endogenous metabolome refers to the collection of "natural" compounds that our cells need to live and which they routinely synthesize or catabolize. The exogenous metabolome refers to the "un-natural" synthetic, exotic or plant-derived chemicals that we deliberately or accidentally consume and which are not needed for basic metabolism. Obviously no single human (unless they regularly consume all known drugs, eat all known foods and live in a toxic chemical dump) will have the full complement of known or detectable exogenous compounds in their body. On the other hand, if a large population of humans is studied, certainly many of these exogenous metabolites will be observed – albeit at relatively low levels. A further source of variation in the human metabolome comes from the metabolites generated by the nearly 400 different microbial species that live in the human gut [28]. In humans, the gut microflora weigh between 1 and 2 kg and constitute a metabolically essential multicellular organ [28]. Each human has their own unique gut microflora and these microbe (along with our diet) contribute significantly to our own, unique metabolic phenotype or "metabotype" [29].

The issue of exogenous versus endogenous metabolites is not the only complication associated with describing the human metabolome. Humans have more than 200 different cell types, several dozen different organs and many highly compartmentalized biofluid systems. Each of these cell types, tissues or organs is metabolically specialized in some fashion or another, often producing a handful of unique metabolites that are not found in other cells or organs. The same metabolic specialization is true for many human biofluids as well. These biofluids include blood, milk, cerebrospinal fluid (CSF), bile, saliva, mucus, lung exudates, lachrymal secretions, semen, lymph and more. As a result, specific cell, tissue, biofluid and organ variations also make the human metabolome hard to define. So too does the wide range of metabolite concentrations found in humans. The concentration of human metabolites can range from the low picomolar level (i.e. exogenous chemicals, certain hormones and many signaling molecules) to as high as near-molar levels (urea). These values may vary with diet, gender, time of day, age, health and genetic background [30]. Therefore the human metabolome is actually defined by when, where and how it's measured.

A further complication with regard to describing the human metabolome is the fact that it consists of both a "known" component and an "unknown" or theoretical component. Figure 11.2 describes the size of the known human metabolomes. It does not indicate the size of the "theoretical" human metabolome. Certainly if all possible combinations of lipids, di- and tripeptides and di- or trisaccharides were considered, the number of endogenous human metabolites could easily exceed 200,000 molecules [31, 32]. However, the vast majority of these theoretical metabolites have not been detected. Either they exist too transiently or are at such low abundance that they cannot be seen with today's technologies. We also know from numerous metabolomic experiments that only 1/3–1/2 of the metabolite signals detected by UPLC-MS experiments correspond to known metabolites. Consequently there are perhaps tens of thousands of hitherto unknown human metabolites for which the chemical structure is not known or has yet to be described.

So unlike the human genome, which is well defined and largely invariant, the human metabolome is an ill-defined, highly individualized, ever-growing entity that is profoundly affected by the genome, the environment and the available technologies used to measure it. These features make the study of the human metabolome both compelling and challenging.

4 The Human Metabolome Database

The Human Metabolome Database (HMDB) is the largest and most comprehensive, organism-specific metabolomic database assembled to date [33]. It contains spectroscopic, quantitative, analytic and molecular-scale information about (mostly) endogenous human metabolites, their associated enzymes or transporters, their abundance and their disease-related properties. The HMDB currently contains more than 8,500 human metabolite entries that are linked to more than 45,000 different synonyms. These metabolites are further connected to 3,360 distinct enzymes,

which in turn, are linked to nearly 100 metabolic pathways and more than 150 disease pathways. More than 1,000 metabolites have disease-associated information, including both normal and abnormal metabolite concentration values. These diagnostic metabolites or metabolite signatures are linked to more than 500 different diseases (genetic and acquired). The HMDB also contains experimentally acquired metabolite concentration data (normal and abnormal) for nearly 5,000 compounds from most biofluids. The entire database, including text, sequence, structure and image data occupies nearly 30 Gigabytes of data – most of which can be freely downloaded.

The HMDB is fully searchable database with many built-in tools for viewing, sorting and extracting metabolites, biofluid concentrations, enzymes, genes, NMR or MS spectra and disease information. Users may through the HMDB compound by compound through a series of hyperlinked, synoptic summary tables. These metabolite tables can be rapidly browsed, sorted or reformatted in a manner similar to the way PubMed abstracts may be viewed. Clicking on the MetaboCard button found in the leftmost column of any given HMDB summary table opens a webpage describing the compound of interest in much greater detail. Each MetaboCard entry contains more than 100 data fields with half of the information being devoted to chemical or physico-chemical data and the other half devoted to biological or biomedical data. These data fields include a comprehensive compound description, names and synonyms, structural information, physico–chemical data, reference NMR and MS spectra, biofluid concentrations (normal and abnormal), disease associations, pathway information, enzyme data, gene sequence data, protein sequence data, SNP and mutation data as well as extensive links to images, references and other public databases such as KEGG [34], BioCyc [35], PubChem [36], ChEBI [37], PubMed, PDB [38], SwissProt/UniPort [39], GenBank [40], and OMIM [9].

Unlike most other "omics" databases, the HMDB has been designed to facilitate exploring the linkage between genes, diseases and metabolites. This has been done in several ways, including the careful compilation of information about most of the known inborn errors of metabolism (IEMs), detailed data on the chemical biomarkers that can be used to diagnose these diseases and links to OMIM (and other databases) that describe their known or probable genetic causes. An example of how the HMDB may be used to go from raw experimental data, to metabolite lists, to disease identification, to pathway analysis and genetic characterization will be given in Sect. 6 of this chapter.

5 The Exogenous Human Metabolome: DrugBank, T3DB and FooDB

As noted in Sect. 2, the human metabolome consists of both an endogenous component and an exogenous component. The HMDB primarily covers the endogenous human metabolome. The exogenous human metabolomes are handled by three separate databases: DrugBank (for drugs), T3DB (for toxins, pollutants and poisons)

and FooDB (for food additives and food components). The separation of the exogenous human metabolome into three components was done to support the specific needs of each user community. For instance, DrugBank was designed to address the needs of pharmacologists and pharmacists, as well as metabolomics researchers. Consequently DrugBank has a number of data fields that would not be found in HMDB, such as mechanism of action, absorption and pharmacokinetic data. Likewise, T3DB was designed to address the needs of toxicologists and emergency room physicians, as well as metabolomics specialists. As a result T3DB contains data fields such as LD_{50}, treatment option and poisoning symptoms. FooDB, which is still under construction, is being designed to address the needs of food chemists, nutritionists and metabolomics scientists. Consequently FooDB contains data on food types and food composition. Unlike the HMDB (which consists of both literature derived and in-house experimental data), the data in DrugBank, T3DB and FooDB has been compiled entirely through literature review and text mining. To better understand the content and utility of these exogenous metabolome resources, it is perhaps useful to provide a brief description of each database and an explanation of how they can be used to connect the metabolome to the genome.

5.1 DrugBank

DrugBank [21, 41] is a drug database that links structure and mechanistic data about drug molecules with sequence, structure and mechanistic data about their drug targets. Like HMDB, DrugBank presents its data on drugs and drug targets using synoptic DrugCards (in anology to MetaboCards). Currently DrugBank contains detailed information on 1,480 FDA-approved drugs corresponding to ~28,00 brand names and synonyms. This collection includes almost 1,300 synthetic small molecule drugs, more than 125 biotech (mostly peptide or protein) drugs and 70 nutraceutical drugs or supplements. DrugBank also contains information on nearly 1,700 different targets (protein, lipid or DNA molecules) and metabolizing enzymes with which these drugs interact. Additionally the database maintains data on almost 200 illicit drugs (i.e. those legally banned or selectively banned in most developed nations) and more than 60 withdrawn drugs (those removed from the market due to safety concerns). Like HMDB, DrugBank also supports a wide variety of text, chemical formula, mass, chemical structure and sequence searches. In addition to these search features, DrugBank also provides a number of general browsing tools for exploring the database as well as several specialized browsing tools such as PharmaBrowse and GenoBrowse for more specific tasks. GenoBrowse is specifically designed to address the needs of those specialists interested in specific drug-gene relationships. This browsing tool provides navigation hyperlinks to more than 60 different drugs, which in turn list the target genes, SNPs and the physiological effects associated with these drugs.

DrugBank also provides detailed sequence and SNP data on known drug metabolizing enzymes and known drug targets. In particular DrugBank contains detailed

summary tables about each of the SNPs for each of the drug targets or drug metabolizing enzymes that have been characterized by various SNP typing efforts. Currently DrugBank contains information on 26,000 coding (exon) SNPs and nearly 74,000 non-coding (intron) SNPs derived from known drug targets. It also has data on 1,188 coding SNPs and 8,931 non-coding SNPs from known drug metabolizing enzymes. By clicking on the "Show SNPs" hyperlink listed beside either the metabolizing enzymes or the drug target SNP field, the SNP summary table can be viewed. The purpose of these SNP tables is to allow one to go directly from a drug of interest to a list of potential SNPs that may contribute to the reaction or response seen in a given patient or in a given population. In particular, these SNP lists may serve as hypothesis generators that allow SNP or gene characterization studies to be somewhat more focused or targeted. By comparing the experimentally obtained SNP results to those listed in DrugBank for that drug (and its drug targets) it may be possible to ascertain which polymorphism for which drug target or drug metabolizing enzyme may be contributing to an unusual drug or metabolite profile.

DrugBank also includes two tables that provide much more explicit information on the relationship between drug responses/reactions and gene variant or SNP data. The two tables, which are accessible from the GenoBrowse submenu located on DrugBank's Browse menu bar, are called SNP-FX (short for SNP-associated effects) and SNP-ADR (short for SNP-associated adverse drug reactions). SNP-FX contains data on the drug, the interacting protein(s), the "causal" SNPs or genetic variants for that gene/protein, the therapeutic response or effects caused by the SNP-drug interaction (improved or diminished response, changed dosing requirements, etc.) and the associated references describing these effects in more detail. SNP-ADR follows a similar format to SNP-FX but the clinical responses are restricted only to adverse drug reactions (ADR). SNP-FX contains literature-derived data on the therapeutic effects or therapeutic responses for more than 70 drug-polymorphism combinations, while SNP-ADR contains data on adverse reactions compiled from more than 50 drug-polymorphsim pairings. All of the data in these tables is hyperlinked to drug entries from DrugBank, protein data from SwissProt, SNP data from dbSNP and bibliographic data from PubMed.

5.2 The Toxin/Toxin-Target Database: T3DB

T3DB [22] is currently the only chemical-bioinformatic database that provides in-depth, molecular-scale information about toxins/poisons, their associated targets (genes/proteins), their toxicology, their toxic effects and their potential treatments. T3DB currently contains over 3,000 toxic substance entries corresponding to more than 34,000 different synonyms. These toxins are further connected to some 1,450 protein targets through almost 35,500 toxin and toxin-target associations. These associations are supported by more than 5,400 references. The entire database, including text, sequence, structure and image data, occupies nearly 16 Gigabytes of data – most of which can be freely downloaded.

As with HMDB and DrugBank, the T3DB is designed to be a fully searchable web resource with many built-in tools and features for viewing, sorting and extracting toxin and toxin-target annotation, including structures and gene and protein sequences. As with HMDB and DrugBank, the T3DB supports standard text queries through the text search box located on the home page. It also offers general database browsing using the "Browse" button located in the T3DB navigation bar. To facilitate browsing, the T3DB is divided into synoptic summary tables which, in turn, are linked to more detailed "ToxCards" – in analogy to the DrugCard concept found in DrugBank [33] or the MetaboCard in HMDB [33]. All of the T3DB's summary tables can be rapidly browsed, sorted or reformatted in a manner similar to the way PubMed abstracts may be viewed. Clicking on the ToxCard button, found in the leftmost column of any given T3DB summary table, opens a webpage describing the toxin of interest in much greater detail. Each ToxCard entry contains over 80 data fields, with ~50 data fields devoted to chemical and toxicological/medical data and ~30 data fields (each) devoted to describing the toxin target(s).

The data included in a ToxCard includes various identifiers and descriptors of the toxin (names, synonyms, compound description, structure image, related database links and ID numbers), followed by additional structure and physico-chemical property information. The remainder of data on the toxin is devoted to providing detailed toxicity and toxicological data, including route of delivery, mechanism of action, medical information, and toxicity measurements. All of a toxin's targets are also listed within the ToxCard. Each of these targets are described by some 30 data fields that include both chemical and biological (sequence, molecular weight, gene ontology terms, etc.) information, as well as details on their role in the mechanism of action of the toxin. In addition to providing comprehensive numeric, sequence and textual data, each ToxCard also contains hyperlinks to other databases, abstracts, digital images and interactive applets for viewing the molecular structures of each toxic substance.

T3DB's sequence searching utility (SeqSearch) allows users to search through T3DB's collection of 1,450 known (human) toxin targets. This service potentially allows users to identify both orthologous and paralogous targets for known toxins or toxin targets. It also facilitates the identification of potential toxin targets from other animal species. With SeqSearch, gene or protein sequences may be searched against the T3DB's sequence database of identified toxin-target sequences by pasting the FASTA formatted sequence (or sequences) into the SeqSearch query box and pressing the "submit" button.

T3DB's structure similarity search tool (ChemQuery) can be used in a similar manner as its SeqSearch tool. Users may sketch a chemical structure or paste a SMILES string of a query compound into the ChemQuery window. Submitting the query launches a structure similarity search that looks for common substructures from the query compound that matches the T3DB's database of known toxic compounds. Users can also select the type of search (exact or Tanimoto score) to be performed. High scoring hits are presented in a tabular format with hyperlinks to the corresponding ToxCards (which, in turn, links to the targets). The ChemQuery tool allows users to quickly determine whether their compound of interest is a known

toxin or chemically related to a known toxin and which target(s) it may act upon. In addition to these structure similarity searches, the ChemQuery utility also supports compound searches on the basis of chemical formula and molecular weight ranges.

5.3 FooDB: The Food Metabolome Database

FooDB, which is still under construction, is a database intended to capture key information on food-related chemicals. When completed in late 2012, FooDB will be the world's largest and most comprehensive resource on food constituents and their health effects. Food chemicals and secondary food metabolites actually constitute a significant part of the human metabolome. Indeed, it is widely believed that most unknown or unidentified peaks in metabolomic studies of human urine, saliva and plasma are derived from food-derived compounds [33]. Knowledge of the food metabolome important not only for advancing the field of metabolomics, but also for understanding the adverse or beneficial effects of food chemicals, for discovering novel drugs or drug leads and for guiding studies in nutrition and nutrigenomics [24, 42]. To date more than 28,000 food constituents, food additives and food metabolites have been identified and catalogued. Many (>80%) of these are phytochemicals or chemicals of plant origin such as polyphenols, phytosterols, alkaloids, quinones, terpenes, phenylpropanoids, etc. Another 2,500 compounds are approved food additives such as synthetic coloring, aroma and flavoring agents. The challenge in building FooDB has been to annotate these food compounds to the same level as found in the HMDB, DrugBank and T3DB. When completed, FooDB will provide information on both macronutrients and micronutrients, including their role in foods (flavor, color, taste, texture and aroma) and their role in human physiology. Each chemical entry in the FooDB will contain data on the compound's nomenclature, a description, information on its structure, chemical class, physicochemical data, food source(s), physiological effect(s), presumptive health effects or health claims, protein targets, biosynthesis or synthesis pathways, breakdown products, known metabolites, biochemistry, concentrations in various foods, metabolic breakdown products and concentrations in human biofluids. Users will be able to browse or search the FooDB by food source, name, descriptors, function or concentrations. Depending on individual preferences users will be able to view the content of FooDB from the "FoodView" (listing foods by their chemical composition) or the "ChemView" (listing chemicals by their food sources). An example FooDB entry can be viewed at http://foodbs.org/example.

6 SMPDB: The Small Molecule Pathway Database

Unlike HMDB, DrugBank, T3DB or FooDB, which are metabolite databases, SMPDB is primarily a picture resource. More specifically, SMPDB is a pathway database specifically designed to facilitate clinical "omics" studies, with a specific

emphasis on clinical biochemistry and clinical pharmacology. Currently SMPD consists of more than 450 highly detailed, hand-drawn pathways describing small molecule metabolism or small molecule processes that are specific to humans. These highly hyperlinked pathways can be placed into four different categories: (1) metabolic pathways; (2) small molecule disease pathways; (3) small molecule drug pathways and (4) small molecule signaling pathways. All SMPDB pathways explicitly include the chemical structure of the major chemicals in each pathway. In addition, the cellular locations (extracellular, intracellular, membrane, cytoplasm, mitochondrion, nucleus, peroxisome, etc.) of all metabolites and the enzymes involved in their processing are explicitly illustrated. Likewise the quaternary structures (if known) and cofactors associated with each of the pathway enzymes or transporters are also shown. If some of the metabolic processes occur primarily in one organ or in the intestinal microflora, this information is also illustrated. The inclusion of explicit chemical, cellular and physiological information is one of the more unique and useful features of SMPDB. SMPDB is also unique in its inclusion of significant numbers of metabolic disease pathways (>100) and drug pathways (>200) not found in any other pathway database. Likewise, unlike other pathway databases, SMPDB supports a number of unique database querying and viewing features. These include simplified database browsing, the generation of protein/metabolite lists for each pathway, text querying, chemical structure querying and sequence querying, as well as large-scale pathway mapping via protein, gene or chemical compound lists.

The SMPDB interface is largely modeled after the interface used for DrugBank [7, 24], T3DB [22] and the HMDB [33], with a navigation panel for Browsing, Searching and Downloading the database. Below the navigation panel is a simple text query box that supports general text queries of the entire textual content of the database. Mousing over the Browse button allows users to choose between two browsing options, SMP-BROWSE and SMP-TOC. SMP-TOC is a scrollable hyperlinked table of contents that lists all pathways by name and category. SMP-BROWSE is a more comprehensive browsing tool that provides a tabular synopsis of SMPDB's content with thumbnail images of the pathway diagrams, textual descriptions of the pathways, as well as lists of the corresponding chemical components and enzyme/protein components. This browse view allows users to scroll through the database, select different pathway categories or re-sort its contents. Clicking on a given thumbnail image or the SMPDB pathway button brings up a full-screen image for the corresponding pathway. Once "opened" the pathway image may be expanded by clicking on the Zoom button located at the top and bottom of the image. An image legend link is also available beside the Zoom button.

At the top of each pathway image is a pathway synopsis contained in a yellow box while at the bottom of each image is a list of references. On the right of each pathway image is a grey-green Highlight/Analyzer tool with a list of the key metabolites/drugs and enzymes/proteins found in the pathway. Checking on selected items when in the SMP-Highlight mode will cause the corresponding metabolite or protein in the pathway image to be highlighted with a red box. Entering concentration or relative expression values (arbitrary units) beside compound or protein names, when in the SMP-Analyzer mode, will cause the corresponding metabolites or

proteins to be highlighted with differing shades of green or red to illustrate increased or decreased concentrations. As with most pathway databases, all of the chemical structures and proteins/enzymes illustrated in SMPDB's diagrams are hyperlinked to other on-line databases or tables. Specifically, all metabolites, drugs or proteins shown in the SMP-BROWSE tables or in a pathway diagram are linked to HMDB, DrugBank or UniProt [39] respectively. Therefore, clicking on chemical or protein image will open a new browser window with the corresponding DrugCard, MetaboCard or UniProt table being displayed.

7 From Experiment to Systems Biology: An Example

In order to understand how the resources generated by the HMP might be useful to researchers studying the human metabolome or to those looking for a broader, systems biology understanding of human metabolism it is perhaps useful to given an example. This particular scenario is intended to illustrate how the HMDB and SMPDB, together, can be used to take relatively raw, untargeted MS data from human serum and facilitate the understanding of the chemistry, biology, genetics and systems biology of a particular disease. The data used here is "synthetic" and the example is somewhat simplified to make the explanations and interpretation easier for the reader.

Here we will assume that two blood or serum samples have been provided. One is from a normal, healthy individual and the other is from a newborn having recurring seizures and exhibiting unusually fair skin and hair. The serum samples have been run through an UPLC-MS system using a higher resolution mass spectrometer (an Orbitrap or an FT-MS instrument) in the positive ion mode. Comparison of the signal intensities of the sick infant's sample with those of an age-matched, healthy control show a number of signals having significantly higher values (Table 11.1).

Using the mass list in Table 11.1 we can go to the HMDB (www.hmdb.ca) and use the menu bar at the top of the home page to select "Search". Under the "Search" menu we should select the "MS Search" submenu item. A web page will appear that should look like the image shown in Fig. 11.3. Once this page appears, select the

Table 11.1 Mass (m/z) values for peaks exhibiting significantly enhanced intensities as collected on blood plasma from a sick infant. The data for this UPLC-MS study was collected in the positive ion mode

Peak number	Mass (m/z) Daltons
1	137.0595
2	159.0420
3	165.0541
4	166.0863
5	167.0705
6	187.0368
7	188.0682
8	189.0525

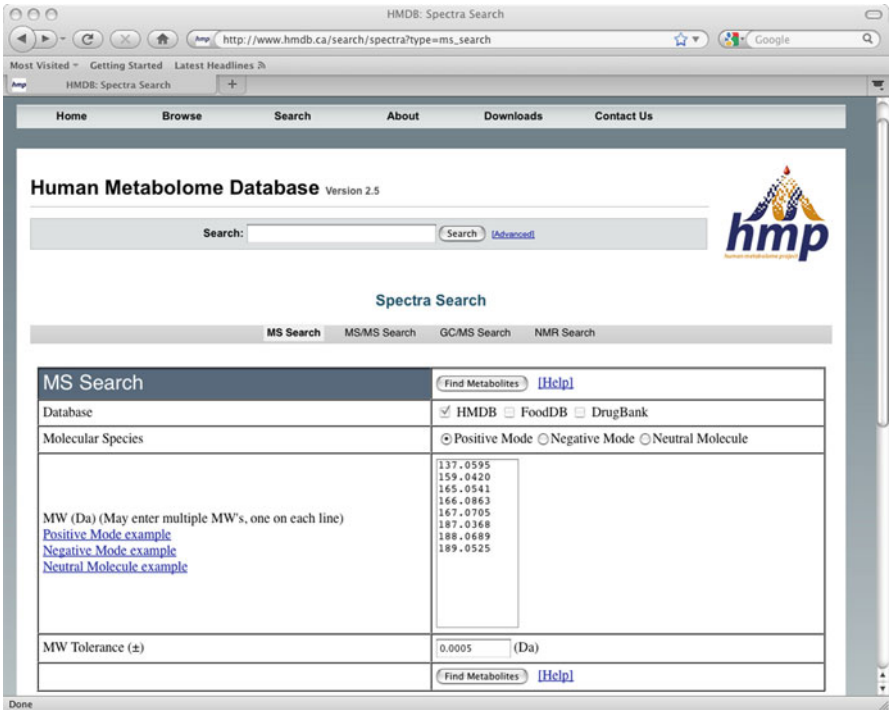


Fig. 11.3 A screenshot of the HMDB "MS Search" web page. Masses can be entered into the text box shown in the middle of this page

"Positive Mode" radio button under the Molecular Species field and then type in the eight masses listed in Table 11.1 in the text box (include all digits). In the MW Tolerance field, enter a mass tolerance of 0.0005 Da. Once these data have been entered, click the "Find Metabolites" button. Within a few seconds the result in Fig. 11.4 should be generated. This table lists the metabolites (HMDB ID, names, chemical formula, adduct molecular weight, mass difference and adduct type) matching to the query mass list we have just entered. Approximately 50 compounds or adducts should be listed (of 8,500 possible compounds in the HMDB). The list is sorted by the mass error difference. Inspecting this list, one might notice a wide range of possible adducts, with potassium, sodium, acetonitrile and various paired ion variants. Not all of these would likely be encountered in an experiment of this nature and certainly many MS platforms now provide software that can help identify the parent ions and their adducts. Nevertheless, this example is intended to show how one could use the HMDB from the very beginning to the very end of a metabolomics experiment.

If one clicks on the "Common Name" data field in the table shown in Fig. 11.4, the list will be re-sorted alphabetically. Re-sorting the data in this way allows one to identify compounds that show up frequently in the list (i.e. those compounds that

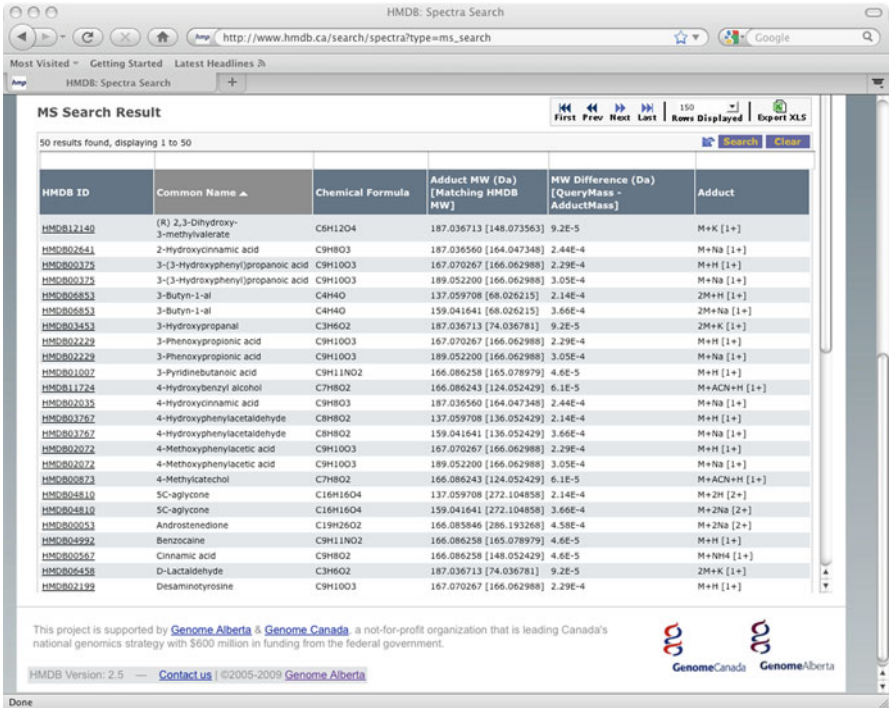


Fig. 11.4 A screenshot of the output generated by an MS Search query using the masses listed in Table 11.1

have several different adducts). Having two or more adducts matching to a given parent compound can provide a greater degree of certainty about what the compound may be. Looking through this re-sorted list, it is quite apparent that there are multiple hits for 3-Hydroxyphenylpropanoic acid (HMDB00375), 3-Butyn-1-al (HMDB06853), 3-Phenoxyprorionic acid (HMDB02229), 4-Hydroxyphenylacetaldehyde (HMDB03767), 4-Methoxyphenylacetic acid (HMDB02072), Desaminotyrosine (HMDB02199), 5 C-aglycone (HMDB04810), Homovanillin (HMDB05175), Phenyllactic acid (HMDB00748/HMDB00779), Phenylalanine (HMDB00159) and Phenylacetic acid (HMDB00209). Further inspection also reveals several occurrences of Phenylpuryvate-like analogs (HMDB00205). The hits to 3-Butyn-1-al and 5 C-aglycone are rather exotic (rare) adducts and so these can probably be ruled out as spurious matches. They are also not reported to occur in blood. Interestingly, the remaining compounds are all tyrosine or phenylalanine intermediates or analogues. A further filtering step can be done to see if these compounds are found (or have ever been reported) to appear in human blood. Clicking on the HMDB identifiers for our hits in the MS hit table shown in Fig. 11.4 and reading about the descriptions of each of these 12 "likely" compounds we will quickly learn that only Desaminotyrosine (HMDB02199),

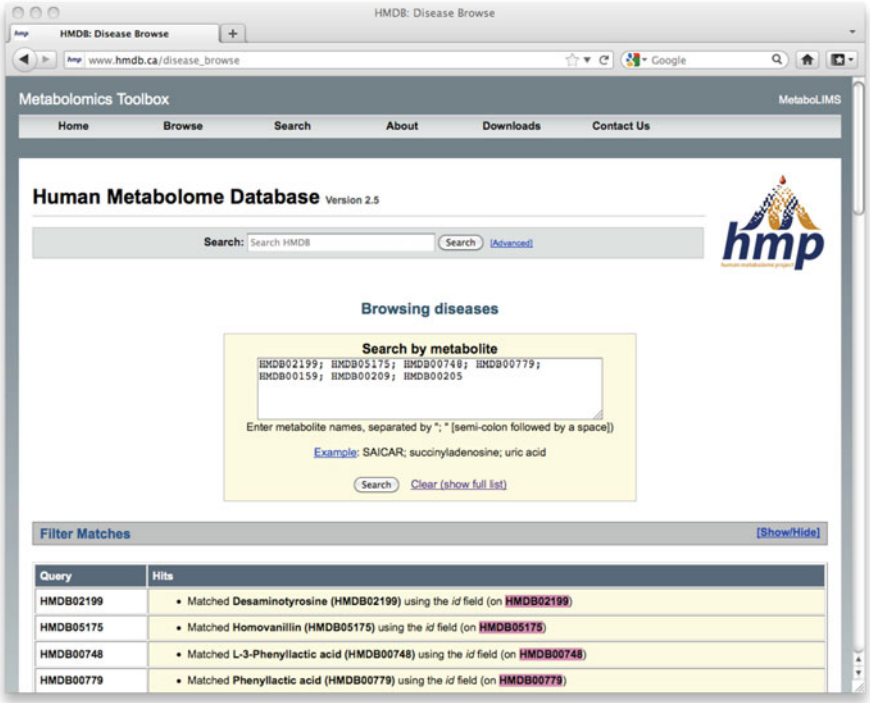


Fig. 11.5 A screenshot of the HMDB's "Disease Browse" web page. The "Search by Metabolite" box allows users to enter the names of metabolites or HMDB ID's to find metabolite matches to diseases

Homovanillin (HMDB05175), Phenyllactic acid (HMDB00748/HMDB00779), Phenylalanine (HMDB00159), Phenylacetic acid (HMDB00209) and Phenylpyruvate (HMDB00205) could be found, or have been previously reported to be, in human blood [26].

While mass matching is not the most reliable approach to compound identification in mass spectrometry [43] the appearance of multiple adducts along with further validation/checking via the HMDB as to whether they have been detected in human blood (via the "Biofluid Browse" option under the "Browse" menu) does add a considerable degree of confidence to their identifications. From this shortened list of candidate metabolites we can start to make use of the HMDB's other utilities, namely "Disease Browse" and "Pathway Browse". If we go to HMDB's top menu bar and click on the Menu item "Browse" and select the submenu "Disease Browse" we will see the following web page (Fig. 11.5). The box "Search by metabolite" allows one to enter the name or HMDB identifiers of a set of metabolites separated by a semi-colon. Enter the list as: HMDB02199; HMDB05175; HMDB00748; HMDB00779; HMDB00159; HMDB00209; HMDB00205 and press the "Search" button. The resulting table will display a number of diseases which exhibit altered levels of these metabolites in blood, urine, CSF and other biofluids/tissues.

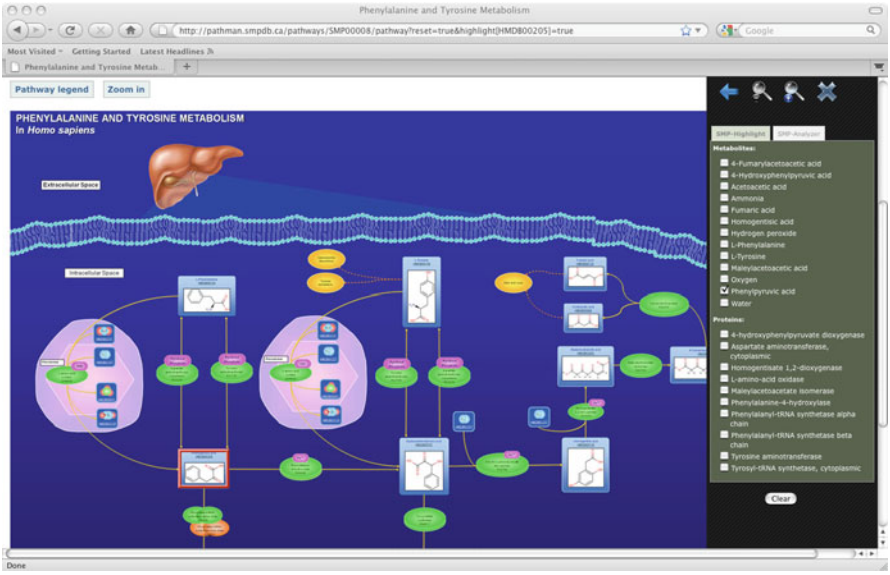


Fig. 11.6 A screenshot of the Phenylalanine and Tyrosine Metabolism pathway from HMDB/SMPDB. This pathway shows the organs (liver), cellular compartments (peroxisome, membrane, extracellular compartments), proteins (including quaternary structure and cofactors) and the chemical structure of each of the compounds in the pathway

Several diseases such as Alzheimer's, Dengue fever, Epilepsy, Bacterial Infections, Kidney disease, Hypothyroidism, Leukemia etc. are listed but almost all just match to one metabolite. The one exception is Phenylketonuria, which has five metabolite matches including Phenylalanine, Phenylacetic acid, Phenyllatic acid, 3-Phenyllactic acid and Phenylpyruvic acid. Clicking on the hyperlink (OMIM: 261600) or on the associated references listed on the right of the table will allow one to learn much more about the chemistry, biochemistry and genetics of phenylketonuria. From the OMIM link we can learn that "*Phenylketonuria (PKU) is an autosomal recessive inborn error of metabolism resulting from a deficiency of phenylalanine hydroxylase, an enzyme that catalyzes the hydroxylation of phenylalanine to tyrosine, the rate-limiting step in phenylalanine catabolism. If undiagnosed and untreated, phenylketonuria can result in impaired postnatal cognitive development resulting from a neurotoxic effect of hyperphenylalaninemia*". Likewise, clicking on the HMDB ID hyperlinks for each of these compounds and reading the compound descriptions will give a clearer idea about how and why these metabolites are generated in the body.

If we now go to HMDB's "Browse" menu and select "Pathway Browse" and enter the same set of metabolites in the "Search by Metabolite" text box we will get matches to three pathways: Phenylalanine and Tyrosine Metabolism, Transcription/Translation and Tyrosine Metabolism. An image of the Phenylalanine and Tyrosine Metabolism pathway is shown in Fig. 11.6. This pathway diagram shows the organs

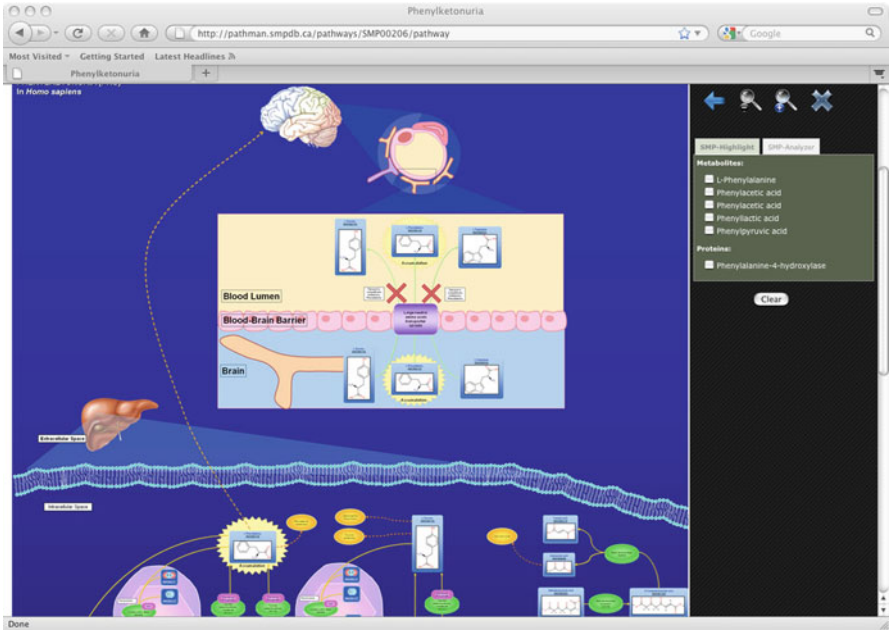


Fig. 11.7 A screenshot of the Phenylketonuria pathway from SMPDB. This pathway identifies the key metabolites in PKU and illustrates how these metabolites target the brain

(liver), cellular compartments (peroxisome, membrane, extracellular compartments), proteins (including quaternary structure and cofactors) and the chemical structures of each compound in the pathway. Each image/structure is hyperlinked to either the HMDB (if it is a chemical) or to UniProt (if it is a protein). Clicking the protein icon corresponding to phenylalanine hydroxylase (PAH or P00439) will generate the UniProt page for this particular protein. Scrolling down through the UniProt page will show the nearly 100 known PAH mutants or variants (see Natural variations) that have been catalogued and which are associated with PKU. The pathway images in HMDB include many of the same images found in SMPDB. By navigating to the SMPDB (www.smpdb.ca) and entering "phenylketonuria" into the search menu, the pathway for this disease (SMP00206) can be selected (by clicking the "Pathway" button) and viewed (Fig. 11.7). The PKU pathway includes all of the compounds identified by our "hypothetical" UPLC-MS experiment and it highlights the affected enzymes, the key transporters and the effects that these metabolites have in the brain (the main site of toxicity due to excess plasma phenylalanine). Once again, all of the images are hyperlinked to more detailed data pages in the HMDB, DrugBank and UniProt.

As simple as this example may be, it is primarily intended to show how the resources developed for the Human Metabolome Project (HMP) can be used to take relatively raw experimental data (MS peak intensities and masses) and generate biologically meaningful or medically useful results. In particular, the HMDB has

been designed specifically to facilitate metabolomic studies "from bench to bedside". This capacity is further enhanced through the HMP's affiliate databases – DrugBank, T3DB, FooDB and SMPDB. Through the use of extensive hyperlinking, cross-referencing to external databases, carefully compiled textual references, hand-illustrated pathway diagrams, detailed image mapping and comprehensive annotations – the HMP's wide-ranging resources allow a remarkable breadth of "omics" queries to be asked and answered. This breadth of "omics" coverage is absolutely crucial since the essence of systems biology is to link all three levels of the "omics" pyramid (Fig. 11.1) and to merge them into a seamless continuum. While this grand vision of integrated biology has yet to be fully realized, it is perhaps fair to say that the products of the HMP are helping to bring this vision a little closer to reality.

References

1. Jakobs C, Schweitzer S, Dorland B (1995) Galactitol in galactosemia. Eur J Pediatr 154:S50–S52
2. Bory C, Boulieu R, Chantin C, Mathieu M (1990) Diagnosis of alcaptonuria: rapid analysis of homogentisic acid by HPLC. Clin Chim Acta 189:7–11
3. Kornberg H (2000) Krebs and his trinity of cycles. Nat Rev Mol Cell Biol 1:225–228
4. Kohler R (1982) From medical chemistry to biochemistry: the making of a biomedical discipline. Cambridge University Press, Cambridge
5. German JB, Hammock BD, Watkins SM (2005) Metabolomics: building on a century of biochemistry to guide human health. Metabolomics 1:3–9
6. Tietz NW (1995) Clinical guide to laboratory tests, 3rd edn. WB Saunders Press, Philadelphia
7. Wishart DS (2008) Metabolomics: applications to food science and nutrition research. Trends Food Sci Technol 19:482–493
8. Mahido C, Ruchirawat S, Prawat H et al (1998) Pure Appl Chem 70:2065–2072
9. Hamosh A, Scott AF, Amberger J et al (2002) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55
10. Katzmarzyk PT, Janssen I (2004) The economic costs associated with physical inactivity and obesity. Can J Appl Physiol 29:90–115
11. Johnson JA, Pohar SL, Majumdar SR (2006) Health care use and costs in the decade after identification of type 1 and type 2 diabetes: a population-based study. Diabetes Care 29:2403–2408
12. Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 279:1200–1205
13. Koren G (1993) Preconceptional folate and neural tube defects: time for rethinking. Can J Public Health 84:207–208
14. French AE, Grant R, Weitzman S et al (2003) Folic acid food fortification is associated with a decline in neuroblastoma. Clin Pharmacol Ther 74:288–294
15. Verster A (2004) Food fortification: good to have or need to have? East Mediterr Health J 10:771–777
16. Venter JC, Adams MD, Meyers EW et al (2001) The sequence of the human genome. Science 291:1304–1351
17. HUPO (2010) A gene-centric human proteome project: HUPO–the Human Proteome organization. Mol Cell Proteomics 9:427–429

18. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117:500–544
19. Wishart DS (2007) Proteomics and the human metabolome project. Expert Rev Proteomics 4:333–335
20. Wishart DS, Knox C, Guo AC et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37:D603–D610
21. Knox C, Law V, Jewison T et al (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39:D1035–D1041
22. Lim E, Pon A, Djoumbou Y et al (2010) T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res 38:D781–D786
23. Scalbert A, Andres-Lacueva C, Arita M et al (2011) Databases on food phytochemicals and their health-promoting effects. J Agric Food Chem 59:4331–4348
24. Wishart DS, Lewis MJ, Morrissey JA et al (2008) The human cerebrospinal fluid metabolome. J Chromatogr B Analyt Technol Biomed Life Sci 871:164–173
25. Guo K, Bamforth F, Li L (2011) Qualitative metabolome analysis of human cerebrospinal fluid by (13)c-/(12)c-isotope dansylation labeling combined with liquid chromatography fourier transform ion cyclotron resonance mass spectrometry. J Am Soc Mass Spectrom 22:339–347
26. Psychogios N, Hau DD, Peng J et al (2011) The human serum metabolome. PLoS One 6:e16957
27. Guo K, Li L (2009) Differential 12 C-/13 C-isotope dansylation labeling and fast liquid chromatography/mass spectrometry for absolute and relative quantification of the metabolome. Anal Chem 81:3919–3932
28. Eckburg PB, Bik EM, Bernstein CN et al (2005) Diversity of the human intestinal microbial flora. Science 308:1635–1638
29. Holmes E, Wilson ID, Nicholson JK (2008) Metabolic phenotyping in health and disease. Cell 134:714–717
30. Slupsky CM, Rankin KN, Wagner J et al (2007) Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Anal Chem 79:6995–7004
31. Bou Khalil M, Hou W, Zhou H et al (2010) Lipidomics era: accomplishments and challenges. Mass Spectrom Rev 29:877–929
32. Smith CA, O'Maille G, Want EJ et al (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27:747–751
33. Wishart DS (2009) Computational strategies for metabolite identification in metabolomics. Bioanalysis 1:1579–1596
34. Kanehisa M, Goto S, Hattori M et al (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354–D357
35. Krummenacker M, Paley S, Mueller L et al (2005) Querying and computing with BioCyc databases. Bioinformatics 21:3454–3455
36. Wang Y, Xiao J, Suzek TO et al (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623–W633
37. Degtyarenko K, de Matos P, Ennis M et al (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36:D344–D350
38. Westbrook J, Feng Z, Jain S et al (2002) The protein data bank: unifying the archive. Nucleic Acids Res 30:245–248
39. Boutet E, Lieberherr D, Tognolli M et al (2007) UniProtKB/Swiss-Prot. Methods Mol Biol 406:89–112
40. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2010) GenBank. Nucleic Acids Res 38:D46–D51
41. Wishart DS (2008) Applications of metabolomics in drug discovery and development. Drugs R&D 9:307–322
42. Wishart DS, Knox C, Guo AC et al (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901–D906
43. Sumner L, Amberg A, Barrett D et al (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3:211–221

Chapter 12
Understanding Cancer Metabolism Through Global Metabolomics

Michael V. Milburn, Kay A. Lawton, Jonathan E. McDunn, John A. Ryals, and Lining Guo

1 Introduction

Otto Warburg, is recognized as one of the key scientific generalists who was able to apply physical chemical data to better understand cancer cell metabolism [1]. In the 1910s he uncovered alterations in the intermediary metabolism of cancer cells that enabled cancer cell growth. [2] His early discoveries led to the idea that cancer cells exhibited a reverse Pasteur effect of glycolysis in which glucose was rapidly metabolized to lactate [3]. Thus, Warburg showed that the metabolism of cancer cells is fundamentally altered relative to normal cells. However, it has only been in the last 10 years that a number of investigators rediscovered that many oncogenes function through alterations in metabolism and that through this new metabolic understanding new cancer diagnostics and cancer treatments might be possible [4, 5] (Box 12.1).

M.V. Milburn, Ph.D. (✉) • L. Guo, Ph.D.
Research and Development, Metabolon, Inc,
617 Davis Drive Suite 400, Durham, NC 27713, USA
e-mail: mmilburn@metabolon.com; lguo@metabolon.com

K.A. Lawton, Ph.D.
Research and Development, Metabolon, Inc, 617 Davis Drive Suite 400,
Durham, NC 27713, USA
e-mail: klawton@metabolon.com

J.E. McDunn, Ph.D.
Oncology Research and Development, Metabolon, Inc,
617 Davis Drive Suite 400, Durham, NC 27713, USA
e-mail: jmcdunn@metabolon.com

J.A. Ryals, Ph.D.
Metabolon, Inc, 617 Davis Drive Suite 400, Durham, NC 27713, USA
e-mail: jryals@metabolon.com

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_12,

Box 12.1 Pasteur Effect

Louis Pasteur first showed that aerating yeasted broth causes yeast cell growth to increase while fermentation decreases the rate of yeast growth. This effect can be explained through two different biochemical pathways. Under low oxygen conditions, pyruvate produced by glycolysis is metabolized into lactate, ethanol and carbon dioxide (fermentation) producing only 2 moles of ATP per glucose molecule, while under sufficient oxygen conditions, fermentation is inhibited and pyruvate produces acetyl CoA which is metabolized by the Krebs Cycle producing 38 moles of ATP per mole of glucose. During fermentation increased glycolysis occurs to compensate for low ATP production. The inhibitory effect of oxygen on glucose flux (glycolysis) and fermentation is termed the "Pasteur effect". The "reverse (or negative) Pasteur effect" refers to the stimulatory effect of oxygen on fermentation. The "Warburg effect" refers to aerobic glycolysis whereby glucose is converted to lactic acid and ethanol in the presence of oxygen.

2 Global Metabolomics

This remarkable new interest in alterations of metabolism and how cancer cells metabolically transform from normal cells is also coinciding with our ability to instantaneously profile 1000s of biochemicals in cells, an approach called metabolomics. The word "metabolomics" (or "metabonomics") first appeared in journal articles in 2000. Only a few metabolomic scientific papers were published that year but by 2009 that number rose to over 1,300 published scientific papers reporting metabolomic results. The major challenge for metabolomics has been to develop a technology that can extract, identify, and quantitate the entire spectrum of small molecules (MW < 1,500 Da) in any biological sample. There are between 2,500 and 3,000 biochemicals synthesized in humans when one disregards complex lipids or peptides. Importantly, in any one sample matrix (i.e. blood, urine, tissue, etc.) there will always be fewer metabolites than the total number synthesized in the entire organism. Unfortunately, many uses of the word metabolomics cover rather limited attempts to study certain classes of molecules such as amino acids rather than the entire repertoire of available species. In this chapter we will refer to metabolomics as a technology to obtain as large a snapshot of biochemicals as possible.

"Global" or "unbiased" metabolomics has been plagued by difficulties stemming from the diverse physical properties of small molecules. These properties can vary greatly, with significant differences in solubilities and molecular weights affecting a small molecules ability to be measured and solublilized. A single chromatography method to separate all of the compounds is very difficult and even more difficult to analyze individual compounds without chromatographic separation. Further complications arise if studies are expected to be completed with a reasonable

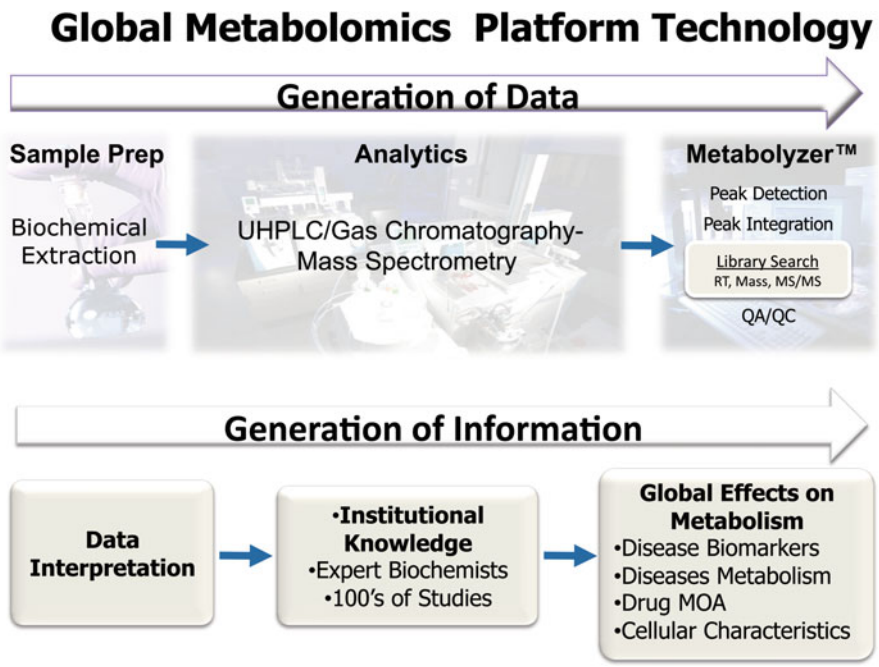


Fig. 12.1 The *top* half of the figure depicts the three steps of a global metabolomics method being applied to a biological sample. These three steps, biochemical extraction, multiple chromatography and mass spectrometry analysis, and then a unbiased global informatics methods to reduce the raw machine data to the biochemicals in the sample and the relative concentration of each biochemical in each sample. The *bottom* half of the analysis involves the data interpretation and statistical analysis that leads to the metabolic understanding available with this method

turn-around time. Methods that can only analyze a few samples per day will simply be impractical from a discovery technology and statistical standpoint. These issues are currently being addressed through advanced multi-system approaches where the best separation and detection instrument technologies are being developed to run in tandem. This approach allows for a comprehensive solution achieved by combining principles offered by various best-in-breed technologies. As this new technology develops and its use in biomarker detection studies increases, it is rapidly becoming clear that metabolomics will represent a high impact technology in various healthcare-related fields such as the diagnosis of disease, identification of drug targets, evaluation of the effects of drugs, and selection of patients most likely to respond to drug therapy (i.e. personalized medicine) [6–8].

One method developed for global metabolomics operates in essentially four steps, as shown in Fig. 12.1. [9] Step one is extraction of the small molecules from the biological sample. Step two is the chromatography coupled with mass spectrometry and data collection. Step three is the automated and manual QC analysis of the data using visual interfaced software [10]. Step four, the final step, is the statistical and biological interpretation of the data itself. In this method a wide range of very

polar to non-polar compounds can be measured from as little as 50 ul of blood plasma [9]. Extracted samples are split into four aliquots for different chromatography and mass spectrometry platforms, two UHPLC methods and one GC method, with one aliquot held in reserve. These three chromatography and MS systems complement each other in the range of biochemicals measured and provide an enhanced biochemical coverage of each sample. Approximately 70–80% of the biochemicals are measured on more than one platform with 30–40% measured on all three platforms. For compounds observed on multiple platforms, the platform with the best analytical characteristics (e.g. fewest interfering peaks or highest signal to noise) is generally used for the analysis of that compound. In general, the GC method provides better separation of molecules that tend to be more difficult to separate using a typical reverse phase LC method (e.g. carbohydrates).

After the raw data has been acquired from the instruments, this method utilizes a suite of software packages that automatically integrates each ion across retention time and then uses that ionic information, which may include additional MS/MS fragmentation information and retention time, to identify the compound [10]. After a compound is identified in a sample, one of the characteristic and stronger ions is used to determine a relative concentration of that compound in each sample. This approach assures that the compound will be represented only once in the subsequent statistical analysis. When the software has finished analyzing the samples, all of the data is loaded into a visual user interface that allows a scientist to curate the data for QC purposes and visually inspect how well each compound was identified and verify only those compounds with the highest degree of confidence for inclusion in the final data set. A variety of statistical approaches can be applied to the final data set at that point, including ANOVA, t-tests, random forest, PCA, etc. The goal of these types of statistical treatments is to identify the biochemicals that best represent the most significant changes in concentration between the groups in the study. One advantage of biochemistry is that multiple compounds in a particular biochemical pathway may often be significantly altered, giving an even higher degree of confidence to the importance of that biochemical change. In this respect, it is important to point out that most statistical treatments assume independent variables when, in fact, we know that certain biochemicals are related to the same or similar pathways. Metabolon is developing a large database of these types of biochemical changes as well as those that result from toxicity, drug mechanism, disease, etc. This knowledge enhances Metabolon's ability to provide a biological interpretation for each study it performs.

3 Understanding Cancer Metabolism

In cancer cells, metabolism is dramatically reprogrammed to support accelerated cell proliferation and adaption to the tumor microenvironment. Untargeted metabolomics is an excellent tool to probe the cancer-altered biochemical pathways to gain insights into pathogenesis and identify biomarkers and/or therapeutic targets. It has

been suggested that cancer cells alter metabolism to serve the following needs: accelerated energy production, biosynthesis of macromolecules, and maintenance of redox status [3]. Here we summarize the core metabolic pathways altered in cancer cells, and which are common to various cancer types.

3.1 Glucose Metabolism, Glutaminolysis, and IDH Mutation

Glucose plays a central role in energy generation. Elevated uptake of glucose and glycolysis is a hallmark of cancer, even under normal oxygen conditions. This phenomenon is referred to as the "Warburg effect" [11]. As shown in Fig. 12.2, glucose utilization in cancer cells has several major branch points: (1) lactate production, (2) citrate production, (3) TCA cycle, and (4) pentose phosphate pathway. Collectively these metabolic pathways contribute to energy production, *de novo* fatty acid and cholesterol biosynthesis, nucleotide biosynthesis, and NADPH generation.

In cancer cells, a majority of glucose (as much as 90% in the case of glioblastoma) [12] from the elevated glycolysis is converted to lactate, which is excreted from the cell. Although the precise mechanism of lactate production in conjugation with the Warburg effect has not been fully elucidated, several hypotheses have been recently proposed in the literature. First, the high rate of glycolysis and lactate may allow faster conversion of glucose metabolites for amino acid biosynthesis and the pentose phosphate pathway to support cell growth [13]. Second, the increased production of glucose-derived acid (mainly lactate) leads to microenvironmental acidosis. Since tumor cells are adapted to be resistant to acidic environments this provides the cancer cells a powerful growth advantage, allowing proliferation and invasion into the extracellular matrix of the surrounding host tissue [14]. Third, many solid tumors contain hypoxic regions, which have limited access to nutrients [15]. It has been demonstrated that glucose is preferentially used by hypoxic tumor cells to produce lactate, which is excreted as the energy source for oxygenated tumor cells. The existence of such "metabolic symbiosis" between hypoxic and aerobic cancer cells allows for more efficient glucose utilization [16].

The high rate of glycolysis is critical to provide the key metabolic intermediate, citrate, for *de novo* lipogenesis. It has been well established that lipogenesis is essential for the growth and proliferation of tumor cells [17]. In fact, due to their importance to tumor growth, glycolysis and lipogenesis have been proposed as cancer therapeutic targets [18]. Pyruvate produced by glycolysis enters the mitochondria where it is converted into citrate. Citrate is then exported out of mitochondria to the cytosol. In the cytosol, citrate acts as a precursor for fatty acid and cholesterol synthesis.

A portion of glucose can be further metabolized through the TCA cycle. However, due to the export of citrate for lipogenesis, TCA cycle intermediates need to be replenished to maintain the full function of oxidative phosphorylation. Cancer cells accomplish this by utilization of glutamine (glutaminolysis). Glutaminolysis also contributes to the production of lactate and NADPH [12].

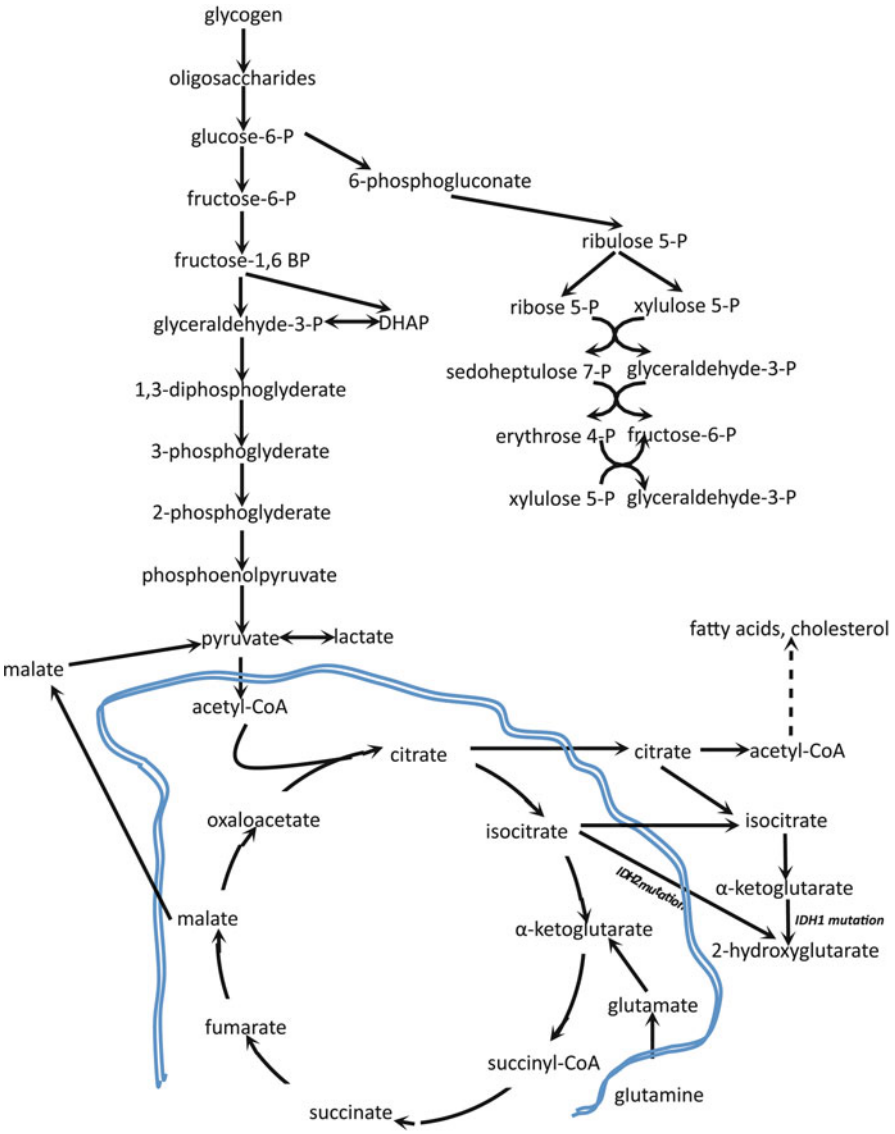


Fig. 12.2 Glycolysis, TCA cycle, glutaminolysis, pentose phosphate pathways, and 2-hydroxyglutarate production in cancer cells

The glycolytic intermediate fructose-6-phosphate can be shunted into the pentose phosphate pathway (PPP). The PPP supplies the majority of cellular NADPH, which is essential for reductive biosynthesis, such as fatty acid biosynthesis. NADPH is also critical to maintain the level of reduced glutathione and mediate oxidative stress. The nonoxidative phase of the PPP produces ribose-5-phosphate for nucleotide

biosynthesis [19]. It was recently found that p53, a tumor suppressor, inhibits the PPP by direct binding to glucose-6-phosphate dehydrogenase, thereby suppressing glucose consumption and NADPH production [20].

In the last couple of years, much excitement has been generated with regard to the presence of elevated levels of 2 hydroxyglutarate (2-HG) in association with human brain cancers. Genetic studies have identified that two (IDH1 and IDH2) of the three isoforms of isocitrate dehydrogenase (IDH) are mutated in a high proportion of gliomas of intermediate malignant grade [21] and acute myeloid leukemia's [22, 23]. An IDH mutation has also been reported in prostate cancer, [24] colorectal cancer, [25] and melanoma [26]. Eukaryotic cells contain two classes of IDH enzymes (NAD^+- and $NADP^+$-dependent) which convert isocitrate to α-ketoglutarate. The two $NADP^+$-dependent forms, IDH1 and IDH2 are located in cytosol and mitochondria, respectively. The NAD^+-dependent IDH3 is located in the mitochondria and is part of the TCA cycle. It has been demonstrated that mutations to IDH1 altered the enzyme function to produce 2-HG instead of α-ketoglutarate [27]. The significance of IDH mutation and the function of 2-HG have not been fully elucidated. The IDH mutation has been reported to be associated with decreased cell proliferation [28] and longer survival rate in low-grade gliomas [21, 29].

3.2 Lipid Metabolism

Another metabolic signature of cancer is accelerated phospholipid biosynthesis, as proliferating cells have a significant need for membrane production. Elevation of fatty acids, as well as precursors for phospholipid head groups, such as choline, phosphocholine, CDP-choline, ethanolamine, and phosphoethanolamine, are often observed in tumor tissues. The increases of fatty acids are accompanied by the triacylglycerol catabolites, monoacylglycerol and glycerol, suggesting that in addition to *de novo* biosynthesis, lipolysis also contributes to elevated free fatty acids. Consistent with the metabolomics observations, monoacylglycerol lipase, an enzyme that hydrolyses, triacylglycerols during lipolysis, has been found to promote cancer pathogenesis [30]. Furthermore, choline kinase, which catalyzes the phosphorylation of choline to phosphorylcholine for phospholipid biosynthesis, has been identified as a target for cancer therapeutics [31, 32].

3.3 Antioxidant and Tryptophan Catabolism

The glutathione pathway is important in tumor development, as it is involved in protection from oxidative stress, a typical characteristic of rapidly growing cells. Glutathione provides a substrate for neutralization of reactive oxygen species, particularly hydrogen peroxide. The glutathione (GSH) content of cancer cells is particularly relevant in regulating mutagenic mechanisms, DNA synthesis, growth, and

multidrug and radiation resistance. In malignant tumors, it has been shown that metastatic cells with high GSH levels survive the combined nitrosative and oxidative stresses elicited by the host tissues [33]. NADPH is required for recycling glutathione, and this is mostly provided by the pentose phosphate pathway and glutaminolysis.

Tryptophan catabolism is another pathway that mediates cancer cells' adaptation to their microenvironment. Tryptophan can be converted to kynurenine by indolamine-2,3-dioxygenase (IDO), which further leads to the production of NAD. In cancer cells, the increases of kynurenine, quinolinate, and NAD are frequently observed, suggesting that the kynurenine pathway is up regulated. The induction of IDO has been found to serve anti-inflammatory functions [34, 35]. IDO has recently gained considerable attention since it has been shown to be expressed in a variety of human cancers [36–39]. It has been suggested that IDO may be exploited by tumors as a mechanism of immune evasion [25, 37, 40, 41]. Numerous studies have identified increased IDO expression as an independent prognostic variable for reduced overall survival in cancer patients [42]. It has been further proposed that selective inhibition of IDO1 may represent an attractive cancer therapeutic strategy [43].

4 Applications of Metabolomics in Cancer Research

Cancer cells are a particularly attractive target for metabolomics since many cellular pathways are up regulated or altered to meet the special needs of cancer cells (e.g. enhanced glycolysis, utilization of energy producing pathways for production of biochemical building blocks, DNA repair, genome stability, telomere maintenance), leading to large changes in endogenous biochemicals that can then be readily quantified. Additionally, since many pathways are altered in cancer cells a global metabolomics approach quantifying both polar and non-polar biochemicals occurring in many pathways, would seem to be particularly well suited. The utility of metabolomics in cancer research has been demonstrated in drug discovery/development and cancer biomarker discovery. Recent examples include the cancer drug GMX1778 where the target and mechanism of action were unraveled using metabolomics and prostate cancer where biomarkers of cancer aggressiveness were identified.

4.1 Mechanism of Cancer Drug Action: GMX1778

One of the more fascinating and important contributions of global metabolomics to understanding cancer metabolism has been the elucidation of a cancer drug target and mechanism of action. GMX1777 is a soluble pro-drug that is rapidly converted in vivo to GMX1778 (CHS828), the active cyanoguanidinopyridine. Although there are many published studies of GMX1778 since its discovery, the actual molecular target of the drug has been elusive. Potent broad spectrum anti-tumor activity has

been demonstrated in several tumor types evaluated in vitro in a large cell panel and in vivo in multiple human xenograft models. The mechanism of action of this small molecule was believed to involve NF-KB inhibition [44]. However, substantial NF-KB inhibition did not occur until 24 h after treatment with GMX1778 suggesting that NF-KB inhibition might be secondary to some other primary drug action. In an attempt to identify the primary mechanism of action and protein target of the drug, a variety of unbiased proteomic techniques and molecular biology were employed to no avail. Therefore, to attempt to discover the anticancer mechanism a global metabolomics analysis was performed to identify intracellular physiological changes over time.

One of the most important steps in a mechanism of action study is to appropriately design the sample collection for the study. For most mechanism of action studies it is critically important to evaluate a time course in order to separate the primary effects of the drug from downstream secondary or tertiary effects. Since GMX1778 has maximum activity on a multiple myeloma cell line in 24 h samples were collected at several timepoints before the maximum activity. IM-9, a sensitive multiple myeloma cell line, was treated with 30 nM GMX1778 or with DMSO (control) for 6-, 13-, 20-, or 27-h in RPMI-1640 media (10% FBS and 0.3 mg/mL L-glutamine), n = 6. Frozen cell pellets (2×10^6 cells) were analyzed using a previously published extraction schema followed by GC-MS and LC-MS metabolomics analysis. The relative standard deviation (RSD) value for a technical replicate of pooled aliquots from the cell samples was 10%. This 10% relative standard deviation represents the total process variation of extraction, chromatography, and quantitation for the biochemicals measured. Using a p-value cutoff of 0.1 and q-value cutoff of 0.2, there were 27, 46, 65, and 65 biochemicals altered relative to the DMSO controls at 6, 13, 20, and 27 h, respectively [45]. Although more biochemicals changed at the later time points a number of significant biochemical pathway alterations occurred as early as the 6 h timepoint. All biochemical changes were interpreted physiologically and in a biochemical pathway context using in-house pathways and a biochemical knowledge data base to expedite this process.

The most significant alteration at the earliest timepoint was in the level of NAD. A 60% decrease in intracellular nicotinamide adenine dinucleotide (oxidized) (NAD+) levels was observed after just 6 h treatment with GMX1778 [45]. At 13 h the levels declined by 91% relative to the control and the later time points were below detection limits. Since this was the most significant early effect of the drug we asked whether other biochemical changes that occurred later could be a result of this primary effect on the levels of NAD.

NAD+ is a cofactor in oxidation-reduction reactions, including ATP generation from glycolysis and oxidative phosphorylation. NAD+ is a substrate for reactions catalyzed by poly (ADP-ribose) polymerase, sirtuins and ADP-ribosyl cyclase. Since NAD is required by three enzymes in the Krebs Cycle we asked whether any of the Krebs Cycle intermediates increased as a result of this NAD inhibition. Indeed, fumarate increased 36%, 79%, 335%, and 244% and malate increased 48%, 87%, 413%, and 268% at the four timepoints measured supporting the idea that the drug inhibited NAD. Additionally pathways involved in glycolysis and alternative glucose metabolism were also supportive of the idea of NAD inhibition.

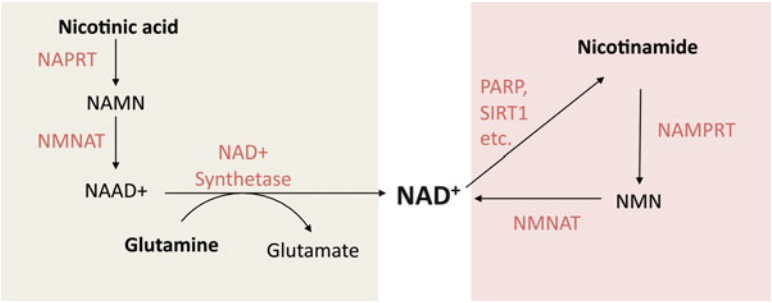


Fig. 12.3 Two independent pathways for synthesizing NAD used in mammals. Either Niacin or Nicotinamide is used as starting biochemicals to synthesize NAD. In the validation experiment the study took advantage of these two independent pathways to validate that the drug inhibited only the Nicotinamide pathway and not the Niacin pathway

The next question proposed was how one could validate in an independent experiment that GMX1778 is indeed an inhibitor of NAD biosynthesis. As shown in Fig. 12.3, in mammals there are two independent pathways for synthesizing NAD using either Niacin or Nicotinamide as substrates. At least three enzymes in these NAD + biosynthetic pathways are potential targets of GMX1778: NAD + synthetase, nicotinamide mononucleotide adenylyl transferase (NMNAT), or nicotinamide phosphoribosyl transferase (NAMPRT). Inhibition of any of these enzymes could account for the observed decreased NAD + levels. Specific experiments were performed to identify whether GMX1778 inhibited the Niacin or Nicotinamide pathway and to determine which of these three targets was most important for drug action. GMX1778-treated cells were rescued by nicotinic acid as seen in Fig. 12.4. [46] Additional experiments demonstrated that GMX1778 inhibits NAMPRT, the rate-limiting enzyme that converts nicotinamide (NAM) to nicotinamide mononucleotide (NMN) in vitro and in vivo [46]. The apparent K_i of GMX1778 for NAMPRT was 1–3 nM. This correlates well with the low nanomolar IC_{50} of this compound in multiple human cell lines. Since the crystal structure of NAMPRT is known, it was also discovered that cell lines resistant to the drug were the result of a single amino acid mutation in the active site of the enzyme and presumably interfered with drug binding.

This study illustrates how a global metabolomic approach found the biochemical needle (NAD+) in the hay stack (metabolome containing hundreds of endogenous biochemicals). The study also demonstrated not only how this technology can uncover the mechanism of a drug but also the importance of potential metabolic targets for new cancer therapies. Tumor cells have elevated NAMPRT and a high rate of NAD + turnover due to high ADP-ribosylation activity required for DNA repair, genome stability, and telomere maintenance, making them more susceptible to NAMPRT inhibition than normal cells [47]. This novel mechanism supports the clinical use of GMX1777 as an anti-cancer agent and further supports these types of anticancer targets.

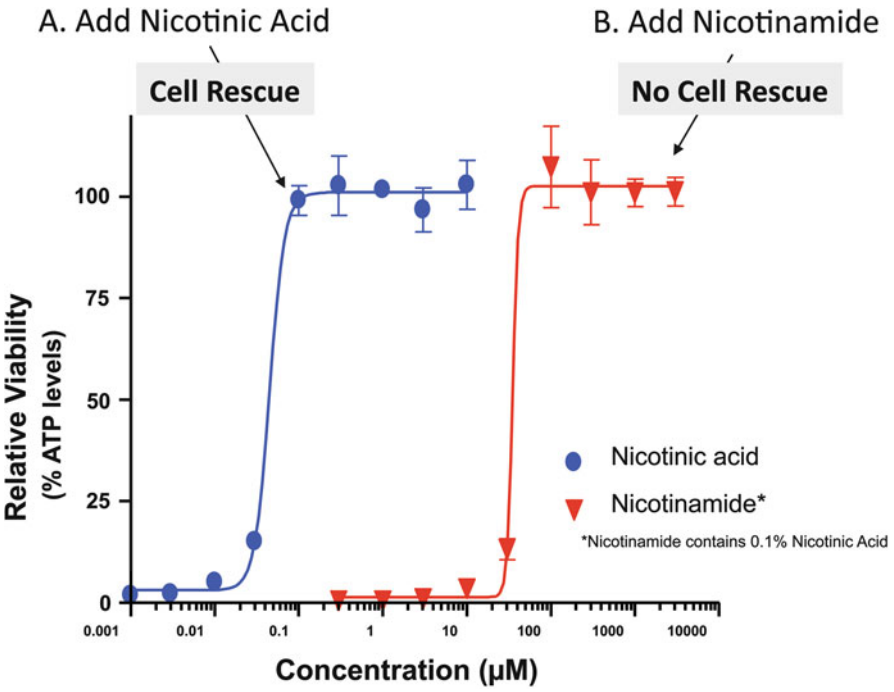


Fig. 12.4 GMX1778-treated cells were rescued only by nicotinic acid

4.2 Metabolism in Aggressive Cancer

A global metabolomic study to understand alterations that characterize neoplastic progression was investigated by profiling over 1,100 metabolites across 262 clinical samples related to prostate cancer from tissue, urine, and plasma [48]. The metabolic differences between benign prostate tissue, prostate cancer, and metastatic prostate cancer were quite striking. However, the differences in the metabolomic profiles of cancer versus non-cancer obtained from plasma and urine were less robust, presumably because of the distal location and dilution effects from the prostate. Between benign tissue and prostate cancer tissue we identified 87 out of 518 metabolites that had p-values of less than 0.05 and corresponded to a 23% false discovery rate. Of these metabolites, 50 were increased in prostate cancer and 37 were decreased in prostate cancer. In the metastatic tumor samples 124 metabolites were significantly altered as compared to localized prostate tumors, with the majority being decreased in concentration. Interestingly, several metabolites that are specifically produced by the prostate tissue, spermine, spermindine, and citrate, are dramactically decreased in prostate cancer suggesting specific metabolism changes from the normal tissue function.

Of primary interested were metabolites that significantly increased with the progression from benign to prostate cancer and then further progressed as the tissue

became metastatic. Six metabolites had this profile including, sarcosine, uracil, kynurenine, glycerol-3-phosphate, leucine, and proline. Of these, sarcosine was the most significantly increased as the disease progressed. Sarcosine is a fairly uncommonly studied amino acid derivative that is simply N-methyl-glycine to which a methyl group has been added to the primary amine of glycine. Sarcosine is fundamentally involved in two pathways, one involves metabolism of choline and the other is the direct synthesis of sarcosine via the methylation of glycine.

To confirm the importance of sarcosine, a highly sensitive and quantitative gas chromatography–mass spectrometry (GC-MS) assay was developed for sarcosine using an isotope dilution method. A completely independent cohort of 89 tissue samples were analyzed to validate the global metabolomics discovery analysis. Not only were the quantitative differences between benign, prostate cancer, and metastatic cancer significant, but, as expected for this more sensitive sarcosine assay the differences between each tissue were more dramatic. Given that the sarcosine data in this independent cohort validated the discovery study using global metabolomics, the paper goes on to investigate the potential mechanistic role of sarcosine in prostate cancer.

To understand the potential mechanistic role of sarcosine various prostate cancer cell lines were tested for their level of sarcosine and cell invasiveness. It was found that not only was sarcosine elevated in these cell lines as compared to benign cells but that it also correlated to the level of cell invasiveness. Furthermore, RNAi experiments that knocked down or increased the levels of sarcosine were shown to manipulate the level of cell invasiveness which tracked with the level of sarcosine. Finally, this paper showed that sarcosine directly added to these cell lines lead to an increase in invasiveness.

Overall this was an excellent example of employing global metabolomics to understand the metabolic differences in prostate cancer progression. From this work, sarcosine was identified as not only a validated tissue marker of increased prostate cancer progression but also mechanistically linked to increasing cell aggressitivity. Further work will be required to understand how this increase in sarcosine leads to increased cell aggressitivity but it clearly demonstrates that changing the metabolism of cells can dramatically alter their cancer state.

References

1. Warburg O, Wind F, Negelein E (1927) The metabolism of tumors in the body. J Gen Physiol 8:519–530
2. Stubbs M, Griffiths JR (2010) The altered metabolism of tumors: HIF-1 and its role in the Warburg effect. Adv Enzyme Regul 50:44–55
3. Cairns RA, Harris IS, Mak TW (2011) Regulation of cancer cell metabolism. Nat Rev Cancer 11:85–95
4. Riefke B, Mumberg D, Kroemer G et al (2007) Preface. In: Keun K, Steger-Hartmann T, Petersen K et al (eds) Oncogenes meet metabolism. From deregulated genes to a broader understanding of tumour physiology. Springer, Berlin

5. Dang CV, Lewis BC, Dolde C, Dang G, Shim H (1997) Oncogenes in tumor metabolism, tumorigenesis, and apoptosis. J Bioenerg Biomembr 29:345–354
6. Zhang Y, Dai Y, Wen J et al (2011) Detrimental effects of adenosine signaling in sickle cell disease. Nat Med 17:79–86
7. Takei M, Ando Y, Saitoh W et al (2010) Ethylene glycol monomethyl ether-induced toxicity is mediated through the inhibition of flavoprotein dehydrogenase enzyme family. Toxicol Sci 118:643–652
8. Barnes VM, Teles R, Trivedi HM et al (2010) Assessment of the effects of dentifrice on periodontal disease biomarkers in gingival crevicular fluid. J Periodontol 81:1273–1279
9. Evans AM, Dehaven CD, Barrett T, Mitchell M, Milgram E (2009) Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal Chem 81:6656–6667
10. Dehaven CD, Evans AM, Dai H, Lawton KA (2010) Organization of GC/MS and LC/MS metabolomics data into chemical libraries. J Cheminf 2:9
11. Scatena R, Bottoni P, Pontoglio A, Giardina B (2010) Revisiting the Warburg effect in cancer cells with proteomics. The emergence of new approaches to diagnosis, prognosis and therapy. Proteomics Clin Appl 4:143–158
12. Deberardinis RJ, Mancuso A, Daikhin E et al (2007) Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. Proc Natl Acad Sci USA 104:19345–19350
13. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324:1029–1033
14. Gatenby RA, Gillies RJ (2004) Why do cancers have high aerobic glycolysis? Nat Rev Cancer 4:891–899
15. Hockel M, Vaupel P (2001) Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. J Natl Cancer Inst 93:266–276
16. Sonveaux P, Vegran F, Schroeder T et al (2008) Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice. J Clin Invest 118:3930–3942
17. Menendez JA, Lupu R (2007) Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. Nat Rev Cancer 7:763–777
18. Hagland H, Nikolaisen J, Hodneland LI et al (2007) Targeting mitochondria in the treatment of human cancer: a coordinated attack against cancer cell energy metabolism and signalling. Expert Opin Ther Targets 11:1055–1069
19. Tong X, Zhao F, Thompson CD (2009) The molecular determinants of de novo nucleotide biosynthesis in cancer cells. Curr Opin Genet Dev 19:32–37
20. Jiang P, Du W, Wang X et al (2011) p53 regulates biosynthesis through direct inactivation of glucose-6-phosphate dehydrogenase. Nat Cell Biol 13:310–316
21. Yan H, Parsons DW, Jin G et al (2009) IDH1 and IDH2 mutations in gliomas. N Engl J Med 360:765–773
22. Ducray F, Marie Y, Sanson M (2009) IDH1 and IDH2 mutations in gliomas. N Engl J Med 360:2248–2249
23. De Carli E, Wang X, Puget S (2009) IDH1 and IDH2 mutations in gliomas. N Engl J Med 360:2248–2249
24. Kang MR, Kim MS, Oh JE et al (2009) Mutational analysis of IDH1 codon 132 in glioblastomas and other common cancers. Int J Cancer 125:353–355
25. Sjoblom T, Jones S, Wood LD et al (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314:268–274
26. Lopez GY, Reitman ZJ, Solomon D et al (2010) IDH1(R132) mutation identified in one human melanoma metastasis, but not correlated with metastases to the brain. Biochem Biophys Res Commun 398:585–587
27. Dang L, White DW, Gross S et al (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 462:739–744

28. Bralten LB, Kloosterhof NK, Balvers R et al (2011) IDH1 R132H decreases proliferation of glioma cell lines in vitro and in vivo. Ann Neurol 69:455–463
29. Houillier C, Wang X, Kaloshi G et al (2010) IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. Neurology 75:1560–1566
30. Nomura DK, Long JZ, Niessen S et al (2010) Monoacylglycerol lipase regulates a fatty acid network that promotes cancer pathogenesis. Cell 140:49–61
31. Janardhan S, Srivani P, Sastry GN (2006) Choline kinase: an important target for cancer. Curr Med Chem 13:1169–1186
32. Glunde K, Serkova NJ (2006) Therapeutic targets and biomarkers identified in cancer choline phospholipid metabolism. Pharmacogenomics 7:1109–1123
33. Estrela JM, Ortega A, Obrador E (2006) Glutathione in cancer biology and therapy. Crit Rev Clin Lab Sci 43:143–181
34. Sorensen RB, Hadrup SR, Svane IM et al (2011) Indoleamine 2,3-dioxygenase specific, cytotoxic T cells as immune regulators. Blood 117:2200–2210
35. Sas K, Robotka H, Toldi J, Vecsei L (2007) Mitochondria, metabolic disturbances, oxidative stress and the kynurenine system, with focus on neurodegenerative disorders. J Neurol Sci 257:221–239
36. Kallberg E, Wikstrom P, Bergh A, Ivars F, Leanderson T (2010) Indoleamine 2,3-dioxygenase (IDO) activity influence tumor growth in the TRAMP prostate cancer model. Prostate 70:1461–1470
37. Leung BS, Stout LE, Shaskan EG, Thompson RM (1992) Differential induction of indoleamine-2,3-dioxygenase (IDO) by interferon-gamma in human gynecologic cancer cells. Cancer Lett 66:77–81
38. Karanikas V, Zamanakou M, Kerenidi T et al (2007) Indoleamine 2,3-dioxygenase (IDO) expression in lung cancer. Cancer Biol Ther 6:1258–1262
39. Prendergast GC, Metz R, Muller AJ (2010) Towards a genetic definition of cancer-associated inflammation: role of the IDO pathway. Am J Pathol 176:2082–2087
40. Macchiarulo A, Camaioni E, Nuti R, Pellicciari R (2009) Highlights at the gate of tryptophan catabolism: a review on the mechanisms of activation and regulation of indoleamine 2,3-dioxygenase (IDO), a novel target in cancer disease. Amino Acids 37:219–229
41. Lee SY, Choi HK, Lee KJ et al (2009) The immune tolerance of cancer is mediated by IDO that is inhibited by COX-2 inhibitors through regulatory T cells. J Immunother 32:22–28
42. Inaba T, Ino K, Kajiyama H et al (2010) Indoleamine 2,3-dioxygenase expression predicts impaired survival of invasive cervical cancer patients treated with radical hysterectomy. Gynecol Oncol 117:423–428
43. Liu X, Newton RC, Friedman SM, Scherle PA (2009) Indoleamine 2,3-dioxygenase, an emerging target for anti-cancer therapy. Curr Cancer Drug Targets 9:938–952
44. Olsen LS, Hjarnaa PJ, Latini S et al (2004) Anticancer agent CHS 828 suppresses nuclear factor-kappa B activity in cancer cells through downregulation of IKK activity. Int J Cancer 111:198–205
45. Watson M, Roulston A, Belec L et al (2009) The small molecule GMX1778 is a potent inhibitor of NAD+biosynthesis: strategy for enhanced therapy in nicotinic acid phosphoribosyltransferase 1-deficient tumors. Mol Cell Biol 29:5872–5888
46. Roulston A, Watson M, Bernier C et al (2007) GMX1777: a novel inhibitor of NAD+biosynthesis via inhibition of nicotinamide phosphoribosyl transferase. American Association of Cancer Research-NCI-EORTC international conference on molecular targets and cancer therapeutics [Online]
47. Beauparlant P, Bedard D, Bernier C et al (2009) Preclinical development of the nicotinamide phosphoribosyl transferase inhibitor prodrug GMX1777. Anticancer Drugs 20:346–354
48. Sreekumar A, Poisson LM, Rajendiran TM et al (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457:910–914

Chapter 13
Genetic and Metabolic Determinants of Fatty Acid Chain Length and Desaturation, Their Incorporation into Lipid Classes and Their Effects on Risk of Vascular and Metabolic Disease

Thomas Kopf, Markus Peer, and Gerd Schmitz

1 Fatty Acid Metabolism

Fatty acids (FA) play an essential role in many cellular processes such as energy production and storage, membrane homeostasis, signalling and metabolic regulation. FA species influence membrane viscosity, anchoring of proteins to membranes, and synthesis of bioactive lipid mediators. FAs are either synthesised de novo or taken up as nutrients and essential FAs. Intestine, liver, adipose tissue and muscle are central organs regulating FA homeostasis, being directly connected to glucose homeostasis and insulin resistance [1, 2]. They depend on the delivery of carbon from the diet, and store excess FAs as acylglycerols or cholesteryl esters in lipid droplets or release them into the circulation.

1.1 Synthesis and Metabolism of FAs

FAs are synthesized de novo through fatty acid synthase (FAS, Fig. 13.1). FAS consists of two identical monomers resembling, complex multifunctional polypeptides encoded by a single gene, with each monomer containing the six catalytic activities necessary for FA synthesis [3]. Only the dimeric form is functionally active [4]. During biosynthesis intermediate products do not diffuse from the complex, ultimately forming palmitate (C16:0) from eight acetyl-CoA building blocks. De novo

T. Kopf • M. Peer • G. Schmitz (✉)
Department of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Franz-Josef-Strauß-Allee 11, Regensburg, Bavaria 93053, Germany
e-mail: Thomas.Kopf@klinikum.uni-regensburg.de; Markus.Peer@klinikum.uni-regensburg.de; Gerd.Schmitz@klinikum.uni-regensburg.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_13,

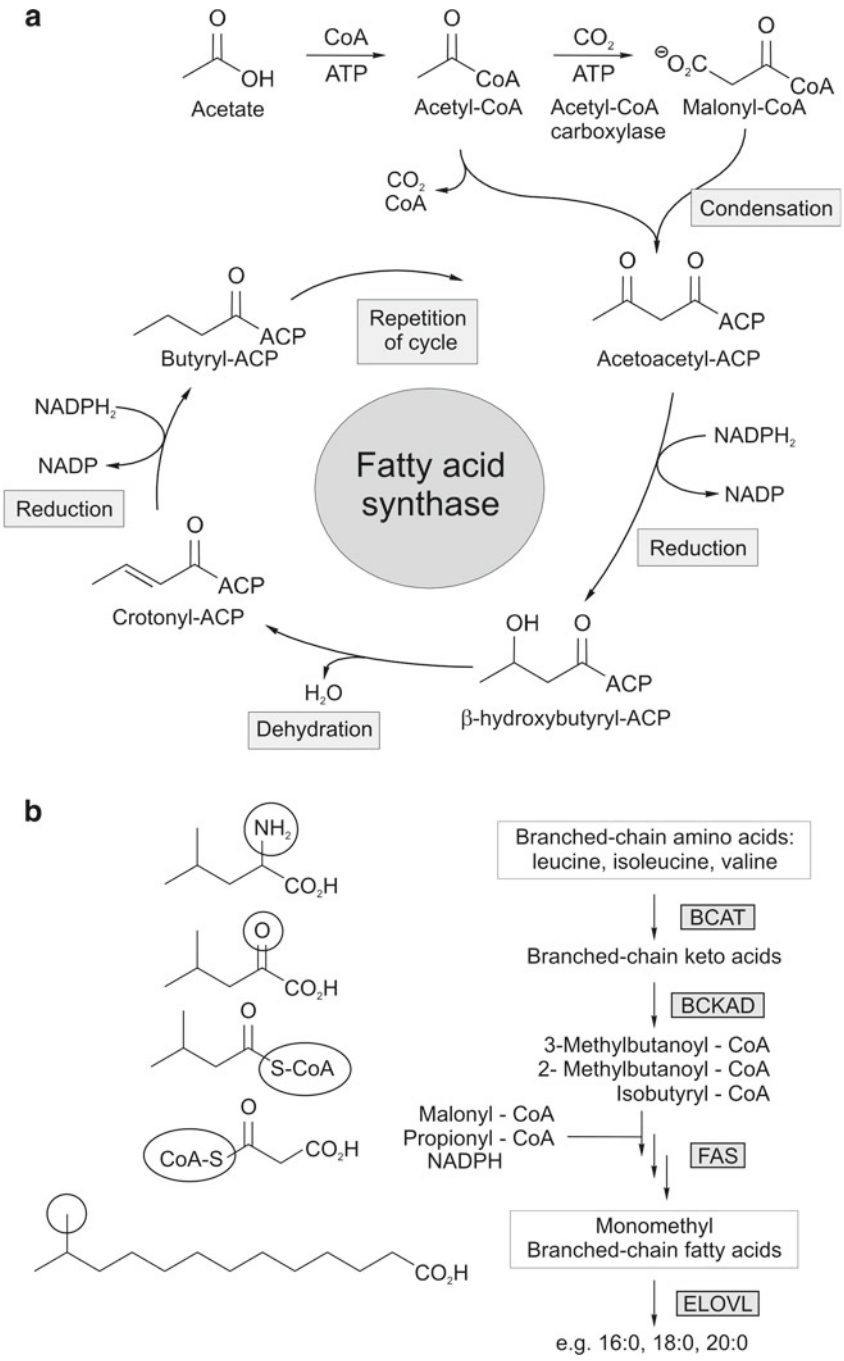


Fig. 13.1 (**a**) Fatty acid synthesis in the cytoplasm. The biosynthesis of fatty acids consists of four reactions per cycle (condensation-reduction-dehydration-reduction) that take place at the multifunctional enzyme fatty acid synthase (FAS), starting with malonyl-CoA. The endpoint of the synthesis is palmitate (C16:0). Fatty acids with longer chains and unsaturated fatty acids are synthesized through subsequent pathways by elongases and desaturases. Malonyl-CoA also regulates

synthesis is especially high during embryogenesis, in fetal lungs, in adult lactating breasts and in the endometrium [5]. Unsaturated and long-chain FAs are taken up through the diet (n-3 and n-6 series) and further processed through elongases and desaturases summarized in Table 13.1.

Increased FA synthesis also has been observed in tumor cells as a result of elevated activity of ATP citrate lyase and fatty acid synthase [6, 7].

1.1.1 Synthesis of Fatty Acids

Under physiological conditions, about 90% of FAs in mammalian cells originate from de novo synthesis. The biochemically important FAs have a chain length between 14 and 20 carbon atoms. De novo synthesis occurs mainly in the liver, where glucose is converted to pyruvate and then to citrate, which is converted to acetyl-CoA by ATP citrate lyase (Fig. 13.1a).

Acetyl-CoA carboxylase (ACC) is a multienzyme-complex that catalyzes the carboxylation to malonyl-CoA. Malonyl-CoA functions as intermediate in FA synthesis and as a regulatory effector controlling FA oxidation in liver and muscle by regulation of the entry of FAs into mitochondria [8]. There are two isoforms, ACC1 expressed primarily in lipogenic tissues, and ACC2 predominating in heart and skeletal muscle. Both ACCs catalyze formation of malonyl-CoA which serves as substrate for FAS, carrying out the first and key regulatory step in FA biosynthesis [9, 10]. There are four steps involved in an elongation cycle of FAs: condensation is the first step, which in case of the primary step of FA synthesis couples acetyl-CoA and malonyl-CoA, while tethered to the enzyme through the pantotheine arm. The second step is the reduction of the β-carbonyl-group through the β-ketoacyl-synthase-condensing enzyme to form the corresponding alcohol, followed by dehydration at this position as the third step to yield the α,β-unsaturated product. The last step is the reduction of the double bond forming the saturated FA. This cycle is repeated seven times, which ultimately releases palmitate (C16:0) from FAS by an intrinsic thioesterase activity (Fig. 13.1a).

Certain tissues, like skeletal and cardiac muscle, lack FAS. Instead, these tissues express ACC2 and malonyl-CoA decarboxylase (MCD), which removes malonyl-CoA by decarboxylation to yield CO_2 and acetyl-CoA. The ACC-catalyzed reaction is regulated by allosteric effectors and phosphorylation by 5′-AMP-dependent protein kinase.

Fig. 13.1 (continued) the fatty acid β-oxidation by inhibition of carnitine transport into mitochondria. *ACP* acyl carrier-protein. (**b**) Branched chain fatty acid biosynthesis with leucine, isoleucine and valine as precursors. Presented on the *left side* are structures of substrates and products, on the *right side* the mechanism of biosynthesis is depicted. *BCAT* branched-chain aminotransferase, *BCKAD* branched-chain keto-acid dehydrogenase, *ELOVL* elongation of very long chain fatty acids. Monomethyl branched chain fatty acids can be elongated with the elongation enzymes ELOVL5 and ELOVL6

Table 13.1 Desaturases and elongases involved in fatty acid and lipid metabolism including synonyms, chromosome location and ENSEMBL-No

ENSEMBL No.	Symbol	Name (ChrLoc)	Synonyms
ENSG00000149485	FADS1 (11q12.2-q13.1)	Fatty acid desaturase 1	D5D, Delta(5) fatty acid desaturase, FADS6, FADSD5
ENSG00000134824	FADS2 (11q12.2)	Fatty acid desaturase 2	D6D, Delta(6) fatty acid desaturase, FADSD6
ENSG00000221968	FADS3 (11q12.2)	Fatty acid desaturase 3	CYB5RP, Cytochrome b5-related protein
ENSG00000172782	FADS6 (17q25.1)	Fatty acid desaturase 6	–
ENSG00000099194	SCD (10q24.31)	Stearoyl-CoA desaturase	Delta(9)-desaturase, FADS5, SCD1, stearoyl CoA desaturase
ENSG00000145284	SCD5 (4q21.22)	Stearoyl-CoA desaturase 5	FADS4, SCD2, SCD4, stearoyl CoA 9-desaturase
ENSG00000066322	ELOVL1 (1p34.2)	Elongation of very long chain fatty acids yeast-like 1	3-keto acyl-CoA synthase, CGI-88
ENSG00000197977	ELOVL2 (6q14.1)	Elongation of very long chain fatty acids yeast-like 2	3-keto acyl-CoA synthase
ENSG00000119915	ELOVL3 (6p24.2)	Elongation of very long chain fatty acids yeast-like 3	3-keto acyl-CoA synthase, CGI30, cold-inducible glycoprotein of 30 kDa
ENSG00000118402	ELOVL4 (10q24.32)	Elongation of very long chain fatty acids yeast-like 4	3-keto acyl-CoA synthase
ENSG00000012660	ELOVL5 (6p12.1)	Elongation of long chain fatty acids like 5	3-keto acyl-CoA synthase
ENSG00000170522	ELOVL6 (4q25)	Elongation of long chain fatty acids like 6	3-keto acyl-CoA synthase, fatty acid elongase 2
ENSG00000164181	ELOVL7 (5q12.1)	Elongation of long chain fatty acids like 7	3-keto acyl-CoA synthase ELOVL7

Branched-chain FAs, including phytanic and pristanic acid, are involved in membrane stability and fluidity as well as anchoring of membrane leaflets and influence gene expression in many cell types [11]. The methyl-substituted amino acids leucine, isoleucine and valine are the substrates for the synthesis of monomethyl branched-chain FAs (Fig. 13.1b). The importance of monomethyl branched-chain FAs has been recently studied in *Caenorhabditis elegans* deficient of LET-767 (Lethal protein 767) which belongs to a family of short chain dehydrogenases/reductases, showing multiple developmental and growth defects [12]. LET-767 deficient worms were rescued with the feeding of triacylglycerides extracted from

other worms. Mass spectrometric analysis showed that odd-numbered monomethyl branched-chain FAs were essential in this rescue. LET-767 is mainly expressed in the intestine of the worms, indicative for the importance of branched-chain FAs in development and maintenance of the epithelial layers and the formation of crypts and villi in the intestinal mucosa.

1.1.2 Fatty Acid Oxidation

Fatty Acid β-oxidation in Mitochondria

The main energy reserve in the body consists of FAs, supplying energy-generating substrates through β-oxidation in mitochondria. Fatty acid β-oxidation (FAO) in mitochondria generates acetyl-CoA and reducing equivalents (NADH and $FADH_2$), which are linked to the Krebs cycle and the mitochondrial respiratory chain, leading to ATP production by oxidative phosphorylation in aerobic tissues. During starvation, FAO provides 80–90% of cellular energy requirements [13]. Almost all tissues rely essentially on FAO for their energy supply during prolonged fasting, but in contrast, cardiac and skeletal muscles derive most of their required energy from long-chain FA oxidation [14].

The transport of long-chain fatty acyl-CoA esters across the mitochondrial membrane is mediated by the carnitine cycle. The enzymes carnitine palmitoyltransferase I (CPT1), carnitine palmitoyltransferase II (CPT2), and carnitine-acylcarnitine translocase (CACT), each with different sub-mitochondrial localisations, are responsible for this transport [15, 16]. The CPT1 protein is located on the outer mitochondrial membrane and exists in two isoforms, a liver type (CPT1A) and a muscle type (CPT1B). They catalyse the formation of long-chain acylcarnitine from free L-carnitine and acyl-CoA esters. The CACT in the inner mitochondrial membrane translocates acylcarnitine into the mitochondrial matrix in exchange for free L-carnitine. CPT2 located in the inner mitochondrial membrane reesterifies fatty-acylcarnitines to fatty acyl-CoA esters, the substrates for ß-oxidation (Fig. 13.2a).

Carnitine acyltransferases (CRATs) catalyze the reversible transfer of acyl groups from acyl-CoA thioesters to carnitine, forming acylcarnitine. They differ in their substrate specificity: carnitine palmitoyltransferase, carnitine octanoyltransferase, and carnitine acetyltransferase. Since CRATs are responsible for the acyl-CoA/CoA ratio in mitochondria, peroxisomes, and ER, they are key enzymes of β-oxidation.

A vital role in mitochondrial β-oxidation of short chain FAs is played by three members of the acyl-CoA dehydrogenase (ACD) family: Short-chain L-3-hydroxyacyl-CoA dehydrogenase (SCAD) catalyzes the reversible dehydrogenation of 3-hydroxyacyl-CoAs to their corresponding 3-ketoacyl-CoAs with eduction of NAD to NADH and has the highest activity towards 3-hydroxybutyryl-CoA. Similar to this, medium-chain acyl-CoA dehydrogenase (MCAD) is involved in the metabolism of FAs with C4-C12-chains, while long-chain L3-hydroxy acyl-CoA dehydrogenase (LCAD) has an important function in mitochondrial β-oxidation of unsaturated FAs [17].

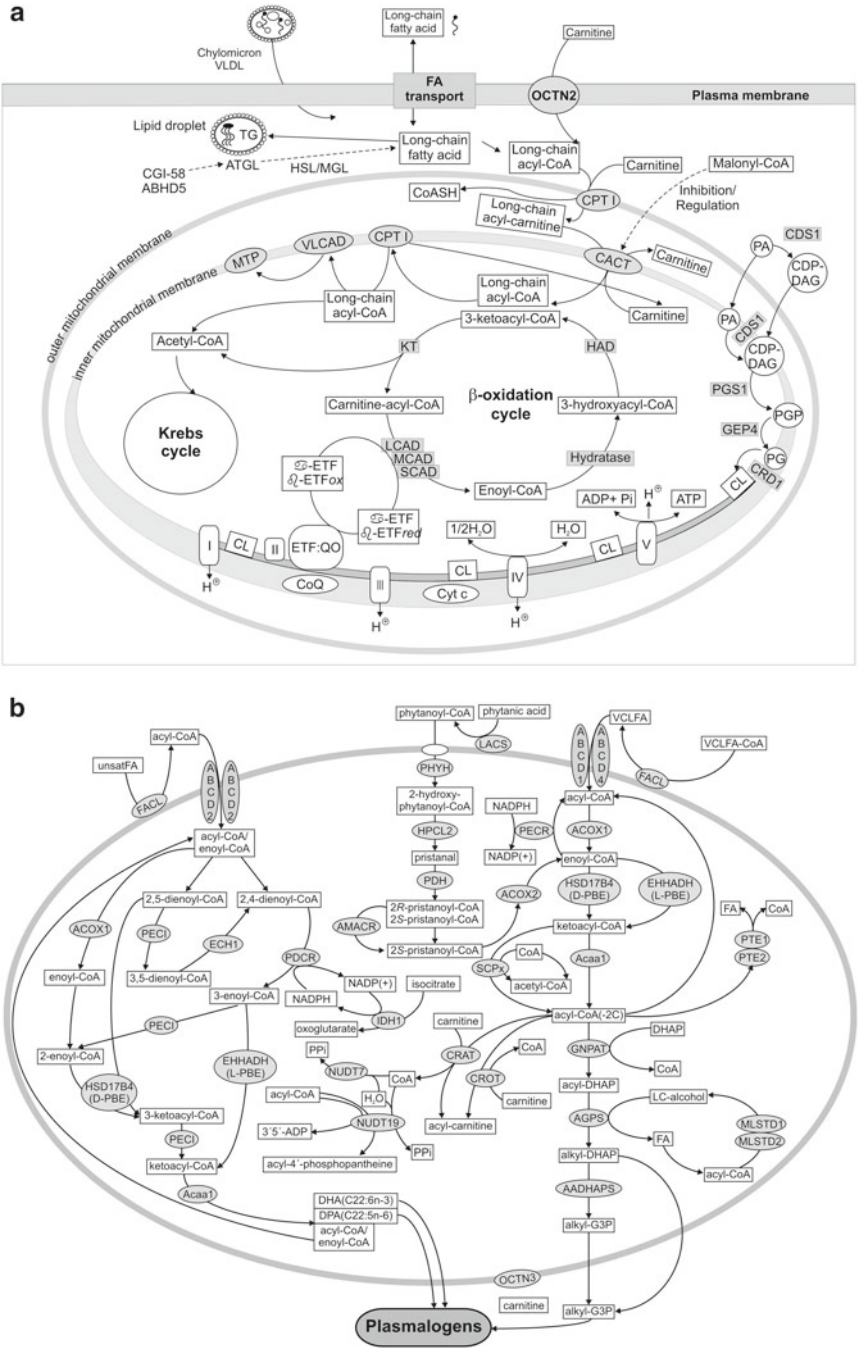


Fig. 13.2 (**a**) Mitochondrial metabolism of fatty acids (Modified with permission from Bruno C et al. 2008). *ABHD5* abhydrolase domain containing 5, *ADP* adenosine diphosphate, *ATGL* adipose triglyceride lipase, *ATP* adenosine-5′-triphosphate, *CACT* carnitine/acylcarnitine translocase, *CDP-DAG* cytidine diphosphate-diacylglycerol, *CDS1* CDP-diacylglycerol synthase 1, *CGI-58* comparative gene identification-58, *CL* cardiolipin, *CoA/CoASH* coenzyme A, *CoQ* coenzyme Q, *CPT I* carnitine palmitoyltransferase I, *CPT II* carnitine palmitoyltransferase II,

Hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase encode the α- and β-subunits of the mitochondrial trifunctional protein (MTP). The heterocomplex contains 4 α-and 4 β-subunits and catalyzes 3 steps in the β-oxidation of FAs, including the long-chain 3-hydroxyl-CoA dehydrogenase step [18].

Mitochondrial FA β-oxidation consists of four sequential reactions, just like their synthesis, catalysed by enzymes with overlapping chain length specificities (Fig. 13.3). This starts with flavoprotein-linked (FAD) dehydrogenation assisted by the acyl-CoA dehydrogenases (ACD), followed by hydration through 2-enoyl-CoA hydratases (ECH), NAD(+)-linked dehydrogenation through L-3-hydroxyacyl-CoA dehydrogenases (HAD), and thiolytic cleavage through 3-ketoacyl-CoA thiolases (KAT), generating acetyl-CoA and an acyl-CoA ester two carbon atoms shorter at the end of each cycle [19]. The electrons generated during FAD-linked oxidation are transferred via electron transfer flavoprotein (ETF) and ETF dehydrogenase (ETFDH) to ubiquinone (Coenzyme Q10), and those from NADH-linked dehydrogenation are passed to complex I in the respiratory chain leading to production of ATP. Mitochondrial β-oxidation also degrades unsaturated FAs with *cis*-double bonds, with pre-existing double bonds being isomerised by auxiliary enzymes such as enoyl-CoA isomerase and dienoyl-CoA reductase [20].

Fig. 13.2 (continued) *Cyt c* cytochrom C, *CRD1* cardiolipin synthase 1, *ETF* electron-transfer flavoprotein, *ETF-QO* ETF:coenzyme Q oxidoreductase, *GEP4* Mitochondrial phosphatidylglycerophosphatase, *HAD* L-3-hydroxyacyl-CoA dehydrogenase, *HSL* hormone-sensitive lipase, *KT* keto-acyl tholase, *LCAD* long-chain acyl-CoA dehydrogenase, *MCAD* medium-chain acyl-CoA dehydrogenase, *MGL* monoglyceride lipase, *MTP* mitochondrial trifunctional protein, *OCTN2* sodium-dependent carnitine transporter, *PA* phosphatidic acid, *PG* Phosphatidylglycerol, *PGP* phosphatidylglycerophosphate, *PGS1* phosphatidylglycerophosphate synthase 1, *SCAD* short-chain acyl-CoA dehydrogenase, *TG* triacylglycerols, *VLCAD* very long-chain acyl-CoA dehydrogenase, *VLDL* very-low-density lipoprotein, I, II, III, IV, V: respiratory chain complex I, II, III, IV, and V, respectively. (**b**) Fatty acid oxidation pathways in peroxisomes. Oxidation of unsaturated fatty acids is depicted on the *left*, oxidation of branched-chain fatty acids in the *center* and b-oxidation of fatty acids is depicted on the *right*. *AADHAPS* dihydroxyacetone phosphate reductase, *ABCD* ATP-binding cassette, sub-family D, *ACAA1* acetyl-Coenzyme A acyltransferase 1, *ACOX* peroxisomal acyl-coenzyme A oxidase, *ADP* adenosine 5′-diphosphate, *AGPS* alkylglycerone phosphate synthase, *AMACR* alpha-methylacyl-CoA racemase, *CoA* coenzyme A, *CRAT* carnitine acetyltransferase, *CROT* carnitine O-octanoyltransferase, *DHA* docosahexaenoic acid, *DHAP* dihydroxyacetone phosphate, *DPA* docosapentaenoic acid, *ECH1* enoyl Coenzyme A hydratase 1, peroxisomal, *EHHADH* enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase, *FACL* acyl-CoA synthetase long-chain, *G3P* glycerol-3-phosphate, *GNPAT* glyceronephosphate O-acyltransferase, *HPCL2* 2-hydroxyacyl-CoA lyase 1, *HSD17B4* hydroxysteroid (17-beta) dehydrogenase 4, *IDH1* isocitrate dehydrogenase 1 (NADP+), soluble, *LACS* acyl-CoA synthetase long-chain, *LC* long chain, *MLSTD* fatty acyl CoA reductase. *NADPH* nicotinamide adenine dinucleotide phosphate reduced, *NUDT19* nudix (nucleoside diphosphate linked moiety X)-type motif 19, *NUDT7* nudix (nucleoside diphosphate linked moiety X)-type motif 7, *OCTN3* organic cation transporter 3, *PDCR* 2,4-dienoyl CoA reductase 2, peroxisomal, *PDH* protein phosphatase 2 C, *PECI* peroxisomal D3,D2-enoyl-CoA isomerase, *PECR* peroxisomal D3,D2-enoyl-CoA reductase, *PHYH* phytanoyl-CoA 2-hydroxylase, *PPi* pyrophosphate, *PTE* acyl-CoA thioesterase 8, *SCPX* sterol carrier protein 2, *VCLFA* very long chain fatty acid

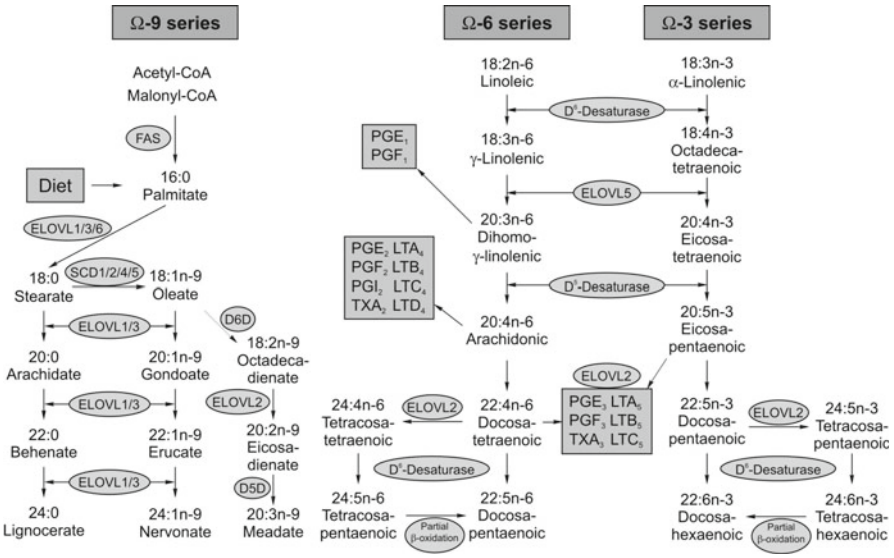


Fig. 13.3 Elongation and desaturation of fatty acids. The n-9-series is endogenously derived from stearate. Starting point for the n-6- and the n-3-series are linoleic acid and a-linolenic acid, respectively, which are not synthesized endogenously and must be taken up as essential fatty acids through the diet. The precursors of eicosanoids are synthesized in these series. *D5D* delta-5-desaturase, *D6D* delta-6-desaturase, *ELOVL* elongation of long chain fatty acids, *FAS* fatty acid synthase, *LTA* Leukotriene A, *LTB* Leukotriene B, *LTC* Leukotriene C, *LTD* Leukotriene D, *PGE* Prostaglandin E, *PGF* Prostaglandin F, *PGI* Prostaglandin I, *SCD* stearoyl-CoA desaturase, *TXA* Thromboxane A

Unsaturated FAs are degraded by β-oxidation as well. With the shortening of the FA chain, the double bonds (usually *cis*-configured) move closer to the CoA-headgroup. Additional enzymes are necessary for further FAO. The first is the dienoyl-CoA isomerase (DCI), catalyzing the isomerisation of the α-γ- to the required α-β-double bond, the second one is the 2,4-dienoyl reductase (DECR) that reduces the α-β-double bond to the saturated acyl-CoA.

The multisubunit NADH:ubiquinone oxidoreductase (complex I) is the first enzyme complex in the electron transport chain of mitochondria where adenosine triphosphate (ATP) is generated via the Krebs Cycle in almost all cells. The protein components include 4 respiratory chain complexes (I-IV) and an ATP synthase. Complex I is the largest with 900 kDa and appears to be the most commonly affected in adult human mitochondrial diseases. Cardiolipins (CL) are synthesized either from phosphatidylglycerol (PG), which is formed from phosphatidic acid (PA), CDP-DAG or alternatively from PC (Fig. 13.2a). CLs are important for mitochondrial electron transfer complex assembly and integrity [21, 22].

An isoenzyme of the long-chain fatty-acid-coenzyme A ligase family is the FA transporter solute carrier family 27, member 2 (SLC27A2). All isoenzymes of this

family convert free long-chain FAs into fatty acyl-CoA esters, differing in tissue distribution, subcellular localization and substrate specificity. They play a vital role in lipid biosynthesis and FA degradation. This isoenzyme activates long-chain, branched-chain and very-long-chain FAs containing 22 or more carbons as CoA derivatives. It is expressed primarily in liver and kidney, and is not present in mitochondria but present in both ER and peroxisomes (Fig. 13.2a) [23].

Peroxisomal Fatty Acid Oxidation

Branched-chain FAs such as phytanic acid cannot be metabolized by usual β-oxidation in mitochondria (Fig. 13.2a, b). Instead, phytanic acid is CoA-activated to phytanoyl CoA at the cytosolic surface of the peroxisome, transported via peroxisomal biogenesis factor 7 (PEX7) into the lumen of the peroxisome, subjected to α-oxidation to pristanic acid, co-activated by very long chain synthase (VLCS) and further degraded via β-oxidation in peroxisomes. Pristanic acid is also CoA-activated at the cytosolic surface of the peroxisome by long chain synthase (LCS), transported into the lumen of peroxisomes and subjected to β-oxidation [24] (Fig. 13.2b).

Very long chain and unsaturated FAs are also substrates of peroxisomal FAO. Acyl-Coenzyme A oxidase 1 (ACOX1) is the first enzyme of the peroxisomal FA β-oxidation pathway, which catalyzes the desaturation of acyl-CoAs to 2-trans-enoyl-CoAs. ACOX1 donates electrons directly to molecular oxygen, thereby producing hydrogen peroxide. Since the peroxisome is unable to generate ATP, the energy that the FAs contain is released as heat. The peroxisomal β-oxidation of FAs follows basically the same mechanism of mitochondrial FAO but uses different, peroxisome-specific enzymes. In mammals, peroxisomal FAO is continued until an acyl-CoA of medium chain-length is reached, which is then shuttled to mitochondria for further degradation and ATP-production via the Krebs cycle.

1.1.3 Elongation and Desaturation of Fatty Acids

The synthesis of FAs longer than 16 carbons occurs at the cytosolic side of the ER, in mitochondria and peroxisomes. The main location of FA elongation is the ER utilizing malonyl-CoA as the carbon source (Fig. 13.3). In contrast, mitochondrial elongation makes use of acetyl-CoA and is basically a reversal of FAO. Peroxisomal FA elongation is closely connected to peroxisomal FAO and produces long chain alcohols serving as precursors for the synthesis of plasmalogens (ether-linked phospholipids).

After formation of palmitate through FAS, a series of chain elongations and desaturations is performed involving ELOVLs (elongation of very long chain FAs) and SCD (stearoyl-Coenzyme A desaturase) generating unsaturated FAs of the n-9- and the n-7-series. Starting from palmitate the elongation by 2 carbons leads to

stearate (18:0), arachidate (20:0), behenate (22:0) and lignocerate (24:0) through ELOVL1/ELOVL3/ELOVL6 (Fig. 13.2). The n-9-FAs are synthesized via desaturation of stearate by SCD1/SCD2/SCD4/SCD5 which yields oleate (18:1 n-9). This desaturation is chemically a reduction, accompanied by an oxidation of water. Consecutive elongation of oleate (ELOVL1/ELOVL3) leads to gondoate (20:1 n-9), erucate (22:1 n-9) and nervonate (24:1 n-9). A second desaturation of oleate by D6D/FADS2 (delta-6-desaturase/fatty acid desaturase) gives rise to the polyunsaturated fatty acid (PUFA) octadecadienate (18:2 n-9) which is elongated to eicosadienate (20:2 n-9) and desaturated once more to form meadate (20:3 n-9) (Fig. 13.3).

Nutritional availability of essential n-3 and n-6 polyunsaturated FAs (n-3 PUFAs, n-6 PUFAs) is critical for many physiological processes [25–28]. PUFAs can serve as intermediates in signal transduction [29, 30], as a source of eicosanoids or docosanoids [31–33], as proinflammatory factors [34] and as neuroprotective agents [35, 36]. They also modulate immune responses [37, 38], and influence human cardiovascular [39] and brain diseases [40]. Western diets are mostly rich in n-6 PUFAs (plant oils). Increased consumption of n-3 PUFAs may therefore counteract some of the proinflammatory, proaggregatory, proliferative and proexcitatory effects promoted by n-6 PUFAs (Fig. 13.3).

Arachidonic acid (AA, 20:4 n–6), eicosapentanoic acid (EPA, 20:5 n-3) and docosahexaenoic acid (DHA, 22:6 n–3) are formed from linoleic acid (LA, 18:2 n–6) and α-linolenic acid, (ALA, 18:3 n–3), respectively, in the liver by a series of alternating desaturation (addition of a double bond) and elongation (addition of a 2-carbon unit) reactions [41, 42] (Fig. 13.3). The essential FAs LA and ALA are formed in plants but not in mammalian cells. This is due to the lack of the D12- and D15-enzymes necessary to insert a double bond at the n (or ω) 6- or 3-position of the FA carbon chain. Once obtained from the diet, LA and ALA are metabolized by D6-desaturation, elongation, and D5-desaturation to form AA and eicosapentaenoic acid (EPA, 20:5 n–3), respectively. The D5-desaturase and subsequent steps in the pathway are found in animal but not in plant cells. AA and DHA, which are also essential FAs, are present in the diet in meat, fish, and eggs but not in fruits, vegetables, nuts, grains, or their products. For some time it was assumed that FA desaturation occurs in the ER and that the final steps in the synthesis of DHA and n–6 DPA (22:5 n–6) involve a D4-desaturation of 22:5 n–3 to 22:6 n–3 and 22:4 n–6 to 22:5 n–6. Now it is known that the pathway forms 24:5 n–3 and 24:4 n–6 through elongation of the 22 carbon chain products of the D5-desaturase.41, 42 The FAs 24:5 n–3 and 24:4 n–6 are desaturated at position 6 to yield 24:6 n–3 and 24:5 n–6, which are translocated to peroxisomes where partial oxidation generates DHA (22:6 n–3) and DPA (22:5 n–6) [41]. Endogenous synthesis of DHA and AA is believed to use the same D6-and D5-desaturase enzymes. The result of this is a competition between LA and LNA as well as inhibition of the enzyme pathway by products of the same and the opposing series of FAs. For example, high dietary intakes of EPA or DHA result in decreased tissue AA and decreased formation of AA derived eicosanoids in favour of n–3 FA derived eicosanoids [43, 44].

1.1.4 Storage of Fatty Acids

The main storage forms of FAs are triacylglycerides (TAG), composed of a glycerol molecule and three FA residues bound via an ester bond. TAGs are core constituents of chylomicrons and VLDL. Their released FAs are reesterified and stored in lipid droplets, incorporated in phospholipids or undergo β-oxidation in muscle cells, hepatocytes and adipocytes. The main organ in FA metabolism is the liver, assembling various apoB-containing lipoproteins, mainly VLDL, which are secreted into the blood, metabolized by lipases and taken up by peripheral cells for energy generation or storage. In contrast, HDL-particles collect cholesterol and phospholipid bound FAs for reverse-cholesterol transport back to the liver.

The precursors of TAGs are FA acyl-CoAs and glycerol-3-phosphate (G3P), synthesized either in liver and intestine from glycerol through glycerolkinase, or from dihydroxyacetone-phosphate (DHAP) through glycerol-3-phosphate dehydrogenase (Fig. 13.4). TAGs can be synthesized in the ER of all cells of peripheral organs through glycerol-6-phosphate acyltransferase (GPAT), which connects two acyl-CoAs to G3P to generate phosphatidic acid. Diacylglycerides (DAGs) are formed in the next step by cleavage of phosphate from the 3-position by PA-phosphatase (LPIN 1-3) and the subsequent addition of another acyl-CoA through diacylglycerol acyltransferase (DPAT) to generate TAGs. In the case of FA overload, monoacylglycerol acyltransferase (MPAT) localized in the intestine and adipose tissue catalyzes the formation of DAGs by adding an acyl-CoA to monoacylglycerides (MAGs, Fig. 13.4).

Energy generation from TAGs occurs through lipolysis. Responsible for the liberation of FAs from TAGs in lipid droplets are three enzymes. Adipose triglyceride lipase (ATGL), coded by ABHD5 and activated by CGI-58, hydrolyses one ester-bond and generates DAG. This is in turn degraded to MAG by hormone sensitive lipase (HSL), liberating a second FA residue. The third hydrolysis step is performed by monoglyceride lipase (MGL), which breaks down MAGs into glycerol and a third FA residue. Glycerol is transported back to the liver by HDL and recycled into TAG-metabolism [45, 46].

Another storage form for FAs are cholesteryl esters (CE), which are a major building block of lipid droplets. CEs in lipid droplets are synthesized through sterol O-acyltransferase (SOAT) which transfers oleyl-CoA onto cholesterol. CEs in the plasma compartment are synthesized from cholesterol and phosphatidylcholine through lecithin: cholesterol acyltransferase (LCAT) which preferentially transfers linoleate to form cholesterylesters and lyso-PC. The PAT protein family (perilipin, adipophilin, TIP47, S3-12, PAT1 or perilipin 1-5) is also important for the metabolism of lipid droplets [47]. They are located on the surface of lipid droplets and regulate the storage and release of lipids through cycles of hydrolysis and esterification driven by phosphorylation and dephosphorylation of the contributing proteins [48]. In this context PA and DAG represent important metabolic hubs that are challenged by chronic metabolic overload ultimately leading to enhanced lipid storage and diabesity (Fig. 13.4). Dysfunction of endolysosomal processing of BMP and mitochondrial cardiolipin dysfunction increase the amount of lipids stored in either endolysosomes or lipid droplets as the basis of vascular and metabolic disease.

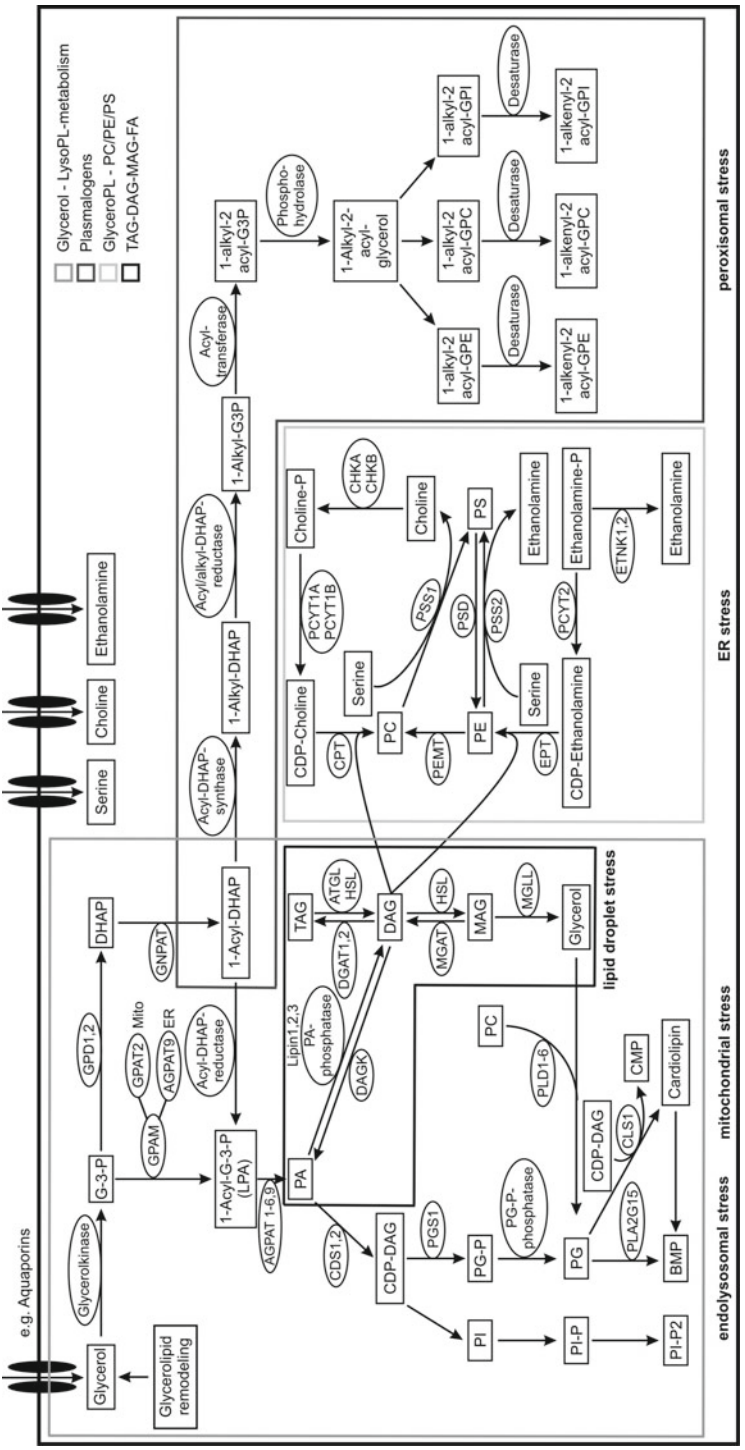

Glycerol - LysoPL-metabolism
Plasmalogens
GlyceroPL - PC/PE/PS
TAG-DAG-MAG-FA
e.g. Aquaporins
Glycerol
Glycerolipid remodeling
Glycerolkinase
G-3-P
GPD1,2
DHAP
GPAM
GPAT2 Mito
AGPAT9 ER
Acyl-DHAP-reductase
1-Acyl-G-3-P (LPA)
AGPAT 1-6,9
GNPAT
1-Acyl-DHAP
Acyl-DHAP-synthase
1-Alkyl-DHAP
Acyl/alkyl-DHAP-reductase
1-Alkyl-G3P
Acyl-transferase
1-alkyl-2 acyl-G3P
Phospho-hydrolase
1-Alkyl-2-acyl-glycerol
1-alkyl-2 acyl-GPE
1-alkyl-2 acyl-GPC
1-alkyl-2 acyl-GPI
Desaturase
1-alkenyl-2 acyl-GPE
1-alkenyl-2 acyl-GPC
1-alkenyl-2 acyl-GPI
Serine
Choline
Ethanolamine
CDP-Choline
CPT
PC
PEMT
Choline-P
CHKA CHKB
Choline
PCYT1A PCYT1B
PSS1
PS
PSD
PSS2
PE
Ethanolamine
Ethanolamine-P
PCYT2
ETNK1,2
CDP-Ethanolamine
EPT
TAG
ATGL HSL
DAG
MGAT
HSL
MAG
MGLL
Glycerol
lipid droplet stress
Lipin1,2,3 PA-phosphatase
DGAT1,2
DAGK
PA
CDS1,2
CDP-DAG
PGS1
PG-P
PG-P-phosphatase
PG
CDP-DAG
CLS1
CMP
Cardiolipin
BMP
PLA2G15
PC
PLD1-6
PI
PI-P
PI-P2
endolysosomal stress
mitochondrial stress
ER stress
peroxisomal stress

Fig. 13.4 Overview of human acylglycerols and glycerophospholipid metabolism. The boxes depict the pathways that are influenced by ER-, mitochondrial-, endosomal-, lipid droplet- and peroxisomal stress. *AGPAT* 1-acylglycerol-3-phosphate O-acyltransferase, *BMP* bismonoacylglycerol, *CDP* cytidine-5′-diphosphate, *CDS* CDP-diacylglycerol synthase, *CHKA* choline kinase alpha, *CHKB* choline kinase beta, *CL* cardiolipin, *CMP* cytidine-5′-monophosphate, *CPT* choline phosphotransferase, *CRLS1* cardiolipin synthase 1, *DAG* diacylglycerol, *DHAP* dihydroxyacetone phosphate, *DAGK* diacylglycerol kinase, *DGAT* diglyceride transferase, *DPG* 2,3 diphosphoglycerate, *EPT* ethanolamine phosphotransferase, *ETNK* ethanolamine kinase, *FA* fatty acid, *G3P* glyceraldehyde-3-phosphate, *GDP* guanosine-diphosphatase, *GNPAT* glyceronephosphate O-acyltransferase, *GPAM* glycerol-3-phosphate acyltransferase, mitochondrial, *GPAT2* glycerol-3-phosphate acyltransferase 2, *GPC* glycerophosphatidylcholine, *GPE* glycerophosphatidylethanolamine, *LPA* lysophosphatidic acid, *MAG* monoacylglycerol, *MGLL* monoglyceride lipase, *PA* phosphatidic acid, *PC* phosphatidylcholine, *PCYT* phosphate cytidylyltransferase, *PE* phosphatidylethanolamine, *PEMT* phosphatidylethanolamine N-methyltransferase, *PG* phosphatidylglycerol, *PG-P* phosphatidylglycerolphosphate, *PGS* phosphatidylglycerol-phosphate synthase, *PI* phosphatidylinositol, *PL* phospholipid, PLA_2 phospholipase A_2, *PLD* phospholipase D, *PNPLA2,3* patatin-like phospholipase domain containing 2, 3, *PS* phosphatidylserine, *PSD* phosphatidylserine decarboxylase, *PSS* phosphatidylserine synthase, *TAG* triacylglycerol

2 Phospholipid Metabolism

Sphingolipids and glycerophospholipids are principal structural components of cellular membranes, with phosphatidylcholine (PC) and sphingomyelin (SPM) being the most abundant. Quantitatively, minor membrane phospholipids include the aminoglycerophospholipids phosphatidylserine (PS) and phosphatidylethanolamine (PE) and the inositol-glycerophospholipid phosphatidylinositol (PI). The above mentioned phospholipids also serve as essential regulators of multiple cellular processes, either directly or by their enzymatic degradation resulting in the formation of specific membrane constituents or bioactive lipid signaling molecules.

2.1 The Kennedy Pathway as the Major Route of PC and PE Synthesis

Mammalian cells derive the majority of their PC from the Kennedy pathway that is located at the cytosolic side of the endoplasmatic reticulum and regulated through protein phosphorylation and translocation of inactive enzymes from the cytosol to active enzyme complexes at the cytosolic side of the ER-membrane (Fig. 13.4).

The first step is the phosphorylation of choline through choline kinase, which is activated by a phosphate cytidylyltransferase that generates CDP-choline. Choline phosphotransferase transfers the choline group of CDP-choline to diacylglycerol (DAG) yielding PC. Similarly, PE is generated via the Kennedy pathway. Newly formed PE can be methylated on its primary amine using S-adenosylmethionine as the methyl donor through PE-N-methyl-transferase (PEMT) to form PC after sequential transfer of three methyl groups. These reactions resemble the PEMT pathway which is most active in hepatocytes (Fig. 13.4).

2.2 PLA2-Remodeling

Fatty acids are stored in the human body as acylglycerols or cholesteryl esters which are incorporated into lipid droplets and constitute the core of lipoproteins. Cell membrane phospholipids can be regarded as another FA storage compartment. This view is supported by the PLA_2-remodeling mechanism. Phospholipase A2 enzymes are members of the PLA-family which catalyzes the hydrolysis of the sn-2-position of membrane glycerophospholipids to generate free FAs and lysophospholipids. The PLA-family consists of at least 19 different enzymes that exhibit PLA-activity [49]. PLAs are classified into three subgroups, the calcium-dependent secretory and cytosolic PLAs (sPLA and CPLA respectively) and the calcium-independent iPLAs. The iPLA2s show a substrate specificity towards the hydrolysis of plasmalogen

species of phospholipids. The FA products of the hydrolysis, like arachidonic acid (AA), are precursors for bioactive components. AA is metabolized to specific prostanoids and leukotrienes that are lipid mediators with a known proinflammatory activity. iPLA is also involved in apoptotic membrane changes such as transbilayer movement of PS [50].

PLA-remodeling fulfills a housekeeping function through generation of lysophospholipids during FA release which can in turn act as acceptors for the integration of AA into phospholipids.

2.3 Synthesis of PS by PC and PE Remodeling and Formation of PE in the Mitochondrial PS-Decarboxylation Pathway

PS is formed through condensation of serine with a phosphatidic acid (PA) moiety. The phosphatidyl donors in mammalian cells are PC or PE, the reaction itself is catalyzed by PS-synthases (PSS I/II) at the cytosolic side of the ER and mitochondria. Newly formed PS is an organelle membrane constituent or is translocated across the mitochondrial membrane and decarboxylated in the lumen of the inner mitochondria to form PE. This reaction is catalyzed by PS-decarboxylase (PSD) (Fig. 13.5b) If this reaction is impaired, PS translocates to the outer plasma membrane, where it binds the apoptotic marker annexin V as a "find me, eat me"-signal of apoptotic cells [51, 52]. Both the PS and "Kennedy pathway" are found in mammalian cells but there are tight restrictions on specific elements of the pathways.

The cytosolic side of the ER and the mitochondrial associated membrane (MAM) compartment have been identified as the principal intracellular localization sites of PC-specific (PSS1) and PE-specific (PSS2) PS-Synthase. The MAM-compartment is an area of transient contact between the ER, mitochondria and peroxisomes where a direct transfer of membrane bound lipids is possible and which can be isolated as a distinct cellular fraction or visualized using electron microscopy [53]. These membrane contact sites (MCS) constitute transient interorganelle assemblies. The MAM-compartment shows significant enrichment in PS-Synthase activity when compared with the total ER-membrane population.

PS-decarboxylase, which reconverts PS to PE, was found as a constituent of the inner mitochondrial membrane and MCS preferentially for the movement of newly synthesized PS to the lumen of mitochondria. The PE generated in mitochondria does not only serve as a structural lipid within mitochondria but is preferentially utilized as a substrate for the methyltransferase reaction to form PC or exported from mitochondria to equilibrate with the cytosolic membrane leaflets to balance the PC-PS-PE species ratio and to function in the biogenesis of other organelles. The molecular mechanism of mitochondrial PE-transport to the plasma membrane is not known, but the process is ATP dependent and insensitive to the Golgi disrupting toxin Brefeldin A. The results with Brefeldin A indicate that the route followed by PE is likely to bypass the Golgi apparatus [54].

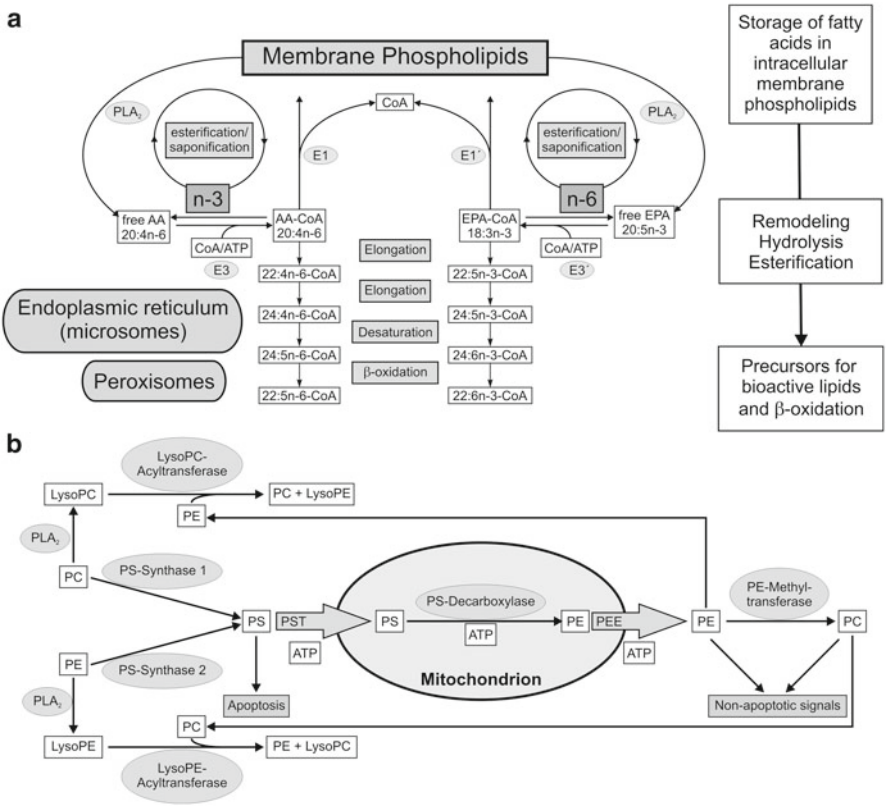


Fig. 13.5 (**a**) PLA_2-remodeling of phospholipids from membranes. Free fatty acids are modified through elongation and desaturation. Cyclooxygenases use released free fatty acids as precursors for the eicosanoids, bioactive lipid messengers that have inflammatory activity. *AA* arachidonic acid, *ATP* adenosin-5′-triphosphate, *CoA* coenzyme A, *EPA* eicosapentanoic acid, *PLA_2* phospholipase A_2. (**b**) Synthesis of phosphatidylserine from phosphatidylcholine and phosphatidylethanolamine and conversion to phosphatidylethanolamine. The interconversion of PS, PC and PE is regulated by several genes. PS is associated with apoptosis while PE and PC are associated with non-apoptotic signals. *ATP* adenosinetriphosphate, *PC* phosphatidylcholine, *PE* phosphatidylethanolamine, *PLA_2* phospholipase A_2, *PS* phosphatidylserine

2.4 Uptake of Glycerophospholipid Precursors and the Interconversion of Glycerophospholipid Species

Recently phosphatidic acid (PA) metabolism attracted significant attention as a metabolic hub upstream of polyglycerophospholipids, phosphatidylinisitols (PI) and acylglycerol storage induced upon metabolic overload (Fig. 13.4).

Hydrolysis of PC by phospholipase D (PLD) and the acylation of lyso-PA by lyso-PA acyltransferases can generate PA. DAG-kinase (DAGK) phosphorylates excess diacylglycerol (DAG) to also yield phosphatidic acid (PA). DAGK isozymes

have different functions and each DAGK isozyme is a critical downstream component of DAG-dependent signaling. The antagonistic enzyme for DAGK action is PA phosphohydrolase (PAP) (Fig. 13.4).

PA is an important metabolite in phospholipid biosynthesis and membrane remodeling. A direct link between the generation of PA and the regulation of endocytosis has been established. PA and other acidic phospholipids affect binding of dynamin to membranes [55]. The role of PA in vesicle trafficking is more general than the modulation of endocytosis. PA has also been shown to affect binding of adaptor protein 2 (AP-2) and clathrin coats to lysosomal membranes [56]. Endophilin A1 plays an important role in the recycling of synaptic vesicles and has a lyso-PA acyltransferase activity [57]. PA also plays a role in Golgi traffic, where it is produced by PLD [58, 59] or by acylation of lyso-PA [60]. In general, PA appears to facilitate fission of vesicles. This function of PA seems to be a consequence of the selective interaction of the lipid with specific target proteins, but given the peculiar structure of PA molecules (a lipid with a small head group and two bulky FA chains attached to the glycerol backbone), it has been proposed that PA may facilitate the formation of local regions of negative curvature on cell membranes [61, 62].

A crucial role in the regulation of several important biological events is played by PA. For instance, PA has been implicated in the regulation of protein phosphorylation [63–65], in the activation of oxidative processes [66, 67], and in the modulation of membrane traffic [68, 69].

Lysophosphatidic acid (LPA) is also a key intermediate in neutral lipid and phospholipid synthesis, implicated in several pathophysiological effects. Glycerophosphate acyltransferase (GPAT) catalyses the formation of LPA by acylation of glycerol 3-phosphate in the ER and mitochondria. LPA may also be synthesized by deacylation of PA. LPA acts through a family of G-protein coupled receptors to modulate cell migration, proliferation and apoptosis. Six GPCRs have been identified, but additional ones may exist: LPA1/Edg2, LPA2/Edg4, LPA3/Edg7, LPA4/GPR23/P2Y9, LPA5/GPR92 and LPA6/P2Y5 [70, 71].

3 Role of Fatty Acid Species, Desaturation and Elongation in Mammalian Sphingolipid Biosynthesis and Metabolism

Sphingolipids (SPs) are a ubiquitous and highly diverse class of lipids. All SPs are characterized by a hydrophobic lipid backbone that consists of a sphingobase (SPH). Amide-linkage to a FA moiety leads to ceramide (Cer), the key intermediate in the SP pathway. Complex SPs are formed by the addition of a headgroup like phosphate, sugar or alcohol and other modifications. Hydrolytic pathways facilitate again the release of individual building blocks.

Not only are SPs essential structural determinants of diverse biological membranes, they also mediate cell interactions, modulate protein functions, are major constituents of lipid microdomains and are also important intra- and extracellular signalling molecules [72–74]. They are widely accepted to be key players in cellular

homeostasis, regulating apoptosis and proliferation processes known as the SP-rheostat, and modulate activities of protein kinases, phosphatases, and phospholipases [75, 76]. As FAs are building blocks of SPs, they share several important biosynthetic and metabolic enzymes connected to elongation and desaturation.

3.1 De Novo Sphingolipid Synthesis

De novo SP biosynthesis is initiated by the condensation of L-serine with palmitoyl-CoA to generate 3-ketosphinganine (3-KS) (Fig. 13.6). This reaction is catalyzed by serine palmitoyltransferase (SPT), the first and rate-limiting enzyme in the de novo pathway, located at the endoplasmatic reticulum (ER).

SPT belongs to the pyridoxal-5′-phosphate (PLP)-dependent α-oxoamine synthase family (POAS). In contrast to the other members of this family, which include 5-aminolevulinate synthase (ALAS), 2-amino-3-ketobutyrate ligase (KBL) and 8-amino-7-oxononanoate synthase (AONS), SPT is not a homodimer but forms a higher organized complex composed of three distinct subunits SPTLC1-3 with a molecular mass of 460 kDa [77]. The two subunits SPTLC1 and SPTLC2 show only about 20% amino acid similarity but a high conservation among species [78]. Knock-out of SPT subunits is lethal and mutations in SPTLC1 cause hereditary sensory and autonomic neuropathy type 1 (HSAN1) [79]. Only recently, Penno et al. could show that HSAN1 is due to altered SPTLC1 amino acid substrate specificity, a gain of function mutation leading to incorporation of alanine or glycine instead of serine and accumulation of two uncommon neurotoxic deoxy-sphingoid bases (DSBs), 1-deoxy-sphinganine (m18:0) and 1-deoxymethyl-sphinganine (m17:0) [80]. Due to the lack of the C1 hydroxyl group, m18:0 and m17:0 cannot be modified by headgroup attachment, but are also not degraded by the classical pathway. Nevertheless, DSBs are subject to Cer and GSL formation and desaturation. Desaturation leads to m18:1 and m17:1 that may resemble bioactive lipids [80]. Together, the bioactive DSBs and their metabolites might also contribute to the well known fumonisin B1 (FB1) associated pathologies [81]. Furthermore associations of DSBs with metabolic syndrome and polyneuropathy in type 2 diabetes (T2D) patients have been reported to correlate with plasma HbA1c-levels joint meeting [77, 78].

SPTLC3 and SPTLC2, exhibiting 68% sequence similarity and both containing a PLP binding motif, are responsible for FA specificity. Instead of the dominating and palmitoyl favouring SPTLC2 subunit, SPTLC3 increases lauryl- and myristoyl-CoA incorporation into SPs, leading to SPH-C16 species. This diverging substrate specificity and also differential gene expression levels observed might contribute to regulatory pathways [78]. Also other FA species can be incorporated eventually into SPs and increasing compatible long chain free FA availability would therefore suggest increasing sphingolipid levels due to the high K_m of this enzyme complex, indeed observed in vitro and in vivo [82, 83].

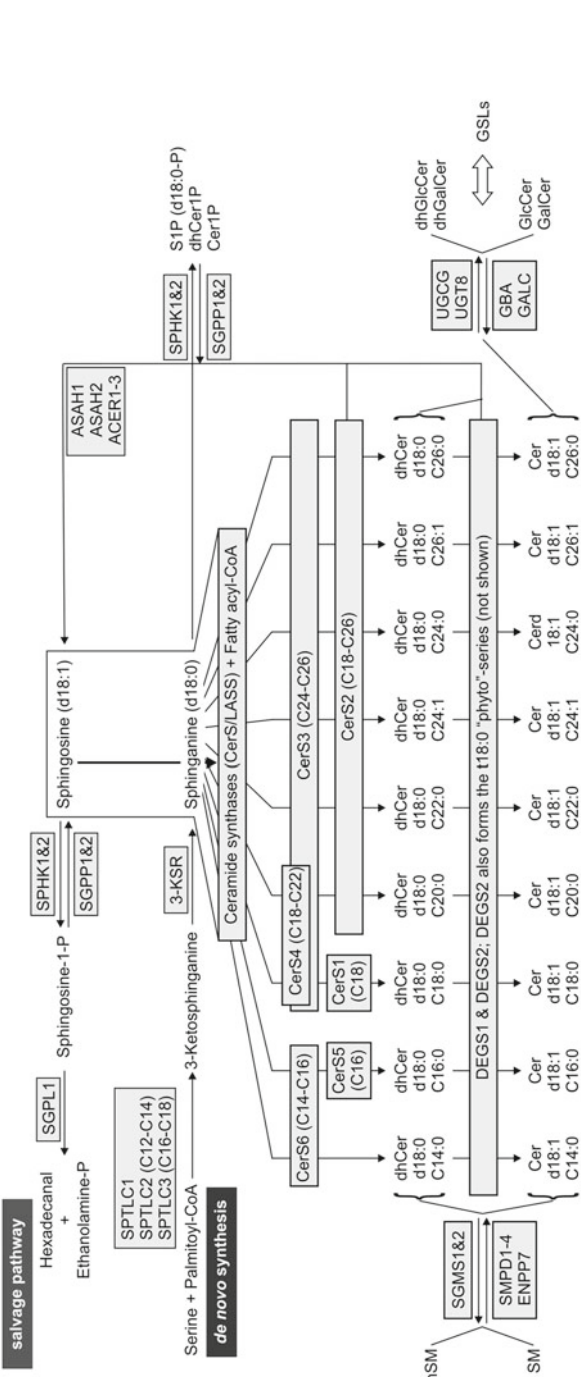


Fig. 13.6 Sphingolipid biosynthesis and metabolism. *De novo* sphingolipid biosynthesis starts with SPTLC, condensating serine and palmitic acid. SPTLC shows substrate specificity depending on fatty acid chain-length. Sphingobases constitute the backbone of all sphingolipids. Fatty acid chain length specificity is also the case for ceramide synthases (CerS), acylating sphinganine. Desaturation of the sphinganine backbone forms ceramide, with sphingosine as backbone. More complex sphingolipids are synthesized through headgroup attachment of alcohols or sugars and other modifications. Complex sphingolipids, ceramides, sphingosine and phosphorylated sphingolipids are interconvertible via salvage or recycling pathways involving several anabolic and catabolic enzymes. *3-KSR* 3-ketosphingoanine-reductase, *ASAH1* acid ceramidase, *ASAH2* neutral ceramidase, *ACER1* alkaline ceramidase, *Cer* ceramide, *Cer1P* ceramide-1-phosphate, *CerS1–6/LASS1–6* ceramide synthase 1–6, *dhCer* dihydroceramide, *dhCer1P* dihydroceramide-1-phosphate, *DEGS1&2* degenerative spermatocyte homolog 1 & 2, *dhSM* dihydrosphingomyelin, *ENPP7* alkaline sphingomyelinases, *GALC* galactosylceramidase, *GalCer* galactosylceramide, *GBA* acid glucosidase beta, *GlcCer* glucosylceramide, *GSL* glycosphingolipid, *S1P* sphingosine-1-phosphate, *SGPP1&2* sphingosine-1-phosphate phosphatase 1&2, *SGMS1&2* sphingomyelin synthase 1&2, *SGPL1* pyridoxal-dependent S1P-Lyase, *SM* sphingomyelin, *SMPD1–4* acid sphingomyelinase 1–4, *SPHK1&2* sphingosine kinase 1&2, *SPTLC1–3* serine palmitoyltransferase, long chain base subunit 1–3, *UGCG* GlcCer synthase, *UGT8*: galactosylceramide synthase

FA condensation is subsequently followed by a reduction of 3-KS by the ER bound NADPH dependent 3-ketosphinganine reductase (3-KSR) to sphinganine (SPHd18:0), the first SPH. Besides palmitoyl-CoA, a second FA is incorporated into SPs by N-acetylation forming dihydroceramide (dhCer). This reaction is catalyzed by the ceramide synthase/LAG1 longevity assurance homolog family (CerS1-6/LASS1-6), also located at the ER. Desaturation to Cer occurs also at the cytosolic side of the ER through lipid desaturases DEGS1 and DEGS2 (degenerative spermatocyte homolog 1 & 2), converting the saturated d18:0 backbone to the trans-Δ^4-monodesaturated sphingosine (SPHd18:1). In mammals formation of the minor SPH species phytosphingosine (t18:0), resembling the predominant SPH species in plants, requires C4-hydroxylation activity of DEGS2 [84].

Cer synthesis is of importance, as it is a key intermediate in SP metabolism and plays an essential role in cellular stress responses. CerS/LASS enzymes utilize several SPHs and show a distinct substrate specificity regarding FA chain length (usually C14-C32), saturation, and hydroxylation as summarized in Fig. 13.6 [85–87]. Tissue specific localization and regulation of CerS/LASS gene expression therefore could be an effective method to regulate biosynthesis of SPs and may also contribute to specific SP patterns on a cellular level, as does FA availability and metabolism. In contrast to cellular triglycerides and cholesterol esters, where predominantly long-chain FAs (LCFAs; C16-C18) are found, SPs show a high rate of very long-chain FAs (VLCFAs; $c \geq 20$) [88, 89]. Saturated, monounsaturated and polyunsaturated FAs (SFAs, MUFAs and PUFAs, respectively) often exhibit completely different physiological and pathological properties, thus SP function is altered by the VLCFA moiety [86, 90]. VLCFA biosynthesis may also determine the cellular or subcellular SP pattern, as C24 FAs are found in the GLSs GM3 and GD3, but not in other gangliosides [88]. There is also a close link between tissue specific FA elongation by ELOVL1-7 and SP synthesis. Elongase disruption in yeast leads to significant alteration in SP composition [91], and the regulation of ELOVL1 by CerS2 has been shown [88]. Especially skin lipids are dependent on VLCFAs, with C28-C36 FAs being of critical requirement for ELOV4 dependent epidermal C28-C36 Cers in animals [92].

All CerSs can also utilize 2-hydroxyl FAs (hFA), modified by the fatty acid 2-hydroxylase (FA2H), which are then incorporated in complex hFA-SPs [93, 94]. Interestingly the UGT8 encoded galactosyltransferase, highly upregulated together with FA2H in neuronal cells, shows a strong preference for hFA-Cer over Cer [93, 95]. Cer formation also occurs via hydrolysis of complex SPs or by reacylation of SPHd18:1 in the salvage or recycling pathway, (see later).

Through the modular architecture of SPs, thousands of possible species can be present in an organism, often distributed tissue specific with only minor concentrations and unusual modifications. For example, major lipid classes in human stratum corneum include cholesterol, FAs and Cers with specific modifications, and a deficient composition is associated with impaired skin barrier function [96]. As the analytical methodologies are getting more sophisticated with higher sensitivity, specificity and accuracy, also new lipid classes and subclasses are discovered, e.g.

the newly described dhCers, amide-linked with an esterified ω-hydroxy FA (EO-Sphinganine) [97].

SPHd18:1 is the quantitatively predominant SPH species in most mammalian tissues. As already mentioned only dhCer is a substrate for DEGS1 and DEGS2, but not free SPH18:0. DEGS1 activity itself is influenced by the alkyl chain length of SPH and the FA, the stereochemistry of the SPH (D-erythro vs. L-threo-dihydroceramides), the nature of the head group and the ability to utilize alternative reductants. This resembles the complexity of the lipid regulatory network. Hu et al. [98] also provided evidence for direct FA-mediated enzymatic regulations, as oleate reverses the affirmative palmitate effect on DEGS1 gene expression, finally giving insight into the mechanisms of metabolic diseases like insulin resistance and diabesity (Box 13.1).

Box 13.1 Educational Box: Lipidomics for Geneticists

Lipidomics describes the extensive analysis of lipid classes and species in biological systems, their pathways and networks. The lipidome therefore not only consists of the complete lipid profile, but also accounts for corresponding genes and the associated transcriptome, epigenome and proteome of any cell, tissue or body fluid sample. In this, the lipidome is one of the four subsets of the metabolome, together with sugars, nucleic acids and proteins/amino acids. Networking of these "omic" approaches constitutes systems health. The recognition of the major structural and regulatory role of lipids in many diseases like diabesity, hypertension, atherosclerosis and stroke has led to rapid expansion of the lipidomics field in preclinical and clinical research. A rapidly evolving, improving and expanding arsenal of analytical techniques, including MALDI and ESI mass spectrometry, fluorometry and NMR spectroscopy have vastly improved our knowledge of the lipidome.

Lipids comprise molecules with wide structural and physicochemical differences, ranging from fatty acid to steroids, leading to an estimated number of about 9,000 lipid species in the human system. Lipidomics is the research of those species and their structures, functions, dynamics and interactions of the lipidome with the transcriptome, proteome and metabolome, including disturbances of these systems through disease, lifestyle and nutrition.

Currently, management of the tremendous amount of data produced by sophisticated high throughput profiling of lipids, transcripts and proteins by mass spectrometry, multiplex affinity binding or arrays is a major challenge in lipidomics. It requires a thorough experimental design, detailed statistical analysis and evaluation of variations detected in the lipidome between different conditions towards systems health to generate novel biomarkers and acquire health information for the benefit of medical care.

DEGS1 and DEGS2 are inhibited by dithiothreitol (DTT) and other thiol reagents, indicative that elevation of cellular thiols could suppress ceramide formation. These characteristics are similar to other desaturases, and inhibitors and redox effectors known to affect Δ^9-stearoyl-CoA desaturase (SCD) and plasmanylethanolamine desaturase severely inhibited dhCer desaturation [99].

To sustain a complex SP pattern, correct trafficking of lipid intermediates to their sites of further modification or recycling is of crucial importance. As principles by which Cers are transported to the Golgi, an ATP- and cytosol-dependent major pathway and an ATP- or cytosol-independent minor pathway have been proposed [100, 101]. At the state of cellular homeostasis, once formed, Cers do not accumulate in the cytosol. Aqueous solubility of Cer is very low, thus a vesicular or protein mediated transfer is necessary to translocate membrane bound Cer to the luminal side of the Golgi apparatus or to the plasma membrane. In the *cis*-Golgi compartment, the sphingomyelin synthase family (SGMS1&2) uses Cer and phosphatidylcholine (PC) as substrates to produce sphingomyelin (SM), one of the major plasma membrane lipids, thereby releasing DAG [102]. In the *trans*-Golgi network (TGN), Cer serves as metabolic precursor for complex SPs such as glycosphingolipids (GSLs) or sulfatides. Recent discoveries of two specific SP transfer proteins, ceramide-transfer protein (CERT) and family A phosphoinositide binding specific member 8 (FAPP2), provided new insights into SP metabolism [103]. CERT transfers Cer from the endoplasmic reticulum (ER) to the Golgi apparatus. On the other hand, FAPP2 and ABCA12 transport complex SPs like GlcCer. Also less specific transport proteins are described, e.g. the glycolipid transfer protein (GLTP) catalyzing the intermembrane lipid transfer of β-glycosylated diacylglycerol (DAG) or of ceramide backbones [104]. Once delivered to the Golgi membrane, Cer needs to translocate to the Golgi lumen through "flip-flop" mechanisms for spontaneous transbilayer movement [105]. SPs additionally serve directly as signalling molecules and therefore have to be distributed intra- and intercellularly. For example, spinster homolog 2 (SPNS2), a sphingosine 1-phosphate transporter plays an important role in developmental processes [106].

SM in mammalian cells has been found to interact and colocalize with cholesterol and GlcCer, mainly in the plasma membrane, in raft microdomains, in lysosomal and Golgi membranes, as well as in the polar surface of circulating lipoproteins. In plasma lipoproteins SM is the second most abundant polar lipid after phosphatidylcholine. Due to their unique physicochemical properties, SMs are enriched in specialized lipid microdomains such as rafts and caveolae and regulate their assembly and dynamics [107].

The importance of proper SM homeostasis has been demonstrated in several experimental designs. SM-deficiency in CHO-cells enhanced ABCA1 dependent cholesterol efflux, and exposure to exogenous SM inhibited this process [108]. In mammalian cells SGMS activation correlates with the activation and nuclear translocation of NFκB, regulated by DAG dependent protein kinase C (PKC) activation. In line with this, exposure of cells to an SGMS inhibitor (D609) or siRNA for SGMS1 and SGMS2 reduces cellular DAG levels, thereby reducing cell proliferation [109, 110]. SGMSs therefore not only regulate SM formation, but also -in a

reciprocal manner- the levels of Cer and DAG, resembling two critical bioactive lipids [111]. Degradation of SM also influences the maintenance of membrane integrity.

Also the chain length and substitution of the GlcCer FA moiety influences membrane properties, function and microdomain-mediated signal transduction [112–115].

The first step in GSL synthesis is the addition of a sugar to the C1 hydroxyl group. In case of glucosylceramides (GlcCer), this is mediated by GlcCer synthase (UGCG), located on the cytosolic leaflet of the Golgi. SPs related to galactosylceramides (GalCer) are formed through galactosylceramide synthase (UGT8), located at the luminal leaflet of the Golgi. More complex sugar attachment is followed by a series of modifications finally leading to the vast class of GSLs, as discussed in detail elsewhere [116–118]. Also, SphinGOMAP© provides a useful overview of GlcCer complexity (www.sphingomap.org).

3.2 Salvage Pathway and Sphingolipid Recycling

SP turnover during salvage and recycling pathways as well as terminal SP degradation are critical processes, as several SPs, metabolic intermediates and degradation products show bioactive properties [72, 119]. Nutritional SPs also have to be metabolized, providing uncommon lipid species [120]. As mammalian SPHd18:1 is only generated through desaturation of dhCer, not by desaturation of SPHd18:0 during the de novo pathway, and SPHd18:1 plays important roles in signalling pathways, e.g. phosphorylation pathways, SP turnover has to be tightly regulated.

SM breakdown is mediated by three classes of sphingomyelinases, named after their individual pH-optimum. Acid sphingomyelinase (SMPD1), intestinal alkaline sphingomyelinases (ENPP7) and neutral sphingomyelinases (SMPD2-4) show distinct SP specificity and are responsible for membrane homeostasis.

GSL degradation occurs in acidic endosomes and lysosomes, mediated by several enzymes, including galacto- and glucosidases (GBA, GLA, GALC, GLB1), sialidases (NEU1-4) and other enzymes (ARSA, ARSB, ARSG, ARSI, ARSJ, GNS, SGSH). Carbohydrates are sequentially released there by glycosidases, finally providing Cer.

Cer catabolism starts with ceramidase activity catalyzing the cleavage of Cer at the amide bond resulting in free SPH and free FA. Ceramidases show an organelle specific distribution. Three types of ceramidases have been described to date and classified according to their pH optima as acid (ASAH1), neutral (ASAH2), or alkaline (ACER1, ACER2 and ACER3).

Many SP-associated diseases are known in humans. In most cases like skin barrier diseases, Alzheimer, dementia, multiple sclerosis and atherosclerosis, specific causes and responsible SP-genes are still not known [121]. But several monogenetic diseases have been identified, and disturbed SP degradation is a wide source of lipid associated diseases, including lysosomal storage disorders like inherited sphingolipidoses, reviewed elsewhere [115, 122, 123].

Sphingosine-1-phosphate (S1P) is a bioactive SP generated by sphingosine kinases (SPHK1&2) and considered as a unique lipid mediator acting both internally and externally. As a second messenger S1P is implicated in the regulation of Ca^{2+} mobilization and in exerting mitogenic and anti-apoptotic effects like cellular growth, proliferation and survival induced by platelet-derived growth factor, nerve growth factor and serum [76, 124, 125]. Through its high affinity G protein-coupled receptors (S1P1-5), S1P acts as an extracellular physiological mediator regulating heart rate, coronary artery blood flow, blood pressure, endothelial integrity and most recently it has been shown to regulate the recirculation of lymphocytes [126].

SGPP1 expression increased the incorporation of sphingosine into several SPs, enhancing C16:0, C18:0 and C20:0Cers and downstream GlcCers and SMs [127] (Fig. 13.6).

Degradation of S1P is either mediated by a pyridoxal-dependent S1P-Lyase (SGPL1) with irreversible cleavage to ethanolamine-phosphate and hexadecanal, the only way to terminally degrade any SP. The other possibility is dephosphorylation by specific S1P-phosphohydrolases (SGPP1&2), thus increasing the level of free SPHd18:1 in the cytosol and cell membranes. SGPP1 and SGPP2 show SPH chain length specificity and localize at the cytosolic side of the ER where they degrade S1P to terminate its actions.

4 Genetic Diseases Related to FA Species and FA Related Lipid Class Metabolism and Processing

There are several genetic defects that influence FA metabolism especially related to FA processing, transport, incorporation into other lipid classes (e. g. acylglycerols, glycerol-PL, Sphingolipids or cholesterylesters) or FA oxidation diseases (Table 13.2), with either impaired energy production or altered FA accumulation. Another consequence of disorders of FA metabolism is the influence on cell and organelle membranes. The fluidity or rigidity of membranes as well as their formation and function are strongly dependent on the degree of saturation and chain length of the FA residues as constituents of membrane lipid species. The higher the degree of desaturation, the more fluid a membrane becomes as a consequence. This is the basis for the importance of desaturases and elongases in metabolic and vascular diseases.

FA transport disorders affect the carnitine shuttle responsible for the transport of FAs into mitochondria for further degradation. They include primary carnitine deficiency (PCD), carnitine-acylcarnitine translocase deficiency (CACTD), carnitine palmitoyltransferase I deficiency (CPT1D) and carnitine palmitoyltransferase II deficiency (CPT2D). The symptoms include hypoketotic hypoglycemia, hyperammonemia, hepatomegaly and cardiomyopathy, as well as sudden infant death [128].

Deficiency or impaired function of elongases leads to an accumulation of AA and DHA and subsequently increased channelling of these precursors to the sites of prostaglandin and thromboxane synthesis, being responsible for proinflammatory and procoagulent responses [129].

Table 13.2 Genetic disorders connected to fatty acid metabolism

Disease	Affected gene	Effect
Mitochondrial FA disorders		
Very long-chain acyl-coenzyme A dehydrogenase deficiency (VLCADD)	VLCAD	Very long-chain fatty acids cannot be metabolized, especially during fasting
Long-chain 3-hydroxyacyl-coenzyme A dehydrogenase deficiency (LCHADD)	LCAD	Very long-chain fatty acids cannot be metabolized, especially during fasting
Medium-chain acyl-coenzyme A dehydrogenase deficiency (MCADD)	MCAD	Impaired FAO, reduced energy production, especially during fasting
Short-chain acyl-coenzyme A dehydrogenase deficiency (SCADD)	SCAD	Impaired FAO, reduced energy production, especially during fasting
3-hydroxyacyl-coenzyme A dehydrogenase deficiency (HADHD)	HADH	Impaired FAO, reduced energy production, especially during fasting
2,4 Dienoyl-CoA reductase deficiency (DECR1D)	DECR1	Impaired FAO of unsaturated fatty acids
Malonyl-CoA decarboxylase deficiency (MCDD)	MCD	Malonic acid is produced by SCAD, Krebs cycle is inhibited, Glycolysis is increased
Mitochondrial trifunctional protein deficiency (MTPD)	HADHA, HADHB	Impaired FAO, reduced energy production, especially during fasting
Barth syndrome	Tafazzin	Abnormal cardiolipin profile due to impaired CL-remodeling
Peroxisomal defects		
Refsum disease	AMACR	Phytanic acid is not degraded and accumulates in plasma
Zellweger Syndrome	PEX	Impaired peroxisomal function, VCLFA and BCFA accumulate
X-linked adrenoleukodystrophy	ABCD1	Accumulation of VCLFA due to impaired transport into peroxisomes
D-bifunctional protein deficiency	HSD17B4	Accumulation of LCFA due to impaired peroxisomal FAO
FA processing and transport disorders		
Primary carnitine deficiency	SLC22A5	Impaired FAO due to loss of carnitine
Carnitine-acylcarnitine translocase deficiency	CACT	Impaired FAO due to lack of transport of acyl-carnitines into mitochondria
Carnitine palmitoyltransferase I deficiency (CPT)	CPT I	Impaired FAO due to lack of transport of acyl-carnitines into mitochondria
Carnitine palmitoyltransferase II deficiency (CPT)	CPT II	Impaired FAO due to lack of transport of acyl-carnitines into mitochondria

(continued)

Table 13.2 (continued)

Disease	Affected gene	Effect
Hepatic lipase deficiency	LIPC	Elevated HDL levels
lipoprotein lipase deficiency	LPL	Hypertriglyceridemia
Tangier disease	ABCA1	Impaired HDL production
Chanarin-Dorfman syndrome	ABHD5	Accumulation of triglycerides due to lack of lipase ATGL
Cholesteryl ester storage disease	LAL	Accumulation of cholesterol and triglycerides due to lysosomal acid lipase (LAL) deficiency,
Majeed syndrome	LPIN 2	Inflammatory disorder characterized by recurrent bouts of osteomyelitis, dyserythropoietic anemia, and cutaneous inflammation
Amino acid disorders related to FA-metabolism		
Maple syrup urine disease	BCKDHA BCKDHB, DBT, DLD	Accumulation of leucine, valine and isoleucine due to branched-chain alpha-keto acid dehydrogenase complex (BCKDC) deficiency
Propionic acidemia	PCCA, PCCB	Accumulation of propionic acid due to propionyl CoA carboxylase (PCC)-deficiency

Chanarin-Dorfman syndrome is a neutral lipid storage disease, caused by a defect in abhydrolase domain containing 5 (ABHD5) which activates adipose triglyceride lipase (ATGL). Symptoms include abnormal storage and accumulation of TAGs in liver, skin, muscles, intestine, eyes, and ears. As a consequence patients suffer from hepatomegaly, ichthyosis, cataracts, ataxia, hearing loss, short stature, myopathy, nystagmus and mild intellectual disability [130].

Another genetic disease connected to lipid storage is cholesteryl ester storage disease (CESD) and its more severe form, Wolman's disease (WD), which is caused by partial and complete lysosomal acid lipase (LAL) deficiency. Symptoms include hepatomegaly, associated with hepatic steatosis and elevated transaminases, leading to chronic liver disease, periportal fibrosis and cirrhosis. WD-patients die within the first years of life from adrenal calcification and consecutive insufficiency. There are also CESD cases reported, where the symptoms were reduced and even unsymptomatic patients have been diagnosed with CESD. On the other hand, chronic alcohol consumption or viral infections of the liver severely influence the progression of chronic liver disease [131].

Lipoprotein lipase (LPL) deficiency leads to excessive hypertriglyceridemia caused by failure of TAG-hydrolysis in chylomicrons and VLDL and the patients may die from pancreatitis, but there are no signs for vascular disease [132]. Overexpression of LPL has been implicated in diabetes mellitus, obesity and tissue specific insulin resistance [133]. Hepatic lipase (LIPC) deficiency is a rare condition resulting in high levels of atherogenic remnants from TAG-rich lipoproteins and elevated HDL_2-levels. Animal models of LIPC reveal a strong connection to atherosclerosis [134].

Several studies have examined the metabolism of TAG and lipid droplets in mice. It has been shown that MPL-deficient mice have impaired lipolysis and attenuated insulin-resistance [135]. DGAT-deficient mice are protected from obesity and have increased insulin sensitivity [136]. The same has been found for HSL-ko mice, which were not obese with reduced white adipose tissue. In contrast to that, ATGL-ko mice show reduced hydrolysis of TAGs from lipid droplets in white adipose tissue and a massive accumulation of TAGs in all tissues, but most prominent in cardiac and skeletal muscle, testis, pancreas and kidney, eventually leading to cardiac arrest [137].

Genetic diseases of FA dependent energy metabolism in mitochondria affect FA dehydrogenases, including very long-chain acyl-coenzyme A dehydrogenase deficiency (VLCADD), long-chain 3-hydroxyacyl-coenzyme A dehydrogenase deficiency (LCHADD), medium-chain acyl-coenzyme A dehydrogenase deficiency (MCADD), short-chain acyl-coenzyme A dehydrogenase deficiency (SCADD) and 3-hydroxyacyl-coenzyme A dehydrogenase deficiency (HADHD). These disorders of FA oxidation have several phenotypes including hypoglycemia, lethargy, muscle weakness, and, in infants or small children failure to gain weight and poor feeding. Patients also suffer from nausea, vomiting and diarrhea. MCADD has been linked to infant sudden death syndrome [138]. These disorders manifest especially during periods of fasting, because the production of energy from triglycerides is severely impaired (Table 13.2). 2.4 Dienoyl-CoA reductase is responsible for the metabolism of FAs with even-numbered double bonds [139, 140] and deficiency has been diagnosed only in a few newborn patients with a small body habitus, microcephaly, symptoms of sepsis, hypotonia, decreased feeding and intermittent vomiting. Patients ultimately die within 6 months.

Another mitochondrial FA metabolism disorder is Barth syndrome, caused by defects in the tafazzin gene, which leads to impaired cardiolipin (CL) synthesis and remodelling, resulting in an unnormal mitochondrial CL-profile. Barth syndrome is clinically characterized by myopathy, neutropenia, growth delay, exercise intolerance, cardiolipin abnormalities and 3-methylglutaconic aciduria [141].

Majeed syndrome is caused by defects of LPIN2, the gene encoding a member of the phosphatidate phosphatase family (PAP, lipin 1–3). Majeed syndrome is a rare condition characterized by recurrent episodes of fever and inflammation in the bones and skin, known as chronic recurrent multifocal osteomyelitis (CRMO). A blood disorder called congenital dyserythropoietic anemia can also occur in Majeed syndrome [142].

There are also four disorders in FA metabolism attributed to peroxisomes. Defects in the ACOX1 gene, which is the first enzyme in the peroxisomal FAO of unsaturated and saturated FAs, result in pseudoneonatal adrenoleukodystrophy, a disease that is characterized by accumulation of very long chain FAs [143].

Zellweger syndrome is a peroxisome biogenesis disorder characterized by impaired neuronal migration, neuronal positioning, and brain development, due to the accumulation of VCLFA and BCFA that cannot undergo β-oxidation in the defective peroxisomes and whose aberrant incorporation into myelin destabilizes that myelin and in turn the neuronal sheath architecture [144]. Another peroxisome biogenesis disorder is X-linked adrenoleukodystrophy (X-ALD). A mutation in the

ABCD-transporter localized in the peroxisome membrane. ABCD defects lead to an accumulation of VCLFA in the brain and adrenal gland due to impaired transport of VCLFA into peroxisomes. Symptoms include loss of previously acquired neurologic abilities, seizures, ataxia, Addison's disease, and degeneration of visual and auditory function. The onset is usually at age 4–10 and X-ALD is present almost exclusively in males [145, 146].

Refsum's disease is an inherited disorder of branched chain lipid metabolism caused by a defect in phytanic acid catabolism. Causal are mutations in the two genes PHYH (Phytanoyl-CoA 2-Hydroxylase) and PEX7, responsible for the metabolism and transport of phytanic acid [147, 148]. Due to these defects, accumulation of phytanic acid reaches toxic levels in plasma and in several tissues, mostly in adipose tissue, liver, kidney, muscle and neuronal tissues. The disease usually begins in late childhood with increasing night blindness due to degeneration of the retina (retinitis pigmentosa) and loss of the sense of smell (anosmia). Other symptoms include deafness, problems with balance and coordination (ataxia), weakness or peripheral neuropathy, numbness dry and scaly skin (ichthyosis) and cardiac arrhythmias [147, 148].

Connected to branched-chain FA metabolism is the maple syrup urine disease (MSUD) or branched-chain ketoaciduria. It is caused by a deficiency of the branched-chain alpha-keto acid dehydrogenase complex (BCKDC), leading to accumulation of the branched-chain amino acids and their toxic side-products in blood and urine. Symptoms of the condition include poor feeding, vomiting, dehydration, lethargy, hypotonia, seizures, hypoglycaemia, ketoacidosis, opisthotonus, pancreatitis, coma and neurological decline, it manifests in early infancy and is treated through a diet without branched-chain amino acids [149].

Propionic acidemia is caused by propionyl CoA carboxylase (PCC)-deficiency, an enzyme that converts propionyl-CoA (from branched-chain amino acids and odd-numbered FAs) into methylmalonyl-CoA, it is instead converted to propionic acid, which accumulates. Symptoms of the condition include poor feeding, vomiting, dehydration, lethargy, hypotonia and seizures. The disease manifests almost directly after birth and is rapidly life-threatening. [150]

5 GWAS and Lipid Species

Genome-wide association studies (GWAS) examine interindividual variations of the genome. GWAS led to the discovery of single gene polymorphisms in coding, intergenetic or intronic regions associated with diseases such as diabetes, obesity and vascular diseases [151]. Within a GWAS, individuals are tested for single nucleotide polymorphisms (SNPs) in the genome and different traits are correlated with variations of metabolic abnormalities, enhanced or reduced disease risk (morbidity) and outcome (mortality).

Several GWAS identified correlations between SNPs and lipoproteins, summarized in Fig. 13.7, where SNPs are grouped according to gene functions. Clearly

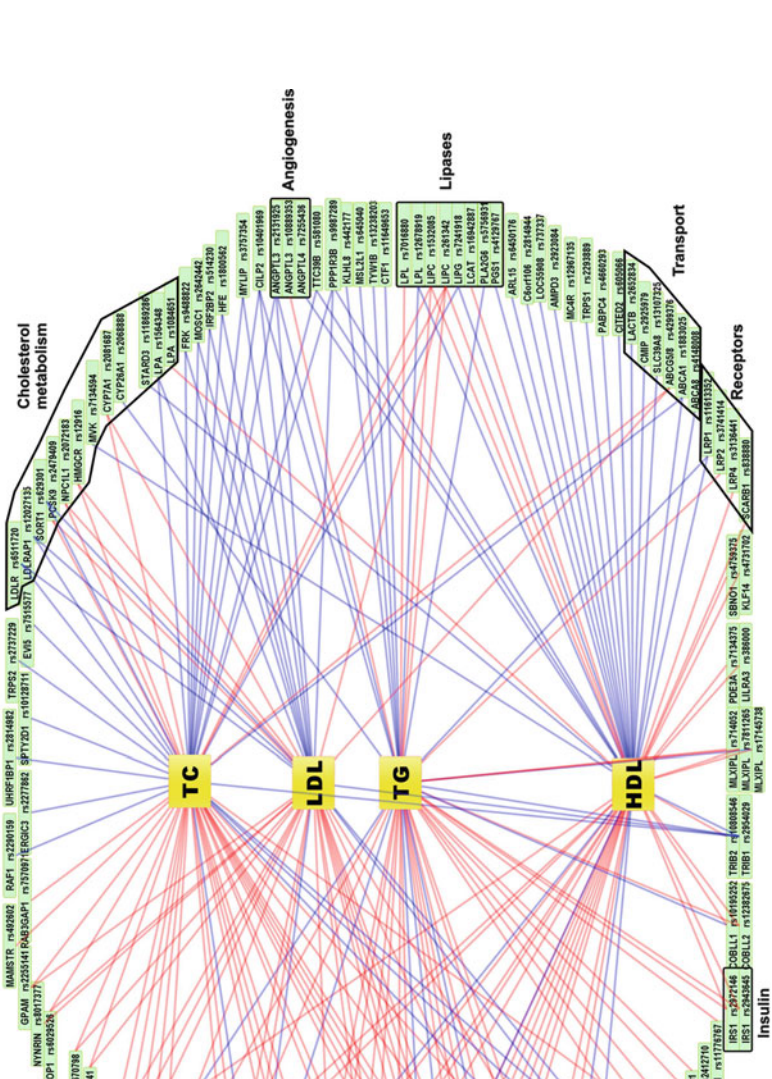


Fig. 13.7 Correlation of lipoproteins and lipid classes with single nucleotide polymorphisms as determined by genome-wide association studies. *Red lines* indicate positive, *blue lines* represent negative correlation. SNPs are grouped together according to genetic regions. *HDL* high density lipoprotein, *LDL* low density lipoprotein, *TC* total cholesterol, *TG* triacylglycerol

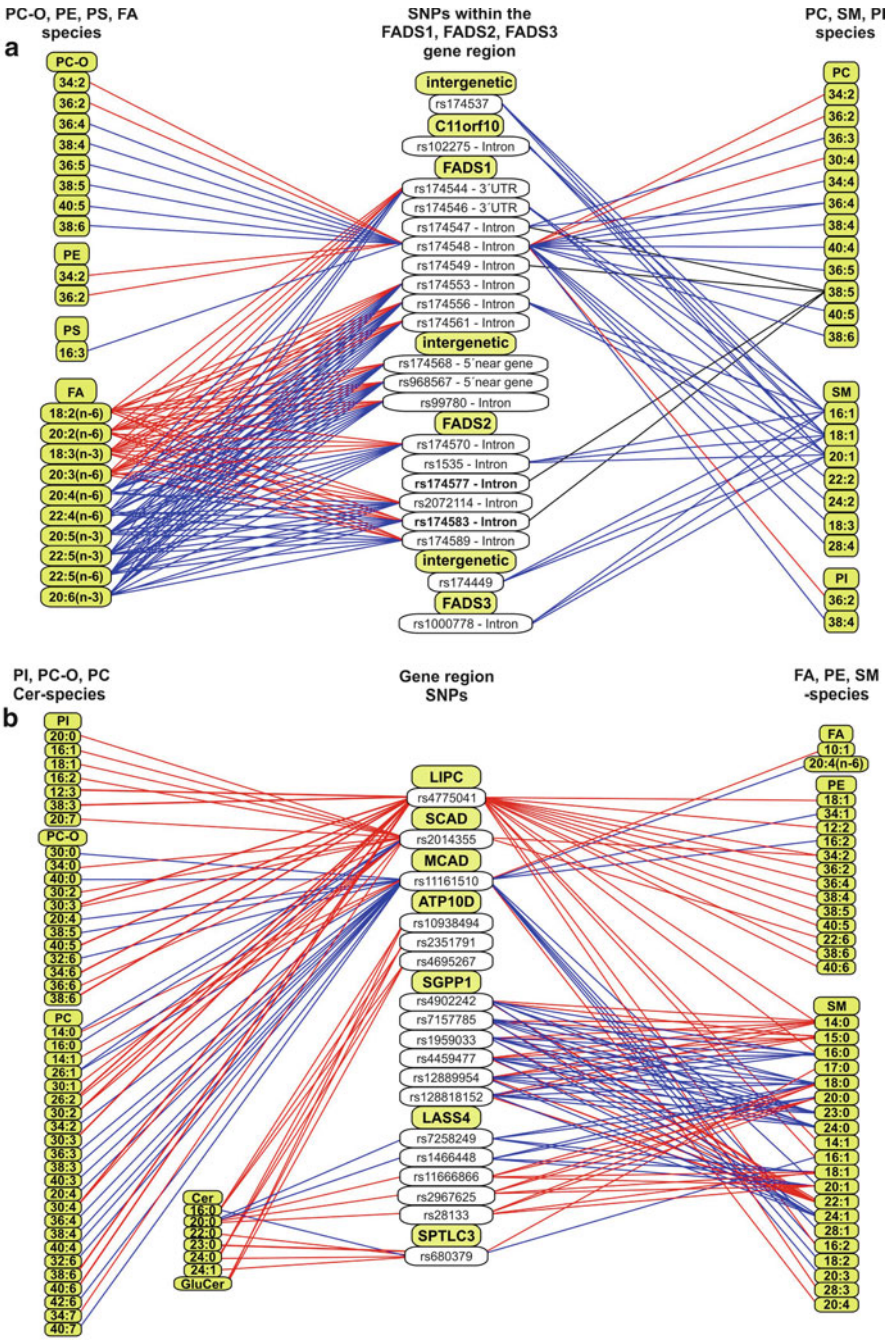


Fig. 13.8 (**a**) Correlation between lipid species and single nucleotide polymorphisms involved in fatty acid desaturation as determined by genome-wide association studies (GWAS). *Red lines* depict positive correlation, *blue lines* negative correlation. *C11orf10* chromosome 11 open reading frame 10, *FA* fatty acid, *FADS1-3* fatty acid desaturase1-3, *PC-O* plasmalogen, *PC* phosphatidylcholine, *PE* phosphatidylethanolamine, *PI* phosphatidylinositol, *SNP* single-nucleotide polymorphism, *SM* sphingomyelin. (**b**) Correlation between lipid species and single nucleotide

visible is the positive correlation of apolipoprotein genes to both LDL and HDL, while lipases correlate to either LDL or HDL and desaturases correlate to HDL, LDL and TG. Interestingly, three polymorphisms in the FADS1-3 region (rs174546, rs174547 and 174601) were found to positively correlate with total cholesterol (TC), LDL, HDL and TG (Fig. 13.7) and only one of those (Fig. 13.8a; FADS1 – rs174547) also shows a correlation to lipid species (negative correlation to PC36:3 and to PC36:4 and to protective PC38:5).

The SNPs in the LIPC region also have a positive correlation to TG, HDL and TC, but only one (rs261342) has a positive correlation to LDL. There are no correlations of LIPC-SNPs to both, lipid species (Fig. 13.8b) as well as lipoproteins (Fig. 13.7). Genes in the region of cholesterol metabolism have positive (PCSK9, NPC1L1, HMGCR, CYP7A) and negative (LDLR, LDLRAP1, SORT1) correlations to TC and LDL. The angiogenesis region shows mainly negative correlations, which is also true for the transporter region, where only ABCG5/8 displays a positive correlation to LDL and TC. In the receptor region there is a negative correlation between LRP1 and TG and HDL and positive correlations of LRP4 and SCARB1 to HDL and LRP2 to TG. The insulin receptor substrate 1 region shows two SNPs with a positive correlation to HDL and TG, while the immune modulation/cell recognition region has positive correlations to the lipoproteins except for one SNP of HLA that negatively correlates to TG [152, 153].

More detailed analysis of FA-species correlations are shown in Fig. 13.8a, b as a summary of all published SNP-lipid species correlations with impact for vascular and metabolic disease [154–157]. This analysis maps the metabolic pathways of lipid species and the influence of different SNPs, exemplified in the up- and down-regulation of FA species incorporated into glycerophospholipids, sphingolipids and acylglycerols. It can be deduced from these correlations that a positive effect between lipid species and SNP corresponds to an increased precursor level of that lipid species and a negative correlation corresponds to decreased product levels of the lipid species concerned.

It is interesting to note that, in the FA cluster shown in Fig. 13.8a there are associations between the FADS-SNPs and the FAs, with each SNP having only a positive or a negative correlation. There is only one SNP (rs174547) that appears in both studies with correlations to both, lipid species as well as lipoproteins (Fig. 13.7). The FAs with a lower degree of desaturation (dienoate and trienoate) revealed a positive association with SNPs in the FADS1-2 region, while FAs with a higher degree of desaturation show a negative correlation to SNPs in this region. Therefore, any genetic variation of a single SNP involved leads to a decrease of desaturation of

Fig. 13.8 (continued) polymorphisms involved in fatty acid metabolism. Note the separation of correlations between sphingolipid/ceramide metabolism-related genes and fatty acid metabolism-related genes. *ATP10D* ATPase, class V, type 10D, *Cer* ceramide, *FA* fatty acid, *GluCer* glucosyl-ceramide, *LASS4* ceramide synthase 4, *LIPC* hepatic lipase, *MCAD* medium-chain acyl-CoA dehydrogenase, *PC-O* plasmalogen, *PC* phosphatidylcholine, *PE* phosphatidylethanolamine, *PI* phosphatidylinositol, *SCAD* short-chain acyl-CoA dehydrogenase, *SGPP1* sphingosine-1-phosphate phosphatase, *SNP* single-nucleotide polymorphism, *SM* sphingomyelin, *SPTLC3* serine palmitoyltransferase3

the FAs, including precursors for either pro-inflammatory (n-6) or anti-inflammatory (n-3) prostanoids. The fact that there is a clear distinction between precursors with a positive correlation and a negative correlation indicates a specific precursor product relationship between pairs of FAs through certain sequence variations of the gene region. This can be exemplified in the following: rs174553, rs174556 and rs174561 are three sequential SNPs in the FADS1-region that have a positive correlation to the FA 20:3(n-6) and a negative correlation to arachidonic acid 20:4(n-6). This is an indication that variations in this gene region raise the level of 20:3(n-6) (precursor) and decreases the level of 20:4(n-6) (product). This can be seen as evidence that this gene region is involved in the desaturation of 20:3(n-6) to AA 20:4(n-6). A SNP in this gene region therefore may lead to a decrease or increase of 20:4(n-6) levels, potentially altering its proinflammatory effects on eicosanoids biosynthesis.

It is also interesting to note that there are negative correlations between SNPs of the FADS-region and phospholipid species with a higher degree of desaturation and a higher number of carbons, indicating a longer chain length of the incorporated FAs. Thus, genetic variations of the corresponding SNPs are connected to a reduction of desaturation.

SNPs in FADS1 (rs174549, rs174548, 174547) and FADS2 (rs174577, rs174583) are associated with PC 38:5, which is made up either of a combination of the FAs 20:4 and 18:1 or of 18:0 and 20:5. Different SNPs in the FADS cluster associate to different glycerophospholipid species, indicating site specific functional alteration within the FADS cluster. Moreover, it is obvious that none of the published SNPs is found in coding sequences, all are intergenic or intronic, indicative for influencing transcription or epigenetic regulation. Generally, any variation in the FADS gene region that has been published in these GWAS is favouring a lesser degree of desaturation in the FAs and has a negative correlation to FAs with a higher degree of desaturation.

Impaired FA desaturation also disturbs the homoeostasis of associated lipid species. Sphingomyelins show only negative correlations to specific FADS-SNPs, indicating the importance of FAs also for SP biosynthesis. In the case of FADS1, rs174556 negatively correlates with medium chain SM16:1, SM18:1 and SM20:1, as also does rs1535 of FADS2. On the other hand, rs174548 negatively correlates with long and very-long chain sphingomyelins SM22:2, SM24:2, SM18:3 and SM28:4. Accordingly, SM synthesis is dependent on specific FA precursors defined by proper FA desaturation. Indeed, FA specificity of the seven individual CerS are known, and Cers are the precursors for sphingomyelin synthases. In a recently published GWAS [158], also clinical relevance for the FA dependent sphingolipid homoeostasis was found, as hypertension and blood pressure regulation was linked to SNPs responsible for de novo ceramide biosynthesis and the sphingolipid rheostat. Fenger et al. thereby emphasised the major importance of externally derived FAs for Cer synthesis, comparing the gene expression of ELOVL3 and free fatty acid receptor 1 (FFAR1), and pointing out the FA specificity of different CerS enzymes [158].

Another interesting correlation is the one for the SNP rs174548. There is a positive correlation to PC-O 34:2 and 36:2 and a negative correlation to PC-Os with a

higher degree of desaturation. This implies a connection between this SNP and the remodelling of PC-Os, thus it should be investigated, whether there is a correlation to the biosynthesis of different PC-Os or rather to the remodelling of preformed PC-Os.

There is a clear distinction between genes involved in phospholipid and sphingolipid/ceramide remodelling, e.g. in FADS1 (rs174546), FADS2 (rs1535) and FADS3 (re1000778). Figure 13.8b shows all published correlations between SNPs in gene regions of lipases and dehydrogenases and lipid species. The SNP (rs4775041) in the LIPC region positively correlates with several phospholipid species. SNPs in these regions also affect phospholipid hydrolysis thus leading to precursor accumulation of certain phospholipid species including PC 30:4, 38:6 and 40:6. The MCAD and SCAD regions show mostly negative and positive correlations to individual phospholipid species, but there is no clear discrimination of the two according to FA chain length, which would actually be expected. The situation in these gene regions is not as clear as it is with the FADS region, because there is not as much data published. SP associated genes, SGPP1, LASS4 and SPTLC3 show species specific correlations, but the overlap with LIPC, MCAD and ATP10 also resembles the interconnectivity of the metabolic pathways.

Analysis of GWAS with this type of graphical illustration provides the possibility to find precursor-product relationships and their connection to SNPs and gene regions. Also, this analysis can give insight into therapeutic and diagnostic target areas for further investigations. Since lipid species, their functions and the effects of metabolic diseases on those lipid levels are well known, effects through the increased or decreased levels of lipid species can be connected to gene regions found through this type of data analysis.

6 Conclusion

The metabolism of FA species and their assembly as acylglycerols, cholesteryl esters, glycerophospholipids and sphingolipids in the human body is regulated by a large number of different genes that are polymorphic in the population and may be involved in vascular and metabolic disease. Fluidity, rigidity and function of membranes are a result of FA chain length and the degree of desaturation of the FA residues that are part of the phospholipids and sphingolipids that make up the membranes. This explains the importance of genetic defects or SNPs in desaturases and elongases for the understanding of vascular and metabolic disease.

Many genetic defects of FA metabolism are either connected to disturbed FA transport, processing or b-oxidation and mostly lead to impaired energy management in the human body that is further challenged by nutrition. Only relatively few genes are affected in these disorders, but their variations have major impact on FA metabolism in humans. One possibility of data assessment from GWAS is the lipid species specific illustration of positive and negative correlations between SNPs and lipid species, which directly depict the “SNP hot spots” that warrant further research.

There are a number of metabolic nodes that constitute the crossroads on this map of pathways. One of those hubs is related to phosphatidic acid, the precursor for several phospholipids, polyglycerophospholipids and diacylglycerides and thus genetical alterations or metabolic overload ultimately lead to storage of FAs as triacylglycerols and diabesity. It can be predicted that these hubs are major targets for lipidomic strategies in diagnostics and treatment, because their increase or decrease is an important clue for any pathology of lipid metabolism pathway no matter whether the underlying cause is genetic, due to metabolic overload or both.

Acknowledgment This work was supported by the "LipidomicNet" (Proposal Number 202272) project, funded under seventh framework program of the EU commission as well as the BMBF network project "Systems Biology Consortium on Metabotypes (SysMBo)".

References

1. Giacco F, Brownlee M (2010) Oxidative stress and diabetic complications. Circ Res 107(9):1058–1070
2. Jump DB (2011) Fatty acid regulation of hepatic lipid metabolism. Curr Opin Clin Nutr Metab Care 14(2):115–120
3. Smith S, Witkowski A, Joshi AK (2003) Structural and functional organization of the animal fatty acid synthase. Prog Lipid Res 42(4):289–317
4. Chirala SS, Wakil SJ (2004) Structure and function of animal fatty acid synthase. Lipids 39(11):1045–1053
5. Swinnen J, Brusselmans K, Verhoeven G (2006) Increased lipigenesis in cancer: new players, novel targets. Curr Opin Clin Nutr Metab Care 9:358–365
6. Turyn J et al (2003) Increased activity of glycerol 3-phosphate dehydrogenase and other lipigenic enzymes in human bladder cancer. Horm Metab Res 35:565–569
7. Kuhajda FP (2000) Fatty acid synthase and human cancer: new perspectives on its role in tumor biology. Nutrition 16:202–208
8. Dowell P, Hu Z, Lane MD (2005) Monitoring energy balance: metabolites of fatty acid synthesis as hypothalamic sensors. Annu Rev Biochem 74:515–534
9. Munday MR (2002) Regulation of mammalian acetyl-CoA carboxylase. Biochem Soc Trans 30(Pt 6):1059–1064
10. Brownsey RW, Boone AN, Elliott JE, Kulpa JE, Lee WM (2006) Regulation of acetyl-CoA carboxylase. Biochem Soc Trans 34(Pt 2):223–227
11. Adida A, Spener F (2002) Intracellular lipid binding proteins and nuclear receptors involved in branched-chain fatty acid signaling. Prostaglandins Leukot Essent Fatty Acids 67(2–3):91–98
12. Kurzchalia TV, Entchev EV, Schwudke D, Zagoriy V, Matyash V, Bogdanova A, Habermann B et al (2008) LET-767 is required for the production of branched chain and long chain fatty acids in Caenorhabditis elegans. J Biol Chem 283(25):17550–17560
13. Mitchell GA, Kassovska-Bratinova S, Boukaftane Y, Robert MF, Wang SP, Ashmarina L et al (1995) Medical aspects of ketone body metabolism. Clin Invest Med 18(3):193–216
14. Neely JR, Rovetto MJ, Oram JF (1972) Myocardial utilization of carbohydrate and lipids. Prog Cardiovasc Dis 15:289–329
15. McGarry JD, Brown NF (1997) The mitochondrial carnitine palmitoyltransferase system. From concept to molecular analysis. Eur J Biochem 244:1–14
16. Kerner J, Hoppel C (2000) Fatty acid import into mitochondria. Biochim Biophys Acta 1486:1–17

17. Lea W, Abbas AS, Sprecher H, Vockley J, Schulz H (2000) Longchain acyl-CoA dehydrogenase is a key enzyme in the mitochondrial beta-oxidation of unsaturated fatty acids. Biochim Biophys Acta 1485:121–128
18. Kamijo T, Aoyama T, Komiyama A, Hashimoto T (1994) Structural analysis of cDNAs for subunits of human mitochondrial fatty acid beta-oxidation trifunctional protein. Biochem Biophys Res Commun 199:818–825
19. Eaton S, Bartlett K, Pourfarzam M (1996) Mammalian mitochondrial beta-oxidation. Biochem J 320(Pt 2):345–357
20. Wanders RJ, Vreken P, den Boer ME, Wijburg FA, van Gennip AH, Ijlst L (1999) Disorders of mitochondrial fatty acyl-CoA beta-oxidation. J Inherit Metab Dis 22:442–487
21. Wenz T, Hielscher R, Hellwig P, Schagger H, Richers S, Hunte C (2009) Role of phospholipids in respiratory cytochrome bc(1) complex catalysis and supercomplex formation. Biochim Biophys Acta 1787(6):609–616
22. McKenzie M, Lazarou M, Thorburn DR, Ryan MT (2006) Mitochondrial respiratory chain supercomplexes are destabilized in Barth Syndrome patients. J Mol Biol 361(3):462–469
23. Knights KM (1998) Role of hepatic fatty acid: coenzyme A ligases in the metabolism of xenobiotic carboxylic acids. Clin Exp Pharmacol Physiol 25(10):776–782
24. Mukherji M, Schofield CJ, Wierzbicki AS, Jansen GA, Wanders RJ, Lloyd MD (2003) The chemical biology of branched-chain lipid metabolism. Prog Lipid Res 42(5):359–376
25. Glaser C, Lattka E, Rzehak P, Steer C, Koletzko B (2011) Genetic variation in polyunsaturated fatty acid metabolism and its potential relevance for human development and health. Matern Child Nutr 7(Suppl 2):27–40
26. Jeffrey BG, Weisingerb HS, Neuringer M, Mitcheli DC (2001) The role of docosahexaenoic acid in retinal function. Lipids 36:859–871
27. Salem N Jr, Litman B, Kim HY, Gawrisch K (2001) Mechanisms of action of docosahexaenoic acid in the nervous system. Lipids 36:945–959
28. Uauy R et al (2001) Essential fatty acids in visual and brain development. Lipids 36:885–895
29. Brash AR (2001) Arachidonic acid as a bioactive molecule. J Clin Invest 107:1339–1345
30. Osumi N (2010) Fatty acid signal, neurogenesis, and psychiatric disorders. Nihon Shinkei Seishin Yakurigaku Zasshi 30(3):141–148
31. Fitzpatrick FA, Soberman R (2001) Regulated formation of eicosanoids. J Clin Invest 107:1347–1351
32. Zhou L, Nilsson A (2001) Sources of eicosanoid precursor fatty acid pools in tissues. J Lipid Res 42:1521–1542
33. Szefel J, Piotrowska M, Kruszewski WJ, Jankun J, Lysiak-Szydlowska W, Skrzypczak-Jankun E (2011) Eicosanoids in prevention and management of diseases. Curr Mol Med 11(1):13–25
34. Toborek M, Lee YW, Kaiser S, Hennig B (2002) Measurement of inflammatory properties of fatty acids in human endothelial cells. Methods Enzymol 352:198–219
35. Lauritzen I et al (2000) Polyunsaturated fatty acids are potent neuroprotectors. EMBO J 19:1784–1793
36. Bousquet M, Gibrat C, Saint-Pierre M, Julien C, Calon F, Cicchetti F (2009) Modulation of brain-derived neurotrophic factor as a potential neuroprotective mechanism of action of omega-3 fatty acids in a parkinsonian animal model. Prog Neuropsychopharmacol Biol Psychiatry 33(8):1401–1408, 13
37. Kelley DS (2001) Modulation of human immune and inflammatory responses by dietary fatty acids. Nutrition 17:669–673
38. Stables MJ, Gilroy DW (2011) Old and new generation lipid mediators in acute inflammation and resolution. Prog Lipid Res 50(1):35–51
39. Dewailly E et al (2001) n-3 fatty acids and cardiovascular disease risk factors among the Inuit of Nunavik. Am J Clin Nutr 74:464–473
40. Martinez M (2001) Restoring the DHA levels in the brains of Zellweger patients. J Mol Neurosci 16:309–316

41. Ferdinandusse S, Jansen GA, Waterham HR, van Roermund CW et al (2001) Peroxisomal fatty acid alpha- and beta-oxidation in humans: enzymology, peroxisomal metabolite transporters and peroxisomal diseases. Biochem Soc Trans 29(Pt 2):250–267
42. Sprecher H, Chen Q, Yin FQ (1999) Regulation of the biosynthesis of 22:5n-6 and 22:6n-3: a complex intracellular process. Lipids 34:S153–S156
43. Broughton KS, Wade JW (2002) Total fat and (n-3:n-6) fat ratios influence eicosanoid production in mice. J Nutr 132:88–94
44. Fischer SM, Cameron GS, Baldwin JK, Jasheway DW, Patrick KE, Belury MA (1989) The arachidonic acid cascade and multistage carcinogenesis in mouse skin. Prog Clin Biol Res 298:249–264
45. Zechner R, Kienesberger PC, Haemmerle G, Zimmermann R, Lass A (2009) Adipose triglyceride lipase and the lipolytic catabolism of cellular fat stores. J Lipid Res 50(1):3–21
46. Watt MJ, Steinberg GR (2008) Regulation and function of triacylglycerol lipases in cellular metabolism. Biochem J 414(3):313–325
47. Kimmel AR, Brasaemle DL, McAndrews-Hill M, Sztalryd C, Londos C (2010) Adoption of PERILIPIN as a unifying nomenclature for the mammalian PAT-family of intracellular lipid storage droplet proteins. J Lipid Res 51(3):468–471
48. Paul A, Chan L, Bickel PE (2008) The PAT family of lipid droplet proteins in heart and vascular cells. Curr Hypertens Rep 10(6):461–466
49. Kudo I, Murakami M (2002) Phospholipase A2 enzymes. Prostaglandins Other Lipid Mediat 68–69:3–58
50. Akiba S, Sato T (2004) Cellular function of calcium-independent phospholipase A2. Biol Pharm Bull 27(8):1174–1178
51. Ravichandran KS, Lorenz U (2007) Engulfment of apoptotic cells: signals for a good meal. Nat Rev Immunol 7(12):964–974
52. Gao S et al (2006) Phospholipid hydroxyalkenals, a subset of recently discovered endogenous CD36 ligands, spontaneously generate novel furan-containing phospholipids lacking CD36 binding activity in vivo. J Biol Chem 281(42):31298–31308
53. Botchkarev VA et al (1997) A simple immunofluorescence technique for simultaneous visualization of mast cells and nerve fibers reveals selectivity and hair cycle-dependent changes in mast cell-nerve fiber contacts in murine skin. Arch Dermatol Res 289(5):292–302
54. Ulmer JB, Donnelly JJ, Liu MA (1994) Presentation of an exogenous antigen by major histocompatibility complex class I molecules. Eur J Immunol 24(7):1590–1596
55. Burger KN, Demel RA, Schmid SL, de Kruijff B (2000) Dynamin is membrane-active: lipid insertion is induced by phosphoinositides and phosphatidic acid. Biochemistry 39:12485–12493
56. Arneson LS, Kunz J, Anderson RA, Traub LM (1999) Coupled inositide phosphorylation and phospholipase D activation initiates clathrin-coat assembly on lysosomes. J Biol Chem 274:17794–17805
57. Schmidt A, Wolde M, Thiele C, Fest W, Kratzin H, Podtelejnikov AV, Witke W, Huttner WB, Huttner HD (1999) Endophilin I mediates synaptic vesicle formation by transfer of arachidonate to lysophosphatidic acid. Nature 401:133–141
58. Bi K, Roth MG, Ktistakis NT (1997) Phosphatidic acid formation by phospholipase D is required for transport from the endoplasmic reticulum to the Golgi complex. Curr Biol 7:301–307
59. Chen YG, Siddhanta A, Austin CD, Hammond SM, Sung TC, Frohman MA, Morris AJ, Shields D (1997) Phospholipase D stimulates release of nascent secretory vesicles from the trans-Golgi network. J Cell Biol 138:495–504
60. Weigert R et al (1999) CtBP/BARS induces fission of Golgi membranes by acylating lysophosphatidic acid. Nature 402:429–433
61. Siddhanta DS (1998) Secretory vesicle budding from the trans-Golgi network is mediated by phosphatidic acid levels. J Biol Chem 273:17995–17998
62. Huttner WB, Schmidt A (2000) Lipids, lipid modification and lipid-protein interaction in membrane budding and fission–insights from the roles of endophilin A1 and synaptophysin in synaptic vesicle endocytosis. Curr Opin Neurobiol 10:543–551

63. Fang Y, Vilella-Bach M, Bachmann R, Flanigan A, Chen J (2001) Phosphatidic acid-mediated mitogenic activation of mTOR signaling. Science 294:1942–1945
64. Ghosh S, Strum JC, Sciorra VA, Daniel L, Bell RM (1996) Raf-1 kinase possesses distinct binding domains for phosphatidylserine and phosphatidic acid. Phosphatidic acid regulates the translocation of Raf-1 in 12-O-tetradecanoylphorbol-13-acetate-stimulated Madin-Darby canine kidney cells. J Biol Chem 271:8472–8480
65. Rizzo MA, Shome K, Watkins SC, Romero G (2000) The recruitment of Raf-1 to membranes is mediated by direct interaction with phosphatidic acid and is independent of association with Ras. J Biol Chem 275(31):23911–23918
66. Erickson RW, Langel-Peveri P, Traynor-Kaplan AE, Heyworth PG, Curnutte JT (1999) Activation of human neutrophil NADPH oxidase by phosphatidic acid or diacylglycerol in a cell-free system. Activity of diacylglycerol is dependent on its conversion to phosphatidic acid. Biol Chem 274:22243–22250
67. Waite KA, Wallin R, Qualliotine-Mann D, McPhail LC (1997) Phosphatidic acid-mediated phosphorylation of the NADPH oxidase component p47-phox. Evidence that phosphatidic acid may activate a novel protein kinase. J Biol Chem 272(24):15569–15578
68. Manifava M, Thuring JW, Lim ZY, Packman L, Holmes AB, Ktistakis NT (2001) Differential binding of traffic-related proteins to phosphatidic acid- or phosphatidylinositol (4,5)-bisphosphate-coupled affinity reagents. J Biol Chem 276(12):8987–8994
69. Williger BT, Ho WT, Exton JH (1999) Phospholipase D mediates matrix metalloproteinase-9 secretion in phorbol ester-stimulated human fibrosarcoma cells. J Biol Chem 274:735–738
70. Choi JW et al (2010) LPA Receptors: subtypes and biological actions. Annu Rev Pharmacol 50:157–186
71. Hama K, Aoki J (2010) LPA3, a unique G protein-coupled receptor for lysophosphatidic acid. Prog Lipid Res 49:335–342
72. Bartke N, Hannun YA (2009) Bioactive sphingolipids: metabolism and function. J Lipid Res 50(Suppl):91–96
73. Merrill AH (2002) De novo sphingolipid biosynthesis: a necessary, but dangerous, pathway. J Biol Chem 277:25843–25846
74. Rao RP, Acharya JK (2008) Sphingolipids and membrane biology as determined from genetic models. Prostaglandins Other Lipid Mediat 85:1–16
75. Hannun YA, Luberto C, Argraves KM (2001) Enzymes of sphingolipid metabolism: from modular to integrative signaling. Biochem J 40:4893–4903
76. Maceyka M et al (2005) SphK1 and SphK2, sphingosine kinase isoenzymes with opposing functions in sphingolipid metabolism. J Biol Chem 280:37118–37129
77. Hornemann T, Wei Y, von Eckardstein A (2007) Is the mammalian serine palmitoyltransferase a high-molecular-mass complex? Biochem J 405(1):157–164
78. Hornemann T, Penno A, Rütti MF, Ernst D, Kivrak-Pfiffner F, Rohrer L, von Eckardstein A (2009) The SPTLC3 subunit of serine palmitoyltransferase generates short chain sphingoid bases. J Biol Chem 284:26322–26330
79. Rotthier A et al (2011) Characterization of two mutations in the SPTLC1 subunit of serine palmitoyltransferase associated with hereditary sensory and autonomic neuropathy type I. Hum Mutat 32:E2211–E2225
80. Penno A et al (2010) Hereditary sensory neuropathy type 1 is caused by the accumulation of two neurotoxic sphingolipids. J Biol Chem 285:11178–11187
81. Zitomer NC, Mitchell T, Voss KA, Bondy GS, Pruett ST, Garnier-Amblard EC et al (2009) Ceramide synthase inhibition by fumonisin B1 causes accumulation of 1-deoxysphinganine: a novel category of bioactive 1-deoxysphingoid bases and 1-deoxydihydroceramides biosynthesized by mammalian cell lines and animals. J Biol Chem 284(8):4786–4795
82. Cowart LA, Hannun YA (2007) Selective substrate supply in the regulation of yeast de novo sphingolipid synthesis. J Biol Chem 282:12330–12340
83. Holland WL, Summers SA (2008) Sphingolipids, insulin resistance, and metabolic disease: new insights from in vivo manipulation of sphingolipid metabolism. Endocr Rev 29:381–402

84. Enomoto A, Omae F, Miyazaki M, Kozutsumi Y, Yubisui T, Suzuki A (2006) Dihydroceramide: sphinganine C-4-hydroxylation requires Des2 hydroxylase and the membrane form of cytochrome b5. Biochem J 397(2):289–295, 15
85. Merrill AH, Sullards MC, Allegood JC, Kelly S, Wang E (2005) Sphingolipidomics: high-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry. Methods 36:207–224
86. Mizutani Y, Mitsutake S, Tsuji K, Kihara A, Igarashi Y (2009) Ceramide biosynthesis in keratinocyte and its role in skin function. Biochimie 91:784–790
87. Van Meer G (2005) Cellular lipidomics. EMBO J 24:3159–3165
88. Ohno Y, Suto S, Yamanaka M, Mizutani Y, Mitsutake S, Igarashi Y, Sassa T, Kihara A (2010) ELOVL1 production of C24 acyl-CoAs is linked to C24 sphingolipid synthesis. Proc Natl Acad Sci 107:18439–18444
89. Sandhoff R (2010) Very long chain sphingolipids: tissue expression, function and synthesis. FEBS Lett 584:1907–1913
90. Iwabuchi K et al (2008) Involvement of very long fatty acid-containing lactosylceramide in lactosylceramide-mediated superoxide generation and migration in neutrophils. Glycoconj J 25:357–374
91. Oh CS, Toke DA, Mandala S, Martin CE (1997) ELO2 and ELO3, homologues of the saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. J Biol Chem 272:17376–17384
92. McMahon A, Butovich IA, Kedzierski W (2011) Epidermal expression of an Elovl4 transgene rescues neonatal lethality of homozygous Stargardt disease-3 mice. J Lipid Res 52:1128–1138
93. Hama H (2010) Fatty acid 2-Hydroxylation in mammalian sphingolipid biology. BBA Mol Cell Biol L 1801:405–414
94. Mizutani Y, Kihara A, Chiba H, Tojo H, Igarashi Y (2008) 2-Hydroxy-ceramide synthesis by ceramide synthase family: enzymatic basis for the preference of FA chain length. J Lipid Res 49:2356–2364
95. Schaeren-Wiemers N, Van Der Bijl P, Schwab ME (1995) The UDP-galactose: ceramide galactosyltransferase: expression pattern in oligodendrocytes and schwann cells during myelination and substrate preference for hydroxyceramide. J Neurochem 65:2267–2278
96. Jungersted JM, Hellgren LI, Jemec GBE, Agner T (2008) Lipids and skin barrier function – a clinical perspective. Contact Dermatitis 58:255–262
97. Van Smeden J, Hoppel L, van der Heijden R, Hankemeier T, Vreeken RJ, Bouwstra JA (2011) LC/MS analysis of stratum corneum lipids: ceramide profiling and discovery. J Lipid Res 52:1211–1221
98. Hu W, Ross JS, Geng T, Brice SE, Cowart LA (2011) Differential regulation of Dihydroceramide desaturase by palmitate vs. monounsaturated fatty acids: implications to insulin resistance. J Biol Chem 286:16596–16606
99. Geeraert L, Mannaerts GP, van Veldhoven PP (1997) Conversion of dihydroceramide into ceramide: involvement of a desaturase. Biochem J 327:125–132
100. Kudo N, Kumagai K, Tomishige N, Yamaji T, Wakatsuki S, Nishijima M, Hanada K, Kato R (2008) Structural basis for specific lipid recognition by CERT responsible for nonvesicular trafficking of ceramide. Proc Natl Acad Sci 105:488–493
101. Yasuda S et al (2001) A novel inhibitor of ceramide trafficking from the endoplasmic reticulum to the site of sphingomyelin synthesis. J Biol Chem 276(47):43994–44002
102. Tafesse FG, Ternes P, Holthuis JCM (2006) The multigenic sphingomyelin synthase family. J Biol Chem 281:29421–29425
103. Yamaji T, Kumagai K, Tomishige N, Hanada K (2008) Two sphingolipid transfer proteins, CERT and FAPP2: their roles in sphingolipid metabolism. IUBMB Life 60:511–518
104. Rao CS, Lin X, Pike HM, Molotkovsky JG, Brown RE (2004) Glycolipid transfer protein mediated transfer of glycosphingolipids between membranes: a model for action based on kinetic and thermodynamic analyses. Biochemistry 43(43):13805–13815
105. Contreras F-X, Sánchez-Magraner L, Alonso A, Goñi FM (2010) Transbilayer (flip-flop) lipid motion and lipid scrambling in membranes. FEBS Lett 584:1779–1786

106. Hisano Y, Kobayashi N, Kawahara A, Yamaguchi A, Nishi T (2011) The sphingosine 1-phosphate transporter, SPNS2, functions as a transporter of the phosphorylated form of the immunomodulating agent FTY720. J Biol Chem 286:1758–1766
107. Gupta G, Surolia A (2010) Glycosphingolipids in microdomain formation and their spatial organization. FEBS Lett 584:1634–1641
108. Nagao K, Takahashi K, Hanada K, Kioka N, Matsuo M, Ueda K (2007) Enhanced ApoA-I-dependent cholesterol efflux by ABCA1 from sphingomyelin-deficient chinese hamster ovary cells. J Biol Chem 282:14868–14874
109. Cerbon J (2003) Diacylglycerol generated during sphingomyelin synthesis is involved in protein kinase C activation and cell proliferation in Madin-Darby canine kidney cells. Biochem J 373:917–924
110. Meng A, Luberto C, Meier P, Bai A, Yang X, Hannun YA, Zhou D (2004) Sphingomyelin synthase as a potential target for D609-induced apoptosis in U937 human monocytic leukemia cells. Exp Cell Res 292:385–392
111. Ding T, Li Z, Hailemariam T, Mukherjee S, Maxfield FR, Wu MP, Jiang XC (2008) SMS overexpression and knockdown: impact on cellular sphingomyelin and diacylglycerol metabolism, and cell apoptosis. J Lipid Res 49:376–385
112. Iwabuchi K, Nakayama H, Iwahara C, Takamori K (2010) Significance of glycosphingolipid fatty acid chain length on membrane microdomain-mediated signal transduction. FEBS Lett 584:1642–1652
113. Mullen TD, Jenkins RW, Clarke CJ, Bielawski J, Hannun YA, Obeid LM (2011) Ceramide synthase-dependent ceramide generation and programmed cell death. J Biol Chem 286:15929–15942
114. Van Echten-Deckert G, Herget T (2006) Sphingolipid metabolism in neural cells. Biochim Biophys Acta 1758:1978–1994
115. Xu Y-H, Barnes S, Sun Y, Grabowski GA (2010) Multi-system disorders of glycosphingolipid and ganglioside metabolism. J Lipid Res 51:1643–1675
116. Hakomori S (2008) Structure and function of glycosphingolipids and sphingolipids: recollections and future trends. Biochim Biophys Acta 1780:325–346
117. Regina Todeschini A, Hakomori S (2008) Functional role of glycosphingolipids and gangliosides in control of cell adhesion, motility, and growth, through glycosynaptic microdomains. Biochim Biophys Acta 1780:421–433
118. Saito T, Hakomori SI (1971) Quantitative isolation of total glycosphingolipids from animal cells. J Lipid Res 12:257–259
119. Kitatani K, Idkowiak-Baldys J, Hannun YA (2008) The sphingolipid salvage pathway in ceramide metabolism and signaling. Cell Signal 20:1010–1018
120. Nilsson Å (2007) Sphingolipids in the gut? Which are the important issues? Eur J Lipid Sci Tech 109:971–976
121. Katsel P, Li C, Haroutunian V (2007) Gene expression alterations in the sphingolipid metabolism pathways during progression of dementia and alzheimer's disease: a shift toward ceramide accumulation at the earliest recognizable stages of alzheimer's disease? Neurochem Res 32:845–856
122. Futerman AH, van Meer G (2004) The cell biology of lysosomal storage disorders. Nat Rev Mol Cell Biol 5:554–565
123. Kolter T, Sandhoff K (2006) Sphingolipid metabolism diseases. Biochim Biophys Acta 1758:2057–2079
124. Itagaki K, Hauser CJ (2003) Sphingosine 1-phosphate, a diffusible calcium influx factor mediating store-operated calcium entry. J Biol Chem 278:27540–27547
125. Sabbadini RA (2011) Sphingosine-1-phosphate antibodies as potential agents in the treatment of cancer and age-related macular degeneration. Brit J Pharmacol 162: 1225–1238
126. Allende ML, Bektas M, Lee BG, Bonifacino E, Kang J, Tuymetova G, Chen W, Saba JD, Proia RL (2010) Sphingosine-1-phosphate lyase deficiency produces a pro-inflammatory response while impairing neutrophil trafficking. J Biol Chem 286:7348–7358

127. Le Stunff H, Giussani P, Maceyka M, Lépine S, Milstien S, Spiegel S (2007) Recycling of sphingosine is regulated by the concerted actions of sphingosine-1-phosphate phosphohydrolase 1 and sphingosine kinase 2. J Biol Chem 282:34372–34380
128. Wilcken B (2010) Fatty acid oxidation disorders: outcome and long-term prognosis. J Inherit Metab Dis 33(5):501–506
129. Innis SM (1993) Essential fatty acid requirements in human nutrition. Can J Physiol Pharmacol 71(9):699–706
130. Yamaguchi T, Osumi T (2009) Chanarin-Dorfman syndrome: deficiency in CGI-58, a lipid droplet-bound coactivator of lipase. Biochim Biophys Acta 1791(6):519–523
131. Chatrath H, Keilin S, Attar BM (2009) Cholesterol ester storage disease (CESD) diagnosed in an asymptomatic adult. Dig Dis Sci 54(1):168–173
132. Lindberg DA (2009) Acute pancreatitis and hypertriglyceridemia. Gastroenterol Nurs 32(2):75–82
133. Wang H, Eckel RH (2009) Lipoprotein lipase: from gene to obesity. Am J Physiol Endocrinol Metab 297(2):E271–E288
134. Karackattu SL, Trigatti B, Krieger M (2006) Hepatic lipase deficiency delays atherosclerosis, myocardial infarction, and cardiac dysfunction and extends lifespan in SR-BI/apolipoprotein E double knockout mice. Arterioscler Thromb Vasc Biol 26(3):548–554
135. Taschler U et al (2011) Monoglyceride lipase deficiency in mice impairs lipolysis and attenuates diet-induced insulin resistance. J Biol Chem 286(20):17467–17477
136. Streeper RS, Koliwad SK, Villanueva CJ, Farese RV Jr (2006) Effects of DGAT1 deficiency on energy and glucose metabolism are independent of adiponectin. Am J Physiol Endocrinol Metab 291(2):E388–E394
137. Zimmermann R, Lass A, Haemmerle G, Zechner R (2009) Fate of fat: the role of adipose triglyceride lipase in lipolysis. Biochim Biophys Acta 1791(6):494–500
138. Hegyi T, Ostfeld B, Gardner K (1992) Medium chain acyl-coenzyme A dehydrogenase deficiency and SIDS. N J Med 89(5):385–392
139. Roe CR, Millington DS, Norwood DL, Kodo N, Sprecher H, Mohammed BS et al (1990) 2,4-Dienoyl-coenzyme A reductase deficiency: a possible new disorder of fatty acid oxidation. J Clin Invest 85(5):1703–1707
140. Miinalainen IJ, Schmitz W, Huotari A, Autio KJ, Soininen R, van Ver Loren TE et al (2009) Mitochondrial 2,4-dienoyl-CoA reductase deficiency in mice results in severe hypoglycemia with stress intolerance and unimpaired ketogenesis. PLoS Genet 5(7):e1000543
141. Houtkooper RH, Turkenburg M, Poll-The BT, Karall D, Perez-Cerda C, Morrone A et al (2003) The enigmatic role of tafazzin in cardiolipin metabolism. Biochim Biophys Acta 1788(10):2003–2014
142. Reue K, Brindley DN (2008) Thematic review series: glycerolipids. Multiple roles for lipins/phosphatidate phosphatase enzymes in lipid metabolism. J Lipid Res 49(12):2493–2503
143. Oaxaca-Castillo D et al (2007) Biochemical characterization of two functional human liver acyl-CoA oxidase isoforms 1a and 1b encoded by a single gene. Biochem Biophys Res Commun 360(2):314–319
144. Sundaram SS, Bove KE, Lovell MA, Sokol RJ (2008) Mechanisms of disease: inborn errors of bile acid synthesis. Nat Clin Pract Gastroenterol Hepatol 5(8):456–468
145. Kemp S, Wanders R (2010) Biochemical aspects of X-linked adrenoleukodystrophy. Brain Pathol 20(4):831–837
146. Ferrer I, Aubourg P, Pujol A (2010) General aspects and neuropathology of X-linked adrenoleukodystrophy. Brain Pathol 20(4):817–830
147. Wierzbicki AS (2007) Peroxisomal disorders affecting phytanic acid alpha-oxidation: a review. Biochem Soc Trans 35(Pt 5):881–886
148. Ruether K et al (2010) Adult Refsum disease: a form of tapetoretinal dystrophy accessible to therapy. Surv Ophthalmol 55(6):531–538
149. Chuang DT, Chuang JL, Wynn RM (2006) Lessons from genetic disorders of branched-chain amino acid metabolism. J Nutr 136(1 Suppl):243S–249S

150. Deodato F, Boenzi S, Santorelli FM, Dionisi-Vici C (2006) Methylmalonic and propionic aciduria. Am J Med Genet C Semin Med Genet 142C(2):104–112
151. Johnson A, O'Donnell C (2009) An open access database of genome-wide association results. BMC Med Genet 10:6
152. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR et al (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42(8):684–687
153. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466(7307):707–713
154. Hicks AA et al (2009) Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet 5(10):e1000672
155. Gieger C et al (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4(11):e1000282
156. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42(2):137–141
157. Schaeffer L et al (2006) Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. Hum Mol Genet 15:1745–1756
158. Fenger M, Linneberg A, Jorgensen T, Madsbad S, Sobye K, Eugen-Olsen J, Jeppesen J (2011) Genetics of the ceramide/sphingosine-1-phosphate rheostat in blood pressure regulation and hypertension. BMC Genet 12:44

Chapter 14
Mapping Metabolomic Quantitative Trait Loci (mQTL): A Link Between Metabolome-Wide Association Studies and Systems Biology

Marc-Emmanuel Dumas and Dominique Gauguier

1 Introduction

Large-scale molecular epidemiology studies have assessed the potential of metabolic profiling and metabolic phenotyping for biomarker discovery [1], which eventually resulted in the introduction of concept of Metabolome-Wide Association Studies [2]. This metabolic phenotyping approach successfully identified metabolic biomarkers related to hypertension. However, these studies point towards the crucial need to estimate genetic variance and heritability for these metabolic phenotypes associated to disease states and to identify genes influencing metabolism in general.

Genome-wide association studies (GWAS) are performed in increasingly large cohorts. Among future steps are the mapping of biomarkers for mechanisms that would predict disease onset and progression. Quantitative genetic analysis of gene expression can generate the required information for genetic analysis of molecular phenotypes [3]. Mapping genome-wide transcriptomic quantitative traits was originally developed in models and applied in humans [4, 5]. It remains mostly based on the analysis of the transcriptome in cell lines, which do not necessarily accurately reflect in situ gene expression and access to biopsies of organs that are central to a pathology (e.g. pancreas in diabetes) is often impossible.

In a genetical genomics context [3], the metabolic complement presents a series of advantages over gene expression products i.e. gene transcripts [5] or proteins [6]. One of these advantages is the fact that metabolic profiles represent hypothesis-free metabolic endpoints at the systems level. Experimental models of human diseases

M.-E. Dumas, Ph.D., M.Eng., M.Sc., B.Eng., B.Sc. (✉)
Surgery and Cancer, Imperial College London,
Sir Alexander Fleming Building Exhibition Road, London SW7 2AZ, United Kingdom
e-mail: m.dumas@imperial.ac.uk

D. Gauguier, Ph.D.
INSERM U872, Cordeliers Research Centre,
15 Rue de l'Ecole de Médecine, Paris 75006, France
e-mail: Dominique.gauguier@crc.jussieu.fr

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_14,

provide several advantages including short generation time, inbred genetic backgrounds, well-conserved metabolic pathways (e.g. glycolysis, gluconeogenesis, etc.…), access to tissues that are usually difficult to collect in humans and simultaneous analysis of multiple tissues that collectively regulate metabolic pathways that, when altered, are the cause of diseases.

Mapping metabolic traits onto the genome and the subsequent identification of quantitative traits loci associated with these phenotypes (mQTL) represents are very active areas of modern genetic research. The first implementation of mQTL mapping was made in plants [7, 8], then in mammalian models [9] and was quickly followed by the development of metabolomic GWAS in humans cohorts [10–12].

This chapter synthesises the individual components required for genetic analysis of quantitative variables of the metabolome, through quantitative trait locus (QTL) mapping in rodent models, and emphasises the importance of complementary expertise and multidisciplinary approaches in this emerging field of research. These include details of:

- Experimental cohorts of hybrids
- A protocol/experimental design and SOPs
- Biological samples (biopsies, biofluids)
- Quantitative phenotypes
- Genetic markers and genetic maps
- Statistical tools such as R/QTL [13], PLINK [14]
- Network biology of complex traits

2 Genetic Crosses and Mapping Panels

Analysis of the genetic basis of complex phenotypes in mammalian species is in theory simplified in inbred models developed in rats and mice, which are genetically homogeneous within a strain and can be intercrossed to produce cohorts of hybrid animals, each carrying alleles that can be traced back to one of the founder strains (Fig. 14.1). This strategy has been particularly successful to map the genetic control of quantitative traits, including primarily phenotypes related to complex diseases such as blood pressure [15]and blood glucose and insulin secretion [16]. Genetic studies of complex traits require the production of hybrid individuals, generally F2 or backcross (BC) cohorts or recombinant inbred (RI) strains, which are used to generate genetic and phenotypic heterogeneity in order to test the co-segregation of alleles and quantitative phenotypic patterns. More elaborated systems (e.g. Heterogeneous stocks -HS) increase the genetic complexity of the hybrid population which is a mosaic of alleles originating for 8 inbred strains randomly bred over many generations [17, 18]. Accurate localisation of the genetic effects relies on recombination rates, which increase in HS when compared to classical F2 of BC, even when large populations are used. As a consequence, large phenotypic effects can be mapped at low genetic resolution in relatively small F2 or BC cohorts

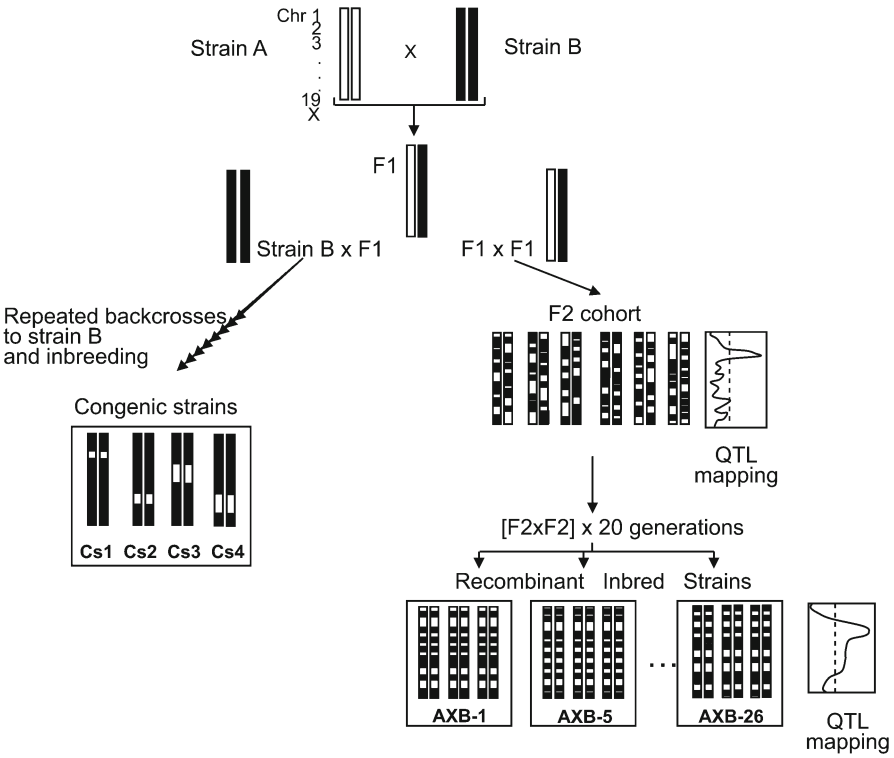


Fig. 14.1 Illustration of genetic crosses generally used to generate QTL mapping panels and to derive congenic strains designed to validate QTLs, fine map the causative gene and characterise its function

(n < 200), whereas high resolution mapping of biological traits can be achieved in HS, which requires much larger populations (n > 500).

Whatever experimental system is used for genetic mapping, the significance and genetic position of variants controlling a phenotype are statistically determined and other experimental systems are required to validate and fine map the genetic effects in genomic intervals, and characterize the biological function of the underlying genetic variants. Congenic strains where segments of chromosomes harboring a genetic locus of interest of one strain are introgressed into a permissive genetic background of another strain currently provide the most reliable way of progressing from genetic mapping of quantitative phenotypic traits to identification of the underlying causative genes (Fig. 14.1) [19]. Multiple sub-strains derived for overlapping introgressed segments are often developed and, by comparing the phenotype in each sub-strain with that of the parental strain, the smallest genomic interval containing the causative gene can be defined.

3 SOP for the Standardisation of Procedures Used in Metabolomic Trait Analysis

As genetic studies of metabolomic phenotypes are carried out with samples from large cohorts of genetically heterogeneous individuals, fully standardised protocols are required for sample collection and preparation. To minimise the effect of confounding variations in experimental conditions on metabolome profiles, biofluid and organ biopsies must be collected at the same time of the day from sex- and age-matched individuals in very specific nutritional conditions (i.e. fasted, free fed). The following sections provide specific technical guidance for the collection of samples that will be used for metabolomic studies.

3.1 Sample Collection Procedures and Nutritional Status

Sample collection protocols have been established and recommendations issued.

Metabolic profiles, unlike genetic material (with the notable exclusion of mRNAs), fluctuate depending on time and physiological status of individuals. Fasting has a strong effect on circulating metabolic levels and has a profound impact on the variation of the metabolome of tissues and on urinary excretion. The "versatility of metabolic profiles" can be advantageously used with a minimum of standardisation of experimental designs, cohort designs and collection protocols for biofluid and tissues.

3.1.1 Biofluids

Urines are usually collected in vessels containing 100 μL of 0.02% NaN3 (w/v) as preservative (for a total volume of about 1 ml). Urinary samples can also be prefiltered using 0.2 μm filters and syringe to remove cellular material during collection. Blood is collected into collection tubes coated with heparin (for blood plasma). Blood samples are then centrifuged for a defined time and clot contact time logged (ideally <30 min). Samples should be stored at −80°C or below and transport on dry ice.

- **Sample requirements:** 200 μL needed for NMR, 50 μL for MS, total 300–350 μL minimum

3.1.2 Tissues

Tissue samples should be snap frozen using liquid nitrogen. The time to freezing should be controlled to minimise the effects of ischemia. Samples are then stored at −80°C or below and transport on dry ice.

About 20 mg minimum of frozen tissue is weighted into an Eppendorf or glass vial. For a 20–30 mg sample 300 μL of cold $CHCl_3$/MeOH (2:1) solution is added before homogenising the tissue using a bead beater. An equivalent volume (300 μL) of HPLC-grade water is added and mixed before centrifuging the homogenate for 10 min at >10,000 g. The lower organic ($CHCl_3$) and upper aqueous (methanol/water) phases are pipetted into separate clean glass vials. Another extraction cycle and pooling with previous fractions will increase extraction recovery. The organic solvents are removed from the samples using a speed vacuum concentrator. The aqueous phase is freeze-dried to remove residual water. All samples to be kept at −80°C until reconstitution with 100 μl water: methanol (1:1) for the organic, and water for the aqueous samples and transferred to glass vials, or 96-well 350 μl plates. For GC-ToFMS analysis, specimens (50 μl) are processed by solid-phase extraction and derivatised using chemistries appropriate to the analyte class.

- **Sample requirements:** 20–30 mg needed for NMR and 20–30 mg for MS, total 50–60 mg minimum.

3.1.3 Cell Cultures and Media

Typically $>10^6$ mammalian cells are required and best results are obtained with 5×10^6 cells in terms of concentrations of metabolites. Cells can be cultured in 75-cm^2 flasks with 12 mL of media, with each flask yielding a single biological replicate. Cultures are removed from incubator and the media are aspirated. The media can be collected in a sterilised tube, centrifuged (4°C, 4 min, 150 × g) to pellet dead cells and the supernatant frozen at −80°C for later analyses. Cells are washed using 1 mL of cold (4°C) PBS to remove media. This step is repeated a couple of times. Cells are then lysed and metabolism quenched by adding 1 mL cold methanol (4°C) to the culture vessel. Cellular material is detached using a cell scraper after 2 min and the resulting suspension is transferred into an Eppendorf or glass tube to dry the sample.

The resulting cell pellet can be extracted using the same procedure as for whole tissue without the need for grinding or sonication. The residual pellet after extraction can be used for sample normalisation. Cell media are analysed using the same procedure are for urine.

- **Sample requirements:** 5×10^6 cells for NMR and MS, 300–350 μL for cell media.

4 Platforms and Tools for High-Throughput Metabolomic Spectral Data Acquisition

A certain number of references have been published for NMR [20, 21] and MS [22].

4.1 Quality Control

To minimise the effects of variation due to instrumental variation, all samples should be run in randomised order. Quality control samples should be inserted for metrological control purpose. High-resolution ^{1}H NMR, GC-MS and UPLC-MS metabolic profiling can be equally performed on tissue samples, plasma and urine.

4.2 NMR

^{1}H NMR spectroscopy is a robust, non-invasive analytical method particularly suited for the qualitative and quantitative analysis of low-molecular weight small molecules and metabolites (see Box 14.1). Untargeted liquid ^{1}H NMR metabolic profiling can be easily performed on biofluids and tissue extracts using 600 MHz NMR spectrometers, which are routinely used for biomedical applications. Standard ^{1}H NMR, spin-echo and diffusion-ordered pulse-sequence experiments are subsequently used to characterise independently the overall metabolic profile, small molecular weight metabolites and lipids components from urine, plasma, serum and tissue samples [21]. For semi-solid samples, such as biopsies or small-organism cultures (*Caenorhabditis elegans*, *Drosophila*), an alternative to sample extraction consists in using ^{1}H High Resolution Magic Angle Spinning (HR-MAS) NMR spectroscopy [20, 23, 24].

4.3 MS

Untargeted UPLC-MS metabolic profiling is ideal for urine and tissue extracts using a time-of-flight (ToF) instrument to provide a broad spectrum, high sensitivity profile. Polar (C_{18} and HILIC columns) and non-polar (C_{18} column) fractions are analysed using positive and negative electrospray modes. Within this global profiling analysis, bile acids, small organic acids and phospholipids can also be assayed through a more targeted approach. Selected discriminatory metabolites may be quantified through the use of ^{13}C-labelled or deuterated standards. This untargeted profiling approach is completed by GC-ToFMS analysis covering short chain fatty acids and free fatty acids.

5 From Spectra to Estimating Metabolite Concentrations

In this section, we describe specific methods applied to spectrum quantification, peak alignment and recoupling, spectral deconvolution and decomposition and eventual dimension reduction of phenotypic measures that can reduce the impact of

Box 14.1 Nuclear Magnetic Resonance Spectroscopy

High Resolution NMR spectroscopy is a robust, quantitative, non-invasive analytical method used to simultaneously measure a wide range of low-molecular weight molecules in complex biological matrices such as tissues or biofluids to provide a metabolic "snapshot" of the sample. High-field NMR spectrometers allow detection of metabolites present at few micromoles per litre.

NMR spectroscopy uses basic properties of the atomic nucleus, made of protons and neutrons. These particles have an intrinsic kinetic moment, or spin. All nuclei having non-zero spin can be observed by NMR - the most important nuclear isotopes used in biological applications are those with a spin =1/2 i.e.,^{1}H, ^{13}C, ^{15}N, ^{31}P.

When a vertical magnetic field (B_0) is applied, spins orient themselves parallel or anti-parallel to this field, resulting in the net magnetization of the sample, and move at a frequency dependant on the nucleus and the magnetic field strength. When a pulse of radiofrequency radiation is applied at the resonant frequency, spins adopt a higher energy level, leading to a new orientation of the magnetization of the sample. When the pulse stops, spins relax to equilibrium. During this relaxation process, the horizontal component of magnetization releases an oscillating voltage, producing the NMR signal, also called Free Induced Decay (FID). A Fourier Transform is then applied to convert the time domain signal of FID in the frequency domain and obtain NMR spectrum. The intensity of the NMR signal is directly related to the number of spins detected over seven orders of magnitude, and therefore metabolite concentrations.

Due to their chemical environment of the *nuclei*, different atoms in a molecule will resonate at different frequencies (known as chemical shift) and spin-spin interactions between different neighboring spins will lead to splitting the NMR signal, a phenomenon known as nuclear coupling. Chemical shift and nuclear coupling information are used to unequivocally assign NMR signals to chemical structures, allowing structural identification of unknown signals (Fig. 14.2).

multiple testing on the summary statistics. Absolute NMR-based quantifications can be derived using an internal standard or an internal radiofrequency standard. For Mass Spectrometry, absolute quantification can be obtained through multiple reactions monitoring (MRM) of several stable isotopes, or by isotopic mass dilution.

Metabolomic spectral data obtained from tissues and biofluids from rodents and human cohorts are usually processed to generate quantitative metabolic phenotypes (>20 K datapoints per sample) for each sample, which are computationally intensive. A few methodological advances in chemometrics have recently been developed in the field of NMR-based and MS-based metabolomics that are particularly pertinent to mQTL mapping.

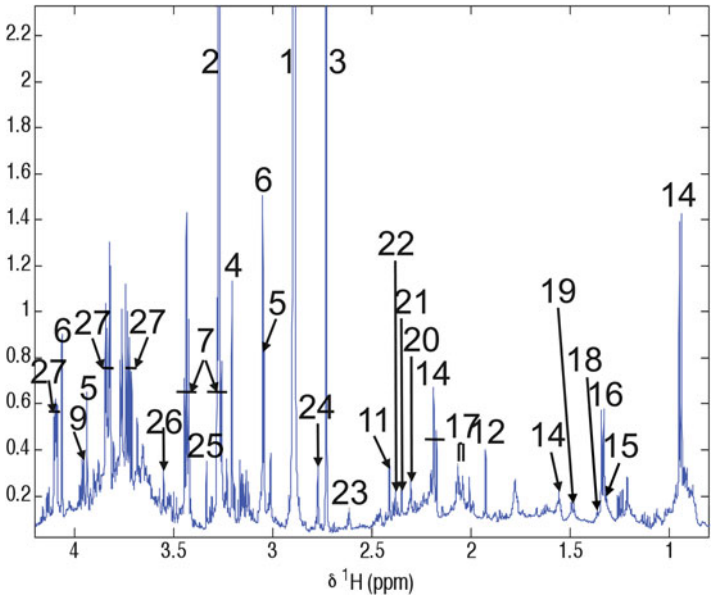


Fig. 14.2 Aliphatic region from a typical 600 MHz ^{1}H NMR spectrum of mouse urine (Adapted from [55]. The horizontal axis represents the NMR frequencies, expressed in chemical shift independent from the spectrometer field (noted *ppm*) rather than in Hz, and the vertical axis corresponds to the intensity of the signal, which is linear over seven orders of magnitudes. Legend: (*1*) TMA (*2*) TMAO (*3*) DMA (*4*) choline (*5*) creatine (*6*) creatinine (*7*) taurine (*8*) allantoin (*9*) hippurate (*10*) citrate (*11*) succinate (*12*) acetate (*13*) formate (*14*) isovalerate (*15*) fucose (*16*) lactate (*17*) *N*-acetyl-glycoprotein(*18*) alpha-hydroxyisobutyrate (*19*) alanine, (*20*) acetoacetate, (*21*) pyruvate (*22*) oxaloacetate (*23*) methylamine (*24*) dimethylglycine (*25*) glycerophosphocholine (*26*) glycine (*27*) glycerate

5.1 Data Preprocessing and Dataset Resolution

MS data are usually pre-processed using instrument manufacturer software such as MarkerLynx™ and XCMS [25] for peak detection, integration and reporting of peak identity as retention time (RT) and m/z pairs. NMR data are pre-processed at a resolution of 10^{-3} ppm, thereby promoting the capacity for biomarker identification [26].

5.2 Normalisation, Alignment and Statistical Recoupling

Mapping mQTLs is intrinsically computer-intensive. With tens of thousands of variables in both the genotyping and the metabolic phenotyping dimensions, the multiple testing correction issue is one of the parameters that needs controlling in order to enhance and maximise mQTL//metabolite association identification.

In order to yield a robust modelling of the relationship between the genomic markers and the metabolic phenotypes, NMR and MS data can be normalised using probabilistic normalisation and recursive segment-wise peak alignment [27] to align peaks from calibrated NMR spectra. The RSPA algorithm can be advantageously applied prior to dimension reduction by statistical recoupling [28] on NMR and MS data, to reliably identify metabolic signals in the spectra. Finally, variance-stabilizing variable transforms has been proven to drastically enhance subsequent modelling [29]. One of the key properties of untargeted signal identification is the reduction of the number of tests to be performed by a factor of 100, whilst recovering >99% of metabolic signals [28].

5.3 Modeling by Partial Least Square (PLS) Methods

PLS methods present a series of advantages for a reliable and flexible analysis of NMR-based and MS-based metabolomic data [30]. Typically, NMR and MS spectra consist of multiple overlapping signals, causing rank deficiency (n variables > p individuals) in classical linear regression methods. PLS methods perform well in such multicollinear context. The recent development of Orthogonal Partial Least Square regression (OPLS) allows the computation of PLS regression with orthogonal signal correction (OSC), leading to orthogonal scores. Further details on standard OPLS implementation in metabonomic studies have been given previously [26]. Owing to the mathematical relationships between O-PLS model coefficients and correlation coefficients (see Fonville et al. for a detailed discussion), O-PLS-based mQTL mapping is identical to correlation-based mQTL mapping. This property is particularly interesting for QTL mapping in congenic and recombinant inbred lines, (where the haplotypes are easily identified), which can advantageously be implemented within a canonical O2-PLS framework. This approach offers the possibility to build a single multivariate statistical model of the metabolic effects related to a series of *loci*.

5.4 Safe Structural Assignment and Biomarker Identification

For structural assignment of untargeted discriminatory metabolites, various analytical and statistical structural identification strategies can be applied.

5.4.1 Statistical Spectroscopy

^{1}H, NMR and MS metabolomic datasets on cell extracts, organ and biofluid profiles are analysed using Statistical TOtal Correlation SpectroscopY (STOCSY) [31], focussing on the spectral signals correlated to the biological variation of interest.

Other statistical structural assignment approaches include combining NMR and MS on the same sample-set to aid structural assignment [32], like in Statistical HeterospectroscopY (SHY) [33]. Targeted profiling can also be achieved by fitting pure compound spectra [34] or by using self-modelling curve resolution (SMCR).

5.4.2 Multidimensional Spectroscopy

Safe structural assignment and quantification are always is achieved using a range of homonuclear and heteronuclear 2D NMR experiments, including ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, J-resolved, ^{1}H-^{13}C HSQC and ^{1}H-^{13}C HMQC [35] experiments. NMR resonances from putative markers are compared to existing internal and external databases and confirmed by spiking samples with authentic standards. Tandem mass spectrometry (MS/MS) experiments are performed using UPLC-QToF-MS, for high accuracy (<2 ppm) mass measurements on both parent and fragment ions, thus facilitating identification of empirical formulae.

5.4.3 Databases

Structural assignment is usually coordinated with spectral compound libraries, i.e. Chemspider, the Human Metabolome Database (HMDB), Lipidmap, Lipidbank, the Madison Consortium Database, METLIN, Golm, PRIMe, and NIST, in addition to in-house databases. Finally, further mechanistic studies using ^{13}C and ^{2}H isotopes confirm putative signals and their associated metabolic pathways.

6 Genetic Map Construction

Genetic markers that show evidence of allele variation between the parental strains are required to define genotypes at genetic loci regularly spaced across the genome of each individual hybrid animal in an experimental cohort (Fig. 14.1). Until recently, microsatellites, which differ in length of short repeated sequences (2–10 nucleotides), were the main source of markers for genetic studies. Single nucleotide polymorphisms (SNP) are now the preferred genetic markers as they are much more frequent across the genome and automated high throughput genotyping platforms have been developed to speed up genotype data acquisition in increasingly large cohorts. These platforms have the advantage of improving the reliability of genotype reads and making possible genetic studies in hybrids derived from closely related strains characterised by low polymorphism rates.

The purpose of genetic mapping quantitative traits is to localise the causative genes using a genetic scale (cM) established on the basis of largely random recombination events that have occurred in each hybrid of the population specifically used for phenotype and genotype analyses. Even though genetic markers are

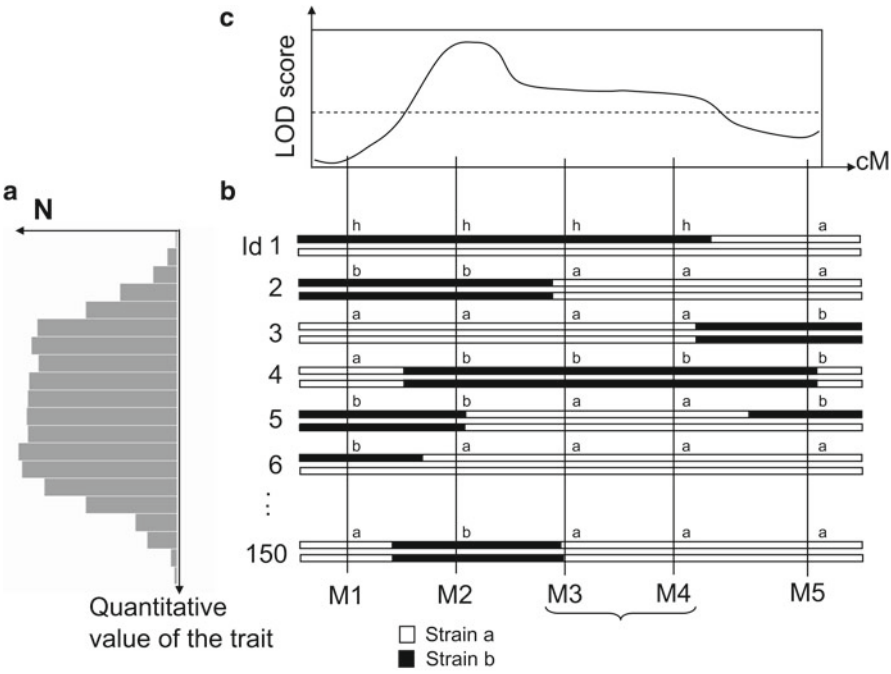


Fig. 14.3 Illustration of genetic mapping applied to map the genetic control of a quantitative phenotype (**a**) using markers showing evidence of allele variability in the population and applied to construct a genetic map using resolved recombination events (**b**) and to test for evidence of linkage to the phenotype (**c**). In this example, the size of the cohort does not allow sufficient recombination events to occur in order to separate markers 3 and 4, which give identical statistics for the all phenotypes quantified in the cohort

precisely localised (Mb) in sequenced genome, construction of genetic maps in the hybrid population is highly recommended to verify the quality of genotype data and the correct position of the markers, and to take into account possible chromosomal rearrangements that may have occurred in the parental inbred strains. In practice, markers that have different genomic locations may be mapped to the same genetic location when lack of recombinant events in the genomic interval between them prevents their actual genetic separation (Fig. 14.3). In this condition their respective contribution to the linkage to the phenotypes cannot be distinguished.

User-friendly softwares have been developed for genetic map construction (e.g. JoinMap, MapMaker) [36], which allow the calculation of genetic distances between markers and the identification of problematic genotypes (e.g. indicating unlikely double recombination or mapping markers beyond chromosome ends) that must be verified prior to linkage mapping. Genetic maps can also be constructed with tools developed in R[13]. This software can be applied for marker imputation obtained through the calculation of conditional genotype probabilities for observed and missing genotypes using hidden Markov models [37].

7 Techniques for Genetic Mapping High-Density Phenotypic Datasets in Large Cohorts

Owing to the continuous distribution of phenotypes characterizing complex genetic disorders in humans (e.g. plasma glucose in type 2 diabetes mellitus, blood pressure in essential hypertension) genetic linkage analysis of quantitative traits in experimental crosses has received much attention by statistical geneticists [38]. The utilization of quantitative values of biological variables represents a key strategy for statistical analysis of data derived from genetic studies of complex traits in order to overcome arbitrary classification of groups of individuals based on a normally distributed phenotype in an experimental cohort. The concept of quantitative trait locus (QTL) mapping (Box 14.2), which uses the continuous values of phenotypic variables in each individual of a population, has been developed and remains an important methodological development in genetic research [39, 40]. Historically only few disease-related phenotypes were considered for QTL analysis [15]. Dissection of a phenotype in discrete physiological variables was later carried out to investigate a possible common etiology between phenotypes contributing to a disease (e.g. glucose intolerance and impaired insulin secretion in type 2 diabetes) [16, 18].

The development of statistical methods designed to use quantitative phenotype variables in genetic linkage studies has had a profound impact in genetic investigations of complex traits. Prior to genetic mapping, all phenotypes in the population must be normalized and tested for correlations. The property of the normal distribution is that 68% of all its observations fall within a range of ±1 standard deviation from the mean, and a range of ±2 standard deviations includes 95% of the values. Obviously, the shape of the sampling distribution becomes normal as the sample size increases and normal distribution is usually obtained for most physiological phenotypes. Validation of normality and evidence of correlation between traits can be easily tested with standard statistical softwares. Several different techniques have been developed for QTL mapping that are based on the utilisation of maximum likelihood techniques to calculate LOD scores at many selected positions in an interval between markers and plotted versus genetic map location. They test by analysis of variance the linkage between genotypes at a succession of marker loci and quantitative variation of a given phenotype. The most popular programs for QTL analysis in the past were Map Manager QT, MAP-MAKER/QTL, JoinMap/QTL and MultiQTL, which still remain valid options for genetic analyses. However, the preferred suite of programs is now R/QTL [13]. Thresholds of statistical significance are determined for each phenotype by permutation testing [41].

Progress in genotyping technology allowing increasingly large panels of genetic markers to be typed and the application of functional genomic platforms as high density molecular phenotype generation tools have provided major advances in genetic and biomedical research for phenotype dissection and disease biomarker discovery [5]. These platforms generate quantitative information on the abundance of transcripts corresponding to over 30,000 genes (Affymetrix or Illumina

Box 14.2 Quantitative Trait Locus (QTL) Mapping in an Experimental Cohort

A QTL is a genetic region containing one or several DNA variants that control the variability of biological traits measured quantitatively. Genetic mapping of QTLs is based on statistical analysis of linkage between the distribution of a phenotype in a cohort of genetically heterogeneous individuals and genotypes at marker loci determined across the genome of the same individuals. QTLs are classically analysed in hybrid individuals characterised by a phenotypic continuum ranging from healthy to disease status. A specific type of QTLs (modifiers) can be detected when linkage analysis is carried out with data from a selected subgroup of affected hybrids in order to investigate the existence of loci controlling disease severity. In the case of polygenic inheritance several independent QTLs are linked to the same biological trait. Pleiotropy is evidenced when a QTL affects different phenotypes, thus providing information on the multiple biological consequences of DNA variants at a single locus. Epistasis is a phenomenon caused by alleles at independent QTLs that interact to affect a phenotypic outcome.

Primarily focused on discrete physiological variables (body weight, blood pressure, glycemia) documenting a disease, QTL analysis has also more recently been applied to investigate the genetic control of individual molecular phenotypes derived from high-density functional genomic datasets, which provide quantitative information on genome expression at the metabolic (metabolome), protein (proteome) and gene transcription (transcriptome) levels. The enormous amount of quantitative traits that can be generated and mapped to the genome raises important issues in statistical genetics, but provides an opportunity to investigate the genetic control of multiple levels of gene expression towards a systems biology approach deciphering causal relationships between gene variants and physiological and molecular phenotypes.

transcriptomes) and the concentration of over 15,000 molecular compounds characterised by specific NMR or MS spectral peaks. Even though this represents an enormous input of biological information for genetic analyses, linkage analysis of increasingly dense genotype and molecular phenotype datasets also has a profound impact on statistical issues, including probabilistic estimates of false positive linkages. Multiple testing, which takes into account the entire phenotype and genotype datasets, is the preferred option to determine thresholds of statistical significance. Considering metabolomic datasets, highly correlated variables in the experimental cohort, such as multiple NMR or MS spectral peaks characterising a single metabolite, artificially increase the number of phenotypes to be tested for linkage and can lead to an overestimation of the threshold of statistical significance.

8 Integrative Systems Biology of mQTL: Metabolite Associations

Associations between metabolites and genetic loci are often very complex (Fig. 14.4). Transforming this complexity into knowledge can be achieved by using integrative systems biology strategies, which describe and help defining functional modules. Metabolic networks share common properties with other biological networks, such as a scale-free topology with embedded modularity [42]. Metabolic networks can be approximated as small worlds (i.e. with few connections between random metabolites [43]. Although our current knowledge of the metabolic networks may not be complete, several network biology strategies can be applied to understand the relationship between mQTL – metabolome associations, from the genome to the metabolome via interaction networks [44].

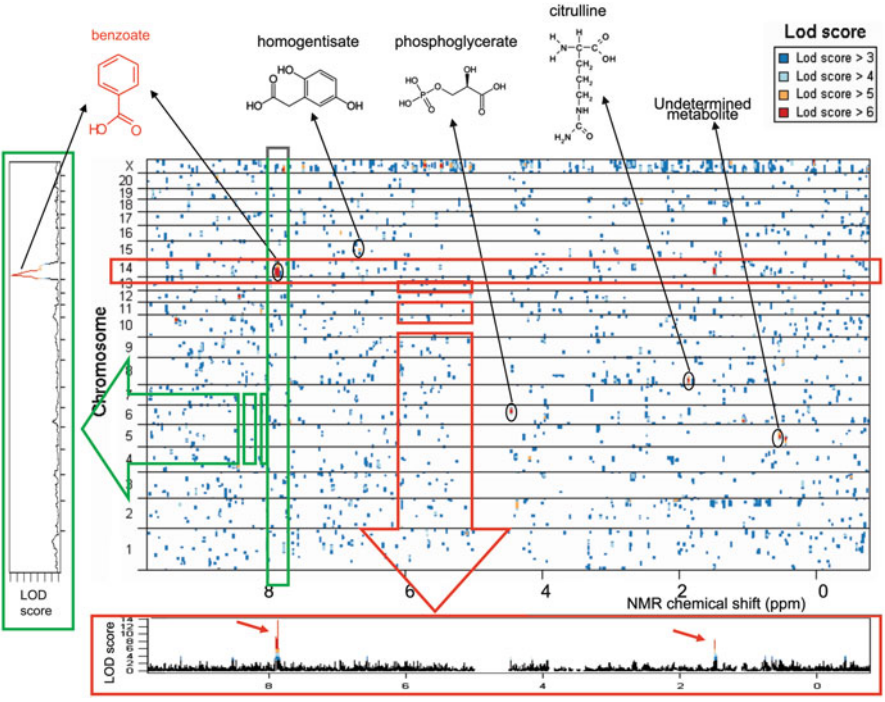


Fig. 14.4 An example of blood plasma mQTL mapping in a F2 cross between Goto-Kakizaki spontaneously diabetic and Brown Norway normoglycemic rats. The highest LOD score is associated with a *cis*-acting mQTL in the sequence of UGT2b7, which encodes a UDP-glucuronosyl transferase directly involved in glucuronidation of benzoate, which generated the signal (Adapted from [9])

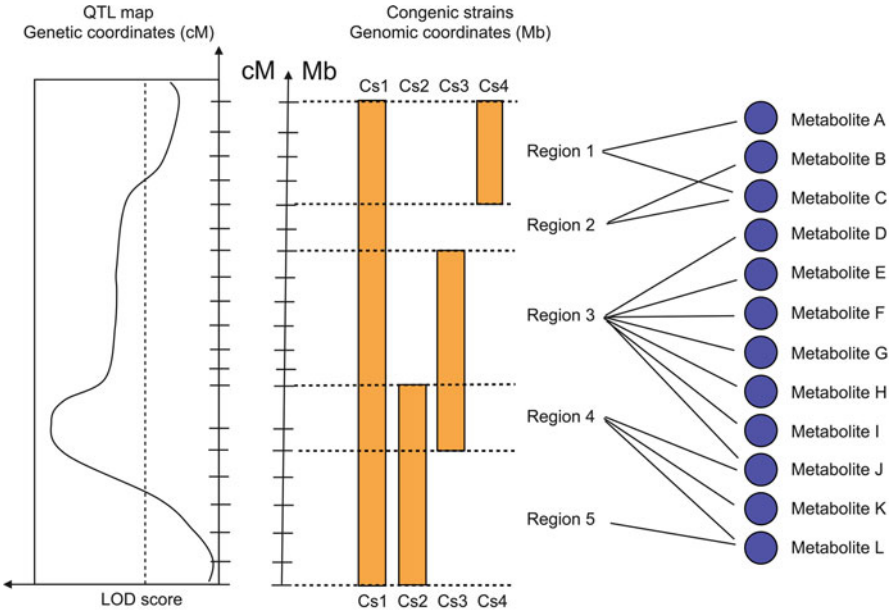


Fig. 14.5 Principle of haplotype – metabotype association networks. The bipartite graph shows how gene variants are connected to their metabolic signature. *Orange bars* indicate the genomic regions a genetically defined QTL transferred from an inbred strain onto the genetic background of another strain in different congenics lines (Cs1–4). Each congenic line has a unique genomic profile and a unique metabolomic profile. Through the use the congenic lines, unique intervals are defined on the genome and each genomic interval will generate a unique metabolic signature

8.1 Haplotype-Metabotype Networks

Mapping mQTLs identifies numerous significant associations between *loci* (or haplotype) and metabolic phenotypes. These associations in both genomic and metabolomic dimensions of biological organisation can be analysed using a systems biology framework. Properties of the "association network" are analysed using graph analysis. The network of genotype-metabotype associations can be mathematically formalized as an undirected bipartite graph G = (Vm,E) composed of two node types Vm – where m = (locus, metabolite) and functional E – where a significant association between a *locus* and a metabolite. The double entry list of significant genotype and metabotype associations corresponds to the adjacency matrix of the bipartite graph (Fig. 14.5).Using graph analysis tools, it is possible to analyse the topology of the network, such as the distribution of connectivities describing which *loci* are poorly, or highly connected to metabolic endpoints.

8.2 *Pathway-Based Network Biology Methods for mQTL Association Post-Processing*

If a *locus* or a gene variant is significantly associated to the concentration of a metabolite, it is also possible that it may also be associated with the concentrations of several neighbouring metabolites in one or more surrounding metabolic pathways, as metabolism does not operate in isolation, but rather in cooperation.

8.2.1 mQTL-Based Metabolite-Set Enrichment Analysis of Metabolic Signatures

A single *locus*, typically in the case of collections of congenic or recombinant inbred lines, can determine the concentration of several metabolites. This multivariate pattern reflects information about biological molecules (i.e. individual metabolites), but also about biological processes when taken altogether as a signature (i.e. metabolic pathways). To test whether a given pathway is affected by a locus, the metabolic signature is compared to the metabolic pathway entries in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] and an enrichment test is performed to identify which pathways are affected [46–48]. This approach was coined metabolite-set enrichment analysis (MSEA), as it shares the same methodological framework as gene-set enrichment analysis (GSEA) [49], but instead of testing enrichment of gene lists, MSEA tests enrichment of metabolite lists.

A direct consequence of performing a MSEA on a metabolic signature related to a particular mQTL locus is the possibility to identify mQTLs affecting entire metabolic pathways, rather than a single metabolite.

8.2.2 Matching mQTLs and eQTLs at the Metabolic Pathway Level

A complementary experiment to mQTL mapping then consists in comparing matching gene expression (or protein) profiles to validate that genetic variants influencing metabolite levels also influence the expression of genes and proteins in the very same pathway.

Such availability of gene-expression and metabolic profiles allows comparing the outputs of MSEA and GSEA that aims at testing the consistency of the biological perturbation throughout the network, based on two (or several) observation modalities, and provides a powerful pathway-level QTL co-localisation approach.

8.2.3 Topological Analysis of Joint Gene-eQTL and Metabolite-mQTL Networks

The above described methods are usually sufficient in the case of *cis*-acting QTL associations influencing both metabolite and gene profiles. *Cis*-acting QTL

associations provide a simple chain of causality: a genetic variant directly influences the expression of an enzyme and the product or substrates are directly affected by this variant, as well as other metabolites or genes in the same pathway. However, when the genetic variant is located near an enzyme or a transporter, the chain of causality between genetic variation and effect at the metabolic level becomes less obvious, and other network biology methods are required to reveal the deeper biological meaning of such associations.

Distances separating pairs of metabolites and enzymes were computed using shortest path lengths within the whole metabolic network. This visualization procedure requires the modeling of the whole metabolic network, which then allows the extraction of the perturbed metabolic network in an unbiased, automated, statistical manner. The complexity of cellular metabolism can be depicted using a network approach. The whole metabolic network is modelled as an undirected multi-labelled graph G = (Nodes, Edges). The set of multi-labelled nodes is defined in [M;E], with M and E corresponding to metabolites and enzymes respectively. The set of edges corresponds to reactions connecting the enzymes with their substrates or products. Due to the partial annotation available for animal models, it is possible to use the more comprehensive Human metabolic network available in the KEGG [45]. Once unspecific hub molecules (H_2O, CO_2, H^+, H_2O_2, NH_3, etc.…) are removed to focus primarily on the exchange of hydrocarbon structures [50], the resulting metabolic network is made of 1,793 metabolites, 942 enzymes and 4,174 reactions [51]. The distance between observed pairs of significant metabolites and enzymes is derived by computing the shortest path length (*spl*) separating them across the reconstructed network.

8.3 Integrated Metabolome and Interactome Mapping

Network biology strategies based on protein-protein interactions can bring a mechanistic understanding in genotype-metabotype associations (typically for *trans*-acting mQTLs) by functionally connecting causal variants to their downstream associated metabolic phenotypes. Databases and literature datasets are used to reconstruct a molecular interaction network encapsulating protein-protein interaction networks [52], including kinases, as well as metabolic and signaling pathways – such as the KEGG database [45].

This framework complements the mapping and enrichment testing developed in MSEA and its topological analysis, as it connects metabolic phenotypes to their causal genes by the signalling and regulation network. The implementation of integrated metabolome and interactome mapping (iMIM) allows identifying key regulatory proteins in the interaction network(Fig. 14.6): the current iMIM network includes 11,500 human proteins and 2,733 metabolites involved in 71,072 protein-protein interactions (PPI), and 34,716 metabolite-enzyme interactions (MEI) [53].

The topology of the resulting iMIM network is then analyzed to visualize functional paths between causal trans-acting variants and their associated metabotypes

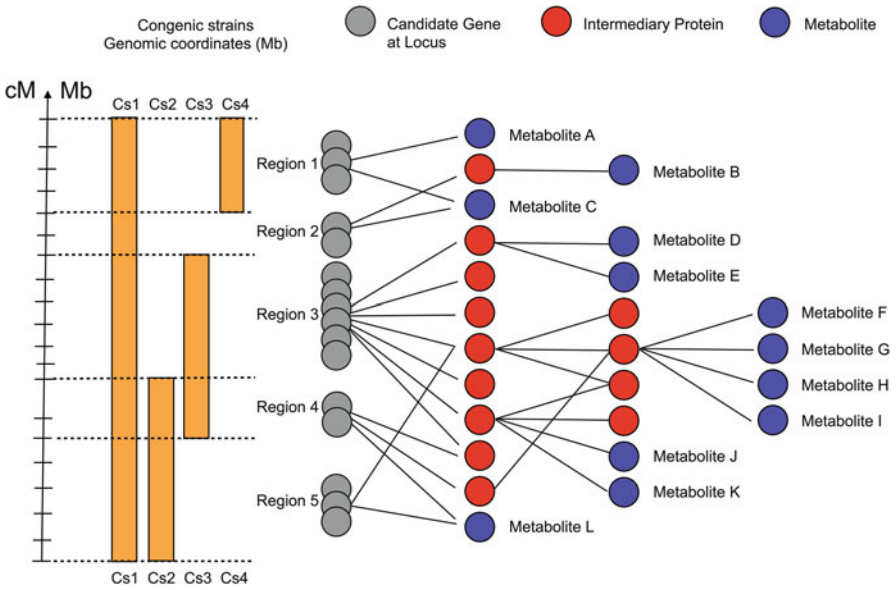


Fig. 14.6 Integrated metabolome-wide intractome-wide mapping. iMIM connects a given causative gene variant to a downstream metabolic phenotype through protein-protein interactions (PPI) and metabolite-enzyme interactions (MEI) networks. In this theoretical example, the minimum number of interactions between a causative gene variant and a metabolite is 1

and to model how genetic variation impacts and propagates through the cellular network until it reaches its metabolic endpoints (Fig. 14.6). Considering a given metabolite M associated with a genetic variant G, the shortest path between G and M through the interactome network is likely to provide a mechanistic link between variants in G and their consequence on metabolite M. In *cis*-acting mQTL cases, the gene-product of G is usually an enzyme, a receptor or a transporter, and the shortest path is 1 (typically, an enzyme catalyzing or producing the metabolite M). When G is not an enzyme or a receptor (i.e. a *trans*-acting mQTL), the shortest path between Gand M becomes: $G \rightarrow$ (interacting protein $\rightarrow$) enzyme $\rightarrow$ metabolite M (Fig. 14.6).

8.4 Topological Analysis of Association Networks: Network Metrics

The topological analysis of KEGG [51], iMIM [53] and haplotype/metabotype networks is the final step of this integration. To perform this topological analysis, several network metrics can be used, by following order of complexity:

– Basic network statistics such as the degree of the network (i.e., the number of edges connected to the most connected node) can be easily derived. This type

of network metrics allows identifying the most connected metabolites or genes (i.e. hubs), as well as the least connected metabolites/genes.

– The scale-free topology of biological networks allows connecting two components of the network in few hops. Deriving the distance between two entries is usually made by computing the length of the shortest path across the network, to derive the shortest path length (spl). Minimizing the *spl* identifies in an unbiased manner the direct connection between an enzyme (or transporter) and its substrates/products. This strategy is often employed to demonstrate the consistency of eQTL and mQTL datasets. Furthermore, this strategy also directly validates the functional relationship between genes and metabolites generating co-located eQTL and mQTL signals.
– Finally, to identify the most likely paths between two sets of entries, deriving measures of centrality of nodes, such as the Betweenness (B), provides a meaningful compilation of all the shortest paths. B reflects the proportion of shortest paths going through a given node, and therefore its centrality to the network [54]. The use of Betweenness-derived metrics such as the pivotal Betweenness (starting from a single node such as a gene variant and connecting it to its metabolic phenotype through the interaction network) identifies key regulatory proteins across the signalling network, which then become ideal candidates for functional validation [53]. This approach is powerful when the functional relevance between a locus and its metabolic response is not direct. This is precisely the case when the locus codes for a signalling protein, which regulates a downstream metabolic pathway.

9 Conclusion

The genetic mapping of quantitative biological variables of the metabolome, which has been optimised in QTL mapping experiments in animal models and plants, is becoming a realistic perspective in human genetics to identify disease predictive biomarkers and in pharmacogenomics. The possibility to carry out genetic studies with time series of biofluid metabolome profiling datasets is particularly promising to also identify specific metabolic signatures underlying the progression of a phenotype. As genotype data in large human cohorts are already available, and with ongoing genome resequencing projects, it is anticipated that GWAS, which have originally analysed the genetic basis of relatively simple phenotypic traits, will soon turn to the mapping of genetic determinants of the human metabolome. This prospect requires the development of improved statistical and analytical tools to deal with the enormous amount of quantitative biological information that a single NMR or MS metabolomic spectrum can generate. Demonstration of causal relationships between co-segregrating QTLs for metabolome and transcriptome variables in experimental cohorts also provides important perspectives in the genetics of systems biology.

Acknowledgements The authors acknowledge support from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° HEALTH-F4-2010-241504 (EURATRANS), the ANR (mQTL ANR-08-GENO-030-02) and the Fondation pour la Recherche Médicale. Marc-Emmanuel Dumas holds a Young Investigator Award from Agence Nationale de la Recherche (ANR-07-JCJC-0042-01) Dominique Gauguier holds a Wellcome Trust Senior Fellowship in basic biomedical science (057733).

References

1. Dumas ME, Maibaum EC, Teague C et al (2006) Assessment of analytical reproducibility of 1 H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. Anal Chem 78(7):2199–2208
2. Holmes E, Loo RL, Stamler J et al (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453(7193):396–400
3. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17(7):388–391
4. Dixon AL, Liang L, Moffatt MF et al (2007) A genome-wide association study of global gene expression. Nat Genet 39(10):1202–1207
5. Schadt EE, Monks SA, Drake TA et al (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929):297–302
6. Klose J, Nock C, Herrmann M et al (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30(4):385–393
7. Schauer N, Semel Y, Roessner U et al (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24(4):447–454
8. Keurentjes JJ, Fu J, de Vos CH et al (2006) The genetics of plant metabolism. Nat Genet 38(7):842–849
9. Dumas ME, Wilder SP, Bihoreau MT et al (2007) Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. Nat Genet 39(5):666–672
10. Gieger C, Geistlinger L, Altmaier E et al (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4(11):e1000282
11. Illig T, Gieger C, Zhai G et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42(2):137–141
12. Suhre K, Wallaschofski H, Raffler J et al (2011) A genome-wide association study of metabolic traits in human urine. Nat Genet 43:565–569
13. Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19(7):889–890
14. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
15. Rapp JP, Wang SM, Dene H (1989) A genetic polymorphism in the renin gene of Dahl rats cosegregates with blood pressure. Science 243(4890):542–544
16. Gauguier D, Froguel P, Parent V et al (1996) Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat. Nat Genet 12(1):38–43
17. Johannesson M, Lopez-Aumatell R, Stridh P et al (2009) A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. Genome Res 19(1):150–158
18. Valdar W, Solberg LC, Gauguier D et al (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38(8):879–887
19. Collins SC, Wallis RH, Wilder SP et al (2006) Mapping diabetes QTL in an intercross derived from a congenic strain of the Brown Norway and Goto-Kakizaki rats. Mamm Genome 17(6):538–547

20. Beckonert O, Coen M, Keun HC et al (2010) High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. Nat Protoc 5(6):1019–1032
21. Beckonert O, Keun HC, Ebbels TM et al (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat Protoc 2(11):2692–2703
22. Want EJ, Wilson ID, Gika H et al (2010) Global metabolic profiling procedures for urine using UPLC-MS. Nat Protoc 5(6):1005–1018
23. Blaise BJ, Giacomotto J, Elena B et al (2007) Metabotyping of Caenorhabditis elegans reveals latent phenotypes. Proc Natl Acad Sci USA 104(50):19808–19812
24. Blaise BJ, Giacomotto J, Triba MN et al (2009) Metabolic profiling strategy of *Caenorhabditis elegans* by whole-organism nuclear magnetic resonance. J Proteome Res 8(5):2542–2550
25. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78(3):779–787
26. Cloarec O, Dumas ME, Trygg J et al (2005) Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1 H NMR spectroscopic metabonomic studies. Anal Chem 77(2): 517–526
27. Veselkov KA, Lindon JC, Ebbels TM et al (2009) Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. Anal Chem 81(1):56–66
28. Blaise BJ, Shintu L, Elena B, Emsley L, Dumas ME, Toulhoat P (2009) Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. Anal Chem 81(15):6242–6251
29. Dumas ME, Debrauwer L, Beyet L et al (2002) Analyzing the physiological signature of anabolic steroids in cattle urine using pyrolysis/metastable atom bombardment mass spectrometry and pattern recognition. Anal Chem 74(20):5393–5404
30. Fonville JM, Richards SE, Barton RH et al (2010) The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. J Chemom 24(11–12):636–649
31. Cloarec O, Dumas ME, Craig A et al (2005) Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1 H NMR data sets. Anal Chem 77(5):1282–1289
32. Dumas ME, Canlet C, Debrauwer L, Martin P, Paris A (2005) Selection of biomarkers by a multivariate statistical processing of composite metabonomic data sets using multiple factor analysis. J Proteome Res 4(5):1485–1492
33. Crockford DJ, Holmes E, Lindon JC et al (2006) Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. Anal Chem 78(2):363–371
34. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem 78(13):4430–4442
35. Dumas ME, Canlet C, Andre F, Vercauteren J, Paris A (2002) Metabonomic assessment of physiological disruptions using 1 H-13C HMBC-NMR spectroscopy combined with pattern recognition procedures performed on filtered variables. Anal Chem 74(10):2261–2273
36. Gauguier D, Samani N (2002) Approaches to the analysis of complex quantitative phenotypes and marker map construction based on the analysis of rat models of hypertension. Methods Mol Biol 195:225–251
37. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69(4):315–324
38. Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. Nat Rev Genet 3(1):11–21
39. Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3(1):43–52
40. Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10(8):565–577

41. Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142(1):285–294
42. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
43. Fell DA, Wagner A (2000) The small world of metabolism. Nat Biotechnol 18(11): 1121–1122
44. Beyer A, Bandyopadhyay S, Ideker T (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 8(9):699–710
45. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
46. Pontoizeau C, Fearnside JF, Navratil V et al (2011) Broad-ranging natural metabotype variation drives physiological plasticity in healthy control inbred rat strains. J Proteome Res 10(4):1675–1689
47. Xia J, Wishart DS (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. Nucleic Acids Res 38:W71–W77
48. Chagoyen M, Pazos F (2011) MBRole: enrichment analysis of metabolomic data. Bioinformatics 27(5):730–731
49. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550
50. Arita M (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101(6):1543–1547
51. Blaise BJ, Navratil V, Domange C et al (2010) Two-dimensional statistical recoupling for the identification of perturbed metabolic networks from NMR spectroscopy. J Proteome Res 9(9):4513–4520
52. Uetz P, Giot L, Cagney G et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623–627
53. Davidovic L, Navratil V, Bonaccorso CM et al (2011) A metabolomic and systems biology perspective on the brain of the fragile X syndrome mouse model. Genome Res 21:2190–2202
54. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41
55. Dumas ME, Barton RH, Toye A et al (2006) Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. Proc Natl Acad Sci USA 103(33):12511–12516

Chapter 15
Metabolic Traits as Intermediate Phenotypes

Florian Kronenberg

1 Introduction

The identification of genes for complex disease has made major progress during the last 5–7 years which was mainly caused by the introduction of hypothesis-free methods such as genome-wide association studies (GWAS) or other systematic screens such as metabolites and the combination of these screens. To be exact, hypothesis-free is not completely accurate since there is indeed a "master-hypothesis" in place meaning that there are genes which are associated with a particular disease or with a particular phenotype or there are gene products and metabolites of these products that are related to the involved pathways related to the phenotype of interest. The hypothesis-free approach has broken new ground which opened the researchers' eyes for pathways one would not necessarily have connected with a certain phenotype if a conservative hypothesis-driven approach would have been followed.

An important misbelief we got rid of over the last decade is that a single genetic effect on a complex phenotype such as cardiovascular disease (CVD), obesity or type 2 diabetes mellitus is strong. In earlier times we expected that certain genetic variants double or triple the risk. Due to the findings in GWAS we have learned that most genetic variants increase the risk for these multifactorial diseases by 5–40%. Relative risks above that range can be considered as exceptionally "low hanging fruits" and are quite rare. The low relative risks associated with certain alleles require very large sample sizes of 10,000 and more study subjects if a GWAS approach is used.

To study the intermediate phenotypes which are associated with the hard endpoints is an interesting complementary approach which might help to elucidate

F. Kronenberg, M.D. (✉)
Division of Genetic Epidemiology, Innsbruck Medical University,
Schöpfstr. 41, Innsbruck 6020, Austria
e-mail: florian.kronenberg@i-med.ac.at

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_15,

the genotype-phenotype associations in a promising way. This chapter explains how this is performed, discusses the advantages and disadvantages and shows also two examples related to lipid metabolism and bilirubin.

2 Characteristics of Intermediate Phenotypes

Intermediate phenotypes are parameters which are considered to be involved in the development of an endpoint of interest such as cardiovascular disease. They represent an important aspect in the pathogenesis of the disease. For the present genetically driven considerations they should be inherited. As Fig. 15.1 illustrates, several of the various intermediate phenotypes contribute to the endpoint of interest. Examples concerning the cardiovascular outcome might be genes influencing cholesterol levels and their pre-products, bilirubin levels or inflammatory processes to mention only a few. By focusing the search for genes to one of these intermediate phenotypes we might increase the chance to find as a first step one or more genes which influence this particular intermediate phenotype. In this phase we are not yet focused on the hard clinical endpoint (e.g. CVD) but on this intermediate (end-point-related) phenotype. By focusing the search on this intermediate phenotype we might considerably decrease the heterogeneity of the phenotype which dramatically increases the power to detect a gene influencing the intermediate phenotype. The closer the gene is related to the investigated intermediate phenotype the clearer the estimation of the influence the investigated genetic variant has on this phenotype. In a second step, after we know the genetic effect size on the intermediate phenotype we are far better able to plan a sufficiently powered study to investigate an association with the clinical endpoint of interest. That means, if by the first step we already detect that a particular genetic variant has only a very small effect on the intermediate phenotype and this intermediate phenotype is not one of the stronger risk factor for the clinical endpoint, we are warned to consider the association study between this genetic variant and the clinical endpoint in a very large number of study subjects.

3 Definition of Intermediate Phenotypes

As we discussed recently [1], intermediate phenotypes have to fulfill the criteria according to Prentice and colleagues [2], Baron and Kenny [3] as well as Gottesman and Gould [4] which is adjusted here for final phenotypes such as cardiovascular disease (see also Fig. 15.2):

1. The intermediate phenotype is associated with the “final” endpoint which could be a disease or a clinical endpoint. In a best case scenario this information should come from a prospective population-based study which supports the prognostic value of the intermediate phenotype of interest.

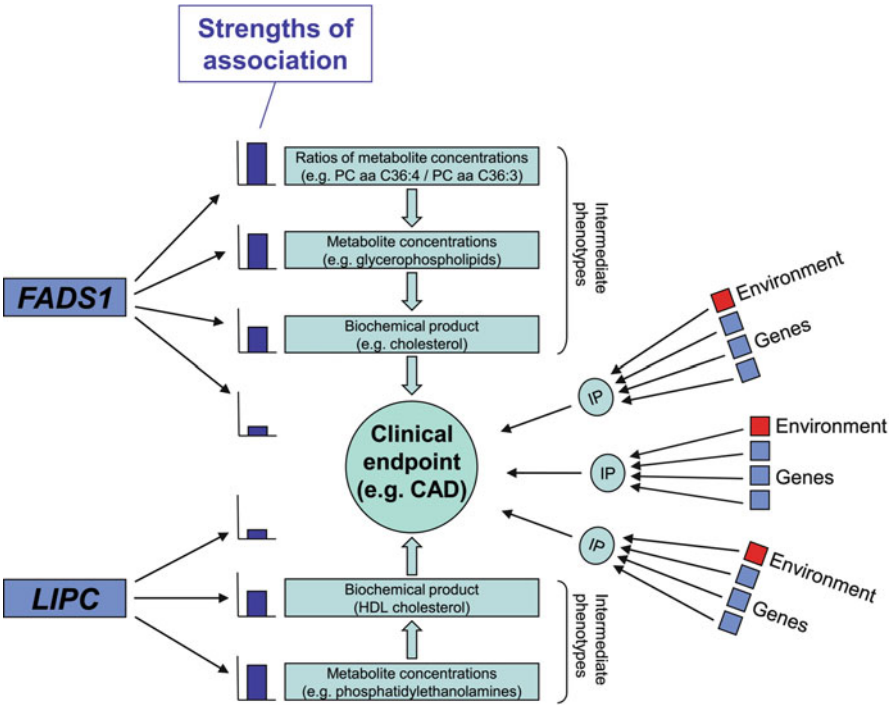


Fig. 15.1 Schematic illustration of the role of intermediate phenotypes (IPs), such as metabolic traits, demonstrated at the examples of two genes that code for major enzymes of the long-chain fatty acid metabolism (*FADS1* and *LIPC*). This demonstrates that new information on the functional basis of the observed associations can be inferred from the biochemical properties of the affected metabolites. Moreover, both genes were previously reported to be associated with common clinical phenotypes, *FADS1* in an extent, which would not attract immediate attention for follow-up in a genome-wide context. Since several genes and pathways are involved in the development of a clinical endpoint, the IP focuses on one pathway (e.g., cholesterol or a given metabotype) which is already known to be involved in the clinical endpoint (e.g. coronary artery disease (CAD)). It is much easier to identify the genes which are associated with the intermediate phenotype since the associations of genetic variation with the intermediate phenotype is much stronger than with the clinical endpoint. Environmental factors interact at different levels with the intermediate phenotypes and thereby add to the variability in the system. The closer the intermediate phenotype is related to the genetic polymorphism, the stronger the association is expected to be. In our case the association reflects enzymatic activity of *FADS1* and *LIPC* which results in very strong effect sizes of the genetically determined metabotype (Reprinted from [10])

2. The intermediate phenotype has to have a heritable component which is supported by twin studies or segregation analysis.
3. The intermediate phenotype lies within the pathway between genetics and disease and is focused on one aspect of the pathogenesis.
4. The intermediate phenotype should be largely unaffected from the disease status. That means this intermediate phenotype should already be present long before the disease develops and should not be changed by the disease. This point is

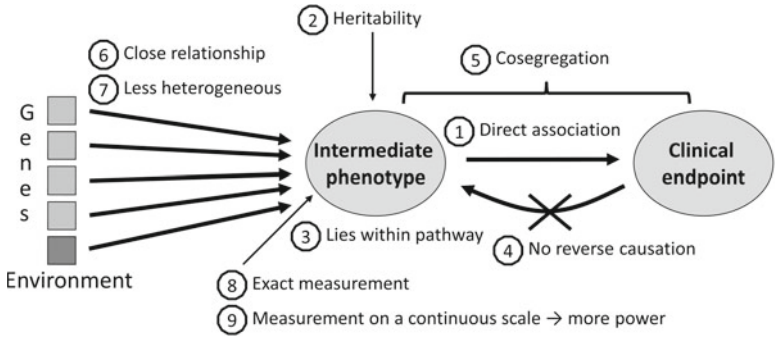


Fig. 15.2 Definition of intermediate phenotypes. For explanation see text

strongly supported if the association was proven in a long-term prospective study with a measurement of the intermediate phenotype long before any disease signs are present.

5. The intermediate phenotype and the disease co-segregate within a given family, which supports point #3 on this list. Some researches request that the intermediate phenotype be observed more often in the unaffected family members compared to the general population.
6. Intermediate phenotypes are related to a certain part of the pathway from the gene or a network of genes to the final phenotype. They should be closer to the gene product and not close to the final phenotype. This is in contrast to the surrogate phenotype which should be as close as possible to the final phenotype. The surrogate phenotype might be easier or earlier to measure than the final phenotype. To give some examples, a surrogate phenotype for CVD could be the measurement of carotid atherosclerosis [5, 6] or the ankle-brachial index [7]. Both are easy to measure, but are limited since they do not fully reflect clinical endpoints of CVD.
7. The number of genetic and non-genetic factors (e.g. lifestyle factors) that influence the intermediate phenotype are easier to identify than those which influence the final pathogenetically heterogeneous phenotype.
8. Under optimal conditions intermediate phenotypes are to be measured exactly, easily, objectively and reproducible. This is usually the case for laboratory parameters. They can be measured in affected patients as well as in healthy controls. Preferably, this is done in long-term prospective population-based studies long before the final phenotype of interest develops. If the intermediate phenotype is measured in diseased populations one has to keep in mind that the intermediate phenotype could be influenced by counter-regulatory mechanisms caused by the disease or its treatment. Blood pressure is a very obvious example and is strongly influenced by antihypertensive medication that is in place in many or most of the affected hypertensive patients. If not considered this creates an underestimated intermediate phenotype.

9. A quantitative intermediate phenotype such as blood glucose or cholesterol concentrations has a further advantage since it is measured on a continuous scale. In contrast, a qualitative parameter (e.g. type 2 diabetes mellitus) does not consider whether a person has blood glucose just above the normal value or whether a highly abnormal value is present. There is no doubt that there is a major difference between a fasting blood glucose of 128 or 320 mg/dl. In a qualitative analysis both values will be treated in the same way (diabetes mellitus present = yes). Therefore, an intermediate phenotype on a continuous scale uses the entire range of a phenotype from completely normal to highly abnormal with all color gradients from black to white. This kind of quantitative data analysis is usually statistically more powerful than a qualitative analysis.

4 Examples for Intermediate Phenotypes

4.1 Metabolites as the Most "Neighboring" Intermediate Phenotype

As mentioned above, ideally the intermediate phenotype should be as close as possible to the gene product. How this can be improved was recently shown by using ratios of metabolites instead of each metabolite itself to find the genes which have an influence on the metabolites (Fig. 15.3). If we consider that in a ratio of two metabolites, one metabolite is the substrate and the other is the product, we expect that a GWAS performed with all three phenotypes (substrate, product and substrate/product ratio) would result in a pronounced gain in information of the GWAS when performed with the substrate/product ratio. This was indeed observed for many of these ratios [8–10]. *FADS1* is one example, as shown in Fig. 15.1; *FADS1* codes for the enzyme fatty acid delta-5 desaturase and is considered a key enzyme in the

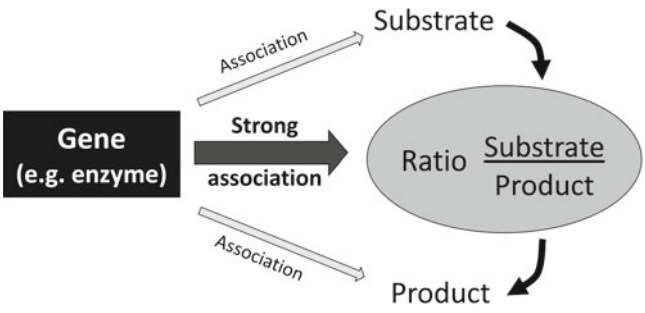


Fig. 15.3 Advantage of using a substrate/product ratio compared the substrate and product alone to identify the gene coding for e.g. the enzyme which is responsible for the particular metabolic step

metabolism of long-chain polyunsaturated omega-3 and omega-6 fatty acids. A certain frequently occurring variant in this gene has a strong effect on the activity of this enzyme. If this variant is analyzed with single species of phosphatidylcholines one can observe that this variant explains at most only a few percent of these phosphatidylcholines. If the various ratios of these phosphatidylcholines are analyzed one can observe a pronounced strengthening of the association as soon as the enzyme is analyzed with the substrate/enzyme ratio, which is the most obvious one the enzyme is catalyzing. This step dramatically increased the variance explained by this variant to more than 28% and improved the p-value several thousand-fold and provides thereby a close insights in the metabolic pathway [9, 10]. Figure 15.1 schematically illustrates the pronounced increase in information by using the most closely related intermediate phenotype. For *FADS1* it was quite impressive that the association with the single metabolite was already observed in only 284 study participants of a GWAS with a p-value of 4.5×10^{-8} which dramatically improved to 2.4×10^{-22} when the ratio of two metabolites was used for analysis. In contrast, it required almost 9,000 individuals in a GWAS on lipids to find a p-value of 1.5×10^{-4} of the *FADS* gene complex with total cholesterol which is still far away from genome-wide significance [11]. Moreover, it required more than 20,000 study participants to get this gene genome-wide significant for total cholesterol concentrations [12, 13]. This can be explained by the fact that despite total cholesterol as an intermediate phenotype and in the same pathway as phosphatidylcholines and their ratios, total cholesterol is far away from the action of *FADS1*. This becomes even more pronounced when it comes to coronary artery disease (CAD) which showed in the WTCCC Study with more than 2,000 CAD cases and 3,000 controls with only a p-value of 0.021 [14].

4.2 Bilirubin an Intermediate Phenotype for the Development of CVD

The antioxidative and cytoprotective properties of bilirubin made this product of the heme metabolism an interesting candidate for investigation of CVD outcomes. Many studies showed an association between low bilirubin levels and CVD [15]. Early segregation analyses suggested a major gene controlling bilirubin levels [16]. Linkage analysis in two independent studies identified the gene *UGT1A1* on chromosome 2q37 as the most probable gene with a strong influence on bilirubin levels [17–19]. It encodes UDP-glucuronosyltransferase, the major enzyme of bilirubin glucuronidation, which mainly determines bilirubin elimination in humans. The activity of *UGT1A1* is significantly influenced by a TA-repeat polymorphism in the promoter region of this gene. Individuals homozygous for 7 TA repeats (7/7) were found to have a lower promoter activity and subsequently higher levels of bilirubin than heterozygous (6/7) or wild type homozygous (6/6) [20]. On the population level this polymorphism explains between 10% and 30% of the bilirubin levels

[21–23] which is exceptionally high compared to other associations between genetic variants and quantitative traits.

Data from the Framingham Heart Study demonstrated that carriers of 6 TA repeats, with their markedly lower bilirubin concentrations, developed significantly more often a cardiovascular event over 24 years of observation than homozygote carriers of 7 TA repeats [21]. However, this was not observed in each study up to now as recently reviewed [15]. It demonstrates also a limitation of the intermediate phenotype approach: as long as no causal relation between the intermediate phenotype and the clinical endpoint of interest is demonstrated, the intermediate phenotype does not receive full attention. It will therefore need much larger studies with a long observation period to prove whether bilirubin is indeed associated with endpoints such as CVD. Therefore many question marks are included in this research field.

5 Mendelian Randomization to Prove Causality

The intermediate phenotype concept is an important corner stone for the idea of Mendelian randomization [24, 25]. During recent years Mendelian randomization projects became quite popular to prove or exclude causality between certain intermediate phenotypes (biomarkers), cardiovascular and other outcomes. A prerequisite for this concept is that the various alleles of a certain genetic polymorphism have an influence on the intermediate phenotype and that the intermediate phenotype shows an association with the outcome of interest (e.g. CVD). Most importantly, it is randomly determined at the time of conception which of the two alleles from the father as well as from the mother will be transmitted to the child. Since the transmitted alleles are of lifelong persistence, these alleles determine to a certain amount also whether a person is exposed to an atherogenic level of the intermediate phenotype and therefore to the CVD risk associated with these levels (Fig. 15.4). Therefore, the association between the polymorphism and CVD is less likely to be influenced by reverse causation or confounding and alleles associated with atherogenic level of the intermediate phenotype should be observed with a higher frequency in patients with CVD in case the intermediate phenotype is indeed causally related to CVD. This concept was used for the first time and quite successfully applied to strongly support causality between lipoprotein(a) [Lp(a)] concentrations and CVD outcomes which is illustrated in Fig. 15.4 [26–28]. Lp(a) is an LDL-like lipoprotein and high concentrations have been shown to be strongly associated with CVD events [29, 30]. Concentrations in the upper tertile of Lp(a) are associated with a doubling of the risk for myocardial infarction [30]. The most obvious difference to LDL is that Lp(a) contains an additional apoprotein called apolipoprotein(a) [apo(a)] that is covalently bound to an LDL particle. Apo(a) shows a high homology with plasminogen and competes with this protein for binding to plasminogen receptors, fibrinogen, and fibrin (for review see reference [31]). A common copy number variation in apo(a) explains a substantial amount of about 50% of the Lp(a) concentrations [32]. Therefore it was much easier to detect an association of this polymorphism with

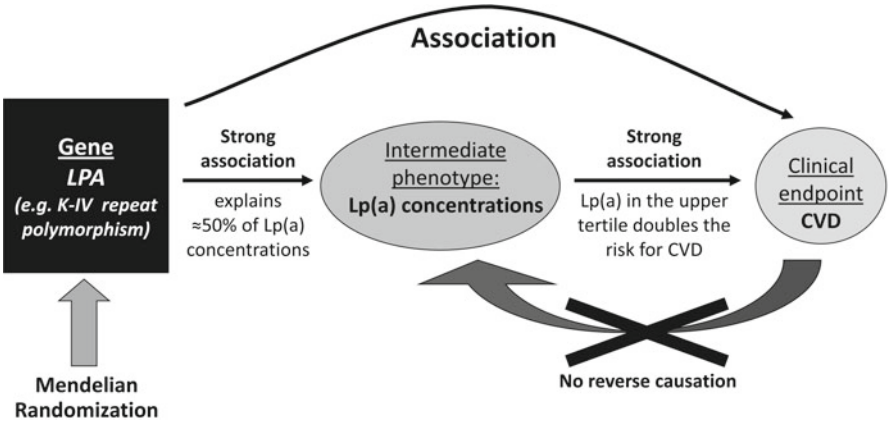


Fig. 15.4 Concept of Mendelian Randomization. For explanation, see text

CVD outcomes [26–28, 33]. A recent meta-analysis including the results from 36 studies yielded a doubling in the relative risk for CVD outcomes for individuals with smaller versus larger apo(a) isoforms which corresponds to approximately 22 or fewer kringle IV repeats vs. more than 22 repeats [28]. This is quite a strong effect compared to other associations of genes with CVD events [34].

The lower the variance of the intermediate phenotype explained by a certain polymorphism is, the higher the required sample size is to show an association between this polymorphism and the CVD outcome. If the intermediate phenotype is not a strong determinant of the clinical endpoint of interest, the required sample size increases further to prove causality. It therefore requires often very large sample sizes to prove causality.

6 Concluding Remarks

The intermediate phenotype concept is indeed an intriguing concept to identify genes associated with clinical outcomes of interest. Intermediate phenotypes are parameters which are considered to be involved in the development of an endpoint of interest. However, identifying the genes associated with the intermediate phenotype is only a first step which can provide an improved insight into metabolic pathways. A further crucial step, however, is to demonstrate an association between the intermediate phenotype and the clinical endpoint. If such an association is present, it does not necessarily mean that the association is of a causal nature since it is also possible that the intermediate phenoytpe changes as a consequence of the clinical endpoint (reverse causation). The causality of this association is strongly supported if also an association between these genetic variants and the clinical endpoint can be

demonstrated. The concept for proving causality between the intermediate phenotype and the clinical endpoint is called Mendelian Randomization and has to show three associations: (1) association between the intermediate phenotype and the clinical endpoint of interest; (2) association between the genetic variant and the intermediate phenotype and (3) and most importantly, an association between the genetic variant and the clinical endpoint. Usually, very well defined study populations with a very large sample size are required to ensure reliability of these results.

References

1. Kronenberg F, Heid IM (2007) Genetik intermediärer Phänotypen. Medizinische Genetik 19:304–308
2. Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431–440
3. Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51:1173–1182
4. Gottesman II, Gould TD (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry 160:636–645
5. Kiechl S, Willeit J, The Bruneck Study Group (1999) The natural course of atherosclerosis. Part I: incidence and progression. Arterioscler Thromb Vasc Biol 19:1484–1490
6. Kiechl S, Willeit J, The Collaborative Study Group (1999) The natural course of atherosclerosis. Part II: vascular remodeling. Arterioscler Thromb Vasc Biol 19:1491–1498
7. Lamina C, Meisinger C, Heid IM et al (2006) Association of ankle-brachial index and plaques in the carotid and femoral arteries with cardiovascular events and total mortality in a population-based study with 13-years of follow-up. Eur Heart J 27:2580–2587
8. Suhre K, Wallaschofski H, Raffler J et al (2011) A genome-wide association study of metabolic traits in human urine. Nat Genet 43:565–569
9. Illig T, Gieger C, Zhai G et al (2010) A genomewide perspective of genetic variation in human metabolism. Nat Genet 42:137–141
10. Gieger C, Geistlinger L, Altmaier E et al (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4:1000282
11. Kathiresan S, Melander O, Guiducci C 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
12. Aulchenko YS, Ripatti S, Lindquist I et al (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41:47–55
13. Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713
14. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
15. Kronenberg F (2010) Association of bilirubin with cardiovascular outcomes: more hype than substance? Circ Cardiovasc Genet 3:308–310
16. Hunt SC, Wu LL, Hopkins PN, Williams RR (1996) Evidence for a major gene elevating serum bilirubin concentration in Utah pedigrees. Arterioscler Thromb Vasc Biol 16:912–917
17. Kronenberg F, Coon H, Gutin A et al (2002) A genome scan for loci influencing anti-atherogenic serum bilirubin levels. Eur J Hum Genet 10:539–546
18. Lin JP, Cupples LA, Wilson PW, Heard-Costa N, O'Donnell CJ (2003) Evidence for a gene influencing serum bilirubin on chromosome 2q telomere: a genomewide scan in the Framingham study. Am J Hum Genet 72:1029–1034

19. Lin J-P, Schwaiger JP, Cupples LA et al (2009) Conditional linkage and genome-wide association studies identify UGT1A1 as major gene for anti-atherogenic serum bilirubin levels – a Framingham Heart Study. Atherosclerosis 206:228–233
20. Bosma PJ, Chowdhury JR, Bakker C et al (1995) The genetic basis of the reduced expression of bilirubin UDP- glucuronosyltransferase 1 in Gilbert's syndrome. N Engl J Med 333: 1171–1175
21. Lin J-P, O'Donnell CJ, Schwaiger JP et al (2006) Association between the UGT1A1*28 allele, bilirubin levels, and coronary heart disease in the Framingham Heart Study. Circulation 114:1476–1481
22. Lingenhel A, Kollerits B, Schwaiger JP et al (2008) Serum bilirubin levels, UGT1A1 polymorphisms and risk for coronary artery disease. Exp Gerontol 43:1102–1107
23. Rantner B, Kollerits B, Anderwald-Stadler M et al (2008) Association between the *UGT1A1* TA-repeat polymorphism and bilirubin concentration in patients with intermittent claudication: results from the CAVASIC Study. Clin Chem 54:851–857
24. Katan MB (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1:507–508
25. Davey SG, Ebrahim S (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32:1–22
26. Sandholzer C, Saha N, Kark JD et al (1992) Apo(a) isoforms predict risk for coronary heart disease: a study in six populations. Arterioscler Thromb 12:1214–1226
27. Kronenberg F, Kronenberg MF, Kiechl S et al (1999) Role of lipoprotein(a) and apolipoprotein(a) phenotype in atherogenesis: prospective results from the Bruneck Study. Circulation 100:1154–1160
28. Erqou S, Thompson A, Di AE et al (2010) Apolipoprotein(a) isoforms and the risk of vascular disease: systematic review of 40 studies involving 58,000 participants. J Am Coll Cardiol 55:2160–2167
29. Erqou S, Kaptoge S, Perry PL et al (2009) Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA 302:412–423
30. Kamstrup PR, Tybjaerg-Hansen A, Steffensen R, Nordestgaard BG (2009) Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. JAMA 301:2331–2339
31. Kronenberg F (2004) Epidemiology, pathophysiology and therapeutic implications of lipoprotein(a) in kidney disease. Expert Rev Cardiovasc Ther 2:729–743
32. Kraft HG, Köchl S, Menzel HJ, Sandholzer C, Utermann G (1992) The apolipoprotein(a) gene: a transcribed hypervariable locus controlling plasma lipoprotein(a) concentration. Hum Genet 90:220–230
33. Kraft HG, Lingenhel A, Köchl S et al (1996) Apolipoprotein(a) Kringle IV repeat number predicts risk for coronary heart disease. Arterioscler Thromb Vasc Biol 16:713–719
34. Schunkert H, Konig IR, Kathiresan S et al (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43:333–338

Chapter 16
Genome-Wide Association Studies with Metabolomics

Karsten Suhre

1 Introduction

The understanding of mechanisms controlling human health and disease, in particular the role of genetic predispositions and their interaction with environmental factors, is a prerequisite for the development of safe and efficient therapies for complex disorders, such as Type 2 Diabetes and cardiovascular disease. Over 100 years ago, Archibald Garrod introduced the concept of "inborn errors of metabolism", which are Mendelian disorders where the loss of function of an individual enzyme or transporter protein typically results in strongly perturbed levels of its reaction substrates and/or products. A textbook example of an "inborn error of metabolism" is phenylketonuria (PKU). PKU is an inherited recessive deficiency of the enzyme phenylalanine hydroxylase and causes an excessive accumulation of phenylalanine in the body. If untreated, PKU results in abnormal and irreversible brain development.

A large number of "inborn errors of metabolism" have been described to date and are today routinely detected by newborn screening tests. However, Garrod also already realized that inborn errors in human metabolism were *"merely extreme examples of variations of chemical behavior which are probably everywhere present in minor degrees"* and that this *"chemical individuality [confers] predisposition to and immunities from the various mishaps which are spoken of as diseases"* [1]. Today, genome-wide association studies with broad panels of metabolite concentrations are identifying common genetic variants in genes coding for enzymes and transporter proteins that induce major differentiations in the metabolic make-up of the human population. In combination with the increasing knowledge about disease

K. Suhre (✉)
Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City – Qatar Foundation, 24144, Doha, State of Qatar
e-mail: karsten@suhre.fr

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_16,

Box 16.1 Definition of "the Genetically Determined Metabotype" (GDM)

The term "Metabotype" is short for "metabolic phenotype". A metabolic phenotype can be a metabolic parameter that is typically obtained from a blood or urine sample using some biochemical measurement methods, such as mass spectrometry and nuclear magnetic resonance spectroscopy. The metabotype of an individual is the result of a complex interplay of environmental factors, life style and genetic predisposition. The ensemble of all metabotypes defines the metabolic individuality of that person; historically also referred to as "chemical individuality" by A. Garrod. As genome-wide association studies with metabolomics are starting to identify the genetic part of human metabolic individuality, the term "genetically determined metabotype" (GDM) was introduced to refer to metabotypes with a genetic contribution. Typically, a GDM would correspond to a genetic association with one or more metabolite concentrations in blood or urine and exhibit a large effect size. Changes in metabolite levels per copy of the minor allele of up to 60% have been observed. GDMs can thus also be viewed as moderate forms of inborn errors of metabolism, which in contrast are rare genetic disorders in which the function of an enzyme or transporter gene is totally lost. GDMs are frequent genetic variants (SNPs), often exhibiting minor allele frequencies of 20% and more. In many cases the leading SNPs of these associations are located in or near enzyme or transporter coding genes that match the metabotype of the metabolic phenotype, such as SNP rs6558295 in the 5-oxoprolinase gene (*OPLAH*), which associates with 5-oxoproline concentrations in serum (see Fig. 16.1 and Table 16.1 for more examples).

associated genetic loci, these so-called *"genetically determined metabotypes"* (see Box 16.1 and Fig. 16.1) are now used to uncover new complex risk factors of common diseases and reveal new insights into the pathophysiology of related disorders, thereby confirming Garrod's 100 year-old prediction.

2 Genome-Wide Association Studies with Metabolomics in Human Blood

When the human genome was published in 2003, expectations were high that this achievement would provide a comprehensive understanding of human biology and disease. With the introduction of microarray based genotyping arrays by Affymetrix and Ilumina, genome-wide association studies (GWAS) with complex diseases became possible. They initiated a new era of genetic research, allowing for the first time the ability to screen a large portion of the natural variation human genome for

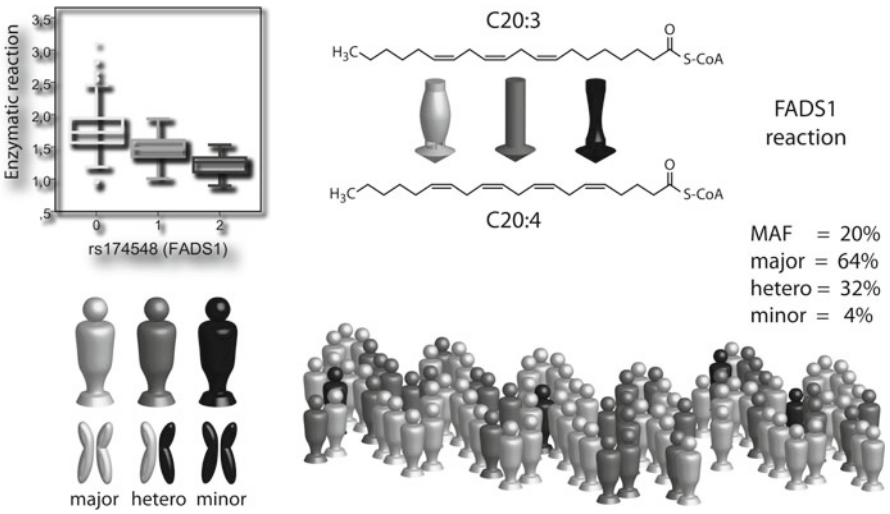


Fig. 16.1 Example of a "genetically determined metabotype." The product of the Fatty Acid Desaturase 1 (*FADS1*) gene catalyses the delta-5 desaturation of polyunsatured fatty acids, such as the omega-6 fatty acid C20:3, to form (in this case) arachidonic acid C20:4. The rate of the enzymatic reaction that is catalyzed by FADS1, approximated by the ratio between C20:3 and C20:4, displays a strong genotype-dependence (*box-plot*). The high minor allele frequency of this genetic variant results in highly variant metabolic capabilities in the human population (See Box 16.1)

association with major diseases, such as bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 and type 2 diabetes mellitus [2]. By 2011, over 1,300 GWAS for over 220 traits have been published and catalogued in the NHGRI GWAS catalog http://www.genome.gov/gwastudies/ [3]. However, from these studies it became clear that often the effects of the discovered genetics variants are modest and that they account for only a small part of the heritable part of a complex trait. For instance, one of the largest GWAS to date, enrolling tens of thousands of participants, has identified about 50 variants that are associated with height. Twin studies have estimated body height heritability to within 68–84% in women and 87–93% in men [4], but taken together the discovered genetic variants explain only 5% of the phenotypic variance. This raises the question of what is referred to as the "missing heritability" [5]. One possible explanation is that complex traits are determined by a larger number of SNPs, where each individual SNP contributes only a small portion to the overall effect. These SNPs shall be hard to detect using classical GWAS with clinical endpoints due to limitations in statistical power [6].

One way to overcome this problem is to investigate associations with intermediate traits, such as impaired levels of blood cholesterol, triglyceride, glucose, bilirubin, and vitamin B12, which are known risk factors of disease. It turned out that associations with such intermediate phenotypes often exhibit much larger effect

sizes, which may be attributed to the fact that these traits are determined by a smaller number of genetic variants and also that they are functionally more directly related to the genetic variants (see also Chap. 15 in this book). Moreover, associations with quantitative traits rather than with binary clinical endpoints tend to be more robust and readily detectable. Recent advances in NMR and mass spectrometry permit to quantitatively map out large parts of the metabolome in human blood and urine. Increased instrument sensitivity and sensibility, together with reduced measurement times, allow for high-throughput metabolomics experiments to be conducted in large numbers of blood and urine samples (see also Chap. 3). These technological advances open the possibility to conduct GWAS with comprehensive metabolomics panels, and thereby to identify the heritable part of human metabolic individuality. In the following we shall present an overview of the GWAS with metabolomics that were conducted to date. Table 16.1 presents a selection of genetically determined metabotypes that were identified by these studies.

Gieger et al. [7] conducted the first GWAS with metabolomics using quantitative measures of 363 metabolites in serum of 284 male participants of the KORA study. They identified associations of frequent single nucleotide polymorphisms (SNPs) with large differences in homeostasis metabolic in or near genes coding for the enzymes FADS1, LIPC, ACADS, and ACADM. Using ratios of metabolite concentrations as a proxy for enzymatic activity, they could explain up to 28% of the overall observed variance at p-values ranging between 10^{-16} and 10^{-21} for these SNPs. What is more, the corresponding metabotype clearly matches the biochemical pathways in which these enzymes are active. This first GWAS with metabolic traits still had relatively low power and did not include any replication. Actually none of the associations with metabolite concentrations attained genome-wide significance per se. However, the hypothesis-free testing of all possible pairs of metabolite concentration ratios lead to associations that were highly significant, even after correction for the large number of additional tests induced by using metabolite ratios. The results from this first study already indicated what was later confirmed by more highly powered studies, that is, common genetic polymorphisms induce major differentiations in the metabolic make-up of the human population.

Three GWAS with lipidomics-centered panels were published next (see Chap. 13 for details on lipidomics). Tanaka et al. [8] conducted a genome-wide association study of plasma levels of six omega-3 and omega-6 fatty acids in 1,075 participants in the InCHIANTI study on aging. They confirmed the association of *FADS1* with arachidonic acid (AA). Minor allele homozygotes had lower AA compared to the major allele homozygotes and rs174537 accounted for 18.6% of the additive variance in AA concentrations. They also identified a new association of eicosapentanoic acid (C20:5) with a genetic variant in *ELOVL2* (elongase of very long fatty acids 2). Hicks and co-workers [9] performed a GWAS with 33 sphingolipid species, based on 4,400 participants from five diverse European populations. They identified associations at five genomic regions in or near genes functionally involved in ceramide biosynthesis and trafficking: *SPTLC3*, *LASS4*, *SGPP1*, *ATP10D*, and *FADS1–3*. They conclude that concentrations of several key components in sphingolipid metabolism are under strong genetic control, and that variants in these loci

Table 16.1 Selected examples of genetically determined metabotypes from recent GWAS with metabolomics. Details on the associations can be found in the cited references. Note the correspondences between the metabolic roles of the associating genes and the associating metabotypes

Genetic locus	Metabolic role of the associating gene	Associating metabotype	Biomedical phenotype	GWAS reference
FADS1	Desaturation of polyunsaturated omega-3 and omega-6 fatty acid metabolism (produces arachidonic acid C20:4)	C20:4 to C20:3 ratio	Association with LDL cholesterol, HDL cholesterol and triglycerides, fasting glucose and homeostatic model assessment B (HOMA-B), Crohn's disease and resting heart rate	[7]
ELOVL2	Elongation of polyunsaturated omega-3 and omega-6 fatty acids	Docosahexaenoate (DHA; 22:6n3) to eicosapentaenoate (EPA; 20:5n3) ratio		[8]
SCD	Fatty acid delta-9 desaturase	Fatty acid C16:1 to C16:0 ratio		[10]
LIPC	Hepatic lipase	Phosphatidyl ethanolamines	Association with blood levels of HDL cholesterol and triglycerides	[7]
ACADS, ACADM, ACADL	Beta oxidation of short, medium, long chain length fatty acids, resp.	Short, medium, long chain length acylcarnitines, resp.	Loss of function of these genes leads to inborn errors of metabolism	[7, 10]
AGXT2	Metabolism of beta-aminoisobutyrate	Beta-aminoisobutyrate	Hyper-beta-aminoisobutyric aciduria; this is an inborn error of metabolism, the genetic basis of which was detected by a GWAS	[13]
NAT2	N-acetylase, metabolism of xenobiotics	1-methylxanthine to 4-acetamidobutanoate ratio in blood; formate:succinate, formate:urea and formate:acetate ratios in urine	Association with triglyceride levels and CAD; bladder cancer and toxicities to docetaxel and thalidomide treatment	[12, 13]
NAT8	N-acetylase, metabolism of xenobiotics	N-acetylornithine in blood, unidentified N-acetylated species in urine	Risk locus for chronic kidney disease	[11, 12]
GCKR	Glucokinase (hexokinase 4) regulator	Glucose:mannose ratio	Pleiotropic diabetes risk locus	[12]

(continued)

Table 16.1 (continued)

Genetic locus	Metabolic role of the associating gene	Associating metabotype	Biomedical phenotype	GWAS reference
ACE	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	The dipeptide aspartylphenylalanine is a product of ACE	ACE is a target of anti-hypertension drugs; association with angiotensin converting enzyme activity	[12]
ENPEP	Amino-terminal amino peptidase	Amino-terminal cleaved fibrinogen A-alpha peptides	Association with blood pressure	[12]
ABO	Determines blood group	Fibrinogen A-alpha phosphorylation	Association with venous thromboembolism	[12]
SCL6A20, SCL7A9, SLC16A9, SLC22A4	Solute carrier proteins	Associating metabotype matches the transporters' known or predicted substrate specificity	Diverse associations (see references)	[10, 13]
SLCO1B1	SLCO1B1 is an organic anion transporter	Eicosenoate:tetradecanedioate ratio	Pharmacogenomics locus; associated with an increased risk of statin-induced myopathy	[12]

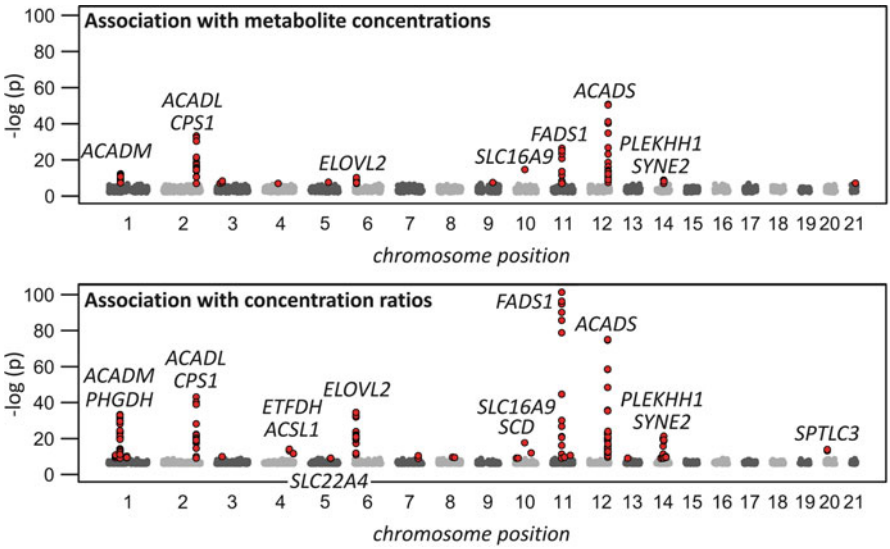


Fig. 16.2 Genome-wide association plot (Manhattan plot). This plot shows the locations of the genetic polymorphisms and their strength of association (expressed as $-\log_{10}$ (p-value)). The *upper plot* is for associations with metabolite concentrations, the *lower plot* is with concentration ratios. The strengthening of the association signal when using ratios can be discerned (Figure reproduced from [10])

can be tested for a role in the development of common cardiovascular, metabolic, neurological, and psychiatric diseases. Illig et al. [10] presented a genome-wide association study with a kit-based panel of 163 metabolic traits, including amino acids, acyl-carnitines and many phospholipid species (BIOCRATES Life Sciences, Innsbruck, Austria). Using samples from over 1,800 participants of the KORA population and samples from 420 participants of the TwinsUK cohort, they identified 14 associations at genome-wide significance in the KORA study (Fig. 16.2). Eight out of nine associations that were replicated in the TwinsUK cohort are linked to one of the enzyme or solute carrier coding genes *FADS1*, *ELOVL2*, *ACADS*, *ACADM*, *ACADL*, *SPTLC3*, *ETFDH*, and *SLC16A9*. Five loci that attained genome-wide significance in the discovery study, but were not fully replicated in the TwinsUK cohort (*SCD*, *SLC22A4*, *PHGDH*, *CPS1*, and *SYNE2*) were later replicated by Nicholson et al. [11] Note here that the *SYNE2* locus of Illig et al. [10] is most likely identical to the *SGPP1* locus reported by Hicks et al. [9] The metabolic traits of most of these associations also match the related genes' function (Fig. 16.3). Many of the implicated proteins actually control rate limiting steps of important enzymatic reactions. Due to the higher statistical power and the use of metabolite concentration ratios, Illig et al. [10] obtained p-values of association as low as 6.5×10^{-179} and explained observed variances up to 36.3%.

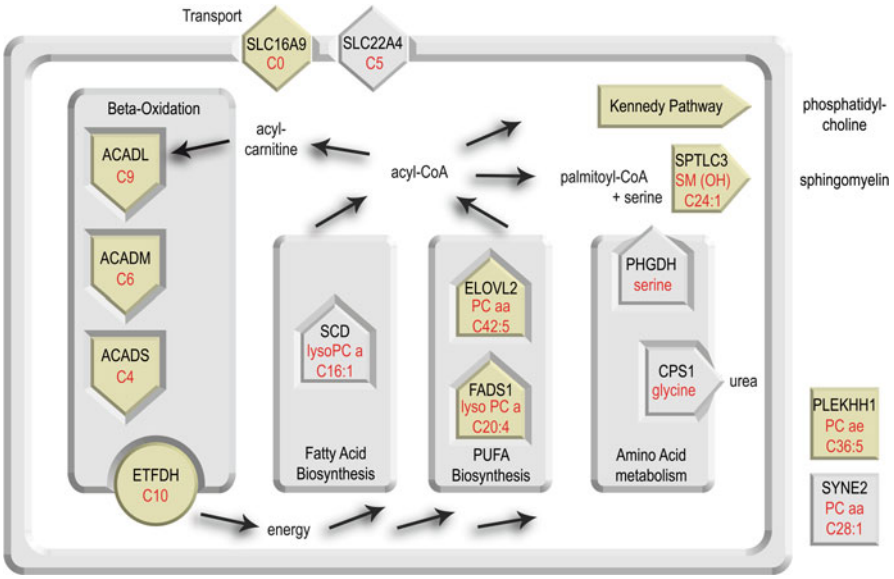


Fig. 16.3 A genome wide perspective of genetic variation in human metabolism. Twelve out of fourteen genetic polymorphisms are located in or near genes encoding enzymes or transporter genes that are central to the different processes in human lipid metabolism: β-oxidation (*ACADS*, *ACADM* and *ACADL*), polyunsaturated fatty acid biosynthesis (*FADS1* and *ELOVL2*), fatty acid synthesis (*SCD*), breakdown of fats and proteins to energy (*ETFDH*), biosynthesis of phospholipids (*SPTLC3*), metabolite carrier proteins (*SLC22A4* and *SLC16A9*), amino acid metabolism (*PHGDH* and *CPS1*). The *SYNE2* locus is most likely identical to the *SGPP1* locus reported by [9]. Only for the genetic variant in *PLEKHH1* does the attribution of a metabolic function remain elusive. For each locus, the most strongly associating single metabolite is indicated in *red* (Figure reproduced from [10])

The studies described so far all deployed targeted metabolomics assays, thereby limiting the scope of these studies to a limited set of metabolic pathways. Suhre et al. [12] then conducted the first comprehensive analysis of genotype-dependent metabolic phenotypes with an extensive, non-targeted and metabolome-wide panel of small molecules, analyzing over 250 metabolites from 60 biochemical pathways (Metabolon Inc., Durham, USA) in serum samples from 2,820 individuals from two large population-based European cohorts. They identified 37 genetic loci associated with blood metabolite concentrations (Fig. 16.4), of which 25 show effect sizes that are unusually high for GWAS and account for 10–60% differences in metabolite levels per allele copy in 25 loci. In the majority of cases, a protein that is biochemically related to the associated metabolic traits is encoded at these loci. 23 of these loci describe new genetic associations with metabolic traits, and 14 replicate and extend the present knowledge about known GDMs. This study also replicates a series of findings from previous GWAS with quantitative traits, including serum levels of fasting glucose, bilirubin, urate and dehydroisoandrosterone sulphate. By reporting only associations that are supported by two independent studies at

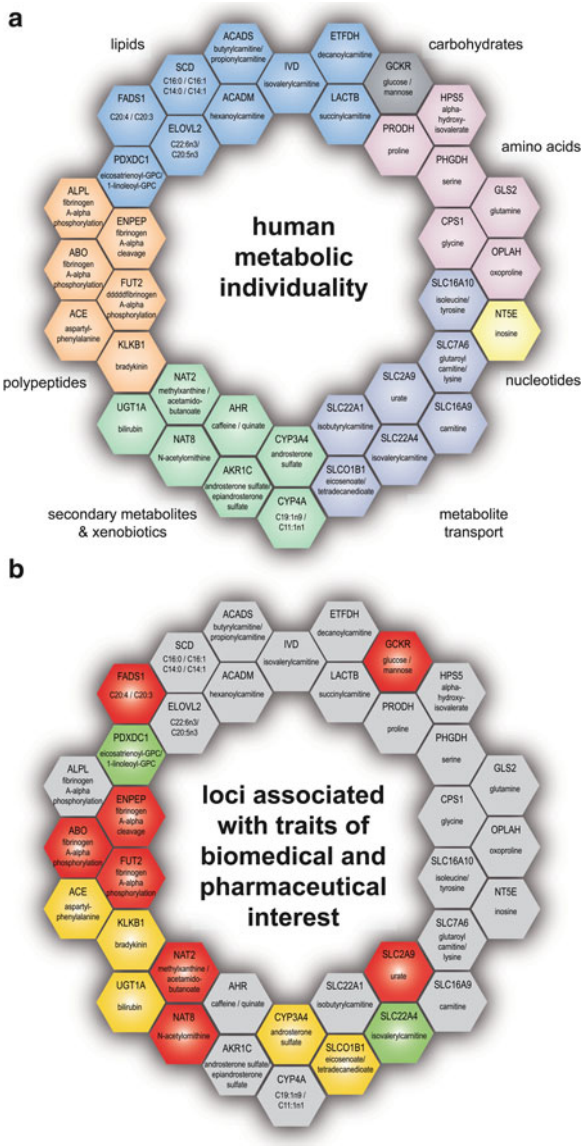


Fig. 16.4 Genetic basis of human metabolic individuality and its overlap with loci of biomedical and pharmaceutical interest. 37 genetically determined metabotypes (GDMs) that explain a highly relevant amount of the total variation in the studied population. These genetic loci contribute substantially to the genetic part of human metabolic individuality. (**a**) GDMs are color by general metabolic pathways. (**b**) GDMs colored by their overlap with associations in previous GWAS with disease (*red*), intermediate risk factors for disease (*yellow*) and other traits (*green*) (Figure adapted from [12])

genome-wide significance, the authors have taken a very conservative approach. They estimate that more than 500 loci with signals of association below that conservative threshold may be confirmed as GDMs in more highly powered studies in the future. Association data for loci that did not reach genome-wide significance level is available via a web-server at http://www.gwas.eu.

3 Genome-Wide Association Studies with Metabolomics in Human Urine

While GWAS with metabolic traits in blood are likely to detect genetic variance in homeostatic processes in human metabolism, analysis of urine may allow for the investigation of genetic variants associated with the detoxification capacity of the human body. Suhre et al. [13] report the first genome-wide association study of metabolic traits in human urine. Using NMR spectroscopy followed by manual peak annotation and quantification (CHENOMX Inc., Edmonton, Canada), they tested 59 metabolites in urine from 862 male participants of the population-based SHIP study for association. For replication and verification of robustness 1,039 additional samples of the same study, including a 5-year follow-up, and 992 samples from the independent KORA study were used. They identify five loci with joint p-values of association ranging from 3.2×10^{-19} to 2.1×10^{-182}, three of which are known to associate with important clinical outcomes: *SLC7A9* is a risk locus for chronic kidney disease, *NAT2* for coronary artery disease and genotype-dependant response to drug toxicity, and *SLC6A20* a contributing factor of iminoglycinuria. Moreover, they identify a coding SNP in *AGXT2* as the potential genetic basis of hyper-beta-aminoisobutyric aciduria. Nicholson et al. [11] also performed a GWAS with ^{1}H NMR in human urine, analyzing samples from 142 female twins' samples of the MolTWIN cohort and 69 participants of the MolOBB cohort. In contrast to the manual automation procedure used by Suhre et al., here the NMR data sets were passed through a semiautomated preprocessing pipeline: phasing, alignment, denoising, baseline correction, manual bin selection, normalization, quality control, peak extraction, and logarithmic transformation. They confirm the *AGXT2* association and report associations at two other loci, *NAT8* and *PYROXD2*, which the authors deem as good candidates for mediating the corresponding associations.

As new GWAS with metabolomics are published, they not only replicate previously reported loci, but also often provide additional and new information on the functional background of the underlying genetic association. For instance, using ^{1}H NMR in urine, Nicholson et al. [11] found that the NAT8 locus associates with a compound termed N-ACu. Although they were unable to attribute N-ACu to a single metabolite, it is clear that N-ACu corresponds to one or more N-acetylated compounds (X.NH.CO.CH3, with X unknown). Using LC-MS/MS, Suhre et al. [12] showed that NAT8 associates with N-acetylornithine in blood serum. It is noteworthy that both studies differed in experimental methods (^{1}H NMR versus LC-MS/MS), as well as in analyzed biomaterials (urine versus blood serum) and study populations.

The NAT8 locus also associates with glomerular filtration rate [14] and chronic kidney disease [15]. Although causality cannot be inferred from these kinds of association studies, the role of ornithine acetylation in the aetiology of CKD warrants further exploration. Moreover, the identity of the N-acetylated compound N-ACu in urine and its relationship to N-acetylornithine in blood should be investigated as it may be developed into a potential CKD biomarker. This case exemplifies how results of different studies may be combined in order to generate new hypotheses of biomedical interest.

4 Genetically Determined Metabotypes and Disease

As outlined above, SNPs identified by GWAS with metabolic traits allow identifying new associations in GWAS with clinically relevant parameters. As an example, in their initial study Gieger et al. [7] suggested *FADS1* to be a risk locus for perturbed blood lipid parameters. This hypothesis was supported by the observed associations with different phospholipids, key components of serum lipids, and the fact that two published GWAS investigating lipid levels reported associations for the *FADS1* locus with low-density lipoprotein (LDL), high-density lipoprotein (HDL) and total cholesterol. However, these associations had not been included in the list of potential candidates for replication in those studies, as their signal of association was not strong given the need to correct for multiple testing in classical genome-wide association studies, where several million SNPs are tested in parallel. More recent GWAS on lipid parameters with significantly increased sample sizes have now confirmed the prediction of FADS1 being a lipid risk locus, thereby proving that a combination of a GWAS using metabolomic profiles with data from large GWAS with clinical parameters can identify new candidate SNPs associated with known phenotypes of relevance to human health.

Associations with metabolic traits can also contribute to the in-depth characterization of already identified disease-related loci and reveal new functional information about previously reported associations to related traits. For instance, a polymorphism in the apolipoprotein cluster *APOA1-APOC3-APOA4-APOA5* was already known to strongly associate with blood triglyceride levels. Illig et al. [10] found that the same SNP associates with ratios between different phosphatidylcholines. These lipid compounds are biochemically connected to triglycerides by the intermediary of only a few enzymatic reaction steps. The specific identities and properties of these lipid species may now be used in order to better understand the pathways that are impacted by this genetic variant and their role in lipid-related disorders. A second example from that study is a SNP in the glucokinase regulator *GCKR* gene. Genetic variance in *GCKR* modulates fasting glucose and triglyceride levels and has an impact on type 2 diabetes risk. This locus associated in the Illig et al. [10] study with different ratios between plasmalogens and phosphatidylcholines concentrations, suggesting new avenues for investigations into the functional background of the GCKR association with diabetes. Thirdly, a genetic variant of the

gene encoding melatonin receptor (*MTNR1B*) was found to associate with fasting glucose and type 2 diabetes risks. The same SNP associated in the Illig et al. [10] study with tryptophan to phenylalanine ratios. As phenylalanine is a precursor of melatonin, this association indicates a functional relationship between the melatonin pathway and the regulation of glucose homeostasis.

Interestingly, GWAS with metabolomics even allow for the identification of the genetic basis of inborn error of metabolism. Given the extreme effect size of the *AGXT2* association with 3-aminoisobutyrate (BAIB) described by Suhre et al. [13] and Nicholson et al. [11] this variant represents in all likelihood the genetic basis of hyper-beta-aminoisobutyric aciduria. *AGXT2* is a mitochondrial aminotransferase expressed primarily in the kidney, and BAIB is a substrate of AGXT2. Recently, it has been shown that AGXT2 also metabolizes asymmetric dimethylarginine (ADMA), and that this pathway could represent an alternative route of ADMA regulation. Elevated plasma concentrations of ADMA are found in association with hypertension, congestive heart failure, progression of chronic kidney disease and atherosclerosis. Hyper-beta-aminoisobutyric aciduria may hence represent a risk (or protective) factor for these diseases.

5 GWAS with Metabolomics in Functional Genomics, Systems Biology and Pharmacogenomics

GWAS uncover statistically significant associations, but causality generally cannot be inferred from such studies. Therefore these studies are mostly hypotheses-generating by nature. Associations of specific metabotypes with genetic variants in only coarsely-characterized enzymes and transporters indicate possible substrates or products of the proteins and create openings for further experimental and functional characterization. For instance, experiments using isotope-labeled derivatives of the associated metabolites as putative target substrates may lead to new insights into the specificity of an enzyme or transporter. As a proof-of-principle, Suhre et al. [12] experimentally validated the predicted function of SLC16A9 as a carnitine transporter using labeled carnitine and transgenic SLC16A9-expressing *Xenopus* oocytes. They show that this hitherto uncharacterized monocarboxylic acid transporter is indeed a carnitine pump, possibly responsible for carnitine efflux from absorptive epithelia into the blood. Another example is SLC2A9. Following its association with blood urate levels in a GWAS [16], SLC2A9 was shown to be a urate transporter, and not a glucose transport, as initially predicted by homology [17]. These examples show how GWAS with metabolic traits may inform functional genomics studies and advance our general understanding of the human genome.

Using a systems biology approach, by integrating DNA-variation and gene-expression data with other complex trait data in segregating mouse populations, Schadt et al. [18] identified three new genes in susceptibility to obesity, including *LACTB*. Chen et al. [19] validated *LACTB* as a previously unknown obesity gene and Yang et al. [20] demonstrated that transgenic *LACTB* mice showed increased fat

and muscle growth. Using liver expression signatures of these animals, the authors predicted that *LACTB* is involved in butanoate metabolism. In addition, the identification of *LACTB* as an HDL cholesterol risk locus [21] suggests a functional link between succinate-related pathways and HDL metabolism. In their GWAS with metabolic traits, Suhre et al. [12] report a genetic variant in *LACTB* in the human population that results in an 8.5% increase per minor allele copy in succinylcarnitine concentrations ($p = 7.2 \times 10^{-27}$). Succinylcarnitine, which is a transport form of the free fatty acid succinate, is located on the butanoate pathway. What is more, a positive association in of succinylcarnitine concentrations with body mass index (BMI) was also observed ($p = 1.0 \times 10^{-12}$ in KORA and $p = 5.3 \times 10^{-5}$ in TwinsUK, with covariates age and gender). Together these findings support Schadt et al's hypothesis that *LACTB* may represent a new potential therapeutic anti-obesity target. This example shows how results from GWAS with metabolomics may be used to confirm and extend hypotheses generated by systems biology studies.

Pharmacogenomics, the field that studies how genetic variants affect the body's response to medication, holds the promise of better, safer, and more efficient drugs. Cross-referencing new loci from GWAS with metabolomics with databases of disease-related and pharmaceutically relevant genetic associations may uncover hitherto unknown links and provide new hypotheses for the functions of these loci. For instance, the family of cytochrome p450 enzymes (CYP) controls the metabolism of a large part of current drugs. Several genetic variants in CYP genes have been associated with slow, intermediate and fast metabolizers of certain drugs. Similarly, the organic anion–transporter SLCO1B1 has been shown to regulate the hepatic uptake of statins. A SNP in SLCO1B1 is strongly associated with an increased risk of statin-induced myopathy [22]. The Pharmacogenomics Knowledge Base [23] provides an extensive and regularly updated list with genes of pharmacogenetic interest. In their GWAS with metabolic traits, Suhre et al. [12] identified three associations at CYP-related loci genes where the associating metabolic traits are closely related to the respective p450 enzymes' substrates, that is, a CYP3A locus with androsterone sulphate, a CYP4A locus with molecules biochemically related to omega-hydroxylated C10 fatty acids, and AHR, which is a transcription factor for CYP1A1, with caffeine. Of further notable interest is the association of SLCO1B1 with a series of fatty acids, including tetradecanedioate and hexadecanedioate. These associations all provide new insights into the metabolic pathways that are impacted by the genetic variants. This information can for instance be used to support the redesign of the respective drug molecules in order to avoid adverse reactions or in the identification of genotype dependant drug side-effects.

6 Conclusion

The GWAS studies with metabolomics that are published so far have already demonstrated the exciting potential of metabolomics to unravel the genetics of human metabolism. In this approach, the concept of the "*genetically determined*

metabotype" as a complex intermediate phenotype is central. The investigation of the genetically determined metabotypes in their biochemical context may help to better understand the pathogenesis of common diseases and gene–environment interactions. Such findings can result in a step towards personalized prevention, health care and nutrition based on a combination of genotyping and metabolic characterization. The SNPs identified in these studies can now be used in clinical studies for association with response to drug treatment, or the development of particular complications during the course of a disease or treatment.

With new GWAS to be published in the future, the number of genetic loci that display parallel associations of clinically relevant parameters with metabolic traits shall increase steadily. Future GWAS that combine multiple 'omics' technologies in a single study, including transcriptomics, proteomics, metabolomics and recent technologies for determining epigenetic modifications and microRNAs on a genome-wide scale, are likely to present the next big step towards a full understanding of the interaction between genetic predispositions and environmental factors in the development of complex chronic diseases, their diagnosis, prevention and safe and efficient therapy. To close this chapter we refer to Motha and Hirschhorn [1], who expect that *"In the future, more comprehensive versions of the profiling technologies could be coupled to perturbations (for example, dietary challenges, drug treatments, aging) or used in combination with isotopic tracers to more directly infer the influence of genetic variation on in vivo reaction biochemistry and homeostasis. Just as Garrod's study of inborn errors of metabolism helped write a generation of textbooks on human biochemistry, so, potentially, could comprehensive studies of inborn variation of metabolism inform the next generation."*

References

1. Mootha VK, Hirschhorn JN (2010) Inborn variation in metabolism. Nat Genet 42:97–98
2. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
3. Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367
4. Silventoinen K, Sammalisto S, Perola M et al (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6:399–408
5. Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21
6. Yang J, Benyamin B, McEvoy BP et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569
7. Gieger C, Geistlinger L, Altmaier E et al (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4:e1000282
8. Tanaka T, Shen J, Abecasis GR et al (2009) Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. PLoS Genet 5:e1000338
9. Hicks AA, Pramstaller PP, Johansson A et al (2009) Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet 5:e1000672
10. Illig T, Gieger C, Zhai G et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42:137–141

11. Nicholson G, Rantalainen M, Li JV et al (2011) A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet 7:e1002270
12. Suhre K, Shin SY, Petersen AK et al (2011) Human metabolic individuality in biomedical and pharmaceutical research. Nature 477:54–60
13. Suhre K, Wallaschofski H, Raffler J et al (2011) A genome-wide association study of metabolic traits in human urine. Nat Genet 43:565–569
14. Kottgen A, Pattaro C, Boger CA et al (2010) New loci associated with kidney function and chronic kidney disease. Nat Genet 42:376–384
15. Chambers JC, Zhang W, Lord GM et al (2010) Genetic loci influencing kidney function and chronic kidney disease. Nat Genet 42:373–375
16. Doring A, Gieger C, Mehta D et al (2008) SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet 40:430–436
17. Caulfield MJ, Munroe PB, O'Neill D et al (2008) SLC2A9 is a high-capacity urate transporter in humans. PLoS Med 5:e197
18. Schadt EE, Lamb J, Yang X et al (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717
19. Chen Y, Zhu J, Lum PY et al (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–435
20. Yang X, Deignan JL, Qi H et al (2009) Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 41:415–423
21. Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713
22. Link E, Parish S, Armitage J et al (2008) SLCO1B1 variants and statin-induced myopathy–a genomewide study. N Engl J Med 359:789–799
23. Klein TE, Chang JT, Cho MK et al (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics research network and knowledge base. Pharmacogenomics J 1:167–170

Chapter 17
Systems Biology Meets Metabolism

Jan Krumsiek, Ferdinand Stückler, Gabi Kastenmüller, and Fabian J. Theis

In the preceding chapters many aspects of metabolite quantification and relation to trait and disease phenotypes have been described, in particular the linkage of intermediate metabolic traits to genetic heterogeneities. Although many analyses start on the genome-wide level, they end up picking out single polymorphisms or other variations and study these in detail. This reductionist approach is very common in molecular biology and has proven hugely successful over the past decades. In recent years however, a second paradigm has become increasingly popular, namely that of integrating multiple such analyses into larger ones commonly called 'models'. This paradigm, nowadays, is known as systems biology and is expected to penetrate many classical molecular analyses.

The aim is to gain a systems-level understanding of the studied processes, and it requires a shift in notion of what to look for in biology [1]. The identification and quantification of genes, proteins and metabolites in an organism is equivalent to generating lists of all parts of the system, which by itself is not sufficient to understand the complexity of the underlying organism. Knowledge about the assembly of these parts is necessary to understand the formation and regulation of the observed objects; this knowledge is often compiled into a biological network. According to Kitano, a resulting systems-level understanding can then be derived from insight into system structures, system dynamics, system control and system design. System structure refers to first assembling known or estimated interactions between the system's parts in order to build an interaction network. Dynamics then implies studying the system's behavior over time. Control is defined by the study of mechanisms that control the system's state and its modulation. Finally, design describes the notion of

J. Krumsiek, Dipl. Bioinf. • F. Stückler, M.Sc., Biochemistry
G. Kastenmüller, Dipl.-Chemiker, Dipl.-Informatikerin
F.J. Theis, Dipl.-Mathematiker, Dipl.-Phys (✉)
Helmholtz Zentrum München, Institute of Bioinformatics and Systems Biology,
Ingolstädter Landstraße 1, Neuherberg 85764, Germany
e-mail: jan.krumsiek@helmholtz-muenchen.de; ferdinand.stueckler@mytum.de;
kastenmueller@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0_17,

building novel or modifying existing biological systems with desired properties, e.g. bacteria, for fuel production from biological waste; nowadays this has spawned the separate discipline of synthetic biology. The central paradigm in systems biology is the integration of the above approaches with experiments in an iterative research cycle. Based on current experimental data and additional literature information, an initial model is compiled; the model is then used to derive novel possibly competing predictions, which can then be tested or invalidated by a second round of experiments. This process of model inference and experimental design can then be iterated to gain more detailed system's knowledge.

In this chapter, we want to briefly review some systems approaches in the field of metabolic modeling. We will see that many ideas are based on the structural and, in parts, the dynamic principle as defined above. We start with the structure approach, namely the compilation of metabolic networks, their reconstruction and representation in the computer. Using a very common computer representation, namely the stoichiometry matrix, we then ask what we can learn about the system even without access to transition rates; for this we review approaches such as metabolic pathway analysis and extreme pathways. From the flux analysis, we proceed to show how to include metabolite concentrations into the model in order to determine system rates. This results in a dynamical system of metabolites, which we can compare across multiple samples, conditions or patients. By describing this heterogeneity as stochastic effect, we can study correlation patterns and even remove indirect correlations, arriving at graphical models describing statistical metabolite associations. We show that these contain pathway signatures of known metabolic interactions. Moreover, we can use this modeling approach to include genetic variations in a concise and systematic fashion.

1 From Genomes to Metabolic Networks

In its most general meaning, metabolic network reconstruction refers to the compilation of a network comprising ideally all metabolites and biochemical reactions (including transport processes) relevant for a biological system (e.g. compartment, cell, organism). In order to reconstruct a metabolic network, we have to ask two major questions: "Which reactions can be accomplished by the system?" and "How are these reactions interconnected to support basic functions such as growth?" The most reliable information on the presence or absence of biochemical reactions derive from laborious experiments that individually test whether the respective biochemical conversions are observable in the system. For extensively studied model systems or organisms such as the human hepatocyte or the bacterium *Escherichia coli*, biochemists have been collecting this type of experimental data for decades. The biochemical data augmented by genetic data and phenotypic observations formed the basis for the manually curated, high-quality metabolic models that are available for these well-studied systems to date [2, 3].

For less studied organisms, typically very few metabolic reactions are experimentally verified. The presence of metabolic reactions is therefore mostly

inferred from genome sequences, which are available for a rapidly growing number of species. During the past 10 years, various protocols and software frameworks have been developed and implemented in order to improve and automate genome-based metabolic reconstruction [4–7]. Though less detailed and less reliable than reconstructions of extensively studied model systems, such draft reconstructions can provide new metabolic insights and facilitate the prediction of phenotypes as well as the interpretation of high-throughput experiments in the context of metabolic pathways [8, 9]. For most sequenced species, reconstructed genome-scale metabolic models are available through metabolic databases such as KEGG [10], BioCyc [11], and The SEED [12]. Most reconstructions therein have been mainly derived from genome sequences by processing them through automated reconstruction pipelines (KAAS [5], Pathway Tools [6], RAST [7]).

Despite major advances in automating the complete process of metabolic reconstruction, the quality of metabolic models still largely depends on manual intervention during each step taken for the metabolic reconstruction from genome sequences. Intensive manual validation using primary literature, consistency checks, and new experiments transforms automatically derived metabolic networks into high-quality networks suitable for metabolic modeling by mathematical means. However, the high manual effort limits the number of high-quality metabolic models available to date to a small fraction of all sequenced species. The BiGG database, which provides systematic representations of such high-quality networks, currently comprises 30 metabolic reconstructions [13].

In this section, we briefly review the major steps required to reconstruct metabolic networks based on genome sequences. More detailed descriptions are provided in [8, 14–19]. Figure 17.1 gives an overview of the reconstruction process and the mathematical representations frequently used for metabolic models. Table 17.1 lists selected databases and tools relevant for metabolic reconstruction.

1.1 Step 1: From Genome Sequences to Gene Functions

Starting from the assembled genome sequence, *genome annotation* is the first mark on the trail towards genome-scale metabolic networks. Genome annotation divides into two major steps, namely gene prediction and functional gene annotation.

First, we identify protein coding regions in the genome sequence using dedicated gene prediction tools (e.g. GLIMMER [20], GeneMark [21]) or gene prediction as implemented in reconstruction pipelines. Typically, such tools determine a list of open reading frames (ORFs). Subsequently, the list is filtered for protein-coding sequences by estimating the coding potential of the ORFs based on intrinsic properties of the sequence (e.g. length, promoter signals, CpG islands).

As a next step, we annotate the coding genes with the biological functions of the encoded proteins such as specific enzymatic, transport, or structural functions. Thereby, annotation mainly relies on transferring functional annotations from reference proteins with known functions to the new proteins based on their sequence similarity. Genome annotation therefore typically involves BLAST [22] or FASTA

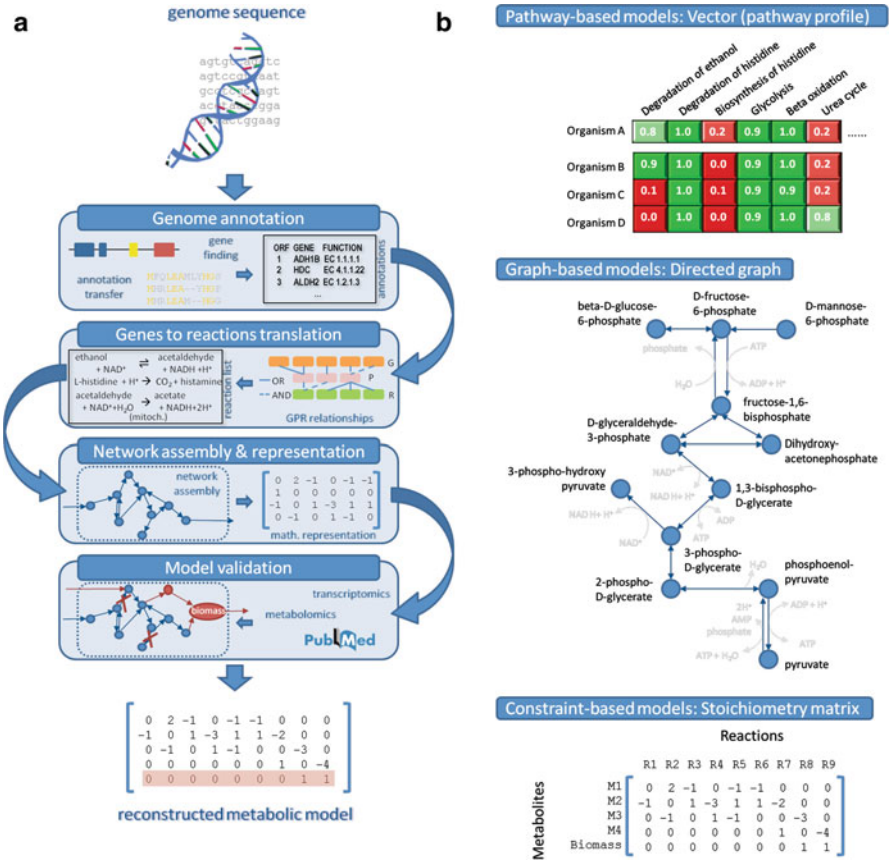


Fig. 17.1 Overview of the reconstruction and mathematical representation of metabolic models. (**a**) Starting from a genome sequence, metabolic reconstruction of a genome-scale metabolic model can be divided into four major steps: (1) Genome annotation provides a list of the identified protein coding genes along with their enzymatic or transport functions. These functions are often described in a structured way using ontologies such as the enzyme classification. (2) For translating genes into reactions, Gene-protein-reaction (GPR) relationships are derived and used to generate a list of the reactions present in the organism along with the (experimental or predicted) reaction directionalities and localizations. This process is based on the gene annotations and on information provided by reaction databases. (3) The reactions are assembled to pathways or complete metabolic networks via shared metabolites. The resulting preliminary metabolic model is translated into a mathematical representation. We can distinguish pathway-based, graph-based, and constraint-based approaches for the assembly and the corresponding mathematical representation. (4) The preliminary model must be corrected, extended, and validated by applying consistency checks and by testing it against experimental and physiological data from literature and high-throughput experiments. (**b**) A metabolic model can be mathematically represented by a (1) vector (pathway profile) describing the completeness or presence/absence of predefined reference pathways (<- pathway-based models), (2) graph describing the connectivity of the reactions (<- graph-based models), and (3) stoichiometry matrix describing the connectivity of reactions along with the reaction stoichiometries (<- constraint-based models)

Table 17.1 Selected databases and tools relevant for metabolic reconstruction

Focus	Database	URL	Content	Tool	Focus
gene/protein sequences	EMBL	www.ebi.ac.uk/embl/	nucleotide sequences/annotations	GeneMark	annotation
	GenBank	www.ncbi.nlm.nih.gov/genbank/	nucleotide sequences/annotations	GLIMMER	
	DDBJ	www.ddbj.nig.ac.jp/	nucleotide sequences/annotations		
	UniProt	www.uniprot.org/help/uniprotkb	protein sequences/annotations		
	Integr8	www.ebi.ac.uk/integr8/	complete genomes (DNA & proteins)		
metabolites	PubChem	pubchem.ncbi.nlm.nih.gov/	small molecules (bioactivities)		
	ChEBI	www.ebi.ac.uk/chebi/	small molecules		
	HMDB	www.hmdb.ca/	small molecules (human metabolism)		
	LipidMaps	www.lipidmaps.org/	lipid molecules		
enzymes/ reactions	BRENDA	www.brenda-enzymes.org/	enzymes/reactions (substr. specif., kin.)	PSORT	localiz. pre-diction
	ENZYME	enzyme.expasy.org/	enzymes/reactions	TargetP	
	TransportDB	www.membranetransport.org/	transport proteins/processes	SignalP	
pathways	KEGG	www.genome.jp/kegg/	pathway-based reconstructions	KAAS	reconstruction pipelines
	BioCyc	biocyc.org/	pathway-based reconstructions	Pathway Tools	
	The SEED	www.theseed.org	pathways (subsystems)/reconstructions	RAST	
	IMG	img.jgi.doe.gov/cgi-bin/w/main.cgi	pathways/reconstructions (microb. seq.)	metaSHARK	
	UM-BBD	umbbd.msi.umn.edu/	pathways for xenotbiotics	IdentiCS	
	Reactome	www.reactome.org	pathways/reconstructions (human, euk.)	Model SEED	
metabolic models	Model SEED	seed-viewer.theseed.org/ seedviewer.cgi?page=ModelView	genome-scale reconstructions with biomass rct. (constraint-based)	GapFind GapFill	validation
	BiGG	bigg.ucsd.edu/	high-quality reconstructions (constr.-b.)	GrowMatch	

[23] searches of the new protein sequence against manually curated reference protein datasets such as UniProtKB/Swiss-Prot [24] in order to identify orthologs [25]. Besides sequence similarity, annotation tools additionally analyze genomic structure conservation and predict functional protein domains to increase the number and reliability of functional assignments [26, 27].

Though functionally characterizing a large fraction of a new genome's proteins, genome annotation by annotation transfer is always restricted to already known sequences and functions. Moreover, contradicting the assumption of annotation transfer, sequence similarity does not necessarily imply functional similarity and vice versa [28, 29]. As a consequence, automatically generated genome annotations usually contain gaps and errors.

Today, sequence databases (e.g. GeneBank [30], IMG [31]) provide automatically derived functional annotations for most sequenced genomes. Tools for genome-scale metabolic reconstruction either build on the annotations therein (e.g. Pathway Tools) or include proprietary genome annotation processing (e.g. RAST, metaSHARK [32], IdentiCS [33]).

1.2 Step 2: From Gene Functions to Chemical Reaction Equations

In the next step, we must translate the annotated gene functions into biochemical reactions and transport processes. In the simplest case, a gene encodes a single protein which catalyzes a distinct reaction. The *gene-protein-reaction (GPR) relationship* becomes more complex if different genes encode the subunits of a protein, which, as such, is needed to catalyze a reaction. Vice versa, a single gene can code for various distinct proteins due to alternative splicing and multiple distinct genes can code for the same protein. Similarly, we can distinguish different cases for mapping proteins onto specific reactions. Again, multiple proteins can form a protein complex needed for the catalysis of a single reaction. In contrast, a protein can catalyze various reactions and a single reaction can be catalyzed by multiple distinct proteins (isozymes). Formally, these relationships are often described using Boolean logic [13]. The evaluation of the GPR logic expression belonging to the reaction R results in either 'reaction R is present' or 'reaction R is absent'. As a side note, the accurate formulation of GPR relationships not only facilitates revealing the set of reactions for the initial reconstruction but also facilitates simulating the metabolic effects of gene knockouts based on the final metabolic model.

Yet, how can we actually infer gene-protein and protein-reaction links from an annotated genome in order to obtain the organism-specific reactions with their chemical equations? While the link between genes and proteins can mostly be extracted directly from genome annotation, protein-reaction links and the chemical equations of the reactions are derived from reaction databases. For metabolic reconstruction, the databases ENZYME [34], BRENDA [35], and TransportDB [36], as well as reaction databases that are part of metabolic databases (e.g. LIGAND

from KEGG, MetaCyc reactions from BioCyc) play an important role. These datasets are collections of chemically well described reactions that have been observed in at least one organism. In order to infer GPR relationships for new genomes, reaction databases often provide links to reference proteins known to catalyze the reaction. Furthermore, formalized descriptions of catalytic functions such as Enzyme Commission (EC) numbers [37] and KEGG orthology numbers (KO) [38] are widely used to link annotated genes and proteins to the chemical equations stored in reaction databases.

Due to the availability of comprehensive reaction databases, the translation of an enzymatic protein into a chemical equation is uncomplicated if the catalytic function has been transferred from the protein of a related species and if the enzyme is very specific for the conversion of a distinct metabolite. However, many enzymes and transporters act on a broad spectrum of substrates. Moreover, the substrate specificity of enzymes often differs between species and isoforms and is difficult to predict from the protein sequence. As an example, the enzyme alcohol dehydrogenase (ADH) catalyzes the oxidation of a variety of primary and secondary alcohols into the respective aldehyde or ketone. In reaction databases, these cases are covered by abstract reactions, in which generic molecules (e.g. alcohol) describe a whole class of substrates. For their correct incorporation into metabolic networks, these abstract reactions must be instantiated. For instance, in order to correctly link a reaction producing ethanol to a reaction consuming acetaldehyde in the presence of ADH, the reconstruction process must first derive the specific chemical equation for the oxidation of ethanol to acetaldehyde based on the abstract reaction for ADH.

Besides their chemical equations, metabolic reconstruction needs information on the reversibility and the localization of reactions. Biochemical reactions can be irreversible under physiological conditions which restrict the way a network can be traversed following a linear pathway. If not described in reaction databases, the reversibility of reactions can be estimated from the reaction Gibbs energy, if known, or based on energy equivalents (e.g. ATP, NADH) in the reactions [39, 40]. In addition to reversibility, restrictions of catalytic activities to specific compartments (e.g. mitochondria, endoplasmatic reticulum, periplasm) play an important role for linking reactions correctly. A metabolite produced in one compartment can only be consumed in a different compartment if the system is able to transport the metabolite accordingly. The localization of an enzyme can be inferred from protein sequence features using tools such as PSORT [41] or TargetP/SignalP [42, 43].

1.3 Step 3: From Chemical Equations to Metabolic Networks and Their Mathematical Representation

After determining the list of organism-specific reactions in the previous step, we can now assemble the entire metabolism of the organism based on this parts list. To this end, we first have to complete the list by a set of spontaneous reactions (i.e. reactions

that do not depend on enzymatic catalysis) that have been frequently observed in biological systems. Most reaction databases provide such reactions.

Having the complete parts list at hand, we must choose among three fundamentally different types of techniques for the assembly of the metabolic network, namely pathway-based, graph-based, and constraint-based approaches. The methods for compiling the network and the network's mathematical representation are intrinsically tied to the technique and objectives of the network analysis itself. In other words, the best choice largely depends on the type of biological questions that we would like to answer by analyzing the reconstructed metabolic model. While pathway-based approaches are well-suited for large-scale comparative analyses across genomes, graph-based and constraint-based approaches allow for more detailed metabolic analyses and predictions for only a few or a single organism.

1.3.1 Pathway-Based Approaches

Historically, metabolism as a whole has been partitioned into functional modules that re-occur in various organisms such as the glycolysis and the citrate cycle. Such a module is referred to as a *metabolic pathway* and comprises all reactions that are necessary to fulfill the respective metabolic function. Most metabolic databases provide a huge set of predefined reference pathways collected from multiple species (e.g. KEGG PATHWAYS and modules, MetaCyc pathways, SEED subsystems).

Given a comprehensive set of known reference pathways, reconstructing the metabolism of an organism corresponds to predicting the presence or absence of each pathway based on the presence of the related reactions in the organism. It is to be noted that pathway-based approaches provide lists of pathways present in an organism rather than metabolic *networks* as such. Mathematically, the pathway-based reconstruction of an organism can be represented as a single high-dimensional vector. Each entry of the vector corresponds to a known pathway and contains a completeness score describing the (predicted) availability of the pathway in the organism [44–46].

Low computational costs make pathway-based approaches well-suited for the reconstruction and comparison of metabolic capabilities across hundreds of species [6, 12, 46–49]. As a consequence, tools designed for the rapid metabolic reconstruction of numerous genomes such as KAAS, Pathway Tools, and RAST mainly rely on pathway-based methods. For far more than 1,000 sequenced genomes, the resulting reconstructions are stored in the metabolic databases that are tied to these tools, namely KEGG, BioCyc, and SEED.

However, on the downside, pathway-based metabolic reconstructions are restricted to known metabolic processes with predefined boundaries. The discovery of previously unknown pathways or pathway alternatives therefore requires more sophisticated metabolic reconstruction techniques.

1.3.2 Graph-Based Approaches

In graph-based approaches, we simply link the reactions that are present in a system via their shared metabolites. Thereby, we directly generate a graph that represents the system's complete metabolic network without any restrictions to reference pathways. Mathematically, the resulting metabolic network can be described as a directed graph, which is defined as an ordered pair (V, E), where V represents a set of vertices (here: metabolites) in the graph and E is a set of ordered pairs (a, b) with $a, b \in V$, called edges (here: reactions connecting the metabolites a and b). Through representing a metabolic network as a mathematically well-defined graph, a huge number of established graph algorithms can be applied for visualizing and traversing the network as well as for analyzing its local and global topological properties. As an example, we can apply such algorithms for calculating the pathway distance (i.e. number of reactions on a linear path) between two metabolites or for revealing alternative biochemical pathways between them.

However, if metabolic graphs are built naively as described above, traversing the graph will yield a plethora of biochemically unrealistic pathways [39]. The main reason is the connection of reactions via abundant cofactors such as ATP or water, which occur in many biochemically otherwise unrelated reactions. For instance, allowing connections via ATP, the shortest path from glucose to pyruvate would involve two reactions instead of the nine reactions that are actually needed for this conversion in the glycolysis pathway. The strategies for avoiding this problem span from deleting or penalizing edges with the most abundant, so-called *side metabolites* [39, 50–52], to the tracing of chemical structure similarities of the source metabolite when traversing the metabolic graph for path finding [53–56]. Besides side metabolites, we must consider two further aspects for the construction of the graph: First, reversible reactions must be split into forward and backward reactions. Second, the localization of reactions must be taken into account, e.g. by representing the same metabolites in different compartments as distinct nodes.

1.3.3 Constraint-Based Approaches

In contrast to pathway-based and graph-based approaches, constraint-based methods make use of reaction *stoichiometries*. The stoichiometries describe the relative quantities of molecules in a reaction following the law of conservation of matter. Constraint-based models rely on the assumption that all metabolites that are neither imported from an external pool nor excreted or accumulated must be *stoichiometrically balanced* (i.e. must be produced and consumed to the same extent) over all (valid) fluxes through the network (= mass balance constraint). Section 2 provides a detailed introduction to the basic principles and applications of constraint-based metabolic models. Here, we only focus on reconstruction issues of constraint-based approaches.

Constraint-based metabolic models generally rely on the so-called *stoichiometry matrix*. This matrix comprises the stoichiometries of all reactions present in a

system and mathematically defines the metabolic network. The columns of this matrix correspond to reactions while rows correspond to metabolites (Fig. 17.1b). The cell (i,j) of the matrix contains the stoichiometric coefficient for the metabolite i in the reaction j. A negative value describes the consumption of the respective metabolite in the reaction while a positive value indicates the production of the metabolite. Reversible reactions are split into a forward and a backward reaction as for the graph-based model. Similar as in graph-based models, we can handle the localization of reactions by defining distinct formal metabolites (e.g. Amitoch, Acytopl) for the same metabolite in different compartments. In contrast to the graph-based reconstruction approach, abundant metabolites such as cofactors do not require any special treatment but are balanced as other metabolites. Instead, we must identify metabolites that are supplied to or excreted from the system and add external copies and transports for them in order to ensure the mass balance constraint inside the system.

1.4 Step 4: From Draft to High-Quality Metabolic Networks

After compiling a network and translating it into a mathematical representation, we are in principle ready to use the model for analysis. However, each of the steps described above is error-prone, especially in a fully automated reconstruction setup. The model thus still contains gaps and errors, most of which originate from incomplete or erroneous annotations and incomplete or erroneous translation of annotated genes into chemical equations.

However, based on the draft model and its mathematical representation, we can, to some extent, systematically identify such gaps and errors by searching for dead ends or inactive reactions (e.g. reactions that cannot carry flux). Various algorithms were proposed and implemented for gap finding and gap filling in metabolic reconstructions (e.g. GapFind/GapFill) [57, 58]. In addition to these consistency checks, the reconstructed metabolic model can be validated by comparing model-based predictions and experimental observations for phenotypes such as growth under various conditions. To this end, metabolic models are often extended by biomass as an artificial metabolite, which is "synthesized" from the combination of biomass precursors such as cofactors, amino acids, lipids, and nucleotides [59–61]. For further validation, we can map high-throughput experimental data (e.g. metabolite concentrations and gene expressions) onto the model in order to identify inconsistencies [62].

Though tools such as Model SEED [60] already support automated reconstruction, correction, and validation of metabolic models, high-quality metabolic models as stored in the BiGG database are still created by an iterative refinement process involving time-intensive manual intervention.

2 From Networks to Metabolic Fluxes: Intrinsic Properties of the System

The assembly of all compounds, enzymes and reactions of the system under investigation described in the previous section is a first, mandatory step in the analysis of its metabolism. Having the stoichiometry matrix as a convenient, computer-readable representation of the system at hand, we can now ask questions of intrinsic properties of the metabolic system; even without experimental data of metabolite concentrations or reaction rates. Pioneering work in this field has been conducted in the 1990s and early 2000s by Palsson and colleagues at UCSD [63]. In this section we will briefly explain the concepts of metabolic pathway analysis, mass balance, extreme pathways and flux balance analysis.

2.1 Basic Concepts

As an example, consider the three-metabolite toy network display in Fig. 17.2a. It consists of a single incoming boundary reaction v_1 which brings A into the system, a splitting reaction v_2 producing both B and C from A, a conversion v_3 from C to B and two outgoing boundary reactions v_4 and v_5. Note that in this framework we only use irreversible reactions, i.e. a reversible reaction must be replaced by two irreversible ones. Interestingly, the mere wiring of the network structure alone already constrains the possible mass-flow through the system. For instance, if each reaction *fires* exactly once (i.e. transports a single molecule), we will end up with an unchanged number of A molecules, an increase of B by 1 and, subsequently, a decrease of C also by 1 (Fig. 17.2b). Formally, if we multiply the stoichiometry

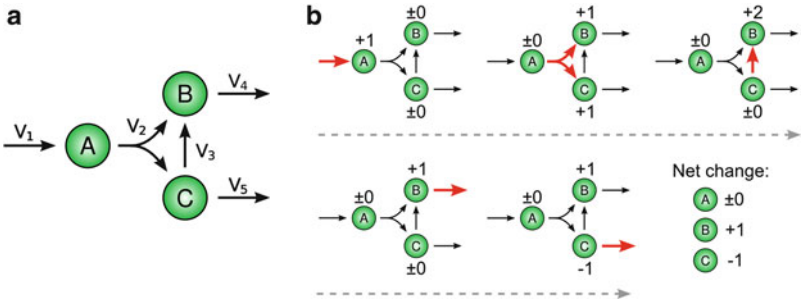


Fig. 17.2 (**a**) Metabolic toy network consisting of three nodes A–C and five reactions v_1–v_5. (**b**) Total metabolite concentration changes after each reaction fired exactly once. Note that the actual reaction order is not relevant for the net change; this figure merely represents a visualization scheme

matrix with the number of times each reaction fires as a column vector, we obtain the net concentration change of each metabolite:

$$\dot{x} = S \cdot v$$

Where $\dot{x}$ is the concentration change vector of metabolites, S again represents the stoichiometry matrix of the metabolic network, v stands for the number of times each reaction fires in the given time step, and represents the matrix multiplication operator. For our example, this results in

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & -1 \end{pmatrix} \cdot (1 \quad 1 \quad 1 \quad 1 \quad 1)^T = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

which is the same result as obtained by manual calculation in Figure 17.2b. The mass passing through the system in a given time step is commonly referred to as *flux* [64], and v thus represents the so-called *flux vector*.

With the matrix multiplication operation above we now have a tool to transform a network with given reaction rates into concentration changes over time. The question of biological relevance now is: What is the actual flux vector in a living cell? In the following we discuss how we can drastically reduce and subsequently interpret the number of possible flux vectors using a simple, biologically motivated constraint.

2.2 Mass Balance and Extreme Pathways

Enzymatic reactions in the metabolic system are considered to be fast in comparison to the physiological or chemical changes that drive the system from the outside. Consequently, we assume the system to be in steady state: despite constant mass flow through the system, the actual metabolite concentrations remain unchanged. Metabolites flowing into the system are processed into output metabolites of the system without changing its internal state (Fig. 17.3). The steady state condition can be formally expressed using the matrix multiplication introduced above:

$$S \cdot v \overset{!}{=} 0$$

We only allow those flux vectors v which are able to maintain mass balance in the system. Intuitively, if the flux acting upon the system does not fulfill this steady state condition for all internal metabolites, the affected compounds will either deplete or grow infinitely over time. Note that the flux vector illustrated in Figure 17.2 above did *not* fulfill the steady state condition, since it yielded non-zero changes for both B and C.

We now need to obtain a solution of the equation above in order to find those fluxes that are biologically feasible with respect to the steady state condition. Simply solving this system of equations by standard linear algebra techniques is mathematically possible, but might result in negative flux values. Since obviously flux values

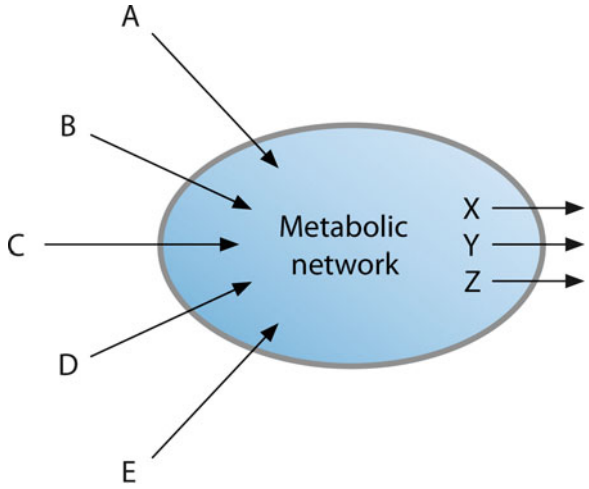


Fig. 17.3 Mass balance. The metabolic network takes input metabolites A–E as a substrate and processes them to the target metabolites X–Z (which could for instance represent compounds required for biomass production). The internal metabolites of the system are required to be constant over time (Adapted from [64])

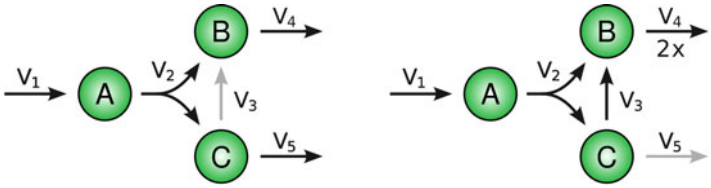


Fig. 17.4 The two extreme pathways for our toy network (*black arrows*). By definition, both flux vectors will maintain steady state, i.e. metabolite concentrations are not changed after the reactions fired

cannot be negative (recall that we always model the system by irreversible reactions), two independent research groups designed algorithms that generate positive solution descriptions for the steady state equation: (a) Elementary flux modes (EFMs) by Schuster and colleagues [65] and (b) Extreme pathways by Schilling and Palsson [66]. Both methods yield identical results if the system only comprises of irreversible reactions [67]. For simplicity, we will here only refer to the extreme pathways approach.

Our toy model from the previous examples gives rise to two extreme pathways (Fig. 17.4):

$$e_1 = (1 \quad 1 \quad 0 \quad 1 \quad 1)^T \text{ and } e_2 = (1 \quad 1 \quad 1 \quad 2 \quad 0)^T$$

It can be easily verified, mathematically but also visually, that both flux vectors will not change the concentrations of A, B and C. Despite the simplicity of this

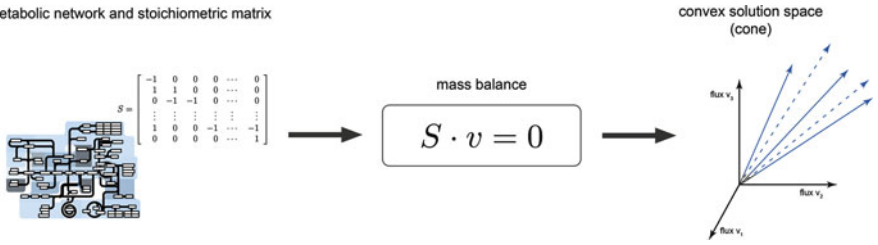


Fig. 17.5 From the stoichiometry matrix over the mass balance constraint to a positive convex cone spanned by the extreme pathways of the system (Figure inspired by [64])

example case, we actually learned something about the intrinsic properties of the system using the extreme pathways. Reactions v_1 and v_2 are absolutely required to keep mass flow through the system, while v_3, v_4, and v_5 have to be in a subtle balance to ensure steady state. Note that any linear combination of these extreme pathways also lies in the solution space of steady states, e.g.

$$v := 2 \cdot e_1 + e_2 = (3 \quad 3 \quad 1 \quad 4 \quad 2) \text{ still yields } S \cdot v = 0$$

Generally, the set of extreme pathways can be interpreted as basis vectors which span a convex positive cone (Fig. 17.5), whose inner volume represents the set of all feasible fluxes that fulfill the steady state condition. In other words, constraining the system to fulfill this condition drastically reduces the number of possible fluxes from the entire space down to this positive cone.

2.3 Applications of Extreme Pathways

In the following, we will present several studies which employed the extreme pathway methodology to address questions regarding correctness of the metabolic reconstruction, minimal growth medium prediction, pathway redundancy and flux balance analysis.

First, extreme pathways can be used to consolidate and refine the metabolic network reconstruction. For example, each reaction contained in the metabolic model should actually be *used*, i.e. be part of at least one extreme pathway and thus contributing to the overall flux [64]. An unused reaction could be an indicator for both reconstruction problems but also biologically relevant mechanisms. In particular, an unused reaction in the model could indicate: (a) There could be incomplete pathways due to evolutionary intermediates, i.e. not all enzymes of the respective pathway are present in the genome. (b) The *in silico* model could be incomplete at this point. The organism indeed contains all enzymes required for the pathway, but the computer model still misses some of them. (c) The functional annotation of the respective enzyme might be wrongly transferred; the protein is expressed but does not fulfill the function it was assigned in the model.

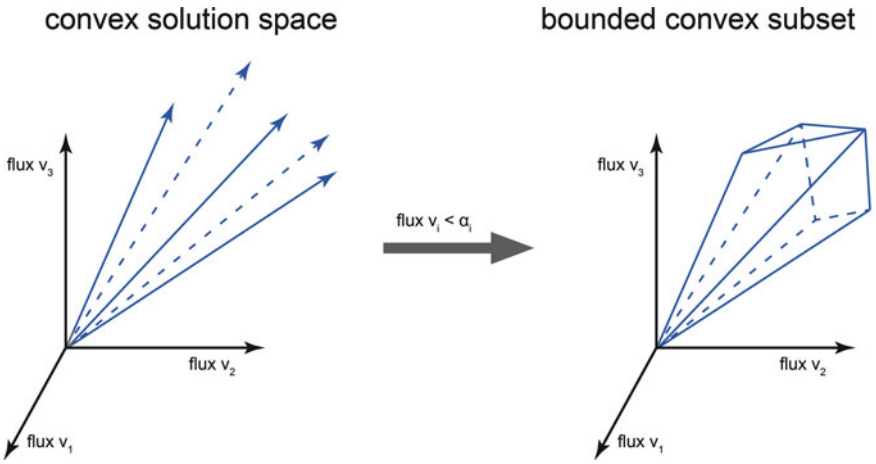


Fig. 17.6 Constraining the convex solution space by *upper* flux boundaries

The second example of extreme pathway applications is the prediction of minimal medium predictions required for the growth of *Haemophilus influenzae* [59]. The idea is to select a minimal subset of possible nutrient metabolites such that the microbe is still able to produce all metabolites required for growth. Again, this question was tackled by the use of extreme pathways, checking which routes through the network are active if a given set of input substrates is present, and whether all required output fluxes (taking biomass compounds from the system) are active. The authors identified a total of 11 metabolites as minimal substrate requirements, which were in accordance with previously published experimental results.

Another biological property which was examined using the extreme pathway approach involves pathway redundancy in *Helicobacterpylori* and *Haemophilus influenzae* [68]. The rationale behind this analysis is the direct relationship between redundancy and robustness. If the system has more possibilities to transform input metabolites into required output metabolites, it will be more robust to disturbances and variations of external conditions. Redundancy of the metabolic system can be assessed using extreme pathways by counting how many pathways connect each input state (possible input metabolite fluxes) with each output state (in this case the production of essential amino acids and other metabolites required to produce biomass). The main finding of this study was the drastically lower robustness of the *H. pylori* metabolic network in comparison to the *H. influenzae* network. Interestingly, this prediction indeed makes sense from a biological point-of-view. *H. pylori* naturally resides in a well-defined metabolic niche, the human stomach, in a wealth of nutrients required for its growth, and thus inherently does not require to cope with strong variations of external states.

Our final case of extreme pathway applications for real-world biological problems is flux balance analysis (FBA). In addition to the extreme pathway framework, FBA adds concrete upper boundaries for each reaction speed (Fig. 17.6) and then

computationally optimizes a given biological function. The goal here is to go beyond a whole set of solutions and pick a specific point from the solution space. For instance, Segrè and colleagues [69] used a genome-scale metabolic reconstruction of E. coli in which they maximized the speed of a bio-mass generating reaction in the model. The biological assumption here is that the bacterium is seeking to arrange its metabolic flux patterns such that it can grow at a maximal rate. Comparing the resulting (computationally) optimal fluxes with experimental data demonstrated good agreement between predicted and measured flux rates.

In summary, in this section we reviewed several analysis techniques which are solely based on the reconstruction of metabolic networks. Interpreting the basis vectors of the steady state solution space, i.e. the extreme pathways, as possible operating modes in the metabolic networks, one obtains an intrinsic definition of a "pathway" as opposed to the rather arbitrarily defined "pathways" in public databases like KEGG or HumanCyc. We discussed several applications of this technique to biological questions like pathway redundancy and reaction flux rate prediction.

3 From Metabolic Fluxes to Reaction Rates: Incorporation of Concentration Data

In the previous part we have seen how to obtain flux properties of networks with given stoichiometric matrices by applying reconstruction techniques resulting in insights into the characteristics of metabolic systems. However, kinetic models of biochemical reactions are essential to understand the dynamic behavior of metabolism under specific conditions. In the following section, we will discuss how to obtain kinetic parameters by incorporating metabolite concentration data into the extreme pathway framework introduced before.

Biochemical in vitro experiments provide a large set of measured kinetic constants, which are collected in freely available databases such as BRENDA [35] and Sabio-RK [70]. However, these collections are mostly incomplete and do not reflect the broad variability and range of the entire metabolism. Moreover, quantitative models incorporating in vitro derived parameters failed to predict in vivo phenotypes in experimental studies [71, 72]. Hence, methods are needed for the construction of metabolic models based on in vivo data in order to describe the dynamics of in vivo reactions on different levels of complexity, including central metabolic pathways, but also the metabolism of a cell or the whole organism.

The recent development of high-throughput metabolomics techniques, mostly based on mass spectrometry and nuclear magnetic resonance spectroscopy, now provide large-scale measurements of chemical compounds in biological samples [73]. This allows for the quantification of in vivo external and internal metabolite concentrations. Several methods have been proposed to integrate metabolite concentrations into quantitative metabolic models. Their common goal is to estimate

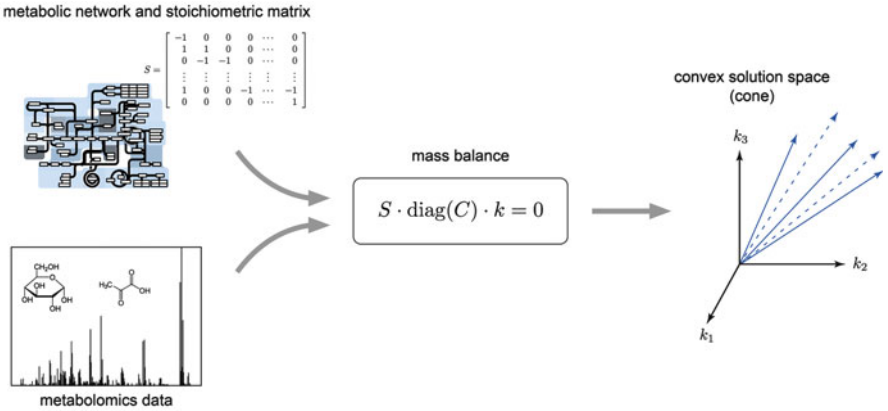


Fig. 17.7 Integrative analysis of metabolomics data – both stoichiometric and concentration data are used to characterize the dynamical behavior of metabolic networks under specific conditions

Fig. 17.8 Linear cascade of metabolic reactions. This topology can be found in metabolic pathways like fatty acid beta-oxidation or glycolysis

kinetic parameters specifically for given cellular conditions such as healthy and disease states.

One such example is *k-cone analysis* [74], which again constructs a convex space (cone) based on metabolite concentrations. This cone then includes all candidate values for kinetic parameters under steady state conditions (see Fig. 17.7). Further extensions of this approach also consider enzymes and their various functional states as compounds and therefore addressing regulatory effects [75, 76].

As an example model, we now investigate a linear reaction cascade with irreversible influx, outflux and enzymatic reactions (Fig. 17.8). Examples for such linear pathways are fatty acid beta-oxidation or glycolysis. The system can be described using mass action rate laws by a system of ordinary differential equations:

$$\dot{A} = k_i - k_A \cdot A$$
$$\dot{B} = k_A \cdot A - k_B \cdot B$$
$$\dot{C} = k_B \cdot B - k_c \cdot C$$

Where A, B and C represent compound concentrations, and $\dot{A}, \dot{B}$ and $\dot{C}$ the respective time differentials. As seen in Sect. 2.1, we can describe the reaction cascade using a stoichiometry matrix **S**. Each row of the matrix belongs to a

compound and each column represents an elementary reaction (denoted as an arrow in the reaction scheme).

$$s = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

A negative entry s_{ij} in **S** states that compound c_i is an educt for reaction v_j. For a product of reaction v_j, the stoichiometric coefficient s_{ij} is positive. If compound c_i is not involved in the reaction v_j, then s_{ij} is zero. Furthermore, we define a vector X containing products of substrate concentrations according to the law of mass action:

$$X = \{x_1, \ldots, x_m\}$$

In addition, we require a vector k containing elementary rate constants for all reactions v_1-v_r. Using **S**, X and k, we can express the system of differential equations in matrix notation:

$$\begin{pmatrix} \dot{A} \\ \dot{B} \\ \dot{C} \end{pmatrix} = \mathbf{s} \cdot \mathrm{diag}(X) \cdot \mathrm{k} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & A & 0 & 0 \\ 0 & 0 & B & 0 \\ 0 & 0 & 0 & C \end{pmatrix} \begin{pmatrix} k_i \\ k_A \\ k_B \\ k_c \end{pmatrix}$$

or by substituting $\mathbf{M} := \mathbf{S} \cdot \mathrm{diag(X)}$

$$\begin{pmatrix} \dot{A} \\ \dot{B} \\ \dot{C} \end{pmatrix} = \mathbf{M} \cdot \mathrm{k} = \begin{pmatrix} 1 & -A & 0 & 0 \\ 0 & A & -B & 0 \\ 0 & 0 & B & -C \end{pmatrix} \begin{pmatrix} k_i \\ k_A \\ k_B \\ k_c \end{pmatrix}$$

Compared to the stoichiometric matrix **S**, matrix **M** combines both information about stoichiometry as well as reaction kinetics.

Under steady-state conditions, the concentrations of all compounds remain constant, i.e. the change in concentration over time is zero. Measured metabolite concentrations often reflect these conditions, as biochemical reactions are assumed to reach equilibrium in the order of milliseconds to seconds. If we know the topology of a biochemical pathway we can describe all reactions using matrix **M**. For a given set of metabolite concentrations we can estimate rate constants by setting the equation above to zero.

Matrix representations of metabolic systems are commonly underdetermined, meaning that we have more unknown kinetic parameters than independent linear equations when only considering non-negative kinetic rates results in a convex cone spanned up by basis vectors in the solution space (see Fig. 17.7). All linear combinations of these basis vectors reside inside the cone and are therefore also solutions satisfying the physiochemical and condition-dependent constraints of the system of

differential equations. Adding additional constraints and information about equilibrium constants can reduce the dimensionality of the solution space. Sampling methods [77, 78] allow for the efficient comparison of feasible solutions between different conditions.

k-cone analysis has been applied to metabolic networks of different biological complexity, such as glycolysis, central biochemical pathways in yeast and human red blood cell metabolism [74–76, 79]. The advantage of such methods is the minimal amount of biological data required. On the downside, such approaches often rely on simplifying assumptions which have to be carefully considered especially when dealing with complex dynamic reactions and interactions in living systems. Therefore, additional methods have been proposed to overcome this issue.

In this section we have discussed methods for the estimation of kinetic parameters. This analysis can be seen as an extension to the methods reviewed in Sect. 2. By incorporating additional concentration data one can deduce kinetic properties of metabolic systems. Since more and more metabolomics data will be available, information can be gathered about distinct dynamic responses of metabolic systems to perturbations such as stress or disease states.

4 From Metabolite Concentrations to Dynamical Systems: Stochasticity

In this part we now further explore the connection between metabolomics measurements and the underlying dynamical systems. In particular, we investigate stochastic fluctuations in both metabolite and enzyme concentrations, which leave detectable footprints of the metabolic pathways in our experimentally measured concentration levels. When generally referring to "variation" in the data, it is important to understand that this term might refer to completely different concepts, each of which requires special attention and treatment. We here distinguish between three types of variation: (a) Technical noise due to inaccuracies in the measurement procedure. (b) Intrinsic noise due to natural fluctuations of molecule quantities inside a living cell. (c) We further distinguish "extrinsic noise" as a third type of noise. This can refer to general differences in reaction speeds and metabolic processes, e.g. between different organisms due to genetic variation, nutritional states, life-style and other external influences.

While technical noise obviously impairs the evaluation and interpretation of measured data, we will see that intrinsic and extrinsic variations contain valuable signals which are directly connected to the underlying metabolic network. The general idea is that metabolite concentrations do not represent independent signals in the data, but display strong correlations which are a direct consequence of the wiring of the underlying metabolic network. Recently, researchers in the field attempted to elucidate the origins of such metabolite-metabolite correlations in metabolomics data. We will briefly outline the findings of two studies in the following. Note that we here specifically focus on cross-sectional steady state data.

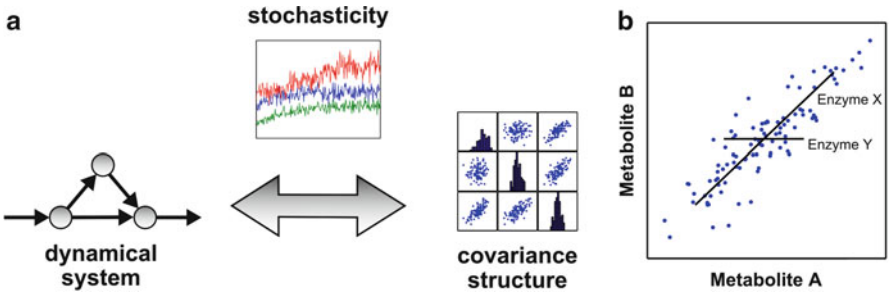


Fig. 17.9 Connecting metabolite level variations and the underlying metabolic system (**a**) Steuer et al. [82] devised a mathematical framework based on stochastic differential equations, which establishes a direct connection between dynamical systems (represented as the corresponding Jacobian matrix) and the observed pairwise covariances. (**b**) In a later study, Camacho and colleagues explained covariance between metabolites using co-response profiles. Each enzyme introduces a specific direction of covariance between metabolites, the overlaying of which results in the finally observed correlation (Figures inspired by [82] and [84])

Other studies focused on the reconstruction of reaction network topologies from time-course perturbation data [80, 81]. While this approach is certainly of great interest as well, it requires a completely different analysis framework which is beyond the scope of this text.

An early study on how to systematically investigate variations in metabolic systems has been published by Steuer and colleagues in 2003 [82]. The authors assumed stochastic fluctuations of metabolite inside and outside the cells which are in identical states otherwise (biological replicates). This corresponds to the "intrinsic variation" scheme we introduced above. Mathematically, such variation can be expressed via stochastic differential equations (SDEs). An SDE is a stochastic extension of a regular ordinary differential equation (ODE), but in addition to the directed *drift* term it contains a *diffusion* term representing white noise-driven fluctuations on the molecule numbers over time [83]. Two cells with an identical internal state will fall into qualitatively the same steady state after a given amount of time, but the actually measured steady state concentrations of biochemical molecules might differ slightly. The main contribution of the study was the derivation of a mathematical relationship between covariance between the measured metabolites and the Jacobian matrix of the dynamical system (Fig. 17.9a). The Jacobian matrix can be understood as a combination of the network topology with specific rates for each reaction. In this framework, given a metabolic network with specified reaction rates, one can immediately derive the covariances between all pairs of metabolites. Moreover, given measured covariance values between metabolites, one can obtain information about the dynamics of the metabolic network acting underneath the metabolite pools. In summary, the paper provided a first link between variation in measured metabolite concentrations and properties of the underlying biochemical system.

A later study by Camacho et al. from 2005 [84] then shifted the focus from intrinsic fluctuations of the metabolite levels to actual differences in enzyme levels,

thus directly affecting reaction rates in the system. This scenario corresponds to what we called "extrinsic" variation above; the states between different cells actually differ and variations are not only due to stochastic fluctuations. The main methodological concept of this study was the investigation of so-called *co-response profiles*. For fixed enzyme concentrations, the system will fall into a single, unique steady state that can be represented as one dot in a 2D phase plane. Varying the concentration of one enzyme at a given time will create a co-response profile for this enzyme in a certain direction in metabolic space (solid lines in Fig. 17.9b). The mixture of co-response profiles of all enzymes in the system then produces the co-variation we see between metabolites (scatter plot in Fig. 17.9b). The study thus provides a systematic definition of the actual origins of pair wise correlations in metabolomics data. Importantly, the paper also describes limitations of correlation-based approaches. For example, if co-response profiles of similar strength are orthogonal, the mutual covariance is canceled out and no correlation will be observed. Such issues have to be kept in mind when attempting to reconstruct metabolic reactions from steady state data in the following section.

5 Unbiased Reconstruction of Metabolic Networks from Metabolomics Data

In this final part we now further focus on the effects of *extrinsic* variation in the systems parameters, i.e. inter-individual variation of the biochemical reaction rates and external conditions. The goal is to reconstruct directly related metabolites from metabolite concentration data *without prior knowledge* of the underlying metabolic pathways. The results presented in the following closely follow a previously published study in BMC Systems Biology [85].

As described above, pairwise correlations between measured variables are usually estimated using Pearson product-moment correlation coefficients. A major drawback of these correlation coefficients, however, is their inability to distinguish between direct and indirect associations. Pearson correlations are generally high in large-scale *omics* data sets, suggesting a plethora of indirect and systemic associations. For example, transcriptional co-regulation amongst many genes will give rise to indirect interaction effects in mRNA expression data [86]. Similar effects can be observed in metabolic systems which, in contrast to genetic networks, contain fast biochemical reactions in an open mass-flow system. Metabolite levels are supposed to be in quasi-steady state compared to the time scales of upstream regulatory processes [87]. That is, metabolites will follow changes in gene expression and physiological processes on the order of minutes and hours, but will appear unchanged on the order of seconds. These properties, even though substantially different from mRNA expression mechanisms, also give rise to indirect, system-wide correlations between distantly connected metabolites.

Gaussian graphical models (GGMs) circumvent indirect association effects by evaluating *conditional* dependencies in multivariate Gaussian distributions [86].

A GGM is an undirected graph in which each edge represents the pair wise correlation between two variables conditioned against the correlations with all other variables (also denoted as *partial* correlation coefficients). GGMs have a simple interpretation in terms of linear regression techniques. When regressing two random variables X and Y on the remaining variables in the data set, the partial correlation coefficient between X and Y is given by the Pearson correlation of the residuals from both regressions. Intuitively speaking, we remove the (linear) effects of all other variables on X and Y and compare the remaining signals. If the variables are still correlated, the correlation is directly determined by the association of X and Y and not mediated by the other variables.

Partial correlation coefficients have previously been applied to biological data sets for the inference of association networks from mRNA expression data [86, 88–90], for the elucidation of relationships between genomic features in the human genome [91], and to investigate genetically determined relations between metabolites [92].

5.1 Computer-Simulated Metabolic Networks

In order to get a general idea of whether GGMs are indeed suitable for the recovery of metabolic reactions from metabolomics data, we first set up a series of computer-simulated reaction systems (Fig. 17.10). Inter-individual variation is modeled by applying a log-normal noise model on the parameters of the system. Such variation might be genetically determined or, more likely, might be the result of distinct regulatory effects and metabolic states between individuals. All reaction systems were implemented as ordinary differential equations with simple mass-action kinetics rate laws. In order to account for the above-mentioned enzymatic variability, a log-normal noise model was applied, which has been previously been described to be a reasonable approximation of cellular rate parameter distributions [93]. For each parameter sample, the steady state concentrations on log scale were derived, and subsequently the GGM was estimated by calculating partial correlation coefficients.

The first network we investigated consists of a linear chain of three metabolites with different variants of reaction reversibility (Fig. 17.10a–c). We observe high pair wise correlations for metabolites in mutual equilibrium due to reversible reactions (Fig. 17.10a). If only irreversible reactions are employed in the chain, neither regular correlation networks nor GGMs can distinguish between direct and indirect effects (Fig. 17.10a). Species A is the only input metabolite into the system, and thus completely determines the levels of both B and C. This leads to generally high and non-distinguishable correlations between the three metabolites. However, if we introduce exchange reactions for all species, the GGM again correctly describes the network connectivity (Fig. 17.10c). Such exchange mechanisms are likely to be present for most intracellular metabolites, which usually participate in multiple metabolic pathways. In addition to linear chains, pathway modules consisting of branched topologies with first-order, reversible reactions are also correctly reconstructed by the method (Fig. 17.10d, e).

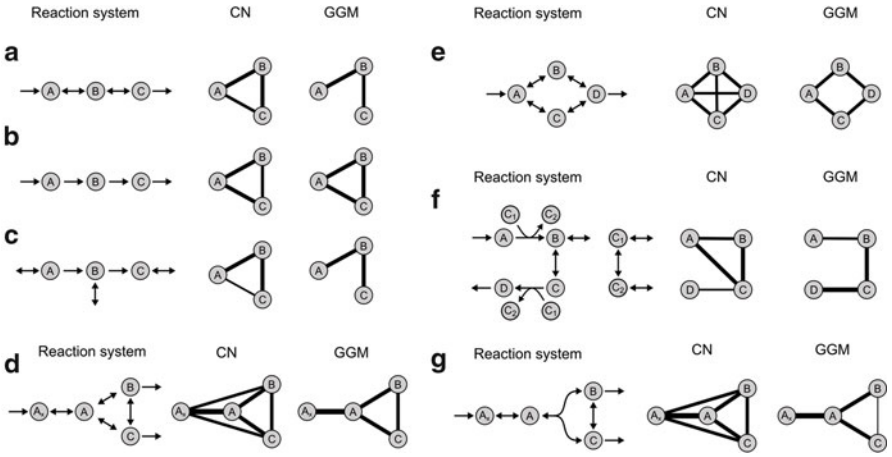


Fig. 17.10 Computer simulated reaction systems. (**a**–**c**): Linear reaction chains are properly reconstructed by the GGM if either reversible reactions or external reactions are present (subnetworks A & C). Regular correlation networks (CN) show high correlations between all metabolites. (**d**–**f**) Branched topologies and co-factor driven systems are also readily recovered by the GGM. (**g**) For this non-linear system with a bi-molecular reaction, the GGM only predicts a weak interaction between B and C. This is due to counter-antagonistic processes of isomerization and substrate participation in the same reaction (Adapted from [85])

Next, we studied the influence of cofactor-driven reactions on the reconstruction. Cofactors are ubiquitous substances usually involved in the transfer of certain molecular moieties or redox potentials [94]. We set up a network resembling the first three reactions from the glycolysis pathway. Again the GGM correctly describes metabolite connectivity in the system, whereas a regular correlation graph leads to false interpretations of the network topology (Fig. 17.10f). Finally, we investigated the effects of rate laws with non-linear substrate dependencies in the absence of cofactors. We modeled a reversible, bimolecular split reaction with isomerization of the two substrates (Fig. 17.10g). An example of such a reaction network can be found in the glycolysis pathway between *fructose-1, 6-bisphosphate, glyceraldehyde-3-phosphate and dihydroxyacetone phosphate*. In this scenario, the GGM only detects a weak association between B and C. As mentioned earlier in Sect. 4, such problematic cases have to be kept in mind when interpreting partial correlations on real data in the next step.

5.2 A GGM on Metabolomics Data

In the following we estimated a Gaussian graphical model using targeted metabolomics data from the German population study KORA [95] (“Kooperative Gesundheitsforschung in der Region Augsburg”). We used a subset of the data set

previously evaluated in a genome-wide association study [96], containing 1,020 targeted metabolomics fasting blood serum measurements with 151 quantified metabolites. The metabolite panel includes acyl-carnitines, four classes of phospholipid species, amino acids and hexoses. Both regular Pearson correlation coefficients and partial correlation coefficients (inducing the GGM) were calculated on the logarithmized metabolite concentrations. All edges corresponding to correlation values significantly different from zero now induce the networks displayed in Fig. 17.11a, b.

Pearson correlation coefficients show a strong bias towards positive values in our data set (Fig. 17.11c); a typical feature of high-throughput data sets, also observed, e.g. in microarray expression data, which can be attributed to unspecific or indirect interactions [86]. We obtain 5,479 correlation values significantly different from zero with $\hat{\alpha} = 8.83.10^{-7}$ ($\alpha = 0.00$ after Bonferroni correction), yielding an absolute significance correlation cutoff value of 0.1619. In contrast, the GGM shows a much sparser structure with 417 significant partial correlations after Bonferroni correction (Fig. 17.11d). Most values center around a partial correlation coefficient of zero, whereas we observe a clear shift towards positive significant values.

The GGM displays a modular structure with respect to the seven metabolite classes in our panel, while the class separation in the correlation network appears rather blurry (Fig. 17.11e, f). We observe a clear separation of the amino acids and acyl-carnitines from all other classes. The four groups of phospholipids (diacyl-PCs, lyso-PCs, acyl-alkyl-PCs, and sphingomyelins) still showed locally clustered structures, but are strongly interwoven in the network. This is probably an effect of the dependence of all phospholipids on a similar fatty acid pool and, subsequently, the biosynthesis pathway acting on this substrate pool. In order to get an objective quantification of this observation, we calculated the group-based modularity Q on all significantly positive GGM edges according to [97]. The same measure was calculated for 10^5 randomized GGM networks (random edge rewiring). For the original GGM we obtain a modularity of $Q = 0.488$, and the random networks yield $Q = 0.118 \pm 0.016$, resulting in a highly significant z-score of $z = 23.49$.

Taken together, partial correlation remove a plethora of correlations from the original dataset, and the resulting network displays a strongly modular structure.

5.3 *Investigation of High-Scoring Subnetworks*

The next step in our analysis is the manual investigation of metabolite pairs displaying strong partial correlation coefficients. Clear-cut signatures of the desaturation and elongation of long chain fatty acids can be seen for various sphingomyelins and lyso-PCs (Fig. 17.12a). For example, SM C18:0 and SM C18:1 strongly associate with a partial correlation of $\zeta = 0.767$, most probably representing the initial $\Delta 9$ desaturation step of the polyunsaturated fatty acid biosynthesis pathway from C18:0 to C18:1-$\Delta 9$ by SCD (*Steaoryl-CoA desaturase*). The similarly high partial correlation between SM C16:1 and SM C18:1 ($\zeta = 0.765$) as well as lysoPC a C16:1 and lysoPC a C18:1 ($\zeta = 0.315$) can be attributed to the ELOVL6-dependent elongation from

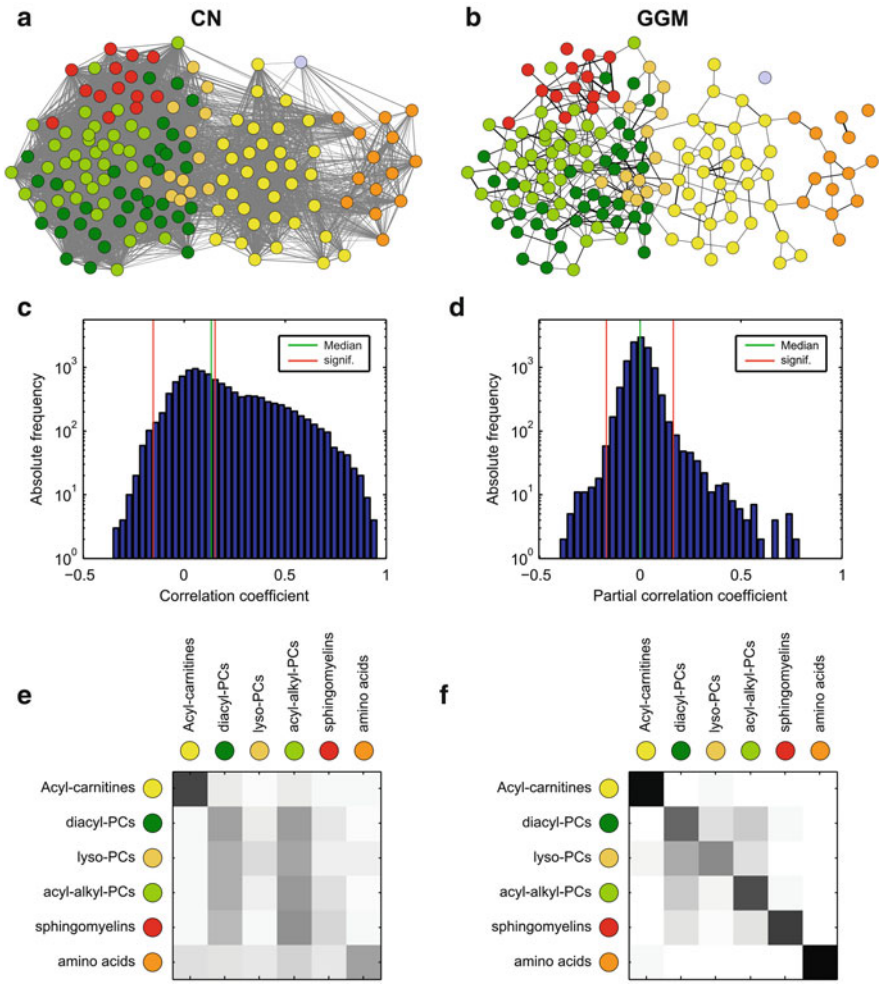


Fig. 17.11 Regular correlation network (CN) and partial correlation network (GGM) reconstructed from metabolomics data. (**a**, **b**) Each *circle* corresponds to a metabolite; edges represent correlations significantly different from zero. (**c**, **d**) Histograms of 11,325 pair wise correlation coefficients (i.e. edge weights) for both networks. *Green lines* indicate the median values; *red lines* denote a significance level of 0.01 after Bonferroni correction. (**e**, **f**) Modularity between metabolite classes measured as the relative out-degree from each class (*rows*) to all other classes (*columns*). The GGM (*right*) shows a clear separation of metabolite classes, with some overlaps for the different phospholipid species diacyl-PCs, lyso-PCs, acyl-alkyl-PCs and sphingomyelins. Values range from *white* (0.0 out-degree towards this class) to *black* (1.0). PCs = phosphatidylcholines (From [85])

C16:1-Δ 9 to C18:1-Δ 11. Interestingly, this reaction was not contained in the public reaction databases but has been previously described by Matsuzaka et al. [98].

We identify a variety of strong GGM edges between diacyl-PC (lecithins, PC aa) metabolite pairs (Fig. 17.12b). For instance, PC aa C34:2 and PC aa C36:2 associate

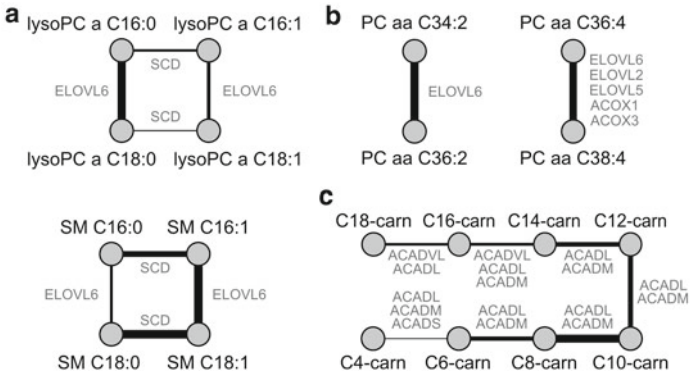
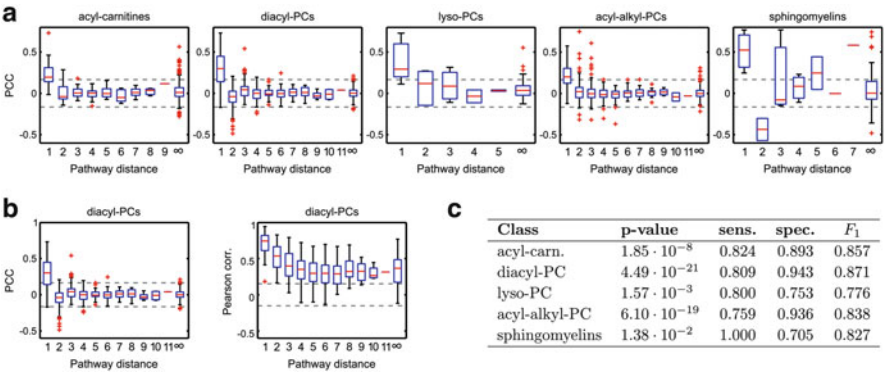


Fig. 17.12 Biochemical subnetworks identified by high-scoring GGM regions. (**a**) Elongation and desaturation signatures for C16 and C18 fatty acids incorporated in lyso-PCs and sphingomyelins, which can be attributed to ELOVL6 and SCD. (**b**) Diacyl-phosphatidylcholine (PC aa) species with elongation and peroxisomal β-oxidation associations. Several combinatorial variants of side chain compositions are possible for C36:4 and C38:4, and thus different enzymes could mediate this connection. (**c**) Recovered β-oxidation pathway from fatty acid chains C18 down to C4. Four enzymes with overlapping substrate specificities catalyze the rate-limiting reactions of this pathway (Adapted from [85])

strongly with $\zeta = 0.735$, and PC aa C36:4 and PC aa C38:4 show a partial correlation of $\zeta = 0.672$. While the first pair can be precisely explained by an elongation from C16:0 to C18:0 by ELOVL6, different combinatorial variants come into play for the PC aa C36:4/PC aa C38:4 pair. The mass-spectrometry technique used in this study only measures the bulk side chain carbon content and total degree of desaturation. Depending on the exact composition of both fatty acid residues in the respective lipids, this association could be caused by long-chain elongations (C14 to C16 and C16 to C18 through fatty acid synthase and ELOVL6, respectively), by very-long-chain elongations (C22:4 to C24:4 through ELOVL2 or ELOVL5) and even by peroxisomal β-oxidation of fatty acids (through ACOX1 or ACOX3).

For the acyl-carnitine group we observe a remarkably high partial correlation of $\zeta = 0.735$ for C8-carn and C10-carn and further acyl-carnitine pairs with a carbon atom difference of two (Fig. 17.12c). These associations can be attributed to the mitochondrial β-oxidation pathway, i.e. the catabolic breakdown of fatty acids [94]. During this degradation process, C_2 units are continuously split off from the shrinking fatty acid chain. Four *acyl-CoA dehydrogenases*, ACADS, ACADM and ACADL, ACADVL, catalyze the rate limiting reactions of β-oxidation for different fatty acid chain lengths [99, 100]. Our interpretation of acyl-carnitine correlations as signatures of mitochondrial β-oxidation is in accordance with [96], who identified associations between C8 + C10, C12 and C4 with genetic variation in the ACADM, ACADL and ACADS loci, respectively.



c

Class	p-value	sens.	spec.	F_1
acyl-carn.	$1.85 \cdot 10^{-8}$	0.824	0.893	0.857
diacyl-PC	$4.49 \cdot 10^{-21}$	0.809	0.943	0.871
lyso-PC	$1.57 \cdot 10^{-3}$	0.800	0.753	0.776
acyl-alkyl-PC	$6.10 \cdot 10^{-19}$	0.759	0.936	0.838
sphingomyelins	$1.38 \cdot 10^{-2}$	1.000	0.705	0.827

Fig. 17.13 Systematic comparison of correlation coefficients and pathway distances. (**a**) Pathway distances from our pathway model against partial correlation coefficients (PCC) for the five lipid-based metabolite classes in our data set. We observe an enrichment of significant partial correlations for a pathway distance of exactly one. (**b**) Comparison of partial correlation coefficients and Pearson correlation coefficients. Pearson correlation coefficients are generally high, independent of the actual pathway distance. (**c**) Wilcoxon rank sum test p-values between the partial correlation distributions of directly and indirectly connected pairs, and sensitivity/specificity/F_1 values measuring the discriminatory power to distinguish direct from indirect pairs (From [85])

5.4 Systematic Evaluation

In addition to the manual analysis of high-scoring GGM regions in the previous part, we now analyze whether this finding represents a systematic signal throughout the entire dataset. In order to assess how GGM edges and pathway proximity between our lipid metabolites are related, we generated a literature-based model of fatty acid biosynthesis and β-oxidation. This model includes reactions from the public databases BiGG (H. sapiens Recon 1) [101], the Edinburgh Human Metabolic Network [102] and KEGG PATHWAY [100]. We then mapped the partial correlation coefficients from the KORA data set onto the minimal number of reaction steps between each pair of metabolites (*pathway distance*).

For all five metabolite classes we observe a strong tendency towards significantly positive partial correlations for a pathway distance of one, i.e. directly connected metabolite pairs (Fig. 17.13a). For instance, for the lyso-PC class (Fig. 17.13a) nearly all partial correlation coefficients for a pathway distance of one are above significance level, whereas most values for a distance of two or larger remain insignificant. Some outliers from this observation, however, require closer inspection: We find 91 of 932 (~9.8%) unconnected metabolite pairs (pathway distance = ∞) with a partial correlation above significance level. These pairs represent potentially novel pathway predictions, missing interactions in the model or effects upstream of the metabolic network like enzyme co-regulation.

A direct comparison of both partial and Pearson correlation coefficients for the diacyl-phosphatidylcholine class is shown in Fig. 17.13b. As described earlier in this chapter, we observe a general over-abundance of significant Pearson correlations independent of the actual pathway distance. Even for the metabolites without a known pathway connection, 1,394 of a total of 1,569 Pearson correlations are significant (88.85%, over all classes), in contrast to 131 out of 1,569 for the partial correlations (8.35%).

The significantly different correlation value distributions between directly and indirectly linked metabolites (Fig. 17.13a, b) barely provide a good quantification of the actual discrimination accuracy of this feature. Therefore, we assessed the discriminative power of partial correlations to tell apart direct from indirect interactions by means of *sensitivity* and *specificity*. The sensitivity evaluates which fraction of directly connected metabolites in the pathway are recovered by significant GGM edges, whereas the specificity states how many of the significant edges actually represent a direct connection. A commonly used trade-off measure between sensitivity and specificity is the F_1 score, which is defined as the harmonic mean of both quantities [103]. Figure 17.13c lists sensitivity, specificity and F_1 for all five metabolite classes along with an evaluation of partial correlation distribution differences between directly and indirectly linked metabolites (determined by Wilcoxon's rank-sum test). F_1 values over 0.75 and significant p-values for the rank-sum test indicate a strong discrimination effect of partial correlation coefficients concerning direct vs. indirect pathway interactions.

6 Outro

In summary, we have shown how to proceed from purely structural studies of the stoichiometric matrix of a system to its quantitative description and estimation from metabolomics data. In particular, we have demonstrated the usefulness of Gaussian graphical models, which are based on partial correlation coefficients, for the unbiased reconstruction of metabolic pathway reactions from cross-sectional blood metabolomics data. Previous studies on blood plasma samples detected similar relationships with cellular processes based on genetic associations [104] or case/control drug trials [105]. The GGM result now demonstrates that metabolite profiles alone are sufficient to capture the dynamics of metabolic pathways. We suggest using GGMs as a standard tool of investigation in future metabolomics studies, utilizing the upcoming wealth of metabolic profiling data to form a more comprehensive picture of cellular metabolism.

We want to finish by noting that Gaussian graphical models closely follow the general concept of systems biology of investigating a biological system as a whole rather than investigating its single parts. Only by utilizing all measured metabolite concentrations at once we are able to specifically recover directly related metabolites in the underlying biochemical pathways.

References

1. Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
2. Gille C, Bölling C, Hoppe A et al (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol 6:411
3. Orth JD, Conrad TM, Na J et al (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. Mol Syst Biol 7:535
4. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121
5. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185
6. Karp PD, Paley SM, Krummenacker M et al (2009) Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11:40–79
7. Aziz RK, Bartels D, Best AA et al (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75
8. Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev 33:164–190
9. Gehlenborg N, O'Donoghue SI, Baliga NS et al (2010) Visualization of omics data for systems biology. Nat Methods 7:S56–S68
10. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40:D109–D114
11. Casp R, Altman T, Dale JM et al (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38:D473–D479
12. Overbeek R, Begley T, Butler RM et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691–5702
13. Schellenberger J (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11:213
14. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143
15. Pitkänen E, Rousu J, Ukkonen E (2010) Computational methods for metabolic reconstruction. Curr Opin Biotechnol 21:70–77
16. Francke C, Siezen RJ, Teusink B (2005) Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol 13:550–558
17. Covert MW, Schilling CH, Famili I et al (2001) Metabolic modeling of microbial strains in silico. Trends Biochem Sci 2:179–186
18. Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320
19. Karp PD, Caspi R (2011) A survey of metabolic databases emphasizing the MetaCyc family. Arch Toxicol 85:1015–1033
20. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641
21. Borodovsky M, Lomsadze A (2011) Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinform Chapter 4:Unit 4.6.1–4.6.10
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
23. Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98
24. The Universal Protein Resource (UniProt) (2009) Consortium, UniProt. Nucleic Acids Res 37: D169–D174
25. Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158:724–736

26. Apweiler R, Altwood TK, Bairoch A et al (2000) InterPro–an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16:1145–1150
27. Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31:6633–6639
28. Seffernick JL, de Souza ML, Sadowsky MJ, Wackett LP (2001) Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J Bacteriol 183:2405–2410
29. Palmer DR, Garrett JB, Sharma V et al (1999) Unexpected divergence of enzyme function and sequence: "N-acylamino acid racemase" is o-succinylbenzoate synthase. Biochemistry 38:4252–4258
30. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Res 39:D32–D37
31. Markowitz VM, Chen I-MA, Palaniappan K et al (2010) The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res 38:D382–D390
32. Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Res 33:1399–1409
33. Sun J, Zeng A-P (2004) IdentiCS–identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. BMC Bioinformatics 5:112
34. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305
35. Scheer M, Grote A, Chang A et al (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39:D670–D676
36. Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35:D274–D279
37. Fleischmann A, Darsiw M, Degtyarenko K et al (2004) IntEnz, the integrated relational enzyme database. Nucleic Acids Res 32:D434–D437
38. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21:3787–3793
39. Ma H, Zeng A-P (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19:270–277
40. Kümmel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics 7:512
41. Gardy JL, Liard MR, Chen F et al (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623
42. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786
43. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971
44. Liao L, Kim S, Tomb JF (2002) Genome comparisons based on profiles of metabolic pathways
45. Hong SH, Kim TY, Lee SY (2004) Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol 65:203–210
46. Kastenmüller G, Gasteiger J, Mewes HW (2008) An environmental perspective on large-scale genome clustering based on metabolic capabilities. Bioinformatics 24:i56–i62
47. Maltsev N, Glass E, Sulakhe D et al (2006) PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways. Nucleic Acids Res 34:D369–D372
48. Haft DH, Selengut JD, Brinkac LM, Zafar N, White O (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21:293–306
49. Kastenmüller G, Schenk ME, Gasteiger J, Mewes HW (2009) Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. Genome Biol 10:R28

50. Croes D, Couche F, Wodak SJ, van Helden J (2005) Metabolic pathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res 33:W326–W330
51. Faust K, Croes D, van Helden J (2009) Metabolic pathfinding using RPAIR annotation. J Mol Biol 388:390–414
52. Blum T, Kohlbacher O (2008) MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics 24:2108–2109
53. Arita M (2003) In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism. Genome Res 13:2455–2466
54. Rahman SA, Advani P, Schunk R, Schrader R, Schomburg D (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics 21:1189–1193
55. Blum T, Kohlbacher O (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. J Comput Biol 15:565–576
56. Pitkänen E, Jouhten P, Rousu J (2009) Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol 3:103
57. Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng 107:403–412
58. Kumar VS, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212
59. Schilling CH, Palsson BO (2000) Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. J Theor Biol 203:249–283
60. Henry CS, DeJongh M, Best AA et al (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28:977–982
61. Kumar VS, Maranas CD (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol 5:e1000308
62. Breitling R, Vitkup D, Barrett MP (2008) New surveyor tools for charting microbial metabolic maps. Nat Rev Microbiol 6:156–161
63. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge
64. Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BØ (2003) Metabolic pathways in the post-genome era. Trends Biochem Sci 28:250–258
65. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18:326–332
66. Schilling CH, Letscher D, Palsson BØ (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theor Biol 203:229–248
67. Llaneras F, Picó J (2010) Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. J Biomed Biotechnol 2010:753904
68. Price ND, Papin JA, Palsson BØ (2002) Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome Res 12:760–769
69. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99:15112–15117
70. Rojas I, Golebiewski M, Kania R et al (2007) Storing and annotating of kinetic data. In Silico Biol 7:S3–S44
71. Rizzi M, Baltes M, Theobald U, Reuss M (1997) In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: II. Mathematical model. Biotechnol Bioeng 55:592–608
72. Teusink B, Passarge J, Reijenga CA et al (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267:5313–5329
73. Blow N (2008) Metabolomics: biochemistry's new look. Nature 455:697–700

74. Famili I, Mahadevan R, Palsson BØ (2005) k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. Biophys J 88:1616–1625
75. Jamshidi N, Palsson BØ (2010) Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. Biophys J 98:175–185
76. Jamshidi N, Palsson BØ (2008) Top-down analysis of temporal hierarchy in biochemical reaction networks. PLoS Comput Biol 4:e1000177
77. Price ND, Schellenberger J, Palsson BØ (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophys J 87:2172–2186
78. Schellenberger J, Palsson BØ (2009) Use of randomized sampling for analysis of metabolic networks. J Biol Chem 284:5457–5461
79. Bakker BM, van Eunen K, Jeneson JA et al (2010) Systems biology from micro-organisms to human metabolic diseases: the role of detailed kinetic models. Biochem Soc Trans 38:1294–1301
80. Arkin A, Shen P, Ross J (1997) A test case of correlation metric construction of a reaction pathway from measurements. Science 277:1275–1279
81. Vance W, Arkin A, Ross J (2002) Determination of causal connectivities of species in reaction networks. Proc Natl Acad Sci USA 99:5816–5821
82. Steuer R, Kurths J, Fiehn O, Weckwerth W (2003) Observing and interpreting correlations in metabolomic networks. Bioinformatics 19:1019–1026
83. Øksendal B (2005) Stochastic differential equations: an introduction with applications. Springer, New York
84. Camacho D, de la Fuente A, Mendes P (2005) The origin of correlations in metabolomics data. Metabolomics 1:53–63
85. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 5:21
86. Schäfer J, Strimmer K, Jos' FF et al (2005) Learning large-scale graphical Gaussian models from genomic data. AIP Conf Proc 776:263–276
87. Lee JM, Gianchandani EP, Eddy JA, Papin JA (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Comput Biol 4:e1000086
88. de la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20:3565–3574
89. Magwene PM, Kim J (2004) Estimating genomic coexpression networks using first-order conditional independence. Genome Biol 5:R100
90. Wille A, Zimmerman P, Vranová E et al (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. Genome Biol 5:R92
91. Freudenberg J, Wang M, Yang Y, Li W (2009) Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. BMC Bioinformatics 10(Suppl 1):S66
92. Keurentjes Joost JB, Fu J, Ric de Vos CH et al (2006) The genetics of plant metabolism. Nat Genet 38:842–849
93. Liebermeister W, Klipp E (2006) Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. Theor Biol Med Model 3:42
94. Berg JM, Tymoczko JL, Stryer L (2006) Biochemistry, 6th edn. W. H. Freeman, Cranbury
95. Holle R, Happich M, Löwel H, Wichmann HE, MONICA/KORA Study Group (2005) KORA–a research platform for population based health research. Gesundheitswesen 67 (Suppl 1):S19–S25
96. Illig T, Gieger C, Zhai G et al (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42:137–141
97. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
98. Matsuzaka T, Shimano H, Yahagi N et al (2007) Crucial role of a long-chain fatty acid elongase, Elovl6, in obesity-induced insulin resistance. Nat Med 13:1193–1202
99. Eaton S, Bartlett K, Pourfarzam M (1996) Mammalian mitochondrial beta-oxidation. Biochem J 320:345–357

100. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
101. Duarte NC, Becker SA, Jamshidi N et al (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104:1777–1782
102. Ma H, Sorokin A, Mazein A et al (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 3:135
103. Van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworth, London
104. Suhre K, Petersen AK, Mohney RP et al (2011) Human metabolic individuality in biomedical and pharmaceutical research. Nature 477:54–60
105. Altmaier E, Ramsay SL, Graber A et al (2008) Bioinformatics analysis of targeted metabolomics–uncovering old and new tales of diabetic mice under medication. Endocrinology 149:3478–3489

Index

K. Suhre (ed.), *Genetics Meets Metabolomics: from Experiment to Systems Biology*,
DOI 10.1007/978-1-4614-1689-0, © Springer Science+Business Media, LLC 2012