TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Embedding metabolism into the omics landscape: Integrated analysis of metabolomics, transcriptomics and proteomics data from cellular to organ level

Jörg Bartel

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. M. Hrabě de Angelis

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. F. J. Theis
2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 28.12.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 27.04.2016 angenommen.

In loving memory of Gerta and Heinz.

Danksagung

An dieser Stelle möchte ich mich bei einigen Personen bedanken ohne deren Unterstützung diese Arbeit nicht möglich gewesen wäre.

Als erstes möchte ich mich bei dir, Jan, bedanken. Danke dass du mich in deinem Team aufgenommen hast, für die großartige Betreuung während meiner gesamten Doktorandenzeit, für die unzähligen fachlichen und auch nebensächlichen Diskussionen und natürlich ganz besonders dafür, dass du mir auch als Freund immer mit Rat und Tat zur Seite gestanden bist.

Mein besonderer Dank gilt natürlich auch dir Fabian. Danke, dass du es mir ermöglicht hast zunächst Teil deiner Gruppe und dann deines Instituts zu sein. Danke für deine immerwährende Fairness, Förderung und uneingeschränkte Unterstützung.

Auch bedanken möchte ich mich bei Professor Dmitrij Frishman für die Zweitbegutachtung meiner Doktorarbeit, sowie bei Professor Martin Hrabě de Angelis für den Vorsitz der Prüfungskommission.

Vielen Dank Susanne, dass du Teil meines Thesis Komitees warst und für die tolle Zusammenarbeit an einigen spannenden Projekten. Auch bei Tom, Önder, Maxie, Martin, Simone W. und den restlichen Kooperationspartnern möchte ich mich für die gute Zusammenarbeit bedanken.

Danke auch an das gesamte Systems Metabolomics Team für die tolle Zusammenarbeit, insbesondere gilt dabei mein Dank Alida, fürs Korrekturlesen meiner Doktorarbeit, sowie Michl und Ferdi für die spannenden Tischtennis Matches! Schön das ich immer gewinnen konnte!

Danke auch an das restliche ICB für die tolle Arbeitsatmosphäre und die vielen gemeinsamen Ausflüge und Aktivitäten. Besonders möchte ich mich noch bei Felix und Carsten bedanken, die es über viele Jahre mit mir in einem Büro ausgehalten haben. Danke auch an Steffen, Ivan, Nikola, Flo, und Michi Str. für unzählige Diskussionen in der Kaffeeküche, die vielen Tischtennis Matches und sonstige Aktivitäten. Durch euch hat die Arbeit immer Spaß gemacht!

Vielen Dank auch an meine restlichen Freunde dafür, dass ihr immer für mich da wart.

Ganz besonders möchte ich mich bei meiner gesamten Familie bedanken. Danke an meine Schwester Sabrina und an meine Eltern, Karin und Jörg, für die grenzenlose Liebe und Unterstützung. Ohne eure unermüdliche Geduld, eure Hilfe und euer Vertrauen schon in frühen Jahren wäre diese Arbeit niemals möglich gewesen. Ein besonderer Dank gilt auch meinen Großeltern, Gerta und Heinz. Ich wünschte ihr hättet die Fertigstellung meiner Doktorarbeit noch miterleben können. Vielen Dank für eure Liebe und Unterstützung!

Abschließend möchte ich mich noch bei dir, Sim, bedanken. Danke nicht nur dafür, dass du mein bester Freund bist und mich fortwährend zu einem besseren Menschen machst, sondern auch ganz besonders dafür, dass du auch in schweren Zeiten niemals aufgibst!

Abstract

Higher order biological systems are extremely complex, at a cellular level built up from different types of macromolecules such as DNA, RNA or proteins, but also metabolites, which are organized into multiple molecular layers. Cells themselves are hierarchically structured into different biological scales, ranging from tissues, over organs to ultimately whole organisms. Modern high-throughput 'omics' technologies nowadays provide the tools to systematically analyze the relationships between the components of biological systems, which however remains a challenging task.

Within the omics landscape, metabolomics display the endpoints of upstream biological processes, and the metabolic profile of an organism, integrating genetic as well as environmental variation, is commonly seen as most closest link to the observable phenotype. Thus the embedding of metabolism into the omics landscape, i.e. investigating the relationships between the metabolome and other molecular levels, is of particular interest to gain a better understanding of both health and disease.

In this thesis, our main focus was on the integration of metabolomics data with varying other omics measurements derived from different biological scales. At a cellular level, we integrated metabolomics data measured on two different platforms with proteomics data to investigate the effect of an environmental pollutant on the metabolism of T cells. To this end, we applied a random walks approach on a genome-scale metabolic model and identified condition-specific metabolic sub-networks most critical to the pollutant-induced changes at both levels.

At a tissue level, we integrated metabolomic, transcriptomic and genetic data derived from a cross-sectional human population study to elucidate general signaling, regulatory and metabolic processes observable in blood. First, we constructed a global correlation network between metabolites and transcripts and evaluated the associations both manually and via systematic comparison with a metabolic pathway model. We thereby revealed systematic signatures of transport and metabolic processes mainly belonging to lipid, energy and amino acid metabolism. Moreover, using Mendelian randomization, we investigated whether the observed associations also represent causal effects. Next, we developed a novel bi-directional enrichment approach using functional annotation-based aggregation, which allowed us to functionally interpret associations between metabolites with missing pathway evidence and non enzymatic transcripts. Using transcription factor binding site enrichment analysis, we found evidences of shared regulatory signatures

between transcripts associated to the same metabolites, suggesting that correlations between transcripts and metabolites not only reflect actual metabolic pathway reactions, but are also of regulatory nature. Finally, we demonstrated how the constructed networks can be used to gain novel insights into molecular mechanisms associated to intermediate clinical traits.

At a multi-organ level, we integrated metabolomics and transcriptomics data from various metabolically active organs with metabolomics data from plasma to investigate cross-organ associations and to identify plasma markers for organ-specific processes. First, a systematic comparison of the organ and plasma metabolomes revealed that metabolic profiles are representative of their source organ and a large fraction of the measured metabolites are only detectable in specific tissues. Calculation of pairwise correlation networks between each organ and plasma both at a pathway and single molecule level showed that metabolites and metabolic processes in kidney are most strongly reflected by plasma metabolites, followed by muscle, liver and adipose tissue. Based on the networks, we inferred different biologically meaningful association categories, for instance between identical metabolites quantified in plasma and organs which indicates an active exchange and a common organism-wide source for these metabolites. In addition, we demonstrated that plasma metabolites also carry regulatory signals of organs. In a final step, we investigated diabetes-related changes in all organs and identified many equally altered organ-plasma association pairs demonstrating the potential of plasma metabolites as proxy markers for organ processes.

In conclusion, we demonstrated how metabolomics data can be integrated with varying omics combinations in order to exploit their complementary information content and reveal an integrated global picture of the wiring between the metabolic, proteomic and transcriptomic molecular levels.

Zusammenfassung

Biologische Systeme höherer Ordnung sind hoch komplex. Auf zellulärer Ebene sind sie aus verschiedenen molekularen Schichten aufgebaut, die aus Makromolekülen wie DNS, RNS oder Proteinen, aber auch Metaboliten bestehen. Zellen wiederum sind in unterschiedliche, hierarchisch aufgebaute biologische Skalen eingeteilt, die sich von Geweben über Organe bis hin zu ganzen Organismen erstrecken. Moderne hoch-durchsatz Omics Technologien stellen heutzutage die Werkzeuge zur Verfügung um die Beziehungen zwischen diesen Bestandteilen biologischer Systeme systematisch zu analysieren. Dies stellt jedoch immer noch eine große Herausforderung dar.

Innerhalb der "Omics-Landschaft" stellen Metabolomics Messungen die Endpunkte vorangegangener biologischer Prozesse dar und das metabolische Profil eines Organismus, welches sowohl genetische als auch Umwelt bedingte Variation wiederspiegelt, wird für gewöhnlich als engste Verbindung zum beobachtbaren Phänotyp angesehen. Folglich ist die Eingliederung des Metabolismus in die "Omics-Landschaft", d.h. die Erforschung der Verbindungen zwischen dem Metabolom und anderen molekularen Ebenen, von besonderem Interesse um ein besseres Verständnis für biologische Prozesse im gesunden sowie im Krankheitszustand zu erlangen.

In dieser Arbeit wurde der Fokus auf die Integration von Metabolomicsdaten mit variierenden anderen "Omics" Messungen auf unterschiedlichen biologischen Skalen gelegt. Auf zellulärer Ebene kombinierten wir Metabolomicsdaten von zwei unterschiedlichen Messplattformen mit Proteomicsdaten um den Effekt eines Umweltschadstoffs auf den Metabolismus von T-Zellen zu erforschen. Zu diesem Zweck wurde eine "Random Walk" basierte Methode auf ein metabolisches Netzwerkmodell angewandt und zustandsspezifische, metabolische Teilnetzwerke identifiziert, welche von maßgeblicher Bedeutung für die durch den Schadstoff hervorgerufenen Veränderungen auf beiden molekularen Ebenen sind.

Auf Gewebe Ebene wurden Metabolomics, Transcriptomics und genetische Daten einer populationsbasierten Querschnittsstudie vereint, um aufzuklären, welche Signal-, regulatorischen und metabolischen Prozesse im Blut generell beobachtbar sind. Im ersten Schritt wurde ein globales Korrelationsnetzwerk zwischen Metaboliten und Transkripten erstellt und diese Verbindungen anschließend durch einen systematischen Vergleich mit einem Stoffwechselprozessmodell evaluiert. Dabei konnten wir systematische Signaturen von Transport und Stoffwechselprozessen ("Pathways") aufzeigen, welche hauptsächlich

dem Fett-, Energie- und Aminosäurestoffwechsel zugeordnet werden konnten. Außerdem wurde "Mendelian Randomization" angewandt um festzustellen ob die gefundenen Abhängigkeiten zwischen Metaboliten und Transkripten kausalen Ursprungs sind. Anschließend entwickelten wir eine neue zweidirektionale Enrichment-Methode, basierend auf der Gruppierung funktioneller Ähnlichkeiten. Diese Methode ermöglichte es uns Abhängigkeiten zwischen Metaboliten mit unbekannter Stoffwechselzugehörigkeit und nicht Enzym-kodierenden Transkripten funktional zu interpretieren. Durch eine Transkriptionsfaktor Bindestellen Analyse konnten außerdem Hinweise für eine gemeinsame Regulation zwischen, mit denselben Metaboliten verbundenen Transkripten, gefunden werden. Diese Beobachtung suggeriert, dass Korrelationen zwischen Metaboliten und Transkripten nicht nur metabolische Reaktionen reflektieren, sondern auch regulatorischer Natur sind. In einem letzten Schritt demonstrierten wir, wie die generierten Netzwerke dazu verwendet werden können um neue Einblicke in die mit klinisch relevanten Messparametern verknüpften molekularen Prozesse zu erlangen.

Auf multi-organer Ebene kombinierten wir Metabolomics- und Transcriptomicsdaten von mehreren metabolisch aktiven Organen mit in Blutplasma gemessenen Metabolomicsdaten, um die Verbindungen zwischen Organen zu untersuchen und um Plasmamarker für organspezifische Prozesse zu identifizieren. Zunächst konnten wir durch einen systematischen Vergleich der Organ Metabolome mit dem Plasma Metabolom aufzeigen, dass die Metabolitenprofile charakteristisch für ihr Ursprungsorgan sind und dass ein großer Teil der gemessenen Metaboliten nur in bestimmten Organen nachweisbar sind. Anschließend wurden paarweise Korrelationsnetzwerke sowohl auf "Pathway" als auch auf Einzelmolekül Ebene zwischen jedem Organ und Plasma berechnet. Durch diese Netzwerke konnte gezeigt werden, dass die Konzentrationen sowohl einzelner Metaboliten als auch ganzer metabolischer Prozesse aus den Nieren am stärksten in den Konzentration von Plasmametaboliten reflektiert werden, gefolgt von Muskel, Leber und Fettgewebe. Weiterhin konnten wir verschiedene, biologisch bedeutsame Assoziationskategorien von den Netzwerken ableiten, wie z.B. zwischen identischen Metaboliten welche sowohl im Plasma, als auch in den Organen gemessen werden konnten. Diese Verbindungen deuten auf einen aktiven Austausch dieser Metaboliten zwischen Plasma und den Organen und auf eine gemeinsame, organismusweite Ursprungsquelle hin. Des Weiteren konnten wir zeigen, dass Plasma Metabolitenkonzentrationen auch Signale regulatorischer Prozesse einzelner Organe in sich tragen. Im letzten Schritt wurden Diabetes bedingte Veränderungen in allen gemessenen Organen untersucht. Dabei konnten viele gleichartig veränderte Paare von Organ-Plasma Verbindungen gefunden werden.

Dies belegt das Potenzial von Plasmametaboliten als stellvertretende Marker für Organprozesse.

Abschließend lässt sich sagen, dass wir zeigen konnten wie Metabolomicsdaten mit variierenden anderen "Omics" Daten kombiniert werden können, um deren komplementären Informationsgehalt auszunutzen. Daraus kann ein globales, ganzheitliches Abbild der Verbindungen zwischen der metabolischen, Protein- und Transkriptebene gewonnen werden.

Scientific publications

The results presented in this thesis are mainly based on previous publications in peer-reviewed journals or contributions that are currently within the publication process. These publications are listed below ordered by the chapters in which they appear:

Chapter 1

⋆ **Bartel, J.**, Krumsiek, J., and Theis, F.J. Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal*, 4(5):e201301009, 2013

Chapter 2

⋆ Baumann, S.*, Rockstroh, M.*, **Bartel, J.**, Krumsiek, J., Otto, W., Jungnickel, H., Potratz, S., Luch, A., Wilscher, E., Theis, F.J., von Bergen, M., and Tomm, J.M. Sub-toxic concentrations of benzo[a]pyrene induce metabolic changes and oxidative stress in non-activated and affect the mTOR pathway in activated Jurkat T cells. *Journal of Integrated OMICS*, 4(1), 2014.

Chapters 1 & 3

⋆ **Bartel, J.***, Krumsiek, J.*, Schramm, K., Adamski, J., Gieger, C., Herder, C., Carstensen, M., Peters, A., Rathmann, W., Roden, M., Strauch, K., Suhre, K., Kastenmüller, G., Prokisch, H., and Theis, F.J. The human blood metabolome-transcriptome interface. *PLoS Genetics*, 11(6): e1005274, 2015

Chapter 4

⋆ **Bartel, J.**, Neschen, S., Fridrich, B., Scheerer, M., Zukunft, S., Irmler, M., Kastenmüller, G., de Angelis, M.H., Adamski, J., Beckers, J., Theis, F.J., and Krumsiek, J. Plasma metabolites as proxy markers for inter-organ processes in diabetic mice. *in preparation.*

* = equal contributions

Further publications

Besides the contributions discussed in this thesis, I was involved in further projects and collaborations during the course of my PhD student time, which lead to the following publications:

⋆ Wahl, S.*, Vogt, S.*, Stückler, F., Krumsiek, J., **Bartel, J.**, Kacprowski, T., Schramm, K., Carstensen, M., Rathmann, W., Roden,M., Jourdan, C., Kangas, A.J., Soininen, P., Ala-Korpela, M., Nöthlings, U., Boeing, H., Theis, F.J., Meisinger, C., Waldenberger, M., Suhre, K., Homuth, G., Gieger, C., Kastenmüller, G., Illig, T., Linseisen, J., Peters, A., Prokisch, H., Herder, C., Thorand, B., and Grallert, H. Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Medicine*, 2015.

⋆ Conlon, T.M., **Bartel, J.**, Ballweg, K., Prehn, C., Krumsiek, J., Meiners, S., Theis, F.J., Adamski, J., Eickelberg, O., and Yildirim, A.Ö. Metabolomics screening identifies reduced L-carnitine to be associated with progressive murine emphysema. *Clinical Science*, 2016, 130 (4) 273-287.

⋆ Krumsiek, J.*, **Bartel, J.***, and Theis, F.J. Computational approaches for systems metabolomics *Current Opinion in Biotechnology*, Volume 39, June 2016, Pages 198206.

Contents

Chapter 1

Introduction

It was in the early sixties when Watson and Crick discovered the molecular structure of DNA, contemporary laying the foundation of modern molecular biology. Based on the identified double helix structure of DNA, they speculated that the precise sequential order of nucleotides encodes the genetic information [1, 2]. Later on, Crick focused his research on the genetic coding problem and, in particular, on the newly emerged role for RNA as intermediate molecule in the flow of information from DNA to proteins. The summary of this idea became famous as the 'central molecular dogma', which essentially describes the interactions between the various molecular levels of a cell [3]. In its original form, the dogma stated that genetic information stored in the nucleotide sequence of DNA is carried into the cytoplasm by a 'messenger' RNA molecule, on the basis of which amino acids are assembled into proteins by ribonucleic protein complexes (Figure 1.1). Several further milestones in molecular biology, such as the discovery of genetic regulation by proteins [4], enzymatic catalysis of metabolic reactions [5], epigenetic- [6, 7] and post-transcriptional regulation [8, 9] further shaped and complemented the central dogma to its recent complex form. This biological complexity becomes even more emphasized given that thousands of genes, proteins or metabolites are co-regulated, physically interacting, or functionally coordinated at - and between - each molecular level (Figure 1.1). On top of that, higher order organisms are further organized into structural levels (biological scales): **Cells** sharing a similar function are organized into **tissues**. At the next level, **multiple tissues** are functionally aligned to constitute distinct **organs** that perform a specific task. Finally, two or more **organs** interact with each other forming **organ systems**, that again perform complex tasks in a concerted fashion to maintain life, ultimately determining what is known as a phenotype (Figure

1.1). Recognizing that biological systems consist of many layers of complexity was a crucial point. From then on it increasingly became apparent that a full understanding of cellular processes, the biological system as a whole, or the etiology of a certain phenotype can never be achieved from the sum of its parts, i.e. by analyzing single molecules or a single molecular level at a time [10]. Instead, a global analysis is required that ideally integrates data across multiple biological scales simultaneously to uncover the complete interaction landscape within and between the different functional levels of a cell or a biological system. This insight and the technical advancement of high-throughput measurement methods, so called *omics*, was the starting shot for a new field of science called systems biology [11]. The main focus of this thesis will be on the integration of multiple large-scale 'omics' measurements, more precisely, the integration of metabolomics with varying other omics data at different biological scales - ranging from *in vitro* experiments at a cellular level (**Chapter 2**), over data from a cross-sectional population study at a tissue level (**Chapter 3**) to data from a multi-organ study (**Chapter 4**). Thereby, our goal will be to achieve a better understanding of the relationships between metabolism and the other functional levels. The following sections will provide an overview of the different molecular levels and corresponding 'omics' data sets relevant to this thesis including the corresponding state-of-the-art analysis methods, recent progress in data integration techniques, including successful example studies from the field of biomedical research, as well as a short introduction to blood as central connective tissue. At the end of this Chapter, we will formulate the specific research questions which we tried to solve, respectively, for each project.

1.1 The omics era

After the discovery of the DNA structure and the formulation of the molecular dogma, there was a rapid advancement in the development of experimental measurement technologies. Nowadays it is possible to quantitatively assess hundreds to thousands of cellular components from multiple functional levels of a biological system simultaneously, including DNA sequence variations, methylation patterns or other epigenetic markers, expression levels of transcripts, proteins and metabolites. Such high-throughput technologies are commonly referred to as 'omics', with the aim to quantify as many of the molecules present in a biological sample as possible at a time. Ideally, the system would be assessed in its entirety, delivering a precise snapshot of the underlying cellular state. All these snapshots can be interpreted as intermediate phenotypes and single molecules

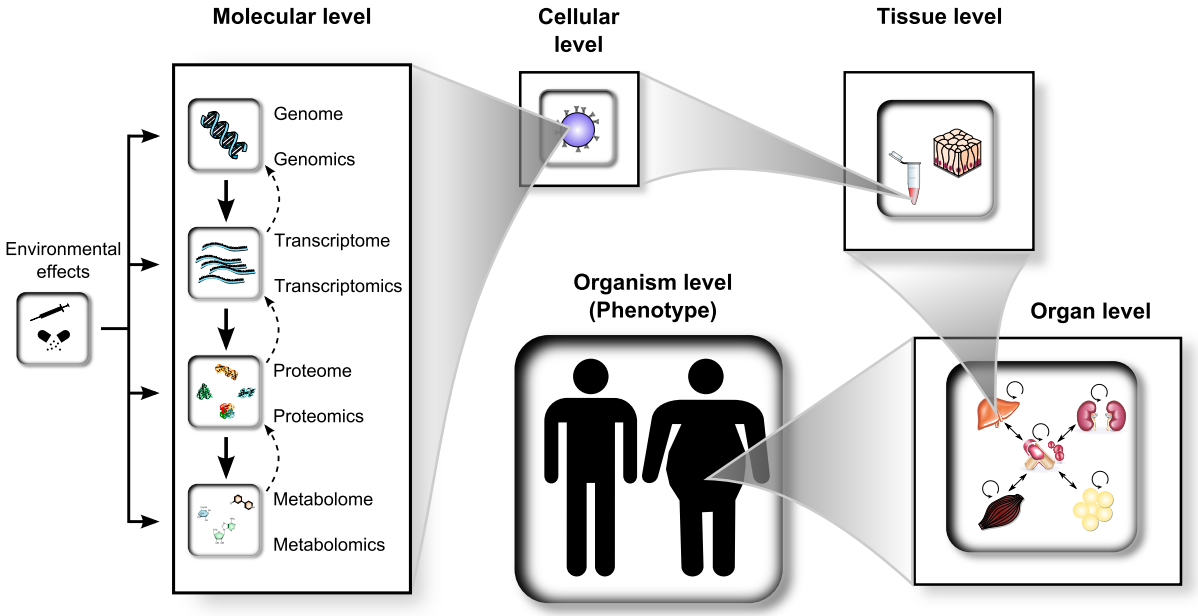


Figure 1.1: Simplified representation of the complexity of biological systems. At a **molecular level**, genes encoded in the DNA are transcribed into messenger RNAs (mRNAs), which are then translated into proteins. Proteins, in turn, are the functional units of a cell, fulfilling many distinct roles in cellular processes, e.g. as metabolic enzymes acting on metabolites (black arrows). The interplay between all these molecules (dashed arrows indicate possible feedback regulation) and compounds organized in different functional layers including environmental influences is determining the **cellular** phenotype. Each type of cell has its own functional or structural role and collections of cells with the same function are organized into **tissues**. Different tissues, in turn, can be organized together to form **organs**, which again have different functional roles in the body. On the next organizational level, organs with the same functionality are grouped into **organ systems** (indicated as interactions between organs). For instance, the digestive system includes organs that are involved in the break down and consumption of food. Lastly, all these distinct organizational levels together determine the observable phenotype on the **organism level**.

can serve as biomarkers to assess different cellular states. Along with the development of 'omics' technologies and the comprehensive large-scale data sets which can now be generated,the necessity arose to develop computer-aided methodologies for the analysis of these data sets. Systems biology provides us with the tools to describe and model the system as a whole, with a focus on understanding how components from various biological scales interact with each other - ranging from sub-cellular, cellular, tissue and organ up to the whole organism level - i.e. how changes in one part of the system affect the other parts [11].

In the following, we will briefly introduce the 'omics' technologies relevant for this thesis.

1.1.1 Genomics

The term *genomics* summarizes all kinds of technologies and methods dealing with systematic sequencing, assembly, functional and structural analysis of a genome (i.e. the entire DNA sequence of a cell) [12]. Soon after the publication of the DNA structure by Watson and Crick [1], early molecular biologists started to develop methods for DNA sequencing. For instance, Sanger et al. [13] developed the "chain-termination" method, which made use of DNA polymerase and radioactively labeled nucleotides, forming the basis for all follow up DNA sequencing techniques. Based on this technology, the first organism *Haemophilus influenzae* was fully sequenced in 1995 [14]. Nowadays, modern next-generation sequencing technologies allowed the full genome sequencing of more than 60 multicellular eukaryotic organisms [15]. The availability of complete genome sequences together with the sequencing technologies further provided important tools in biomedical research, for example to identify genetic variations between individuals. The simplest and most common type of DNA sequence variants are *single nucleotide polymorphisms* (SNPs), in which one single nucleotide differs between individuals of the same biological species. Such polymorphisms can be located anywhere in the DNA sequence, including the coding or regulatory region of a gene, thereby potentially affecting its function or expression. These variations in DNA sequence can be determined using SNP arrays such as the Affymetrix GeneChip array 6.0 that is capable of measuring more than 900,000 SNPs at once[1], or more recent next-generation sequencing technologies. Genotype data is often combined with phenotypic traits or other 'omics' data with applications ranging from pharmacokinetics [16], Mendelian randomization [17] or most commonly in *genome-wide association studies* (GWAS) [18–20]. If not stated otherwise, when using the term genomics in this work, we refer to genetic variation by the means of SNPs (see **Chapter 3**).

1.1.2 Transcriptomics

Transcriptomics refers to the genome-wide quantitative assessment of RNA molecules present in a cell or tissue at a certain time or condition. The transcriptome consists of mRNAs, non-coding, and regulatory RNAs which vary between different tissues, cell-

[1] `http://www.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf`

types and environmental conditions. When dealing with mRNA, the assessment of this variation is called *gene expression profiling*, i.e. globally measuring the activity (in terms of expression/abundance) of all genes at a specific condition in the respective cell or tissue. Several technologies have been developed for transcriptional profiling, for instance sequence-based methods like *serial analysis of gene expression* (SAGE), hybridization-based methods such as the most commonly applied DNA microarrays and more recently next-generation technologies like RNA-seq. The microarray technology makes use of the DNA double-helix structure, i.e. the hybridization of a fluorescent labeled complementary DNA strand to a probe sequence by forming hydrogen bonds between base pairs [21]. Basically, there are two main manufacturing techniques for microarrays: *in situ* synthesis (e.g. via photo-lithography; typical chip: *Affymetrix GeneChip*$^{\text{TM}}$) and direct spotting of DNA fragments onto the chip surface (e.g. PCR amplified cDNA clones; typical chip: *Illumina BeadArray*$^{\text{TM}}$). In comparison to RNA-seq, microarray technologies are relatively inexpensive and straight-forward to analyze [22]. However, due to technical issues such as amplification inequalities, labeling efficiency or cross-hybridization, microarrays are prone to measurement noise which has to be borne in mind. Moreover, microarrays are limited to a pre-defined set of transcripts, as opposed to RNA-seq, which allows the quantification of the complete transcriptome down to a single base resolution [23]. DNA microarrays are typically used to assess differentially expressed genes, i.e. to discover biomarkers between sets of samples belonging to different treatment groups or conditions. For example, the assessment of differentially expressed genes between healthy individuals and diabetes type II (T2D) patients [24], or the classification of breast cancer patients into different tumor subtypes based on the molecular signatures of selected genes [25]. A typical way to asses these changes are univariate statistical tests, such as *Student's t-test* or the analysis of variance (ANOVA), which provide a probability value (p-value) for a tested null hypotheses. Here, the null hypothesis is, that the expression levels of two genes between the two investigated groups are identical. If the observed p-value is below an *a priori* chosen significance level α (standard values are 1% or 5%), the null hypothesis can be rejected and the gene can be considered as significantly changed between groups. However, a problem that occurs when statistical tests are applied many times on the same data set is the increasing probability of false positive test results. To account for this, *multiple testing* correction procedures can be applied, which adjust the resulting p-values for the number of performed tests. The most common approaches for multiple testing correction are the *Bonferroni correction*, where the chosen significance threshold for the hypotheses (i.e. the α value) is divided by the number of performed tests, or the less stringent *false-discovery rate (FDR)* [26],

which will be reviewed in more detail in Section 3.2. After the identification of significantly changed genes, the second question that is frequently attempted to be solved using transcriptomics data is how genes or samples relate to each other, for instance functionally using *gene set enrichment analysis* (GSEA, [27]) or the identification of sample/gene groups using cluster analysis [28]. Classic gene set enrichment approaches use predefined sets of genes as input to a statistical test, such as Fisher's Exact test, in order to find significantly overrepresented processes or pathways among the list of differentially expressed genes when compared to a background set [29]. Later, an improved version was proposed, which uses the whole data distribution for the enrichment analysis, instead of only using the significant genes [27]. In this thesis, more details on the use of transcriptomics data measured on Illumina BeadArrays^{TM} will be given in **Chapter 3** and **Chapter 4**.

1.1.3 Proteomics

The term *proteome* refers to the complete set of proteins produced or modified by an organism or cellular system, which was first coined by Wilkins et al. [30]. Proteins are macromolecules consisting of a sequence of amino acids encoded in the genetic code of the respective gene. Proteins can be seen as functional units of a cell participating in almost all cellular processes. For instance, changes in the cellular metabolic state or external stimuli might trigger a cellular response in order to adapt to the new conditions, which typically comprises the regulation of abundance and activity of the respective proteins. This makes the proteome highly variable, not only differing between certain cell types, but also within a cell type between different developmental stages (time-points) or under varying conditions. Recapitulating the biochemical information flow formalized in the central molecular dogma (Figure 1.1), assessing the proteome can be seen as an intermediate 'snapshot' of cellular activity. The term *proteomics* thus describes the quantitative assessment of a specific proteome in terms of abundance, structure, interactions and modifications. Protein quantification can be roughly divided into antibody based (immunoassays) or mass spectrometry (MS) based measurement techniques. Methods using antibodies for protein detection, such as enzyme-linked immunosorbent assay (ELISA) are commonly used but also face some disadvantages, e.g. the identification and quantification is limited to proteins with known antibodies. More recent technological advancements combined 2D gel electrophoresis with MS, which made it possible to measure protein expression at a systems level in a high-throughput fashion [31]. A typical proteomics work-flow nowadays consists of several steps including sample

preparation, protein digestion and extraction according to the specific needs of the applied measurement technology, separation of proteins by 2D gel electrophoresis or liquid chromatography and finally protein identification using MS-based technologies. However, technical challenges still exist, for instance the assessment of post-translational modifications, the high number of redundant proteolytic peptides, and the relatively low coverage in protein identification [31]. Based on the research question, Klein and Thongboonkerd [32] defined three basic categories for proteomic analyses: *expression proteomics* or *quantitative proteomics*, which aims at identifying and subsequently quantifying the proteins present in a specific cell or tissue; *bioinformatic analyses* dealing with protein structure, motifs and domains, homology, but also protein-protein interactions and networks; *functional proteomics* trying to elucidate the functional roles of the proteins in cellular processes. Similar to other omics based analyses, proteomics are often used in medical studies such as biomarker discovery or diagnostics in cancer [33] and diabetes mellitus [34, 35]. Moreover, proteomics measurements have been integrated with other omics data sets such as transcriptomics or metabolomics data to get a better understanding of an organism's response on varying external conditions [36], or of infectious diseases [37], renal cell carcinoma [38] and microbial infections [39]. In this work, we utilize proteomics data measured with SDS-PAGE and subsequent LC-MS/MS with more details in **Chapter 2**.

1.1.4 Metabolomics

With the advent of *metabolomics*, a new and important milestone in the endeavor to fully measure a biological system has been reached. Metabolomics aims at measuring all endogenous metabolites within a biological system in an unbiased fashion (small molecules $< 2,000$ Da) [40]. The resulting metabolic profiles may be regarded as functional signatures of the physiological state (see Figure 1.1) and have been shown to integrate signatures of genetic regulation as well as environmental factors [41]. This potential to connect genotypic to phenotypic information promises new insights and biomarkers for different research fields, including biomedical and pharmaceutical research [42]. Similar to proteomics, the analytical techniques predominantly used for the quantification are mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy, both having different strengths and weaknesses [43–46]. There exist two main strategies for the quantification and identification of metabolites, the choice of which mainly depends on the experimental question to be answered. *Targeted metabolomics* is the method of choice in a hypothesis-driven experiment, i.e. if the research focus lies on one or more

particular metabolic pathways that are known to play a role in the examined biochemical setting. Only a predefined panel of metabolites - often belonging to a particular pathway of interest - is quantified, allowing for a precise snapshot of the desired physiological context. We utilize this approach to measure amino acids, sugars and fatty acids in Jurkat T cells (see **Chapter 2**). In contrast, *untargeted metabolomics* aims to measure as many metabolites contained in a biological sample as possible providing a global and unbiased picture of a systems metabolism. However, the chemical identification and functional characterization of many yet unidentified peaks measured in untargeted metabolomics approaches remains a substantial challenge [47]. More specific details on untargeted metabolomics measured in a heterogeneous population will be given in **Chapter 3**. Applications of metabolomics can be found in a huge diversity of research fields, including environmental perturbations of biological systems, toxicology, disease diagnostics and biomarker identification. The suitability of metabolites as molecular biomarkers was demonstrated in several recent publications. For instance in genome-wide association studies (GWAS), Suhre et al. [19, 48] and Gieger et al. [49] showed that changes in the concentration levels of biochemically related metabolite pairs are often highly correlated with genetic variation in the general population. Specifically, they report that a SNP in the proximity of the coding regions of genes is frequently associated with variations in the concentration levels of metabolites which the protein/enzyme processes or transports. Jain et al. [50] examined the concentration changes of metabolites from NCI-60 cancer cells along with gene expression data. They reported a strong correlation between glycine consumption, the expression of glycine biosynthetic pathway related genes and the proliferation rate of cancer cells. Further successful applications of metabolomics include nutritional challenge studies [51, 52] and the investigation of molecular cell mechanisms [53, 54]. We previously reviewed several statistical methods, suitable for the analysis of metabolomics data, which will also be shortly discussed in the following sections [55].

1.2 Organisms as complex systems and the necessity of omics data integration

Despite the availability of omics data from all functional levels and the above-mentioned complexity of biological systems (Figure 1.1), most traditional studies focused only on a single dimension to search for relationships with biological processes. However, single omics data sets are inevitably restricted to one molecular level and assessing the vari-

ation on only a single data dimension might lead to only a limited understanding of the underlying biological processes and mechanistic aspects of the system. For instance, processes that require regulation (and thus variation) across multiple functional layers might never be identified. In higher organisms, the additional compartmentalization of many of these processes into distinct organs adds another level of complexity, highlighting that a full understanding of these remains a challenging task. Consider for instance complex diseases, such as diabetes, where multiple organs are involved and dysfunctions in the crosstalk among them are often crucial for disease development [56–58]. In diabetes, chronic hyperglycemia, a persistent state of high blood sugar levels, is the primary characteristic of the disease [59]. The regulation of systemic glucose metabolism is a complex process involving multiple molecular levels in various organs. Dysfunctions can occur at several locations in this process which might lead to a hyperglycemic state (Figure 1.2). Insulin plays an essential role in the control of glucose levels and metabolism in the body, and is the major regulator of glucose uptake from blood by target organs such as liver, skeletal muscle and adipose tissue [58]. For instance, insulin regulates the glucose uptake of the liver via the Glucose transporter type 4 (GLUT4), which increases the expression levels of genes and the activity of enzymes involved in glycogen synthesis, while inhibiting genes involved in glycogenolysis or gluconeogenesis. Already in this small extract of the regulation of systemic glucose metabolism are numerous sources for irregularities that can lead to a hyperglycemic state. For example, it could be caused by a decreased insulin sensitivity of target organs, possibly due to rare variants in the insulin receptor (INSR) leading to decreased levels, or by other processes like the excessive release of adipokines from fat cells. It was shown that the expression of tumor-necrosis factor-α (TNF-α) is increased in adipose tissue of obese humans, leading to a phosphorylation of insulin receptor substrates (IRS) and subsequent insulin resistance [60]. Such complex mechanisms, in combination leading to the development of diabetes, can only be fully understood by an integrated analysis of measurements from multiple omics techniques, ideally derived from various organs. To this end, many established analysis methods and concepts from single omics studies have been adapted for the analysis of multiple omics data sets, but also novel analysis techniques were developed which we will shortly introduce in the following. An overview of some the most commonly applied methodologies for single omics analysis is given in table 1.1.

Method type		Model
Univariate		Student's t-test, Analysis of variance (ANOVA), Mann-Whitney U test
Multivariate	**unsupervised**	Principle component analysis (PCA) [61], Self-organizing maps (SOM) [62], Independent component analysis (ICA) [63], Hierarchical clustering analysis (HCA) [61]
	supervised	Partial least squares (PLS) [64], PLS discriminant anaylsis (PLS-DA) [65], orthogonal PLS (O-PLS) [61], Support vector machines (SVM) [66]
Pathway (Model)-based		Gene set enrichment analysis (GSEA) [27], Metabolite set enrichment analysis (MSEA)[67], Ingenuity pathway analysis (IPA), Metabolites Biological Role (MBRole)[68]
Correlation-based		Gaussian graphical model (GGM) [69], Weighted gene co-expression network analysis (WGCNA)[70]
Network module extraction		BioMet [71], Ingenuity Pathway analysis (IPA), SigArSearch [25], HotNet [72], Network smoothing [73]

Table 1.1: Methodological overview for single omics analysis. Shown are some of the most frequently used analysis techniques for single omics data. If appropriate, references to application examples are provided.

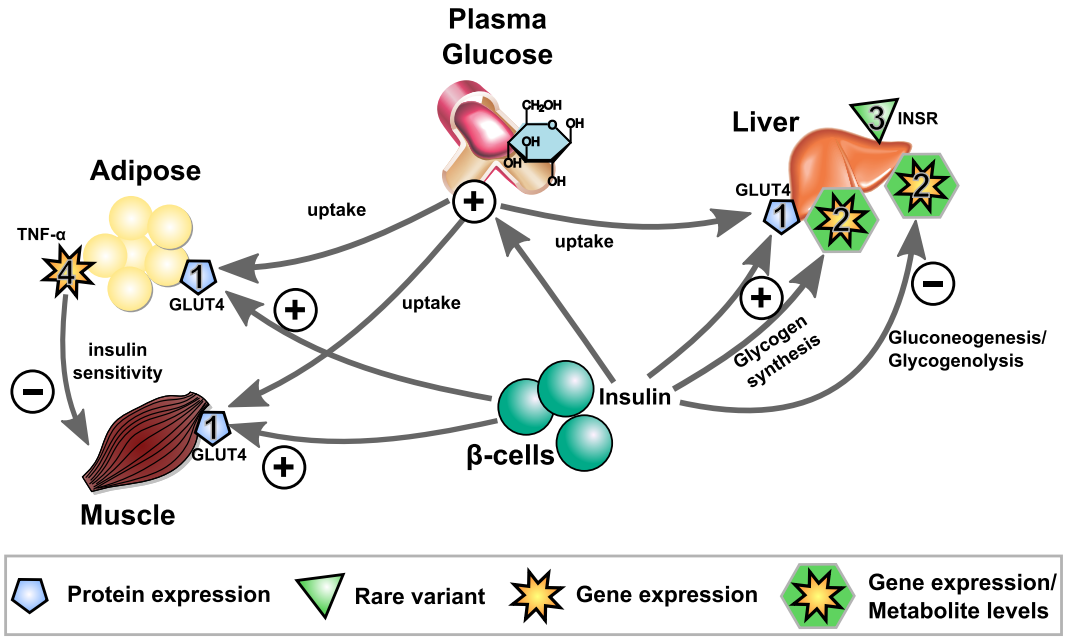


Figure 1.2: Example of variations on multiple omics levels in various organs leading to hyperglycemia. Insulin regulates the level of GLUT4, a glucose transporter responsible for the uptake of glucose from blood. **(1)** Changes in the protein level of GLUT4 can lead to an impaired uptake of glucose by target tissues. Insulin also stimulates the gene expression of genes involved in glycogen synthesis and inhibits the expression of genes in glycogenolysis and gluconeogenesis. **(2)** An impairment in hepatic glucose production and energy metabolism can lead to hyperglycemia. Moreover, a rare variant **(3)** in the insulin receptor gene INSR can cause severe forms of insulin resistance. Finally, the expression of the adipokine TNF-α **(4)** is increased in obese humans, resulting in a reduced insulin sensitivity of target organs. Symbols indicate different sources for irregularities.

1.3 Data integration using network biology

The field of systems biology generally aims to model a biological system by taking into account its systemic nature, ideally combining data from each level of biological information to fully capture the operating principle of the underlying biology [74]. Thus, data integration is an essential part of systems biology and a plethora of integrative methods have been developed, many of them specifically tailored to the requirements of the respective research question and all with their own strengths and weaknesses. For instance, a popular concept to model the information flow between different molecules or molecular levels is *network biology*, where nodes represent molecular entities (genes, proteins, metabolites, etc.) and edges represent direct or indirect interactions between them [75]. Some examples of such network abstractions include gene regulatory net-

works, where edges may reflect inhibitory or activating relationships between two nodes [76, 77], protein-protein interaction networks, where edges might represent direct physical or functional interactions [78, 79], metabolic networks, where kinetic parameters can be assigned to edges simulating the dynamics, and signaling networks, where edge directionality can indicate the actual flow of information within the biological system under study. Sources for metabolic networks across species are the MetaCyc database [80] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [81] or the human specific *Recon 2* [82]. Networks can be generated directly from time-resolved or perturbation experiments, but also from other resources like previously published information using different tools and methodologies. Once these networks are generated, a variety of software tools exist to visualize them: Cytoscape [83], VisANT [84] and yEd (`http://www.yworks.com`) to mention just a few. One of the main advantages of these network representations is their ability to integrate heterogeneous data from multiple sources into one common framework. For example, a metabolic network can be used to integrate transcriptomics, proteomics and metabolomics data [85]. This framework can then provide additional insight into the complex molecular processes of a biological system, for instance to uncover dysregulated modules, pathways or network motifs in disease. As further examples, Chuang et al. [25] utilized sub-networks to classify patients for different cancer types and Hofree et al. [73] used sub-networks of tumor mutations to stratify cancer patients into subtypes. In both cases, the stratification based on network modules yielded better results than using known disease markers clearly emphasizing the information gain due to the use of networks. In addition, the individual characteristics of the nodes and edges can give useful hints as to how all the molecules and the interactions between them determine the function of the underlying system. In the following subsections, we will briefly introduce two network-based approaches relevant for this thesis and describe how they are used in an integrative analysis: (1) contextualization of *a priori* knowledge-derived networks and (2) data-derived (unbiased) networks.

1.3.1 Knowledge-based integration

The concept of knowledge-based integration is also known as *bottom-up* systems biology [86, 87]. In general, a bottom-up approach denotes the generation of a model based on *a priori* known interactions, and subsequent analysis using the network model as additional input with the aim to determine the active parts, so called *active modules*, of the network under specific conditions. As mentioned above, such models, also called static networks [87], can be of different type and format but with the conformity that

they typically encompass all known (or possibly occurring) reactions/interactions within the biological system under study. Pioneering work in this field has been performed by the Palsson group, which published the first genome-scale metabolic reconstruction of *Haemophilus influenzae* [88]. A well-established mathematical tool to analyze and identify active network parts is constraint-based modeling, for example flux balance analysis (FBA), where the flow of metabolites through biochemical reactions is modeled under a steady-state assumption, thereby preserving stoichiometric constraints [89]. Applications of constrained-based methods range from the computation of minimal required reaction or gene sets [90] over the prediction of metabolic phenotypes in model organisms [91] to the process of adaptive evolution [92]. However, it has to be kept in mind that bottom-up approaches rely heavily on previous knowledge, thus results are always biased towards the information contained (or not contained) in databases and previously undiscovered associations might remain undetected.

With the availability of omics data sets, several extensions for constrained-based methods to integrate single high-throughput measurements into the analysis for contextualization were developed. For instance, iMAT by Zur et al. [93] utilizes transcriptomics and/or proteomics data to introduce additional constraints to the metabolic model, i.e. by labeling reactions as active/inactive based on the respective gene/protein levels, thereby improving the accuracy of metabolic flux predictions. Another example is the INIT algorithm developed by Agren et al. [94], which takes protein abundance levels from a public repository as input to construct tissue-specific metabolic models. More recently, Concentration Change Coupling Analysis (CoCCoA) was developed, which integrates transcriptomics and metabolomics measurements with a reaction kinetics model to investigate the relationship between gene expression and metabolite concentration changes [95].

Besides extensions to existing methods like FBA, several novel approaches were developed that combine omics data analysis with *a priori* defined interaction networks. For example, the Nielson group developed a hypothesis-driven method that combines transcriptomics measurements with different biological networks to identify reporter features [96, 97]. The key idea is to map expression changes between conditions onto a network of interest and subsequently use a mathematical algorithm, in this case simulated annealing, to determine regulatory 'hot spots', i.e. areas in the network where affected genes accumulate significantly. This approach was recently applied to elucidate functional differences in the gut microbiome of T2D patients in comparison to healthy controls [98] or to investigate regulatory signatures of T2D and their impact on metabolism [24].

In parallel, Çakir et al. [99] extended this method, allowing a combined analysis of metabolomic and transcriptomic data. Using a similar approach also belonging to the class of bottom-up systems biology, we integrate proteomics as well as metabolomics data with a genome-scale metabolic model to investigate the effect of environmental perturbations on the metabolism of T cells, which is described in more detail in **Chapter 2**. Further integrative approaches utilizing a variety of mathematical algorithms to identify active network regions were recently reviewed by Mitra et al. [100].

1.3.2 Data-driven integration

In contrast to knowledge-based approaches, data-driven or *top-down* systems biology approaches do not necessarily rely on any prior information [87]. Instead, measured omics data - combined with appropriate statistical and bioinformatics methodologies - are directly used to infer associations between molecules in an unbiased fashion. The rationale behind this approach is that molecular measurements inherently comprise information on the structure of the underlying biological network. More precisely, molecules belonging to the same cellular process are usually co-regulated and interact in a concerted fashion to warrant proper function. As a consequence, the concentrations of these molecules display a dependency structure that directly follows the wiring of the underlying molecular network. For example, if two genes are co-regulated they will also correlate, i.e. individuals with high levels of one gene will also exhibit high levels of the other gene, and vice versa. With the increased availability of omics measurements, a systematic inference of the underlying network structure became possible and a number of studies developed methods to reconstruct large-scale biological networks from time-series [101, 102] or steady-state data [103–105]. A very common approach to resolve the underlying network structure in an unbiased way is to apply second order similarity measures such as *Pearson correlation* coefficients [106].

For network construction, pairwise similarity scores are calculated between all possible molecule combinations, which are subsequently tested for statistical significance. If the similarity of a pair exceeds a predefined significance threshold, an edge is drawn between the two molecules, otherwise they remain unconnected. Studies using correlation-based networks commonly identify context-specific functional modules [107], but also global co-expression networks [108] from different organisms [109] and cell types [110]. Similarly, for metabolomics data a variety of studies systematically analyzed interactions between metabolites in various tissues, conditions and species [103, 111, 112].

However, statistical similarity measures have different strengths and weaknesses hence the method of choice should be suitable for the respective experimental design and data. For instance, if the analysis focus lies on the precise reconstruction of the biological network structure, standard Pearson correlation might not be the most suitable method. Instead, partial correlations could be applied which allow to distinguish between direct and indirect effects. For example, Krumsiek et al. [69] utilized a Gaussian graphical model (GGM) approach on serum metabolomics data from a population cohort to reconstruct known metabolic pathways and to successfully remove spurious correlations. Furthermore, they have shown that these data-derived metabolic networks can be useful in a variety of applications, e.g. for the functional annotation of unknown metabolites [113] or to identify sex-specific serum metabolome differences [64]. Another example are higher-order dependencies, possibly occurring in non-linear biological processes. Multivariate methods that have been proven to efficiently detect signatures from such processes in metabolomics data are independent component analysis (ICA) [63], O2-PLS [114] and BL-SOM [62]. Moreover, several methods allow to elucidate causal dependencies, for example Bayesian networks in the analysis of gene interactions [115] or extensions to partial correlation based methods analyzing genetic [116] or gene expression data [117], but also conceptually different methods which rely on genetic data as instrumental variables mimicking randomized controlled trials to assess causality such as *Mendelian randomization* [17], which will be reviewed in more detail in Section 3.2. Lastly, non-linear pairwise dependencies can be determined using methods like mutual information [118] or Spearman's rank correlation [119].

Correlation-based methods have also been successfully applied to integrate pairwise omics data, for instance to analyze the association between transcriptomics and proteomics [120, 121], transcriptomics and metabolomics [122–124] and even some attempts to successively combine more than two omics levels were made [125, 126].

In this thesis, we performed top-down systems biology approaches to integrate transcriptomics and metabolomics data measured in human whole blood samples (**Chapter 3**) and various organs/tissues of healthy and diabetic mice (**Chapter 4**). In both cases, we used standard correlation measures instead of partial correlation coefficients for two reasons: First, due to the curse of *data dimensionality* $n < p$, i.e. a high number of measured transcripts and metabolites as opposed to a few hundred samples, the data matrix does not reach full rank and a direct calculation of full-order partial correlations is impossible. Although several regularization approaches exist that enable an estimation of partial correlations in the $n < p$ scenario [127], resulting correlation coefficients

are almost vanished because of the conditioning against a high number of variables and thus hard to interpret [128]. Second, there is a large (functional) distance between transcripts and metabolites at a molecular level (Figure 1.1) which lets us assume that there is only a small number of direct associations but a high number of indirect associations. In our analysis, we are particularly interested in the general interplay between the functional layers not necessarily in the precise structure of the underlying network. An indirect association still can provide biologically meaningful insights, for instance, a shared coregulation or an involvement in the same biological process. Chapter 3 of this thesis will give concrete evidence of the validity of this approach. However, it has to be kept in mind that these indirect associations have to be interpreted carefully and a thorough functional evaluation is necessary to distinguish between biologically meaningless spurious correlations and biologically relevant associations.

1.3.3 Systems Genetics

As mentioned above, omics measurements implicitly carry information of the underlying network structure. At a population level, this interaction structure is encoded in the pattern of biological variability across individuals, which can be due to, for instance, variations in enzyme levels or their activity, but also due to fluctuations in cellular processes caused by natural genetic variation or environmental effects (see also Figure 1.3). This concept became known as *systems genetics*, a particularly important and new analysis type in the field of systems biology. In contrast to the standard analysis strategy in systems biology, where biological systems are specifically perturbed, systems genetics interprets and utilizes naturally occurring genetic variation in a population as perturbation to the system [129]. More precisely, omics measurements of population samples at inter-individually different steady states (biological replicates) are used to capture these variations and thus determine the underlying network structure. This allows studying a health or disease state in humans in its natural embedding, i.e. multiple genetic perturbations, with the goal to understand the general molecular and genetic architecture of a complex trait. In a sense, variations on the different molecular levels, caused by the underlying genetic variation and environmental factors, such as variation in gene expression, protein levels and metabolite concentrations, can be considered as *intermediate phenotypes* all contributing to the physiological or clinical phenotype under investigation. Statistical or computational methods from systems biology can be used to infer relationships between multiple intermediate phenotypes from patterns of covariation (Figure 1.3). Systems genetics has been successfully applied to investigate

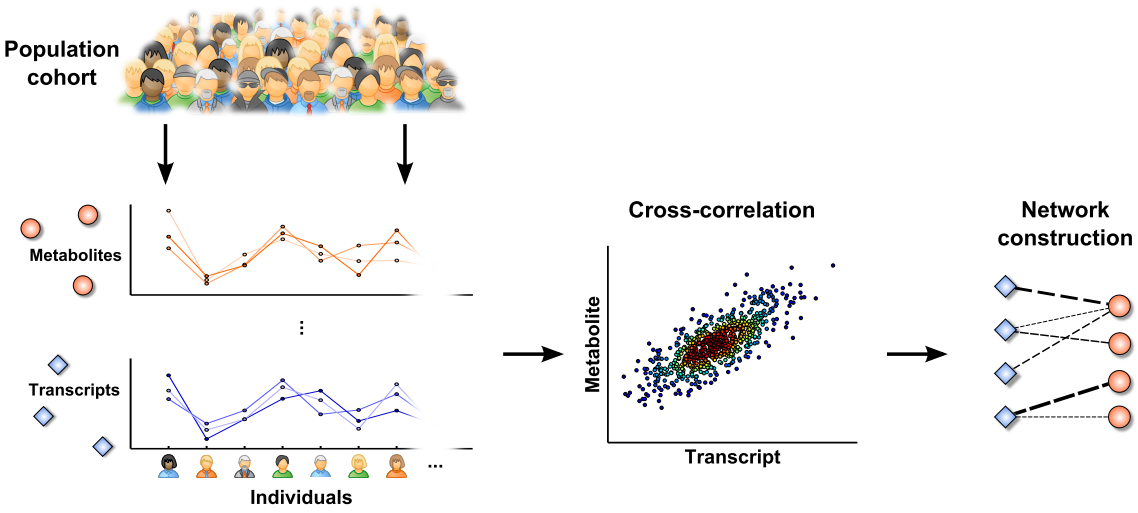


Figure 1.3: Overview of the typical analysis steps in a systems genetics approach. Starting with a population cohort of unrelated individuals that differ in a phenotypic trait of interest, one or more intermediate (molecular) phenotypes can be quantified by omics technologies. In this example, metabolite and transcript levels are measured in each individual simultaneously, leading to profiles of co-responses across individuals (different color shaded lines). The relationships between these traits can then be determined by pairwise cross-correlations. Finally, the determined associations can be modeled as networks where edges are drawn between highly correlated molecules. In this example a bipartite network is shown which only comprises edges between the two molecular levels (not within). Constructed networks can be further functionally characterized using adequate bioinformatics approaches like clustering or enrichment analysis.

the flow of biological information from DNA to phenotype, including complex traits. For instance, Ghazalpour et al. [130] integrated genetic, transcriptomic, metabolomic and clinical data to investigate the genetic regulation of metabolites in the liver of 104 different mouse strains. As another example, Gargalovic et al. [131] studied the phospholipid induced inflammatory response of human endothelial cells by modeling a gene co-expression network. In this work, we apply a systems genetics approach in **Chapter 3** to analyze the relationships between gene expression, metabolite levels and clinical traits measured in the same blood samples of the KORA cohort [132].

1.4 Blood as a surrogate tissue in biomedical research

Blood is a connective tissue, which not only ensures nutrient and oxygen supply of all organs of the human body through the blood vessel system, but also their communi-

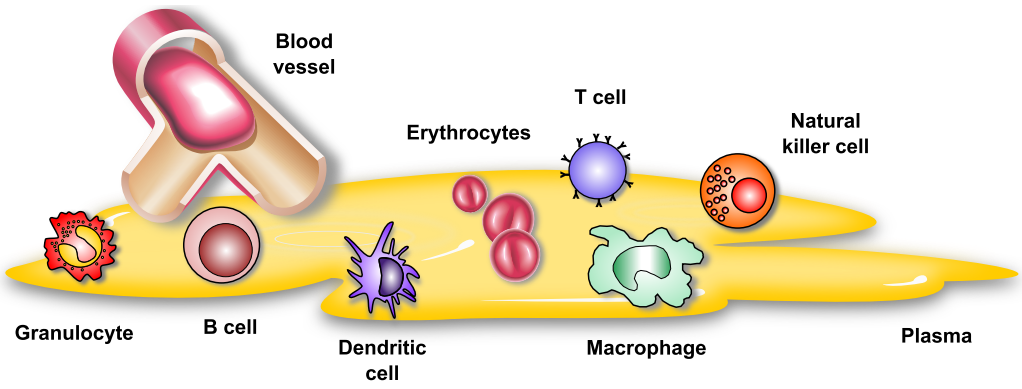


Figure 1.4: Simplified illustration of the human blood constituents. Only mature blood cells relevant for this thesis are shown. The cells can be roughly divided into red (erythrocytes) and white blood cells (leukocytes). White blood cells can be further subdivided into the myeloid lineage (dendritic cell, granulocyte, monocyte/macrophage) and lymphocytes (B cell, T cell, natural killer cell). Blood plasma is a aqueous phase carrying proteins, hormones and nutrients (metabolites) besides other compounds.

cation at a physiological level. Among the variety of key tasks performed by blood are immunological functions mediated through leukocytes (white blood cells). Due to this diverse functionality, blood is heterogeneous and complex in its composition (Figure 1.4). On a coarse level, blood consists of a cellular and a liquid phase. All blood cells originate from hematopoietic stem cells, which are able to differentiate into the functionally distinct blood cell lineages in a stepwise process [133]. These cells can be roughly divided into red (erythrocytes) and white blood cells, the latter of which can be further subdivided into two major types: myeloid leukocytes and lymphocytes. The myeloid lineage, which constitutes $\sim$ 70% of the blood cellular volume, basically consists of granulocytes and monocytes, both of which can be even further subdivided. The remaining 30% are made up of lymphocytes, which mainly include B cells, T cells and natural killer (NK) cells. Besides cellular constituents, which sum up to $\sim$ 45% of the total volume, blood mainly consists of plasma. Plasma represents the aqueous phase containing proteins, peptides, signaling molecules and steroid hormones, but also other metabolites (e.g. carbohydrates, amino acids and lipids) which are taken up and released by the organs.

The investigation and diagnosis of diseases and other physiological processes is often only possible to a limited extend, since the access to relevant tissue for sampling is restricted. For example Alzheimer's disease leads to a loss of neurons and synapses in the brain, or multiple sclerosis damages nerve cells in the brain and spinal cord, both

tissues for which biopsies are accompanied by severe risks for the patient thus making a thorough investigation hardly possible. Apparently, in such circumstances a 'surrogate tissue' would provide a substantial advantage. Blood might be a useful substitute, since it pervades throughout the entire body thereby being exposed to systematically released molecules by different organs, and also due to its easy availability in relatively large quantities. Indeed, a recent study showed that $\sim$ 80% of all genes contained in the human genome are expressed in circulating leukocytes in a micro-environment dependent fashion [134]. Furthermore, these genes were shared to a large extent with 9 different human (organ) tissues, including genes formerly believed to be expressed (non-blood) tissue-specific. Thus, Liew and colleagues suggested the *Sentinel Principle*, denoting that blood cells can act as sentinels of disease, which could be utilized for disease diagnosis and prognosis. Indeed, several recent studies demonstrated that gene expression profiles of circulating blood cells carry signatures of various health and disease states, including expression in human brain tissue [135], multiple sclerosis [136] and neuronal injuries [137], and under various environmental or stress exposures including smoking [138] and exercise [139]. In another study, it was also shown that metabolomics measurements in blood serum reflect the structure of underlying metabolic pathways [69].

Taken together, the unique composition of blood, agglomerating both metabolic and transcriptional variation carrying molecular signatures of system-wide processes, together with its minimally invasive accessibility, makes blood an ideal system for integrative biomedical research [134, 140]. In this thesis, we specifically focus on multiple omics data related to blood (i.e. derived from a blood cell line; see **Chapter 2**), directly derived from blood (i.e. whole blood samples; see **Chapter 3**) or on tissue/organ-specific processes and disease markers reflected in blood (see **Chapter 4**).

1.5 Research questions

Biological systems operate on multiple, intertwined organizational layers that can nowadays be accessed by high-throughput 'omics 'measurement methods. A major aim in the field of systems biology is, to understand the flow of biological information between the different layers at a systems level in both health and disease. Within the omics landscape, metabolomics display the endpoints of upstream biological processes, and the metabolic profile of an organism, integrating genetic as well as environmental variation, is commonly seen as most closest link to the observable phenotype. Thus, embedding

of metabolism into the omics landscape, i.e. investigating the relationships between the metabolome and other molecular levels, is of particular interest. Although more and more studies nowadays generate data from multiple molecular layers in parallel, most existing analytical methods were developed for the analysis of only single omics levels; hence the interpretation of multi-omics data in an integrated fashion remains challenging.

The main focus of this thesis will be on the integration of metabolomics data with (i) proteomics data in cells, (ii) transcriptomics and genetic data in tissues and (iii) metabolomics and transcriptomics data measured in multiple organs, with the aim to contribute to a better understanding of the biological relationships between metabolism and other molecular levels. Within a molecular level, the interpretation of an association between two molecules is often relatively straight-forward, e.g. the observation of a positive correlation between two transcripts could be explained by common regulation. In contrast, less is known about the underlying nature of associations across multiple omics layers at different biological scales. For this reason, we will investigate such cross-omics relationships in the context of different biomedical questions. Due to the lack of generally applicable methods, this involves development of novel analysis techniques that are specifically suitable, but also the extension and application of existing methodologies from distinct systems biology paradigms to integratively analyze multi-omics measurements.

In the first project, we ask how externally induced changes both at the protein and metabolite level can be combined to get an integrated view on the affected biological processes. Due to the underlying experimental design of repeated measurements from cell cultures and the relative low number of replicate samples, a straight-forward statistical integration (e.g. correlation-based) was not possible. Instead, a bottom-up approach is required, which utilizes *a priori* existing knowledge about the associations between two omics levels of interest in order to integrate them.

Another important biological question that will be addressed in this thesis is how transcriptional regulation assessed in blood cells and blood circulating metabolic compounds interact with each other. It is known that the uptake of nutrients, or altered levels of some blood metabolites can act as signal to the system, thereby affecting the transcriptional regulation of many responsive genes in various cells and tissues. And also vice versa, a change in the expression of enzyme-coding genes could affect the levels of metabolites via cellular metabolism. From a naive perspective one would expect to

observe an association between pairs of metabolites and transcripts representing metabolic reactions or signaling pathways. However, taking into account the heterogeneous nature of blood, it becomes unclear whether such systematic processes are reflected in the relationship between whole-blood derived transcript levels and serum metabolite concentrations. This thesis will demonstrate that systematic signatures of specific metabolic and signaling processes are directly observable from blood.

We will then ask to which extent inter-organ processes are mirrored in blood metabolites of diabetic mice. Specifically, we will systematically investigate how the global metabolomes and transcriptomes of various organs are reflected by plasma metabolites and identify plasma proxy markers for inter-organ processes. As discussed above, blood is fast and noninvasively accessible from humans, rendering it an ideal tissue for biomedical research. Moreover, the inherent characteristic of blood, acting as a pipeline that continuously transports (immune) cells, and exchanges nutrients between tissues and organs, thereby responding to (i.e. the immune cells) or mirroring (i.e. the exchange process) local changes that might occur due to injury or disease, highlights it a potential surrogate tissue for diagnosis of diseases where relevant tissues and organs are not easily accessible.

In this thesis, we systematically exploit the complementary information content of different molecular layers by integrative analyses of metabolomics, proteomics and transcriptomics data. Our results advance the embedding of metabolism into the omics landscape and contribute to an integrated global picture of the wiring between the metabolic, proteomic and transcriptomic molecular levels.

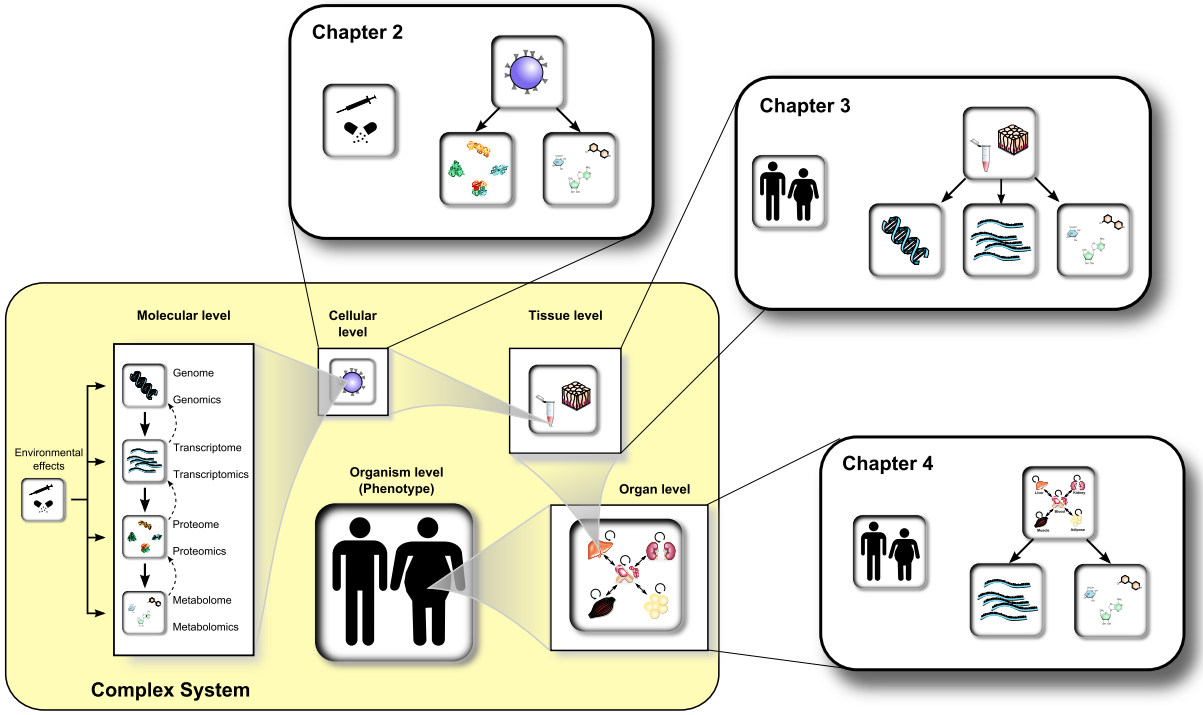


Figure 1.5: Overview of the thesis. Complex biological systems consist of different levels of organization that range from various molecular (omics) levels constituting a cell to ensembles of similar cells forming tissues to groups of tissues forming organs to whole organisms. The emergent properties of biological systems arise from interactions equally among the components of an organizational level, between different organizational levels and with external factors. In this thesis, for a range of biological issues, we systematically study the interactions between changing combinations of omics levels at varying biological scales. Starting at a cellular level, we apply a top-down approach integrating proteomics as well as transcriptomics data to investigate the influence of an environmental pollutant on the immune system (**Chapter 2**). In **Chapter 3**, we perform a systems genetics approach combining genetic with metabolomics and transcriptomics data from a population cohort to investigate their interplay in blood and show how the inferred interactions can be used to gain new insights into the physiology of certain phenotypes. Finally, **Chapter 4** introduces a multi-organ analysis of metabolomics and transcriptomics data in order to investigate to which extent organ processes are reflected in blood in diabetic mice.

1.6 Overview of this thesis

In the following, we provide a short outline of the content of this thesis. Considering three distinct biological scales, various combinations of high-throughput measurements and clinical traits will be analyzed. A graphical representation of the thesis outline is given in **Figure 1.5**.

In **Chapter 2**, we jointly analyze proteomics and targeted metabolomics data at a cellular level. To elucidate the impact of benzo[a]pyrene (B[a]P), a common air-pollutant, on cellular metabolism of Jurkat T cell lines both in activated and resting state, we study the induced concentration changes at the proteomic and metabolic level. By constructing a consensus genome-wide metabolic model out of three major reconstructions and computational integration of molecular profiles with this model, we search for context-dependent *active modules* which highlight network areas with strong changes in molecular activity. Our results provide novel insights on how environmental pollutants affect the metabolism of T cells and demonstrate the complementary property of an integrated multi-omics analysis. For example, we identify pathway signatures of an affected leukotriene metabolism in resting T cells, or signatures of a dysregulated phosphatidylinositol signaling system in activated T cells which have not been reported before.

A systems genetics approach is introduced in **Chapter 3**. Using transcriptomics and untargeted metabolomics data measured in 712 whole blood and serum samples from the KORA F4 population, we globally model the relationships across these two molecular levels in a biological network. We evaluate this network functionally both via manual investigation of pairwise associations and on a systems level using a metabolic network model. Thereby, we not only identify metabolic pathway mechanisms and regulatory signatures, but also cell-type specific signatures of various blood cells. By further integrating genetic variation across the 712 individuals, we investigate potential causal directions. In order to further analyze the cross-talk between both molecular layers at a pathway level, we develop a novel aggregation based approach which allows us to construct a pathway-pathway interaction network. Finally, by integrating intermediate physiological phenotypes, we demonstrate how the network can be used to gain novel insights into underlying molecular mechanisms.

In **Chapter 4**, we expand our analysis from a single tissue to an organ system level. By integrating untargeted metabolomics measurements from five distinct organs of diabetic and wild-type mice, we systematically model inter-organ associations and demonstrate that organ processes are generally reflected by plasma metabolites. By further extending the analysis with transcriptomics data from two organs we additionally show that plasma metabolites can serve as inter-organ proxies even for different omics layers. Based on the observed associations, we define multi-organ and single-organ proxy markers that either displayed associations with all investigated organs or exclusively with a single organ.

Finally, we demonstrate the usability of these newly derived markers in the context of type 2 diabetes.

The final **Chapter 5** will discuss the scientific contributions in the context of the field and discuss possible extensions and potential future projects.

Chapter 2

Integrated analysis of proteomics and metabolomics data uncovers a global environmental influence on T cell metabolism

2.1 Background

In this chapter, we present a combined analysis of proteomics and metabolomics data measured in Jurkat T cell line and investigate the impact of an environmental pollutant on the adaptive immune system represented by T cells in activated and resting state. Various studies investigated the effect of environmental toxic compounds, such as polycyclic aromatic hydrocarbons, on the human system and showed that exposure even at low concentrations compromises the immune system [141]. For example, benzo[a]pyrene (B[a]P) is such a polycyclic aromatic hydrocarbon, which is ubiquitous in the atmosphere and primarily produced during the combustion of any organic substance, e.g. fossil fuels, tobacco or charcoal broiled food.

B[a]P was shown to impair a variety of immune and regulatory processes, including the maturation of B cells, cytokine production and the cytotoxic activity of T cells [141–143]. The activation of T cells, a crucial immune system process inducing rapid T cell division and secretion of cytokines, is known to be mainly driven on a metabolic

level [144]. Moreover, a recent study reported that B[a]P initiates its own enzymatic degradation process in the body into various compounds, which ultimately results in the detrimental effects of benzo[a]pyrene [145]. This observation of an altered endogenous metabolism, together with changes on the enzyme level, indicates an involvement of multiple levels of regulation in response to a B[a]P-induced perturbation. However, only very few studies focused on the effect of an environmental pollutant like B[a]P on the activation process of T cells. Thus, the molecular mechanisms by which B[a]P affects the immune system and in particular the activation process of T cells remain largely unknown.

High-throughput measurement technologies like proteomics and metabolomics have become a vital tool to study environmental effects at a molecular level (cf. Chapter 1.1). The most common way to analyze these high dimensional data sets is the use of univariate statistics on each data type individually to detect differential levels of proteins and metabolites. However, a major goal in such an analysis then is to integrate the results obtained from the different molecular levels, which might provide further insights into the underlying biology then analyzing an interpreting the results separately. A conventional tool to further functionally interpret the typically large lists of differentially expressed proteins/metabolites and to identify affected pathways is gene set enrichment analysis (GSEA) [27]. However, most of these methods treat biochemical pathways as disjoint entities, thereby ignoring the interconnectedness and respective overlap between them. For instance, affected transcripts/proteins might be largely dispersed among many processes, leading to many isolated local changes which can not be detected by such a standard GSEA approach. Incorporating network information has been shown to overcome this limitation and was successfully applied using, for example, transcriptomics data in combination with a protein-protein interaction network to seamlessly identify cancer-associated processes, or in combination with a metabolic network to detect pathways involved in the metabolic syndrome [146, 147]. Lastly, at the time this analysis was performed, most of the existing enrichment tools either allowed an analysis at the transcript/protein level or the metabolite level individually (MSEA; [148, 149]), with only a few enrichment methods that allow for an integrated analysis of multiple molecular levels such as Integrated Molecular Pathway-Level Analysis (IMPaLA) [150] or Ingenuity^{TM} pathway analysis (IPA). However, proprietary methods like IPA only allow for a limited adaption to the customers needs. For instance, the underlying databases used for enrichment calculation often cannot be changed and also the mapping possibilities of molecules are rather limited, especially for metabolite species.

An alternative approach to identify affected pathways in an integrative fashion are model-based approaches. For example, metabolic network models are complex and highly interconnected representations of system-wide metabolism, ideally incorporating all known reaction mechanisms in a given species [82]. Several clustering or module-based methods exist to analyze omics data in the context of a biological network [100]. However, given the inherent mechanism of metabolic fluxes in metabolic networks, i.e. the constant mass flow through the system, a more intuitive way to analyze high-throughput data in the context of a metabolic model might be to consider reaction pathways between molecules of interest, rather than modules or clusters. For instance, changes in enzymatic activity or metabolite availability might lead to a changed flux through a metabolic pathway. Possible methods that allow for an identification of such altered paths rather than clusters -thus more naturally reflecting a changed flux through a metabolic pathway- are random walks based approaches [151]. Given a set of query nodes, the method of random walks on graphs has been shown lately to extract biologically meaningful pathways from metabolic networks [152]. Moreover, in a similar approach, random walks have been used in combination with gene expression data to extract relevant sub networks from a protein-protein network [153]. To the best of our knowledge, there is no study so far made use of more than one omics data in combination with random walks to extract sub networks from a metabolic network model and to subsequently use these networks to identify enriched pathways.

In order to investigate the effect of B[a]P on the immune system and in particular the activation process of T cells, we jointly analyzed proteomics and metabolomics data measured in Jurkat T cells using a random walks-based approach. Jurkat cells are an immortalized cell line commonly used to model human native T lymphocytes [154], which were both activated and treated with a sub-toxic concentration of B[a]P. As mentioned above, the activation process of T cells is strongly controlled by metabolic changes, thus our analysis was primarily focused on the effect of B[a]P on metabolic pathways.

In the following, we will shortly describe results from an univariate analysis on both single molecule levels and then show how these results can be combined with a metabolic model to jointly analyze the proteomics and metabolomics data. For this purpose, we used the significantly changed proteins/metabolites identified from the univariate analysis as seed nodes and a combined genome-scale metabolic network model as scaffold for a random walk approach to extract condition-specific metabolic sub networks. The method thereby assigns weights (relevance scores) to the traversed network nodes and edges according to their importance in connecting the seed nodes. These values

can subsequently be used to extract most relevant network modules for the investigated conditions. After extraction, we compare and functionally interpret the derived metabolic sub-networks between the different treatment induced cellular states and evolve new hypotheses how exposure to an environmental pollutant such as B[a]P affects the metabolism and the activation process of Jurkat T cells.

At the time this analysis was performed, three major sources for biochemical pathways existed, namely KEGG [81], the manually curated genome-scale metabolic model *H. sapiens Recon 1* (available from BiGG databases) [155, 156] and the Edinburgh Human Metabolic Network (EHMN) [157]. However, it was shown that these reconstructions share only little overlap and are far from being complete [158]. We therefore decided to assemble a combined metabolic model based on the data of all three databases. Note that, in 2013, using the three aforementioned sources combined with several cell type specific reconstructions, a consensus metabolic model was published which we also used later on (cf. Chapter 3.2 and 3.6, [82]).

This project was performed in collaboration with the research groups of Janina Tomm and Martin van Bergen from the Department of Metabolomics and the Department of Proteomics at the Helmholtz Zentrum Leipzig. The experimental work has been performed by Maxie Rockstroh and Sven Baumann of the Department of Proteomics from Helmholtz Zentrum Leipzig. The results of this study were published in:

⋆ Baumann, S.*, Rockstroh, M.*, **Bartel, J.**, Krumsiek, J., Otto, W., Jungnickel, H., Potratz, S., Luch, A., Wilscher, E., Theis, F.J., von Bergen, M., and Tomm, J.M. Subtoxic concentrations of benzo[a]pyrene induce metabolic changes and oxidative stress in non-activated and affect the mTOR pathway in activated Jurkat T cells. *Journal of Integrated OMICS*, 4(1), 2014.

* = equal contributions

The content of this publication is also part of another thesis by Maxie Rockstroh, who performed parts of the experiments. My contribution to this work was the statistical analysis of both metabolomics and proteomics data, the construction of a metabolic network model and implementation of the random walks based analysis, as well as subsequent analysis and biological interpretation of the extracted subnetworks.

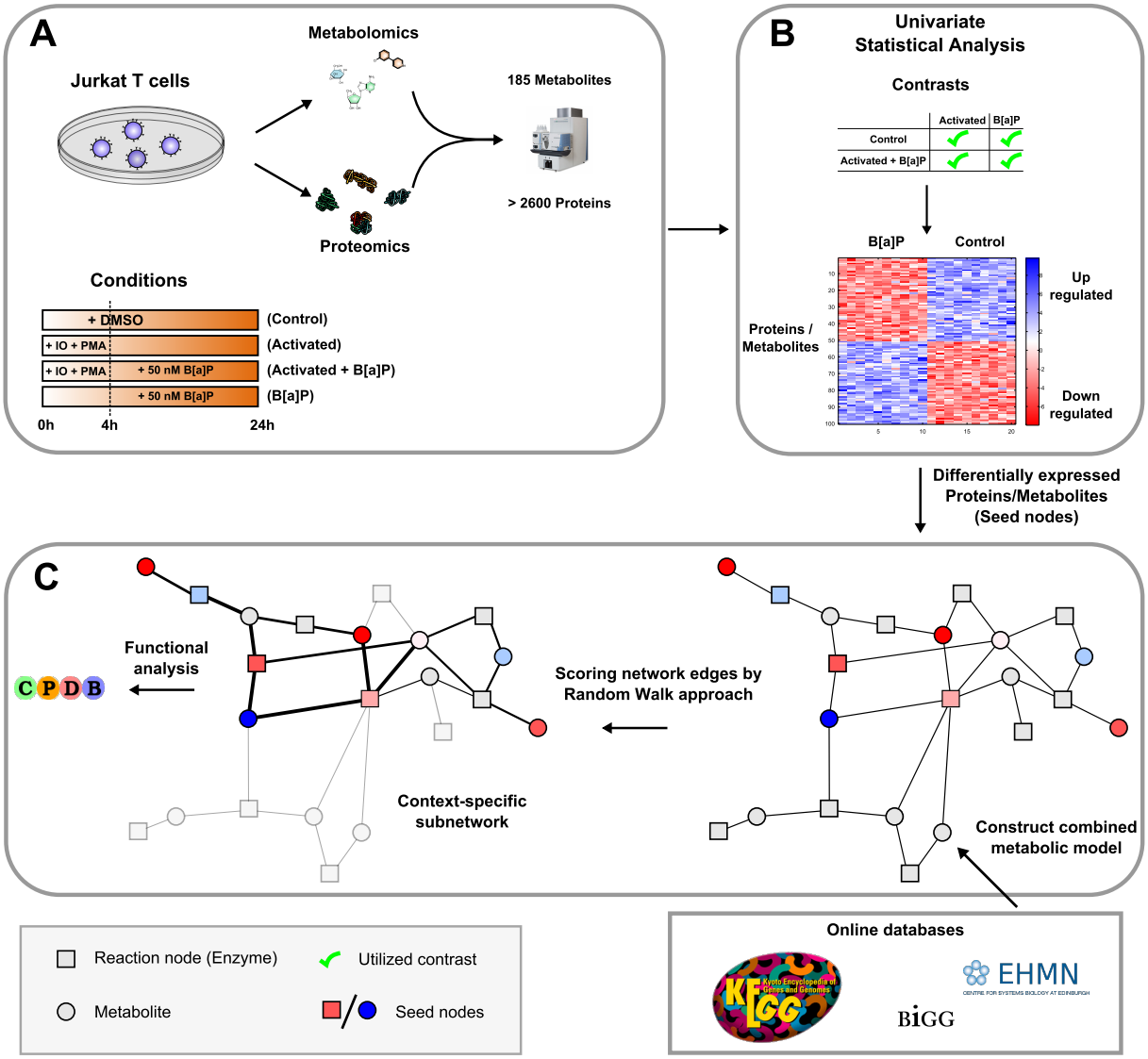


Figure 2.1: Experimental design and analysis workflow. **A:** Jurkat T cells were cultivated on RPMI-1640 medium. After incubation, the cells were treated for 4 hours with 4 different media: DMSO (control), IO + PMA (activated), DMSO + B[a]P (B[a]P), IO + PMA + B[a]P (activated + B[a]P). After 24 hours, both proteomics and metabolomics measurements were conducted using mass spectrometry. **B:** Significantly changed proteins/metabolites identified by univariate analysis from three different analysis contrasts (treatment comparisons) were used as seed nodes for the random walks approach. **C:** To this end, differentially expressed proteins/metabolites were mapped onto a combined genome-scale metabolic network model. An iterative random walks approach (3x) was performed starting in each seed node and ending with a maximum of 50 steps in any other seed node, thereby assigning probability weights to the traveled paths. Based on the resulting scores, the most relevant subnetwork for the respective contrast is extracted, which is then further functionally interpreted. Abbreviations: DMSO, dimethylsulfoxide; IO, ionomycin; PMA, phorbol-12-myristat-13-acetate; B[a]P, benzo[a]pyrene; BiGG, knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions; CPDB, ConsensusPathDB; EHMN, Edinburgh human metabolic network reconstruction; KEGG, Kyoto Encyclopedia of Genes and Genomes

2.2 Methods

In this section we will briefly describe the experimental setup of this study (Figure 2.1A) and introduce the mathematical concept of a random walk on a graph. Metabolomics and Proteomics measurements have been performed by Sven Baumann and Maxie Rockstroh, respectively, both members of the group of Martin van Bergen at the Helmholtz Zentrum Leipzig. Detailed information about the experimental methods for protein extraction (cell fractionation, SDS-PAGE, LC-MS/MS) can be found in Rockstroh et al. [159], and Baumann et al. [160].

Data acquisition

Jurkat T cells were cultured at 1×10^6 cells per ml RPMI-1640 medium (Biochrom AG., Berlin, Germany) containing 10% FBS (Biochrom AG., Berlin, Germany), 1% streptomycin-penicillin (100 U/ml) (PAA, Pasching, Austria) and 1% L-Glutamine (Biochrom AG., Berlin, Germany). The cells were incubated (MCO-18AIC, Sanyo Electric Co Ltd, Gunmaken, Japan) at 37C in an atmosphere of 5% CO2, 95% humidity. Cells were treated with four different media and all experiments were performed in triplicates: incubation with DMSO (control); activation with 750 ng/ml ionomycin (IO) and 10 ng/ml phorbol-12-myristat-13-acetate (PMA) for 4 h and subsequent resuspension in dimethylsulfoxide (DMSO; activated); some of the cells dissolved in DMSO (B[a]P) and of the activated cells (activated + B[a]P) were supplemented with 50 nM B[a]P (all Sigma-Aldrich, Steinheim, Germany). After 24h, B[a]P exposed cells and activated cells were washed with ice cold PBS and supernatants were collected for further analysis. Cell viability and the activation status was analyzed by flow cytometry.

Proteomics

Cells were fractionated using the Qproteome Cell Compartment Kit (Qiagen, Hilden, Germany) as described previously [159]. Briefly, after various centrifugation and incubation steps, three different cellular fractions (cytoplasm, membrane, nucleus) were collected for further analysis. Protein abundance in all fractions was measured using Pierce 660 nm Protein Assay (Jermo Fisher ScientiQc, Bonn, Germany) and following the manufacturer's protocol except that 5 μl of all samples and standards were used. For analysis with GeLC-MS/MS the proteins were digested using sequencing grade trypsin

(Roche Applied Science, Mannheim, Germany) as described in Rockstroh et al. [159]. The resulting peptides were extracted and a minimum of two peptides (with at least one unique peptide) were used for protein identification with MaxQuant (version 1.2.2.5) using the human Uniprot database (version 11/16/2011). Peptide and protein FDR was set to 1%. Protein quantification was performed by label-free quantification with a match between runs time window of 4 min and a minimum ratio count of 1. The overall quantification intensity per sample was normalized to the average quantification intensity of all samples of a particular fraction. Only proteins present in minimum two of the three replicates across all four treatments were considered for further analysis. In total, 2624 unique proteins were unambiguously identified, of which the most were identified in the cytoplasm (1969), followed by membrane (1842) and nuclear fraction (1506). Significant differences in the quantification intensities between the different treatments was determined using a Student's t-test. Proteins with a nominal p-value < 0.05 and a mean linear fold-change of 1.5 were treated as significantly regulated.

Metabolomics

Metabolites from cell lysates were quantified using targeted flow injection analysis-tandem mass spectrometry (FIA-MS/MS) and ion chromatography-tandem mass spectrometry (IC-MS/MS). Targeted profiling of 163 metabolites was performed using the AbsoluteIDQ^{TM} p150 kit (BIOCRATES Life Sciences AG, Innsbruck, Austria) as described in [161]. The metabolite panel mainly covers amino acids, carbohydrates, acylcarnitines, sphingomyelins, phosphatidylcholines besides some other lipid derivatives. In addition, 22 metabolites were measured using IC-MS/MS measured on an ICS-5000 (Jermo Fisher ScientiQc, Dreieich, Germany) coupled to an API 5500 QTrap (AB Sciex). Extracts were separated using an IonPac AS11-HC column (2 x 250mm, Jermo Fisher ScientiQc) with an increasing potassium hydroxide gradient. Finally, MS was done using multiple reaction monitoring and negative electrospray ionization. Statistical significance of differential metabolite concentrations was determined by a Student's t-test and a nominal p-value < 0.05 was considered significant. Note that on the metabolite level, treatment induced changes were much smaller than on the proteome level. Thus, no fold-change was used as selection criteria for metabolites.

Metabolic model construction

In order to integrate the three different metabolic databases BiGG (3311 reactions, 1477 metabolites), KEGG (1827 reactions, 1656 metabolites) and EHMN (2289 reactions, 2307 metabolites), an undirected, bipartite network consisting of metabolite and reaction nodes was compiled. Identical compounds and enzyme coding genes were merged based on common available KEGG identifiers or corresponding Entrez gene IDs, respectively. Database-specific compounds that could not be mapped were integrated into the metabolic network based on the reactions they participate in. To preserve database-specific information about the reaction modifiers (enzymes) in case of completely identical reactions (meaning all metabolic compounds participating in the reaction are equal between two or more databases), the reactions were merged and only the additional enzyme information was added to the reaction. The full merged model consists of 3695 metabolite and 5415 reaction nodes connected by 22897 edges. For further analysis, only the largest connected component of the integrated network was considered, which consisted of 3657 metabolite nodes and 5389 reaction nodes, and 22848 edges between the two node partitions.

Random walks on a graph and k-walks algorithm

In this section, we provide a brief introduction for random walks on graphs in general and the k-walks algorithm in particular, which is based on [152, 162]. Given an undirected, connected graph $G = (V, E)$ consisting of a set $V = v_1, v_2, ..., v_n$ of vertices (nodes) and a set $E = e_1, e_2, ..., e_n$ of edges with $E \subseteq V \times V$, let $\mathbf{A}$ denote the $n \times n$ adjacency matrix of G, whose (i, j)-entry a_{ij} denotes the weight of the edge from node i to node j. The a_{ij} entry is assumed to be zero if and only if there is no edge between node i and node j in the graph G. A random walk of length l in G is then defined as a sequence of vertices and edges $v_1, e_1, v_2, e_2, ..., v_l$ where each node v_{l+1} is a random neighbor of v_l, $\{v_l, v_{l+1}\} \in E(G)$. Such a random walk can be modeled as a Markov chain where the states are associated with each node of the graph [163]. The state of the Markov chain at moment t can be represented by random variable $X(t)$ and the transition probability from state (node) i at time t to state (node) j at time $t + 1$ can be formulated as:

$$P[X(t+1) = v_j | X(t) = v_i] = p_{ij} = \frac{a_{ij}}{d_i} \quad \text{with} \quad d_i = \sum_{j=1}^{n} a_{ij} \tag{2.1}$$

where p_{ij} is the transition probability from node i to node j, a_{ij} denotes the weight of the edge from node i to node j and n is the number of all nodes in the network. Note that from Markov chain theory, it follows that after many steps ($t \to \infty$) the probabilities converge to a stationary distribution. Furthermore, in cases where $\mathbf{A}$ equals a binary adjacency matrix, d_i simply denotes the degree of node i. In matrix form, equation 2.1 can be written as:

$$\mathbf{P} = [p_{ij}] = \mathbf{D}^{-1}\mathbf{A} \quad \text{with} \quad D_{ij} = d_i\delta_{ij} \tag{2.2}$$

where $\mathbf{D}$ is a diagonal degree matrix and δ_{ij} is the Kronecker delta. Note that $\mathbf{P}$ is row-stochastic by construction, where $\sum_i p_{ij} = 1$.

Let us now assume that we have a finite set of $K \subseteq V$ with $k = |K| \geq 2$ nodes of interest and we are searching for the network paths that best describe the relationships between these nodes. Consider, for instance, that the concentrations of isocitrate, succinyl-coA and dihydrolipoamide dehydrogenase (DLD) are significantly changed between two different treatment conditions. Then we might be interested in the relationship between these two metabolites and the enzyme, more specifically, in all the biochemical reactions/pathways connecting them (Figure 2.2). In graph theory, this problem essentially boils down to the extraction of a relevant subgraph describing the relationship between these three nodes.

A method particularly designed for such problems is the k-walks algorithm developed by Dupont et al. [162]. The algorithm starts from a given set of $K \subseteq V$ with $k = |K| \geq 2$ nodes of interest (seed nodes) and uses random walks to assign node and edge relevance values while connecting the seed nodes. This relevance is defined as the relative frequency, or in other words, the likelihood of a specific node or edge to be passed along the walks connecting any two seed nodes. Mathematically, Dupont et al. [162] defined a *node relevance* function $nr_{\mathbf{A},K} : V \to \mathbb{R}^+$ and an *edge relevance* function $er_{\mathbf{A},K} : E \to \mathbb{R}^+$ which assign relevances to any node or edge. Again, this can be modeled from Markov chain theory [163]. Recall that we assume a connected graph and from the theory of absorbing Markov chains follows that a state of the Markov chain is *absorbing* only if any walk reaching this state will stay with probability one.

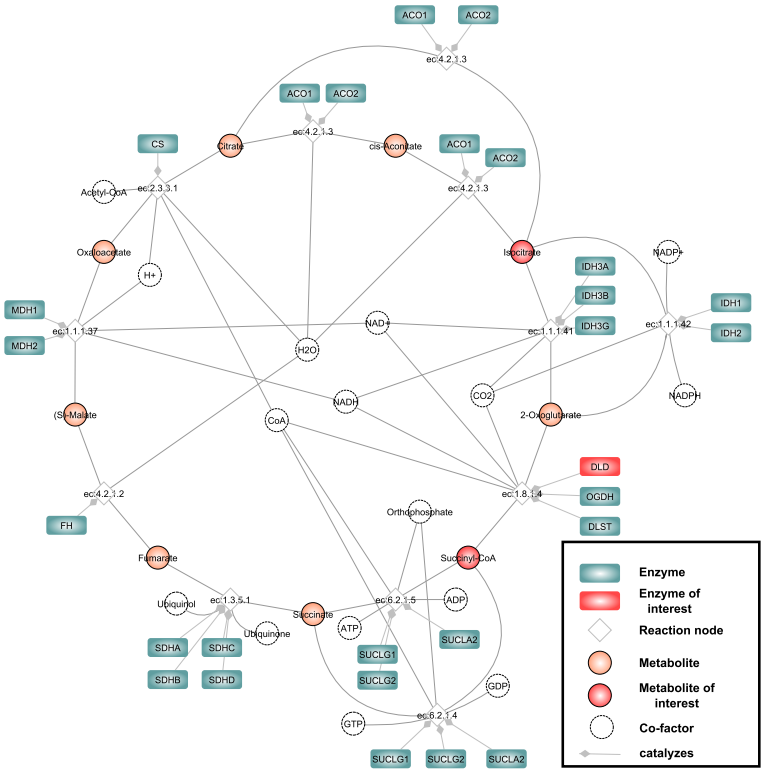


Figure 2.2: Network representation of the TCA cycle. In this toy scenario, the concentrations of isocitrate, succinyl-coA and dihydrolipoamide dehydrogenase are significantly changed between two treatment conditions. Thus, we might be interested in the biological relationships between these molecules. Note that there are several highly connected co-factors (depicted within the circle) offering potential (biologically meaningless) shortcuts between two nodes of interest.

If we consider random walks starting from one particular node $x \in K$, a modified transition matrix ${}^{x}\mathbf{P}$ can be defined from $\mathbf{P}$, where all states $\in K \setminus \{x\}$ have been transformed to be absorbing [162]:

$$ {}^{x}\mathbf{P} = [{}^{x}p_{ij}] = \begin{cases} 1 & \text{if } i \in K \setminus \{x\} \text{ and } i = j, \\ 0 & \text{if } i \in K \setminus \{x\} \text{ and } i \neq j, \\ p_{ij} & \text{non-absorbing nodes.} \end{cases} \tag{2.3} $$

Here, only the states of interest $\in K \setminus \{x\}$ have a probability of one to be absorbed in themselves and zero of being absorbed in another node $\in K$. All other states including x form the set of *transient* states V_T with a strictly positive probability for the walker

to leave these states. Let ${}^{x}\mathbf{Q}$ now denote the $(n-k+1)\times(n-k+1)$ submatrix of ${}^{x}\mathbf{P}$, considering only columns and rows related to nodes $\in V_T$, then ${}^{x}\mathbf{P}$ can be reordered to:

$$ {}^{x}\mathbf{P} = \begin{pmatrix} {}^{x}\mathbf{Q} & {}^{x}\mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{2.4} $$

where ${}^{x}\mathbf{R}$ is a $(n-k+1)\times(k-1)$ matrix denoting the probability of transitioning from a transient state to an absorbing state in one step, $\mathbf{0}$ is an $(k-1)\times(n-k+1)$ zero matrix and $\mathbf{I}$ is the $(k-1)\times(k-1)$ identity matrix. Then, the expected node passage time $n(x,i)$, defined as the number of times a node i is used in a random walk starting in x before being absorbed, can be computed by forming the fundamental matrix ${}^{x}\mathbf{N}$ of an absorbing Markov chain [162]. For this calculation, only the total probability of absorption ${}^{x}\mathbf{R}_i = \sum_r {}^{x}\mathbf{R}_{ir}$ in state i is relevant, which following from the row-stochasticity of ${}^{x}\mathbf{P}$ is equal to $1 - {}^{x}\mathbf{Q}_i$, where ${}^{x}\mathbf{Q}_{xi}$ denotes the probability of transiting from state x to i in a single step. Then, the fundamental matrix can be derived:

$$ {}^{x}\mathbf{N}_{xi} = \sum_{k=0}^{\infty} \left[({}^{x}\mathbf{Q})^k \right]_{xi} = (\mathbf{I} - {}^{x}\mathbf{Q})^{-1} \tag{2.5} $$

where $\mathbf{I}$ is the identity matrix and the entry ${}^{x}N_{xi}$ denotes the expected number of times node i is traversed during a walk, given that the walk starts in x and ends in any other seed node $\in K \setminus \{x\}$. Generalizing these computations to k nodes of interest as starting nodes, the global node relevances can be derived from the mean node passage times $nr : V \to \mathbb{R}^+$ [162]:

$$ \forall i \in V, nr(i) = \begin{cases} \sum_{x \in K} \tau_x \, {}^{x}N_{xi} & \text{if } i \in V \setminus K, \\ \tau_i \, {}^{i}N_{ii} & \text{otherwise} \end{cases} \tag{2.6} $$

where τ is a vector containing a prior probability distribution for the nodes of interest.

Similarly, the expected passage time ${}^{x}E$ of an edge $i \to j$ being used in a random walk starting in x before getting absorbed can be obtained from the above computations:

$$ {}^{x}E(i,j) = \begin{cases} {}^{x}N_{xi}\ {}^{x}p_{ij} & \text{if } i \in V \backslash K, \\ {}^{i}N_{ii}\ {}^{i}P_{ij} & \text{if } i \in K \text{ and } x = i, \\ 0 & \text{if } i \in K \text{ and } x \neq i \end{cases} \tag{2.7} $$

Analogously, the edge relevances can be computed from the mean edge passage times $er : E \rightarrow \mathbb{R}^{+}$ over all seed nodes $x \in K$ according to [162]:

$$ \forall (i,j) \in E, er(i,j) = \begin{cases} \sum_{x \in K} \tau_x\ {}^{x}E(i,j) & \text{if } G \text{ is directed}, \\ \sum_{x \in K} \tau_x\ |{}^{x}E(i,j) - {}^{x}E(j,i)| & \text{if } G \text{ is undirected} \end{cases} \tag{2.8} $$

In a final step, a subgraph can be obtained by keeping only those edges/nodes above a defined minimal relevance threshold. Figure 2.3A provides an illustration of a subgraph extraction using the three above-mentioned molecules (see Figure 2.2) as seed nodes and edge relevances above an manually chosen threshold as selection criterion for the subgraph. Note that also the network is undirected and thus the adjacency matrix $\mathbf{A}$ is symmetric, $\mathbf{P}$ is generally asymmetric (see Figure 2.3A, left panel) which is owed to the fact that two adjacent nodes not necessarily have the same degree.

Extensions to the k-walks algorithm

In this section, we briefly introduce some of the extensions to the k-walks approach developed by [162] that were used in this study. For a precise mathematical description, the interested reader is referred to the original publication of Dupont et al. [162].

Limited k-walks

For large input graphs, the calculation of node/edge relevances quickly becomes computationally infeasible, since the calculation relies on k matrix inversions (cf. equation 2.5) which can be performed in $\mathcal{O}(kn^3)$. By limiting the number of steps of random walks to a fixed maximum length l the relevances can be computed in linear time. Furthermore, limited k-walks have been shown to be a good approximation of the k-walks approach [162, 164]. Besides lower computing times, another advantage of limiting the length of random walks is the additional control over the compactness of walks between the seed nodes, which allows the extraction of smaller subgraphs.

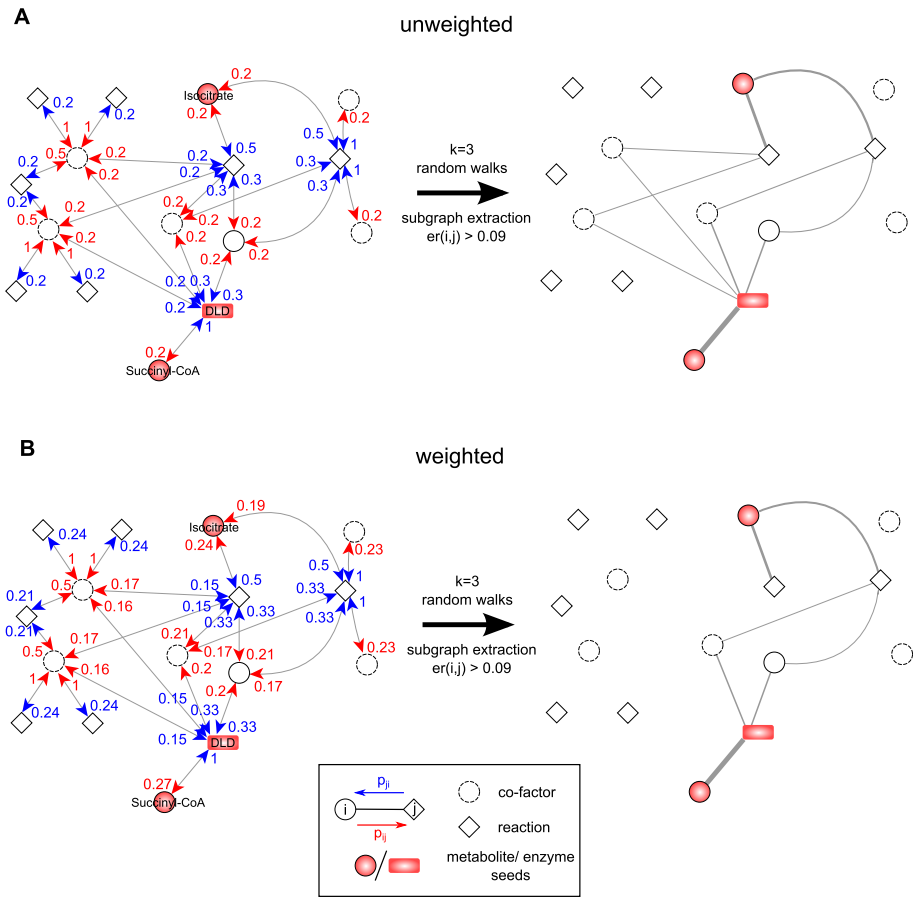


Figure 2.3: Artificial toy example of a random walks-based subgraph extraction. **A:** Transition probabilities between adjacent nodes (see equations 2.1 and 2.3 for calculation) on an unweighted network (each edge has a prior probability of one). Each of the three nodes of interest is used once as the seed node for the random walk. After calculating the mean edge relevance $er(i, j)$ of k=3 random walks for each edge, a manually chosen threshold of $er(i, j) > 0.09$ was applied, resulting in a subnetwork consisting of 9 nodes connected by 11 edges (upper right panel). Note that 3 out of the 4 possible routes between isocitrate and succinyl-coA are biologically meaningless shortcuts via co-factors. **B:** In the second scenario, transition probabilities between adjacent nodes were calculated on a weighted network (see equation 2.9 for calculation of the weights). Using exactly the same seed nodes and edge relevance threshold results in a more sparse subnetwork with two invalid shortcut routes between the nodes of interest being removed.

Iterative k-walks

The k-walks algorithm provides a list of node or edge relevance values as output. These values can be used in a recursive manner as new weights for the edges of the input graph in subsequent runs of the k-walks algorithm. This iterative process leads to an inflation

of the relevance values of important nodes/edges, thus enabling a better discrimination between relevant and irrelevant nodes/edges [162].

Group seed nodes

In metabolic networks, similar reactions can occur multiple times in many different pathways, which might all be catalyzed by the same enzymes (see for instance ACO1 and ACO2 in Figure 2.2). Considering such an enzyme as seed for the k-walks approach would lead to a whole list of reaction nodes as seeds which are obviously not independent of each other, but belong to the same group. Thus, relevances calculated on a list containing many dependent seed nodes hardly describe the true relationship between all the distinct seed nodes. Instead, a strong bias towards large groups would be observable. To account for this, the seed nodes can be separated into groups G_x and the Markov chain can be constructed relatively to each group, i.e. all nodes $\in K \setminus \{G_x\}$ are transformed to be absorbing [162]. Let ${}^{G_x}\mathbf{N}$ denote the fundamental matrix of such a transformed Markov chain, then by replacing ${}^{x}\mathbf{N}$ in equations (2.5) and (2.7) we can define the mean node and mean edge passage times analogously. As a result, the k-walks algorithm only considers relationships between seed nodes belonging to different groups.

Parameter setup

Significantly regulated proteins were pooled over all three cell fractions per treatment, since metabolites where measured from whole cell data. If the same protein was measured in more than one fraction, the respective fold-changes per fraction were checked for consistency and mean fold-changes were calculated. Inconsistently regulated proteins were removed from this analysis. Significantly regulated proteins and metabolites that could be mapped onto the integrated network were further used as seed nodes for the k-walks algorithm. Reaction nodes catalyzed by the same enzyme were treated as 'group seed nodes' as described above. Thus, when mentioning seed nodes later on, these might actually be groups of nodes instead of single entities. The maximal limit of steps for the random walks l, was set to 50 and the number of calculation iterations was set to 3 according to previous evaluations by others [152, 162].

Initial network weights and subnetwork extraction

Random walks in a network are highly dependent on the underlying local network topology, i.e. node visitation frequencies are strongly correlated with the respective node connectivity (degree, see Figure 2.4, upper panels). Thus, a random walk process is strongly biased towards hub nodes in the network. A major issue when dealing with metabolic models are so-called currency metabolites, such as H_2O, ATP and NAD+, which are unspecifically taking part in many reactions (recall the nodes depicted within the circle in Figure 2.2). Such hub nodes usually introduce biologically invalid shortcuts to the metabolic network model. In order to control for this topological bias, a network weighting strategy as proposed by Croes et al. [165] was used. More precisely, the edges between compound and reaction nodes were weighted by the inverse of the mean compound degree:

$$w_{ij} = \left(\frac{d_i + d_j}{2} \right)^{-1} \tag{2.9}$$

Using this weighting strategy, edges attached to a currency metabolite receive a lower probability to be used by the random walker than the other edges (Figure 2.4 lower panels). To obtain a relevant subgraph with maximal information about the relationship between the seed nodes, a dynamic relevance threshold was determined for each treatment contrast separately, such that at least one node of each group of seeds is connected to any seed node of another group. In a final pruning step, uninformative branches ending in non-seed nodes were removed as proposed in Faust et al. [152].

Modularity calculation

Modularity of the extracted weighted subgraphs was assessed as described in [166]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[a_{ij} - \frac{d_i d_j}{2m} \right] \delta\left(c_i, c_j\right) \tag{2.10}$$

where $m = \frac{1}{2} \sum_{ij} a_{ij}$, a_{ij} again denotes the weight of the edge between i and j, $d_i = \sum_j a_{ij}$ is the sum of the edge weights connected to i (see equation (2.1)), c_i is the

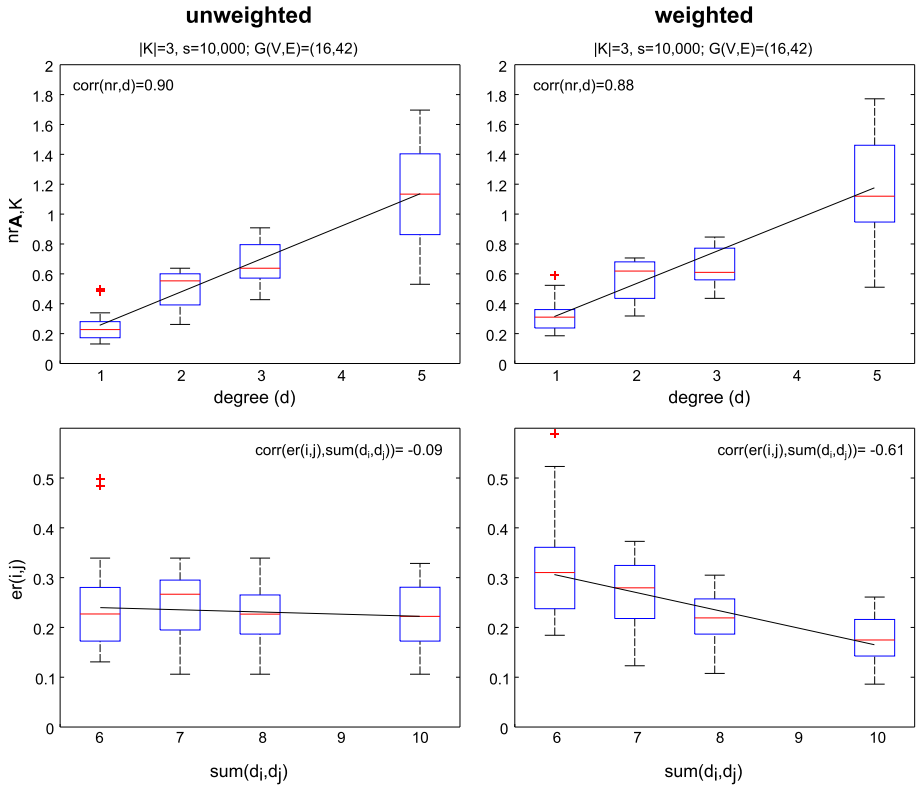


Figure 2.4: Node and edge visitation frequencies using random walks. Based on the same toy network as in Figure 2.3, we performed $k \times s$ random walks keeping two seed nodes fixed, while randomly sampling the third for $s = 10,000$ times from the remaining nodes. The upper panels show the calculated node relevances from each iteration against the respective node degrees. For both the weighted and the unweighted network, a clear correlation between node relevance (node visitation frequency) and node degree is visible. However, while edges receive an equal relevance in the unweighted network irrespective of the degrees of the nodes connected by them, our weighting approach clearly directs the random walks and thus the calculated edge relevances towards edges connecting nodes with lower degrees. Note that there were no nodes with degree 4 and no edges with a summarized degree of 9 included in the toy network (Figure 2.3).

community of node i and the δ-function $\delta(u, v)$ takes on a value of 1 if $u = v$ and 0 otherwise.

Functional enrichment

Functionality assignment of the determined network modules was performed by an over-representation analysis using the gene set and metabolite set analysis function of ConsensusPathDB [149].

If not stated otherwise, all analyses were performed with MATLAB version 7.12.0.635 (R2011a; The MathWorks Inc., Nattick, MA, USA). An implementation of the k-walks algorithm was downloaded from `www.ucl.ac.be/mlg/index.php?page=Softwares`. Graph visualization was performed with the yEd graph editor (yWorks GmbH, Tuebingen; `http://www.yworks.com`).

2.3 Effects of activation and B[a]P exposure at the metabolite and protein level

To investigate the effect of T-cell activation on the proteome and metabolome, and in particular how an environmental pollutant like B[a]P perturbs such a fine-tuned process, we firstly assessed the changes in protein and metabolite levels. To this end, we focused on four treatment comparisons (Figure 2.1B) from which we expected to observe the strongest effects. Note that the results reported here are all based on the pooled proteomics data across all three measured cellular fractions (membrane, cytoplasm, nucleus; cf. Section 2.2). For a more detailed analysis of the treatment effects on the single cellular fractions, the interested reader is referred to the original publication [160].

The largest impact on the proteome and metabolome was observed in activated T cells, with 365 differentially regulated proteins and 103 metabolites, respectively (Figure 2.5A+B). In contrast, B[a]P treatment alone had only a marginal effect on both molecular levels within unactivated T cells (131 regulated proteins, 1 regulated metabolite). However, when activated T cells are additionally exposed to B[a]P, a clearly distinct pattern can be observed. We first compared the combined treatment (activated + B[a]P) with the B[a]P effect on unactivated T cells, which again should resemble the activated

state of T cells if no adverse effects are caused by the interaction between B[a]P and the activation process of T cells. Although the activated + B[a]P versus B[a]P contrast yielded similar total numbers of regulated proteins and metabolites when compared to the activation effect (347 vs. 365; 101 vs. 103, respectively), a considerable amount of proteins and metabolites are exclusively regulated in both treatment contrasts (Figure 2.5A+B, red and green ellipses). A similar observation can be seen for the activated + B[a]P vs. B[a]P contrast when compared to B[a]P treatment of unactivated cells, where many molecules are exclusively regulated in the respective treatment comparisons (Figure 2.5A+B, blue and yellow ellipses). These observations of altered protein/metabolite concentrations suggest that B[a]P exposure specifically influences the activation process of T cells.

Next, we examined which pathways are influenced by activation and B[a]P exposure. The analysis was performed using the commercial Ingenuity^{TM} pathway analysis (IPA) software with predefined parameters (`www.ingenuity.com`; see also supplement of original publication for details). Note that pathways related to cardiovascular signaling as well as neurotransmitter and other nervous system signaling were excluded prior to the analysis.

The identified enriched pathways for all four conditions are shown in Figure 2.5C. Focusing on the activation process of T cells (second panel), we observed a strong enrichment of the EIF2 signaling pathway and also the CD28 signaling in T-helper cells was among the top ten enriched pathways. The latter pathway is known to play an important role in T cell activation, which can be seen as a validation on the proteomic level for a successful activation of Jurkat T cells [160]. Further confirming this observation, a Western blot analysis performed by our collaboration partners revealed an induced expression level of *NFKB1* -an important regulator of the immune response in general and in particular of the expression of cytokines- which is known to be regulated upon CD28 stimulation [167]. The EIF2 signaling controls the initiation of the translation process and is mainly regulated by changes in the phosphorylation state of the translation initiation factors [168]. Our collaboration partners experimentally verified the identification of this pathway by analyzing the phosphorylation state of eIF-2α, which was found to be decreased in activated cells [160]. Finally, two metabolic pathways, purine and pyruvate metabolism, were among the top enriched pathways identified by IPA analysis. This is in accordance to the above-mentioned involvement of metabolic processes in the activation of T cells.

Interestingly, seven out of the top ten enriched pathways identified for the effect of B[a]P exposure on non-activated T cells are metabolic pathways, further indicating that the effect of B[a]P and its derivatives is mainly mediated via the cellular metabolism (Figure 2.5C, first panel). Similar to the observations from univariate analysis, the effect of B[a]P seems to become even more profound in activated T cells, as, for instance, indicated by the identification of an enriched mTOR signaling pathway which was not observed from activation alone. The mTOR signaling pathway plays a crucial role in the integration process of metabolic and environmental influences and is particularly involved in the activation and differentiation of T cells via reprogramming of metabolic pathways [144, 169].

Taken together, the results from univariate analysis of both data sets and also from the Ingenuity^{TM} pathway analysis already indicated that B[a]P exposure has an impact on both molecular levels and that the cellular metabolism plays a central role in mediating the effect. However, IPA relies on a proprietary knowledge base with a particular focus on protein-protein interactions and canonical (signaling) pathways [170], thereby providing only a limited capability to integrate and analyze metabolomics measurements.

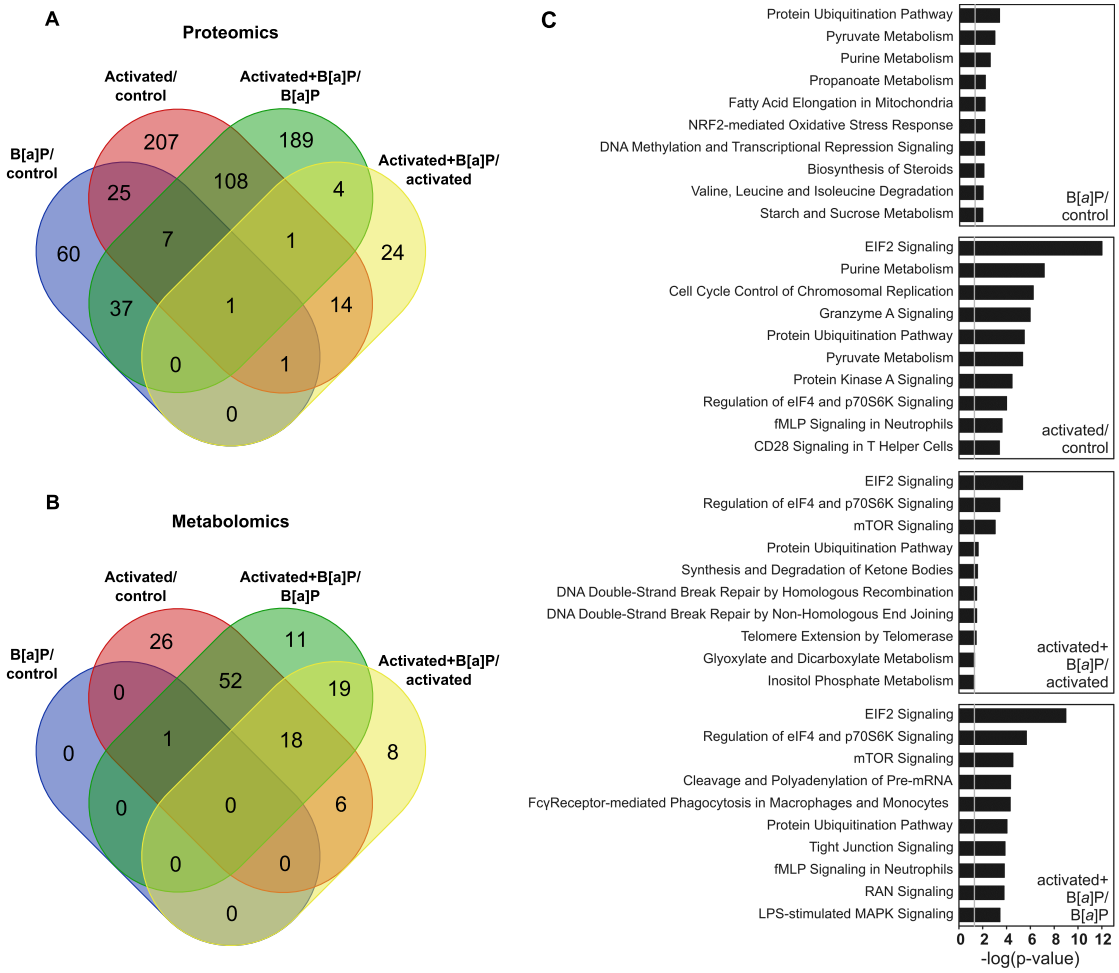


Figure 2.5: Comparison of treatment effects and pathway enrichment analysis. **A:** Comparison of the significantly regulated proteins reveals strong differences between the treatment contrasts. **B:** At the metabolite level, the activation process showed the strongest effect, which however is different to the effect of activated cells additionally treated with B[a]P. **C:** Pathway enrichment of proteins using IPA. Shown are the top ten enriched canonical pathways per treatment contrast. In part reprinted from Baumann et al. [160].

2.4 Metabolic network-based integration reveals the impact of B[a]P exposure on the metabolism of (activated) T cells

Motivated by the observations from the previous section of an effect of B[a]P treatment on both the protein and metabolite level, together with an involvement of metabolic pathways identified by protein-based pathway analysis, we now specifically focus on the effect of B[a]P treatment on the cellular metabolism. This is of particular interest, since metabolism is closely intertwined with many signaling processes and provides a direct interface between the protein and metabolite level, enabling us to integratively study the conditional changes on both molecular levels. To this end, we applied a k-walks approach that takes (treatment) relevant nodes as seeds to randomly traverse a given network and assign edge or node relevances in relation to the given seed nodes. The key idea behind this is, that some edges in the input network are more important than others in order to connect the seed nodes. More precisely, significantly changed metabolites/proteins identified in the previous section, and thus associated to B[a]P exposure and/or activation, can be used as seeds to extract treatment-relevant (affected) metabolic subnetworks [162]. The utility of random walks-based approaches to extract context-specific subgraphs was already shown in previous studies, either from protein-protein interaction networks [171], functional networks assembled from genomics data [153] or metabolic models [152]. Figure 2.1C shows the general workflow of the k-walks approach.

A comparison of the extracted subnetworks between all four conditions is given in Table 2.1. The extracted subnetworks from two conditions, B[a]P treatment of unactivated T cells versus control (B[a]P/Control) and B[a]P treatment of activated T cells versus the activation effect (Act. + B[a]P/Act.), were of particular interest to investigate the impact of B[a]P exposure on the cellular metabolism T cells which will be discussed in more detail in the following. For a discussion and illustration of extracted subnetworks from all four conditions, the interested reader is referred to the original publication [160].

Condition	Seed nodes			Extracted Network	
	Total	Groups ($N \in G$)	Met.	Nodes (%)	Edges (%)
Activated/Control	85	39 (292)	28	1415 ($\sim$ 16%)	1575 ($\sim$ 7%)
B[a]P/Control	30	16 (237)	1	1191 ($\sim$ 13%)	1349 ($\sim$ 6%)
Act. + B[a]P/Act.	18	7 (21)	8	160 ($\sim$ 1.7%)	178 ($\sim$ 0.7%)
Act. + B[a]P/B[a]P	74	38 (336)	21	1613 ($\sim$ 18%)	1892 ($\sim$ 8%)

Table 2.1: Comparison of metabolic subnetworks inferred with the k-walks approach. For each condition, the number of utilized seed nodes as well as the size of the extracted subnetworks are displayed. Note that in all four conditions, almost half of the seed nodes are group seed nodes, each consisting of multiple reactions catalyzed by the same enzyme. The extracted subnetworks can be considered relevant or 'active' for the respective condition. $N \in G$: number of total reaction nodes within the groups; Met.: number of metabolites used as seed nodes.

2.4.1 Effects of B[a]P exposure on cellular metabolism of unactivated T cells

Univariate analysis of the B[a]P effect on non-activated cells (B[a]P/Control) yielded 131 significantly regulated proteins and 1 metabolite, respectively (Figure 2.5 A+B). Out of these, we were able to unambiguously map 30 (29 proteins, 1 metabolite) onto nodes from the metabolic model which were subsequently used as seed nodes for the k-walks approach. The identified subnetwork covered 1191 nodes ($\sim$ 13%) of the global metabolic model connected by 1349 edges ($\sim$ 6%) (Figure 2.6). Given the relatively low number of seed nodes in comparison to the other three conditions, the extracted subnetwork appears rather large, indicating a wide-spread effect of B[a]P exposure on cellular metabolism (Table 2.1). Moreover, the large sizes of the 16 group seed nodes suggesting an involvement of mainly ubiquitously acting enzymes. The B[a]P subnetwork included reaction paths involved in 'purine and starch and sucrose metabolism' as well as in 'valine, leucine and isoleucine degradation' and 'fatty acid elongation' which is in accordance to pathways identified with the IPA pathway enrichment (Figure 2.5 C). In addition to a mainly down-regulated 'fatty acid elongation', we find many adversely regulated reaction paths belonging to the β-oxidation of fatty acids with varying degrees of saturation, but also some belonging to the 'oxidative phosphorylation' process. Together with several other reactions paths, e.g. from 'butanoate metabolism' or 'TCA cycle', this observation indicates a strong impact of B[a]P exposure on the cellular energy metabolism. Interestingly, the subnetwork also contained several reaction paths associated to 'leukotriene metabolism', which is known to be involved in T cell differ-

entiation and recruitment [172]. Thus, we hypothesize that 'leukotriene metabolism' might represent one of the molecular mechanisms by which the effect of B[a]P exposure is mediated in non-activated T cells.

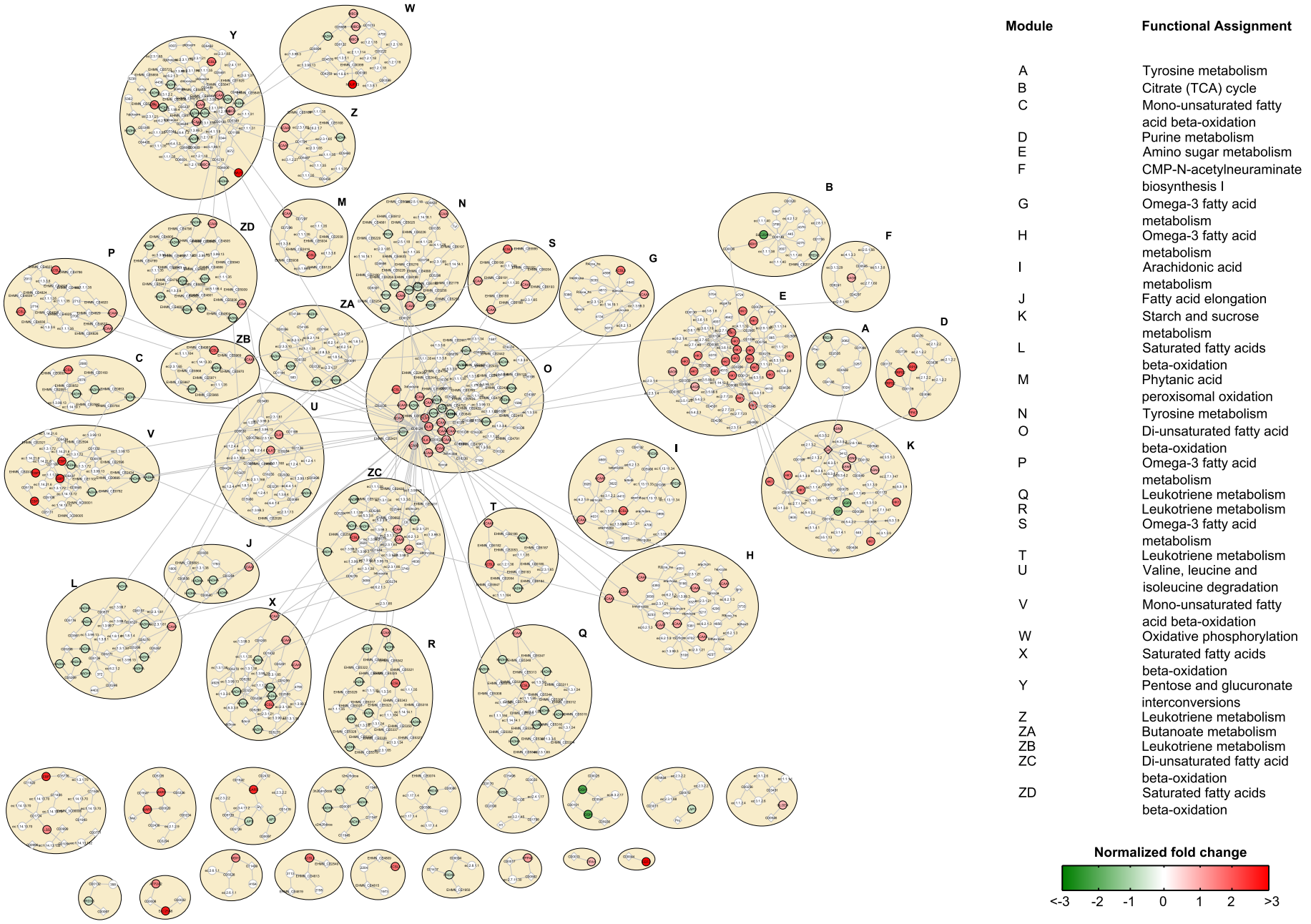


Figure 2.6: See next page for caption.

Figure 2.6: Effects of B[a]P on the metabolism of non-activated T cells. Metabolic subnetworks inferred from significantly changed proteins and metabolites by the k-walks approach. Functional assignment of the modules (highlighted ellipses) was performed by an over-representation analysis using the gene set and metabolite set analysis function of ConsensusPathDB. Predominant pathways are indicated for each module. Diamonds represent metabolites and circles depict reaction nodes. In case of measured proteins catalyzing a reaction, the node is assigned to the respective gene symbol. Otherwise nodes are generically labeled by the associated EC number. Measured metabolites are labeled by the respective name from the experimental platform, all others by an identifier of the database they originate from. Seed nodes used to infer the network are colored according to their fold-change normalized by the standard deviation; non-seed nodes are colored in white. Reprinted from Baumann et al. [160].

2.4.2 Effects of B[a]P exposure on cellular metabolism of activated T cells

The effect of B[a]P on activated T cells was determined based on the Act. + B[a]P/ Act. condition (Table 2.1). 18 out of 96 significantly regulated molecules (45 proteins, 51 metabolites) could be used as seed nodes (Figure 2.5 A+B) resulting in a subnetwork including 160 nodes and 178 edges. In comparison to the B[a]P effect on non-activated T cells, the network seems rather small, an observation that might be explained by overlapping effects between the activation process and exposure to B[a]P. Among the metabolic pathways significantly overrepresented in the extracted subnetwork, we again observed a treatment-induced activity change in the energy metabolism ('butanoate metabolism', 'TCA cycle', 'glycerophospholipid metabolism'), but also an affected phosphatidylinositol and IL-7 signaling system (Figure 2.7). The latter finding is of particular interest, since it is known that IL-7 signaling is involved in T cell development [173], and phosphoinositide lipids are important signaling molecules in T-cell development and function [174]. Hence, a decreased activity in phosphatidylinositol signaling or an induced IL-7 signaling evoked by exposure to B[a]P might indicate another molecular process by which the effects of B[a]P are mediated in activated T cells. In addition, the Act. + B[a]P/Act. subnetwork included several reaction paths assigned to 'vitamin B6 metabolism'. Previous studies reported diverse interactions between B[a]P metabolism and vitamins. For example, Wolterbeek et al. [175] observed a protective effect of vitamin A against B[a]P induced DNA damage and Israels et al. [176] reported an inhibitory effect of vitamin K3 on B[a]P metabolism. Further interactions were reported between B[a]P metabolism and vitamins C, D and E [177, 178], however to the best of our knowledge, no previous studies reported an association between B[a]P

exposure and vitamin B6 metabolism. Vitamin B6 might therefore represent another vitamin interacting with B[a]P metabolism.

Effect of B[a]P exposure on the TCA cycle

The only metabolic pathway signature shared between the extracted subnetworks from all four experimental conditions was associated to the TCA cycle (Figures 2.6 and 2.7 and online supplement of original publication). The TCA cycle, also known as citric acid cycle, belongs to the energy metabolism and is of central importance to numerous other biochemical pathways, i.e. by direct production of chemical energy in the form of ATP, but also by feeding NADH into the oxidative phosphorylation process for subsequent energy production [179]. In its central role, the TCA cycle links carbohydrate, fatty acid and protein metabolism, all of which were also affected by B[a]P exposure. The effects of activation and B[a]P exposure on the TCA cycle are shown in Figure 2.8 in more detail. When comparing the effect of B[a]P exposure on non-activated T cells with changes induced by activation, we observe a relatively similar overall alteration pattern, yet however with substantial differences in some key areas. For instance, not only the concentrations of metabolites like malate and citrate, but also of *malate dehydrogenase 2* (MDH2) and *isocitrate dehydrogenase 1* (IDH1), two key enzymes of the TCA cycle, were much stronger regulated by B[a]P treatment than during the regular activation process. The effect of B[a]P exposure on the TCA cycle becomes even more emphasized in activated T cells. For example, considering the B[a]P + activated condition in comparison with activated T cells, we observed a strong upregulation of citrate, isocitrate, α-ketoglutarate and succinate, while enzymes like IDH1 or MDH1 displayed downregulation. Taken together, these observations might indicate specific points where a substantial reconfiguration of the TCA cycle evoked by both the activation process and the exposure to B[a]P takes place.

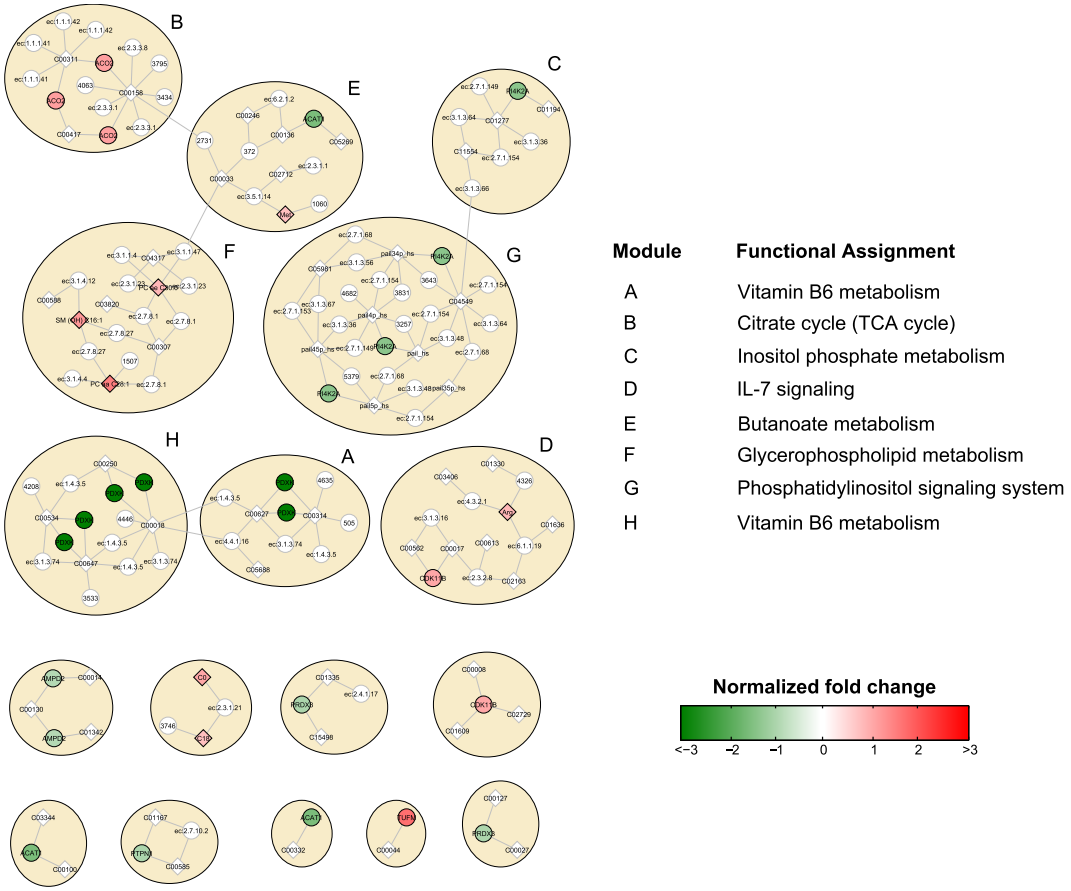


Figure 2.7: Effects of B[a]P on the metabolism of activated T cells. Subnetwork extracted with seed nodes from Act. + B[a]P/ Act. condition. Predominant pathways are indicated for each module (highlighted ellipses). Diamonds represent metabolites and circles depict reaction nodes. In case of measured proteins catalyzing a reaction, the node is assigned to the respective gene symbol. All other nodes are generically labeled by the associated EC number. Measured metabolites are labeled by the respective name from the experimental platform, all others by an identifier of the database they originate from. Seed nodes used to infer the network are colored according to their fold-change normalized by the standard deviation; non-seed nodes are colored in white. Reprinted from [160].

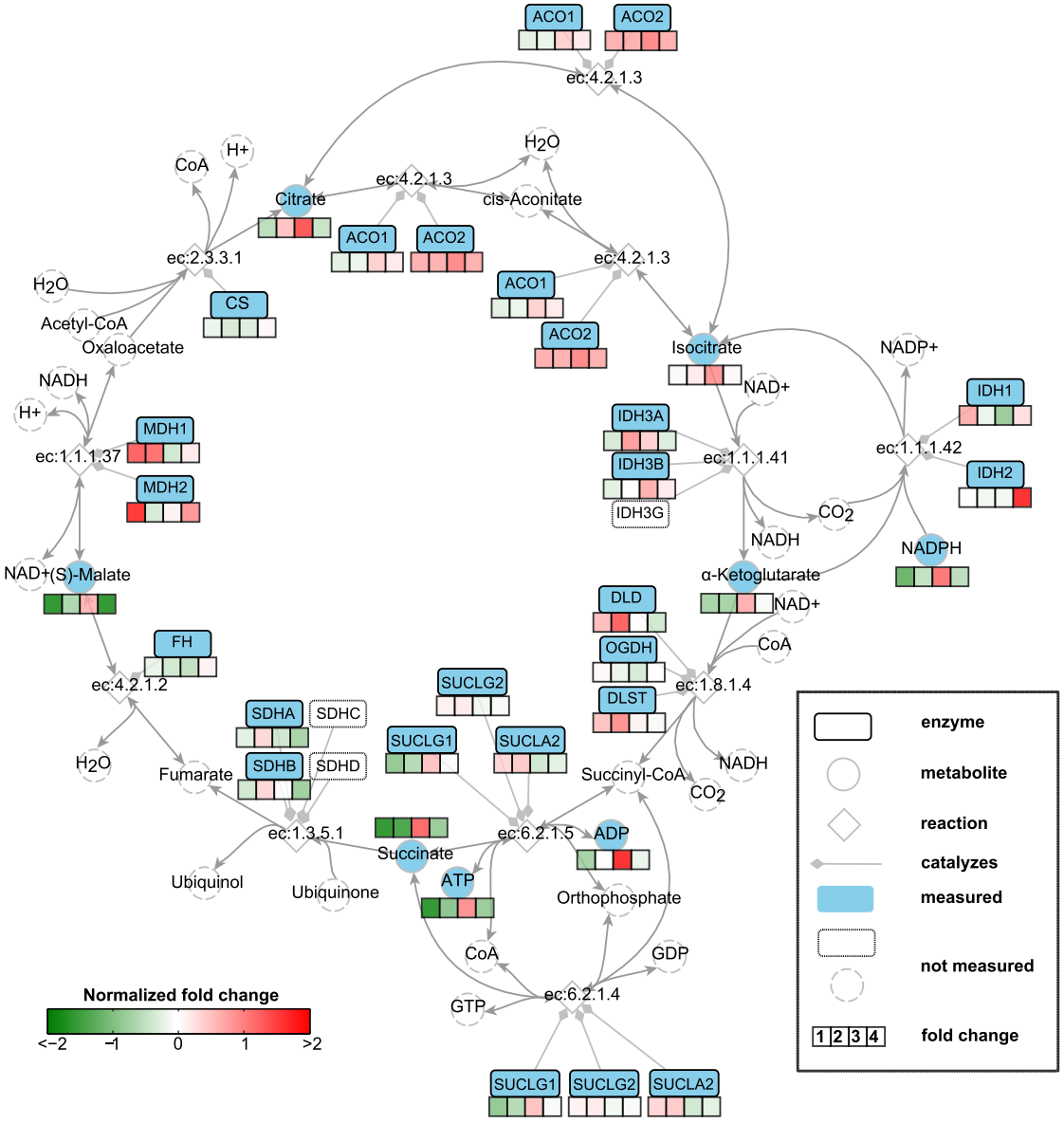


Figure 2.8: Treatment induced effects on the TCA cycle of T cells. Colors below nodes indicate fold-change in protein or metabolite abundances for all conditions. 1 - B[a]P/Control; 2 - Activated/Control; 3 - Activated+B[a]P/Activated; 4 - Activated+B[a]P/B[a]P). We observed many differently regulated metabolites and proteins, indicating an effect propagation through the TCA cycle evoked by the different experimental conditions. Adapted from Baumann et al. [160].

2.5 Discussion

Studying the effect of activation and benzo [a] pyrene exposure in T cells at both the proteome and the metabolome level revealed many cellular alterations. For instance, the statistical analysis at both single molecule levels individually revealed various significantly changed proteins and metabolites between pairwise comparisons of all four conditions (Figure 2.5A+B). In addition, a pathway analysis using IPA showed an affected eIF2 pathway and CD28 signaling in activated T cells and indicated a changed mTOR signaling pathway when additionally exposed to B[a]P (Figure 2.5 C). Interestingly, both the activation process of T cells and B[a]P exposure of non-activated T cells demonstrated an overlapping effect on metabolism in terms of purine and pyruvate metabolism, besides several additional metabolic pathways solely affected by B[a]P exposure, indicating an extensive treatment-induced reconfiguration of the cellular metabolism (Figure 2.5 C). For the further analysis, we particularly focused on the effect of B[a]P exposure on the metabolism of non-activated and activated T cells.

To integratively investigate cellular metabolism for protein and metabolite concentration signatures that might be associated with both treatment conditions, we used a network model-based approach. First, we constructed a global metabolic consensus network combining three major sources for biochemical pathways (see Section 2.2), which was subsequently used as scaffold to extract context-specific metabolic subnetworks. To this end, we mapped significantly changed proteins/metabolites inferred from univariate analysis (Section 2.3) onto the global metabolic model and connected them via the most relevant reaction pathways using a random walks algorithm. A network model-based random walks approach has several advantages for the integrated identification of treatment induced changes of proteins and metabolites. First, considering metabolic pathways as a global network rather than disjoint entities not only provides a holistic (systematic) view on involved molecular mechanisms, but also illustrates the inherent cross-talk between the affected metabolic pathways. Moreover, using random walks for the identification of affected pathways more closely resembles the nature of metabolic pathways in terms of metabolite flux through the system. In particular, connecting two seed nodes by a random walker on the most relevant paths includes the functional neighborhood (local network topology, i.e. the reaction weights), which might lead to more coherent results then taking only a list of significantly changed proteins/metabolites into account. For instance, IPA identified 'propanoate metabolism' to be affected by B[a]P exposure of non-activated T cells which was not included in the B[a]P subnetwork

(Figure 2.5C and Figure 2.6). However, a closer inspection of the proteins that led to the identification reveals an overlap to a large extend with 'valine, leucine and isoleucine degradation', a process which was identified by both methods. One can thus speculate that the 'propanoate metabolism' has no relevance given the metabolic network context and the total amount of changed proteins/metabolites (seed nodes). Hence, this observation might be a false positive that just occurred due to redundancy caused by the pathway overlap. Another important feature of our network-based random walks approach is the easy integration of public, custom-made or even data-generated biological networks. Such specialized, high-granularity networks provide a more detailed view on underlying biological processes and further enable the extraction of the most relevant subnetworks with respect to the data. Lastly, the subnetwork approach directly provides models of affected molecular mechanisms, which might be involved in the mediation of the treatment effect.

In line with the findings from Ingenuity pathway analysis, the extracted subnetwork from B[a]P exposure of non-activated T cells contained signatures of different pathways involved in energy metabolism and components of the fatty acid elongation pathway. Beyond that our subnetwork-based approach additionally detected several specific signatures of lipid metabolism, e.g. from ω-3 fatty acid metabolism, and also mono-, di-unsaturated and saturated fatty acid β-oxidation, thus providing a more fine-grained view on the affected parts of cellular metabolism (Figure 2.6) and complementing the results from Ingenuity pathway analysis. Furthermore, we detect pathway signatures of leukotriene metabolism, known to play an important role in T cell differentiation and development which was not observed using IPA.

As stated above, IPA identified the mTOR signaling pathway to be affected in Activated + B[a]P/Activated condition (Figure 2.5C). Using our network-based approach, we could additionally identify signatures of the phosphatidylinositol signaling system, which is known to act upstream of the mTOR signaling pathway, thereby activating mTORC1 [180]. Interestingly, the mTOR signaling pathway can furthermore be regulated by energy depletion [180] and also vitamins have been shown to affect the mTOR system [181]. For both pathways, we found signatures in the subnetwork extracted for the Activated + B[a]P/Activated condition. Thus, our results suggest that the molecular process by which B[a]P exposure mediates its effect might be located upstream of the mTOR signaling pathway.

In summary, we showed for the first time that a network-based random walks approach can be used for an integrated systematic analysis of proteomics and metabolomics data. As a result, we identify signatures of three pathways affected by B[a]P exposure, leukotriene metabolism, IL-7 and phosphatidylinositol signaling, known to play a prominent role in T cell development and function, that have not been reported before. Thus, our network-based approach not only provides complementary results to classical GSEA approaches or commercial tools like IPA, but also generates novel hypotheses on the molecular mechanisms by which benzo[a]pyrene and its derivatives mediate their effects. A more detailed biological analysis of the obtained results was left for future projects with our collaboration partners.

Chapter 3

The human blood metabolome-transcriptome interface

3.1 Background

In the previous chapter, we demonstrated how a knowledge-based approach can be used to integratively analyze multiple high-throughput data types at a cellular level. Based on a genome-scale metabolic network, we investigated the relationship between protein expression (proteomics) and metabolite levels (metabolomics) of *in vitro* blood cell cultures across varying conditions. The main advantage of such an integration strategy is that a knowledge-guided qualitative model allows the integration of data sets and types measured in different sets of patients or samples with equal experimental conditions (i.e. all patients/samples have the same treatment or phenotype). However, due to the shortage of metabolite-gene/(protein) interaction annotations in existing databases and the limited amount of metabolites with known biochemical identity, the results achieved by this approach are limited and biased by current knowledge [158].

A possibility to overcome this limitation are data-driven approaches (cf. Chapter 1.3.2), which are not dependent on existing knowledge, thus enabling the investigation of yet unknown metabolite-gene interactions even for metabolites with missing annotations or unknown identity. In case of transcriptomics data, an established framework to

systematically investigate the constituents of involved biological processes and their interactions are correlation network-based approaches, where pairwise statistical associations between molecular entities (nodes) are modeled as network edges (see Section 3.2 for more details). When particularly focusing on the blood system, several studies investigated the co-regulation of transcripts either from single white blood cell types or whole blood samples. For example, regulatory networks [182, 183] and global gene co-expression networks [184–186] were constructed from B- and T-cells to investigate pathways and mechanisms involved in the immune response. Further examples using whole blood data include the identification of disease-related gene networks [187, 188] or molecular signatures of distinct human vaccines captured in blood transcriptional modules [189]. Focusing on metabolomics measurements, several studies recently systematically characterized molecular interactions in the blood metabolome [20, 190–192].

However, only few studies with large sample sizes focused on an integrative, data-driven analysis of multiple omics datasets in human blood. One recent example is the work of Inouye et al. [193], who analyzed whole blood transcriptomics and genetics data in combination with blood lipid measurements and metabolites from a Finnish population cohort. In their study, the authors associated a module of highly co-expressed genes with 134 blood metabolic markers in the context of heart disorders and identified a link between the immune system and circulating metabolites. The study by Inouye et al. was among the first to provide clear evidence for this immune system link in blood, suggesting that gene expression in white blood cells is responsive to changing blood metabolite levels. Thus, it can be concluded that even if not cell-specific, the signals derived from whole blood data still reflect organism-wide processes. This is also in line with previous studies conducted on whole blood transcriptomics or metabolomics data separately [134, 194, 195].

In this chapter, we exploit the joint power of metabolomic and transcriptomic profiling to comprehensively characterize the complex interplay between serum metabolomics and whole-blood transcriptomics data in a systems genetics approach (cf. Chapter 1.3.3). While serum metabolomics represent a footprint of whole-body processes, blood transcriptomics data will mainly reflect immune system processes through white blood cells. We analyzed metabolomics and transcriptomics measurements of 712 individuals from the German population study KORA ("Kooperative Gesundheitsforschung in der Region Augsburg"), comprising 440 metabolites and 16,780 genes after filtering. We constructed a global correlation network to elucidate the complex interplay and regulation between these omics layers (Figure 3.1A). The correlation analysis takes advantage of the

naturally occurring variation from individual to individual, which we assume to carry a systematic footprint of the coregulation of metabolites and mRNAs. We deliberately left out an analysis of metabolite-metabolite and transcript-transcript correlations, which were rigorously investigated in the above-mentioned earlier studies. Instead, we specifically sought to assess the interconnection and information flow between the two omics layers.

The chapter is organized as follows: In the first part, we systematically characterize the blood metabolome-transcriptome interface (BMTI) using different strategies. First, we manually investigated the strongest associations and provide evidence from literature wherever possible. Moreover, using a Mendelian randomization (MR) approach, we examined potential causal relationships between metabolites and transcripts. Second, using the most recent genome-wide human metabolic network Recon 2 [82], we systematically analyzed correlations between metabolites and transcripts at pathway level (Figure 3.1B). Third, we developed a novel network clustering approach based on functional annotations, leading to a pathway interaction network (PIN) that allows for fast functional interpretation of the BMTI and furthermore provides insights into the cross-talk among distinct molecular pathways (Figure 3.1C). In the second part of this chapter, we demonstrate how the identified networks can be used as a resource to further investigate the link between metabolism and gene regulation by two different applications. First, we investigated whether a common regulatory signature is observable from transcripts connected to the same metabolite or to metabolites that are part of the same metabolic pathways. For this purpose, we analyzed promoter regions of the respective metabolite-associated genes for overrepresented transcription factor binding sites (Figure 3.1D). Second, we integrated the metabolome-transcriptome and the pathway interaction network with associations to high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C) and triglycerides (TG), which are well-known risk factors for cardiovascular disease [196]. To this end, we mapped the results of linear regressions between these clinical lipid parameters with metabolites and mRNAs onto the networks (Figure 3.1 E). Finally, we demonstrate the potential of our systems genetics approach to generate novel hypothesis by combining results from all separate analysis steps and establish an association between the branched-chain amino acid pathway and the levels of plasma TG and HDL-C.

All results reported in this chapter are part of the following publication:

⋆ **Bartel, J.***, Krumsiek, J.*, Schramm, K., Adamski, J., Gieger, C., Herder, C., Carstensen, M., Peters, A., Rathmann, W., Roden, M., Strauch, K., Suhre, K., Kastenmüller, G., Prokisch, H., and Theis, F.J. The human blood metabolome-transcriptome interface. *PLoS Genetics*, 11(6): e1005274, 2015

* = equal contributions

3.2 Methods

Population cohort and data acquisition

The Cooperative Health Research in the Region of Augsburg (KORA) study is a series of independent population-based epidemiological surveys and follow-up studies of participants living in the region of Augsburg, southern Germany [132, 197]. In this work, cross-sectional data from 712 participants of the KORA F4 population cohort was used for whom metabolite concentration, gene expression data and genotyping information were available. This subpopulation contains combined fasting serum metabolomics and whole blood transcriptomics measurements of 354 males and 358 females aged 62-77 years (mean 68.82 ± 4.31). All participants are residents of German nationality identified through the registration office and written informed consent was obtained from each participant. The study was approved by the local ethics committee (Bayerische Landesärztekammer). Detailed descriptions of blood sample acquisition and experimental procedures for the metabolomics and transcriptomics data, and clinical trait measurements can be found in [48, 198–200]. Briefly, metabolic profiling was performed by Metabolon, Inc. using ultra-high performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry. In total, 517 serum metabolites were measured, thereof 293 with known chemical identity and 224 unidentified metabolites ('unknowns'). All identified metabolites were assigned to one out of eight *superpathways* and one out of 61 *subpathways*, representing two different levels in the metabolic pathway classification hierarchy. Gene expression profiling was performed using total RNA extracts from whole blood samples on Illumina Human HT-12 v3 Expression BeadChips. Genotyping was carried out using the Affymetrix GeneChip array 6.0. A detailed description of the experimental procedures and preprocessing of the genetic data can be found in [48].
Replication of the significant metabolite-mRNA associations identified in the KORA

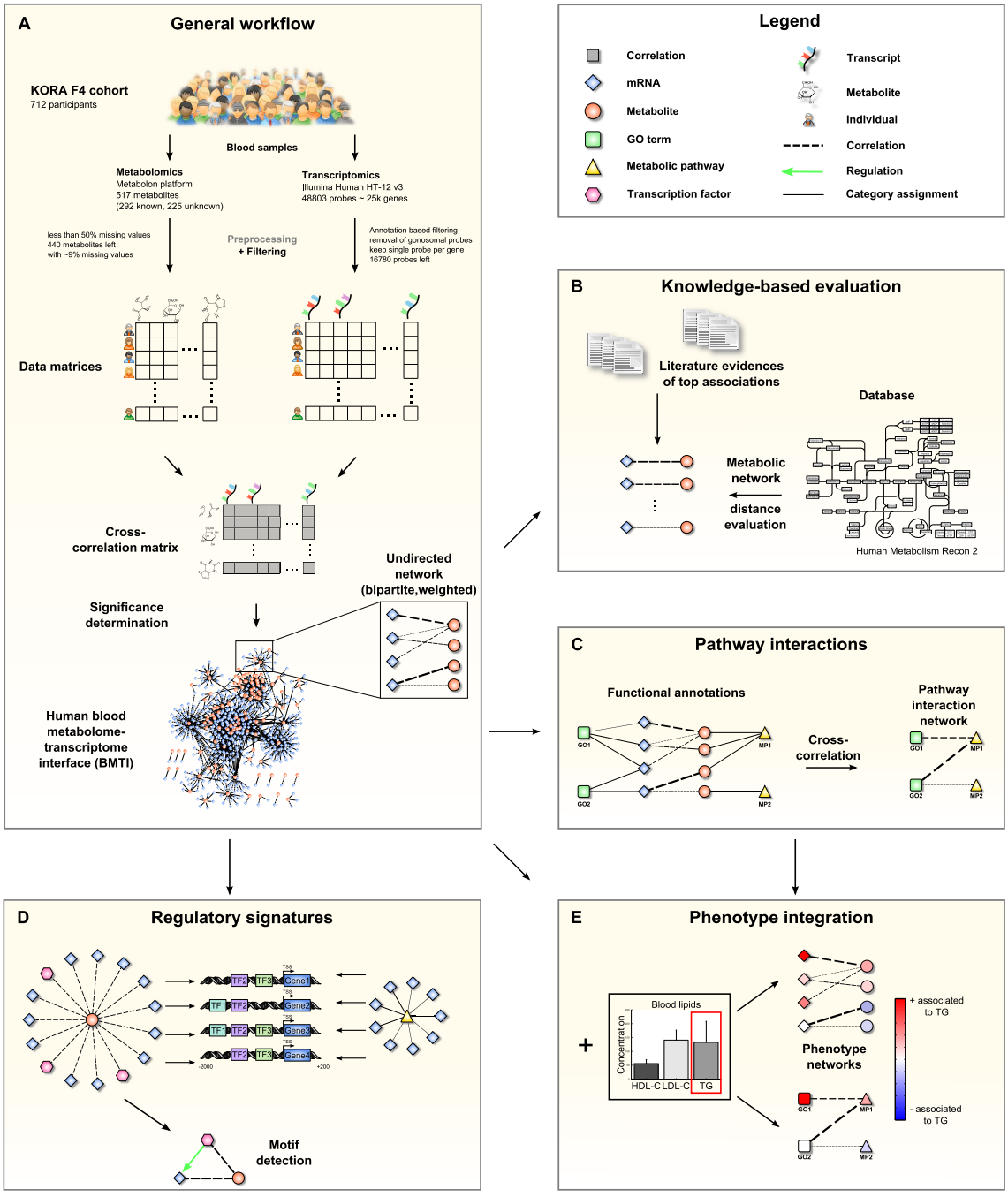


Figure 3.1: Data integration and network analysis workflow for the blood metabolome-transcriptome interface (BMTI). **A:** We analyzed fasting serum metabolomics and whole blood transcriptomics data from 712 samples of the KORA F4 cohort. After preprocessing and filtering, a cross-correlation matrix between 440 distinct metabolites and 16780 unique, gene-mapped probes was calculated. The correlation matrix was transformed into a bipartite network by applying a statistical significance threshold. **B:** Scientific literature was screened for biological evidence for the strongest metabolite-mRNA associations. All correlating metabolite-mRNA pairs contained in the human metabolic model Recon2 were systematically evaluated with respect to their distance in the metabolic pathway network. **C:** Aggregated z-scores for each functional annotation were calculated. A pathway interaction network (PIN) was then constructed via cross-correlation of scores between pairs of functional annotations. **D:** For each metabolite contained in the BMTI, we investigated the promoter regions of associated transcripts for shared regulatory signatures. Similarly, shared regulatory signatures within and between metabolic pathways were examined. As a final step, we identified specific regulatory motifs in the BMTI. **E:** Both BMTI and the PIN were integrated with the results from an association analysis to the three intermediate physiological traits (HDL-C, LDL-C and TG).

dataset was carried out with the Finish DILGOM cohort dataset which included whole blood NMR metabolomics data as well as transcriptomics data for 518 individuals. A detailed description of the sample acquisition as well as data preparation can be found in [188, 193].

Data preprocessing and quality control

To ensure data quality, metabolites with more than 50% missing values were excluded, leaving 440 metabolites (254 knowns and 186 unknowns) for further analysis (Figure 3.2). The remaining metabolite concentrations were log-transformed, since testing for normality indicated that for most cases the log-transformed concentrations were closer to a normal distribution than the untransformed values [113]. For gene expression arrays, quality control and imputation of missing values of the raw intensities was performed as described in [201]. Briefly, the initial preprocessing of the raw intensity data was done with GenomeStudio V2010.1. Raw probe level data was then imported to R and further preprocessed by log transformation and quantile normalization using the lumi package [202] from the Bioconductor open source software (`http://www.bioconductor.org`). To account for technical variation, gene expression intensities data were adjusted for RNA amplification batch, RNA integrity number and sample storage time. Only probes with the annotation flag *QC_COMMENT* "good" as provided in the updated Illumina Human HT-12 v3 BeadChip annotation file were considered for analysis [201]. In addition, probes mapping to gonosomal chromosomes were removed. Out of 48,803 probes on the Illumina Human HT-12 v3 array, 24,818 passed all of the filtering criteria.

Correlation network generation

The metabolite-transcript interface was constructed based on Spearmans correlation coefficients between the concentrations of all possible metabolite-transcript pairs ($24,818 \times 440$) across the individuals of the study cohort according to [119]:

$$r_{xy} = 1 - \frac{6 \sum_{i=1}^{n} \left(Rk\left(x\right)_i - Rk\left(y\right)_i\right)^2}{n\left(n^2 - 1\right)}, \tag{3.1}$$

where Rk is a ranking function converting the raw values of x_i and y_i into corresponding ranks. Correlation calculation was performed separately for each variable pair, only

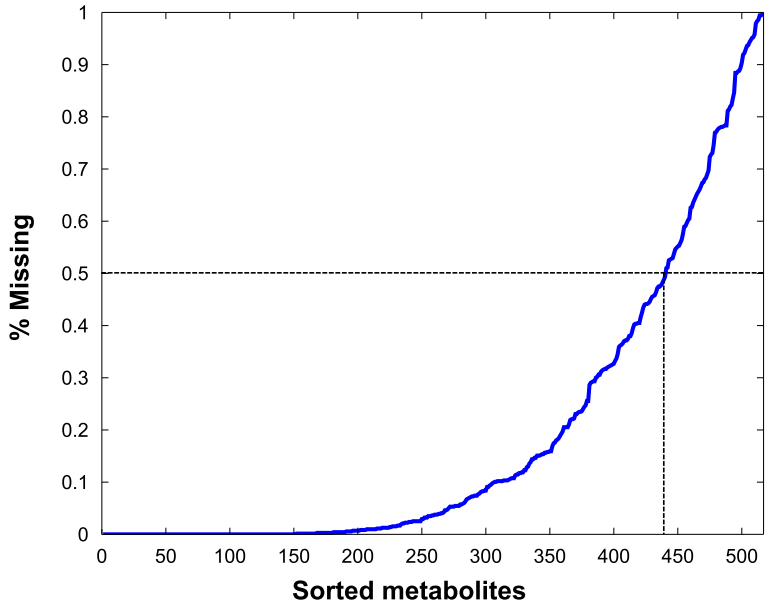


Figure 3.2: Missing values in metabolomics data. The x-axis represents the 517 measured metabolites sorted by percentage of missing values in ascending order. Dashed lines represent the chosen cutoff of $< 50\%$ missing values across the 712 samples.

considering samples without missing values for the metabolites. The Spearman correlation measure was used to account for possible non-linear associations and to ensure robustness against outliers. Note that for this particular dataset, Spearman and Pearson correlations produced almost identical results (3.4A). Moreover, we decided not to use partial correlations, since from a biological perspective, we expect to observe mainly indirect interactions and because of the high dimensionality of the transcriptomics data (cf. Chapter 1.3.2). Statistical significance of correlations was determined at an FDR controlled by the Benjamini-Hochberg procedure at $\alpha = 0.01$ [26]:

1. sort p-values such that $P_1 \leq P_2 \leq \cdots \leq P_m$
2. $R_{BH} = \max\left\{0 \leq i \leq m : P_i \leq \frac{i}{m}\alpha\right\}$
3. Reject H_0 for every test where $P_j \leq P_{R_{BH}}$

where m is the total number of tests, R_{BH} corresponds to the largest i for which the null hypothesis H_0 can be rejected. Applying this corresponded to an absolute correlation value of 0.1816 and an adjusted significance level of 1.07×10^{-06} in our data. To get a unique network node per gene, redundant probes matching the same gene were

removed. One representative probe per gene was chosen based on the maximum correlation strength to any metabolite, leaving 16,781 unique probes for subsequent analysis. It has to be noted that the applied significance level was still calculated on the whole dataset (including multiple matching probes per gene) to properly account for multiple testing. Network density was calculated as described in [203]. Briefly, for a stepwise increasing correlation threshold, the ratio between the total number of observed edges and all possible edges was calculated. Significant correlations between metabolites and transcripts were visualized as a bipartite graph using yEd graph editor (yWorks GmbH, Tuebingen; `http://www.yworks.com`).

Tissue/Cell-type specificity

BMTI genes were mapped to three published lists of tissue- and cell-specificity based on gene expression microarrays from purified cells or tissues. The first two marker gene lists were taken from Palmer et al. [204], who defined markers for B-cells, CD4+ T-cells and CD8+ T-cells, lymphocytes and granulocytes, and from the HaemAtlas as generated by Watkins et al. [205], who reported markers for CD19+ B-cells, CD4+ T-cells and CD8+ T-cells, CD14+ monocytes, CD56+ NK cells, CD66b+ granulocytes, erythroblasts and megakaryocytes. The third marker list was downloaded from the CTen website: `http://www.influenza-x.org/~jshoemaker/cten/db_info.php` and comprised markers for 84 different human tissues/cell types [206].

Mendelian Randomization

Estimation of causal effects within the BMTI was performed using a Mendelian randomization (MR) approach [17] which we will shortly explain in the following. MR can be seen as an instrumental variable method that uses genetic variants to make causal inferences for the effect of modifiable (non-genetic) traits on an outcome of interest [207]. In epidemiology, a major aim is to identify causal relationships between certain risk factors and health or disease-related outcomes, for instance the relationship between smoking and lung cancer or alcohol consumption and liver cirrhosis. However, in many cases these relationships are caused by unobserved (unmeasured) confounders (e.g. other lifestyle factors, pharmacological treatments, etc.) which then can lead to wrong conclusions. The gold standard to avoid such confounded causal associations are randomized controlled trials (RCT), where thoroughly selected subjects are randomly assigned to

different (treatment) groups such that the only observable difference between them is intrinsic to the compared treatments. The drawbacks of RCTs are the high costs, resource intensity and ethical aspects rendering them infeasible in many cases. In these situations, a good alternative are instrumental variable (IV) methods, which make use of an *instrument* to estimate causal relationships even in the presence of (unmeasured) confounding factors.

In order to be a valid instrument, a variable has to meet the following requirements [207]:

- The IV Z is associated with the trait of interest X
- Z is independent of confounding factors U that influence the association of X and the outcome Y
- Z is independent of outcome Y given X and the confounding factors U

A graphical illustration of these requirements is depicted in Figure 3.3. Genetic variants are especially suitable to be used as instruments, since they are generally independent from confounders and also the direction of causation is typically from a genetic variant to a modifiable trait of interest which reduces the risk of reverse causation [17].

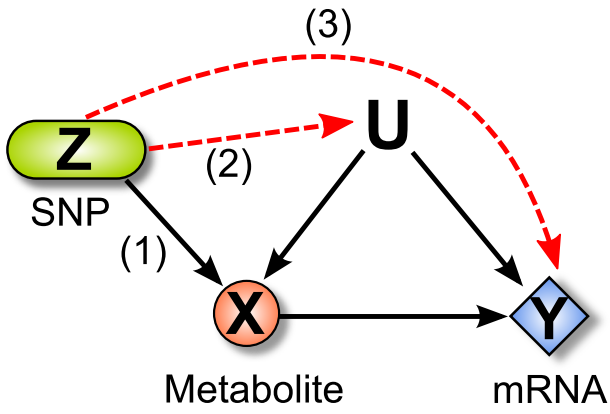


Figure 3.3: Graphical representation of Mendelian randomization. To test the causal relationship between two traits of interest **X** and **Y** in a MR approach, certain criteria have to be fulfilled: **(1)** The instrumental variable **Z**, in our case a SNP, is associated to a modifiable trait of interest **X**, e.g. the concentration of a metabolite. **(2)** The instrumental variable **Z** is independent of potential unmeasured confounders **U**. **(3)** The instrumental variable **Z** is dependent of the outcome of interest **Y**, here the expression of a transcript, only via its association with **X**.

In this study, a total of 224 candidate SNPs reported as lead association signals at genome-wide significance in two recent GWAS studies for 16 metabolites and 186 mRNAs (BMTI contained) were preselected as instrumental variables [113, 200]. To ensure the validity of the instrumental variables, only candidate SNPs that showed a significant association with a trait (metabolite or gene expression level) at an FDR of 0.05 in our data were considered for further analysis (32 SNPs were removed). Associations between SNPs and traits were assessed using linear regressions with age and sex as covariates. To further avoid potential confounding, all candidate SNPs were checked for pairwise linkage disequilibrium using the SNiPA tool [208]. None of the remaining 192 SNPs were in LD. Based on the metabolite-mRNA edges in the BMTI, 550 SNP-metabolite(Met)-mRNA and SNP-mRNA-Met sets were defined, covering 44% of all edges contained in the BMTI. Causal relationships SNP → Met → mRNA and SNP → mRNA → Met were estimated, i.e. whether changes in the metabolite level cause changes in the transcript level and vice versa. Causal effects of both models were calculated using the Wald ratio method [207]:

$$\hat{\beta}_{(\text{Met}\rightarrow\text{mRNA})} = \frac{\hat{\beta}_{(\text{SNP}\rightarrow\text{mRNA})}}{\hat{\beta}_{(\text{SNP}\rightarrow\text{Met})}} \quad \text{and} \quad \hat{\beta}_{(\text{mRNA}\rightarrow\text{Met})} = \frac{\hat{\beta}_{(\text{SNP}\rightarrow\text{Met})}}{\hat{\beta}_{(\text{SNP}\rightarrow\text{mRNA})}} \tag{3.2}$$

where $\hat{\beta}_{(\text{Met}\rightarrow\text{mRNA})}$ and $\hat{\beta}_{(\text{mRNA}\rightarrow\text{Met})}$ are the causal effects given the respective instrumental variable, and $\hat{\beta}_{(\text{SNP}\rightarrow\text{mRNA})}$ and $\hat{\beta}_{(\text{SNP}\rightarrow\text{Met})}$ are regression coefficients of the respective mRNA or metabolite levels on SNPs, under a simple linear model with age and sex as adjustment variables. 95% confidence intervals and p-values of the causal effects were calculated by sample bootstrapping with 10,000 repetitions. Q-values were calculated to control the false discovery rate (FDR). Summary information for the utilized SNPs together with detailed results of the MR approach can be found in the online supplementary of the original publication (`http://dx.doi.org/10.1371/journal.pgen.1005274`).

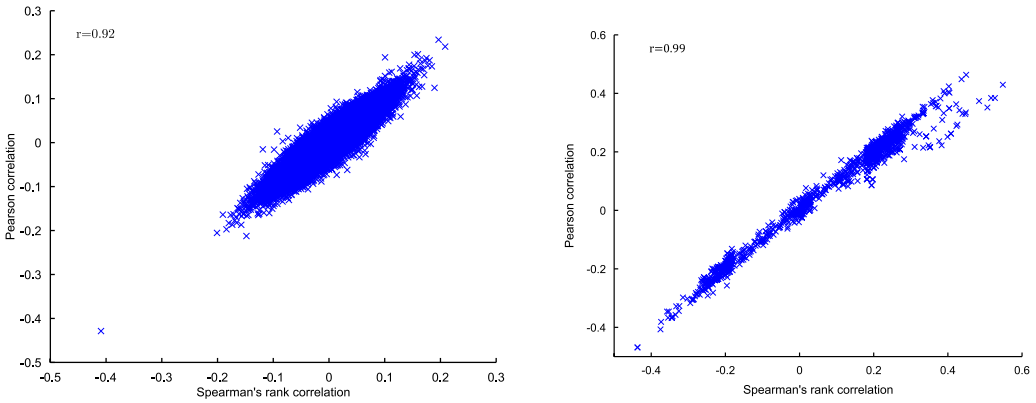
Metabolic pathway model and distance calculation

Metabolic reactions were extracted from the consensus metabolic reconstruction 'Recon 2' (version 2.02) available at `http://humanmetabolism.org` as of October 2013 [82]. Compartmental information was removed by merging shared nodes and reactions between different compartments. To avoid potential biologically meaningless shortcuts

between network nodes, co-factors and currency metabolites were excluded from the metabolic network prior to the distance calculation. Measured metabolites and transcripts were mapped onto the corresponding network nodes based on KEGG IDs or HMDB identifiers for metabolites, and Entrez Gene IDs for transcripts. Note that in contrast to the previous Chapter 2, we were only interested in the minimal distance between two measured entities within the metabolic network in order to evaluate if the calculated correlations reflect metabolic pathways, and not in enriching the most relevant subnetwork consisting of many alternative routes between measurements. Distances between all mapped pairs of metabolites and transcripts were defined as the shortest path in the network, i.e. the minimal number of reaction steps between them. For instance, a distance of zero between a transcript and metabolite indicates that the metabolite is a direct reactant of the reaction catalyzed by the particular enzyme encoded by the transcript. A distance of one indicates that the enzyme-encoding transcript catalyzes a directly connected reaction, which takes a product of the particular metabolite as input, and so on. A distance of infinity (Inf) was assigned if the respective metabolite and transcript were disconnected in the pathway network. Moreover, a 'not mapped' (NM) distance was assigned if either the metabolite or the transcript could not be mapped to Recon 2. Note that the network was treated as undirected, i.e. all reaction directions were ignored.

Annotations, aggregated z-scores and construction of pathway interaction network

Functional annotations were retrieved from two different sources. For transcripts, the generic GO Slim catalogue was downloaded from Gene Ontology (GO, `http://www.geneontology.org/GO.slims.shtml`). Generic GO Slim is a broad and non-redundant subset of the Gene Ontology consisting of 148 unique terms covering all three GO domains (cellular component, molecular function and biological process; [209]). The three root terms *cellular component*, *molecular function* and *biological process* and terms with no annotations for any of the 16,781 transcripts were removed, resulting in 140 terms for further analysis. For metabolites, the 'subpathway' annotations were used (see section 3.2). Metabolic pathways (MP) with less than two metabolites were excluded from the analysis, leaving 48 metabolic pathways. To aggregate the components belonging to a specific annotation term and to derive a score for each of these functional categories, the average of the associated z-score normalized gene expression profiles or metabolite concentrations was calculated according to



(A) All metabolite-mRNA pairs (B) Subpathway $aggZ$-Scores

Figure 3.4: Comparison of Spearman and Pearson correlations at single molecule and pathway level. **(A):** For all metabolite-mRNA pairs in the dataset, we observe a high agreement between the two measures indicating mostly linear relationships, with a correlation (of correlations) of r=0.92. Note that due to the high number of pairs ($440 \times 16,780$), only every 200th correlation of the ordered list of correlation coefficients is plotted. **(B):** Similarly, we observe a even higher concordance of Spearman and Pearson correlation coefficients for the pathway scores (r=0.99).

$$aggZC_j = \frac{1}{|C|} \sum_{i \in C} Z_{i,j} \tag{3.3}$$

where C corresponds to a metabolic pathway or GO term, i enumerates all members in this set, and $Z_{i,j}$ is the z-score of the gene/metabolite with index i in sample j. Spearmans rank cross-correlation between the $aggZ$-Scores of all possible GO-MP combinations was then calculated (note that Pearson correlation yielded similar results, see Figure 3.4B). To account for functionally distinct subbranches of biological processes and biochemically diverse molecules in metabolite classes fulfilling complementary tasks and/or controlled by mutual regulation, only those members of the two categories were considered for $aggZ$-Score calculations which share at least one mutual edge within the reconstructed network for the respective GO-MP combination (see Figure 3.5 for more details). Finally, significant associations between the functional annotation pairs were visualized as a bipartite pathway interaction network (PIN).

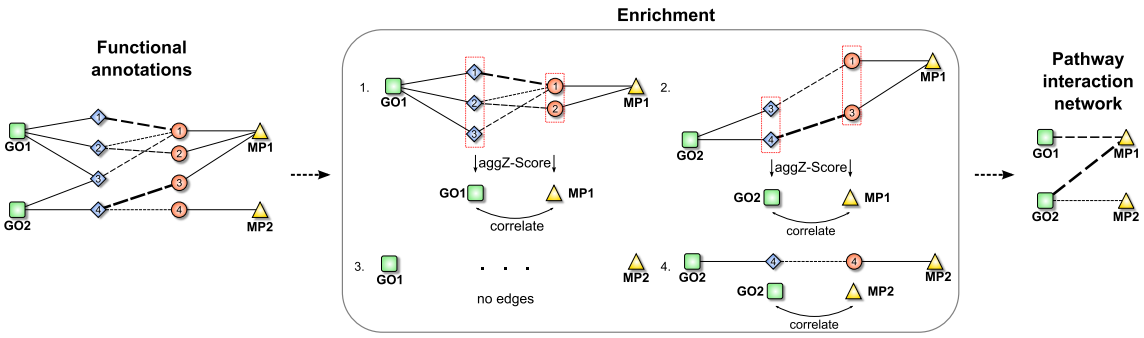


Figure 3.5: Pathway interaction analysis. The left-hand side of the flowchart shows an exemplary set of interactions between 4 transcripts and 4 metabolites, with annotations to two gene ontology terms and two metabolic pathways. For each GO-pathway pair we then determine the transcripts and metabolites which generate the connection between those two (middle panel). For example, GO1 and MP1 are connected through 3 transcripts and 2 metabolites, whereas GO1 and MP2 do not share any connection. Aggregated z-scores are then calculated for each pair based on the shared transcripts and metabolites to generate the pathway interaction network (right-hand side).

Phenotype analysis

Linear regression analysis was performed with age and sex as covariates:

$$y = \beta_0 + \beta_1 * x + \beta_2 * \text{age} + \beta_3 * \text{sex} + \epsilon \tag{3.4}$$

where y is the concentration of HDL, LDL or TG over all individuals, β_0 is the intercept, β_{1-3} are regression coefficients, x is a vector of expression/concentration values of a particular gene/metabolite and ϵ is a independent, normally distributed error term. In the same way, the association of annotations (GO and MP) was tested with all three phenotypic traits using the $aggZ$-Score for the particular annotation of x. Note that for this analysis, $aggZ$-Scores were calculated only on those members of a particular annotation that are also contained in the BMTI. Each network node was then color-coded with the $-\log_{10}(\text{p-value}) \times \text{sign}(\beta_1)$, where the p-value and β_1 were derived from the linear regression with the respective metabolite, gene or annotation category. To assess statistical significance of the determined associations, a Bonferroni-corrected threshold of $0.05/(16,780 \times 440) \approx 2.9 \times 10^{-6}$ was applied. Note that in contrast to the BMTI construction, we use this more stringent multiple testing correction to account for possible inflated p-values due to the BMTI edge-based selection.

Promoter analysis

To investigate regulatory signatures in the BMTI, an enrichment analysis of transcription factor binding sites was performed. Sets of input sequences were created from the neighbors of each metabolite with a degree 3 (at least 3 connected genes). Analogously, the pathway interaction network was used to construct sequence sets based on the neighborhood of a metabolic pathway node. For each set of input sequences, a separate search for overrepresented TFBS was performed with the sequences of all remaining genes as background model. Promoter regions (-2,000 bp to +200 bp relative to the TSS) and TSS positions of all genes were extracted from the UCSC database using the R package BSgenome.Hsapiens.UCSC.hg19 version 1.3.1. Position-specific weight matrices of the transcription factor binding motifs were taken from the vertebrate collection of the Jaspar database version 5.0 alpha [210]. Enrichment analysis was performed with the TFM-Explorer command line tool [211]. The p-value threshold to determine significance of the motifs in all input sets was set to 1.0×10^{-7} which lies in the recommended optimal range given the numbers of input sequences we used in this study (mean number of input sequences: 62) [212]. The authors showed that for a fixed false positive rate of 10%, the optimal p-value threshold was 1.0×10^{-7} for a dataset of 100 input sequences.

3.3 The human blood metabolome-transcriptome interface

We focused on a subset from the KORA F4 cohort with simultaneously available metabolomics and transcriptomics data. After quality control and filtering, the data set comprised 712 human blood samples (354 males, 358 females; see Table 3.1) with gene expression data for 16,780 uniquely mapping gene probes and concentrations of 440 metabolites (Figure 3.1A, see section 3.2 for details). 186 of these 440 metabolites were not chemically identified, which is marked by a metabolite name starting with X- throughout this chapter. Both gene expressions and metabolite concentrations were log transformed and adjusted for age and sex effects. Pairwise Spearmans rank correlations between the measured mRNAs and metabolites were then calculated.

Metabolite-mRNA Spearman correlation coefficients were symmetrically distributed at zero (mean:$-4.5 \times 10^{-4} \pm 0.0433$, Figure 3.6A) with a maximum absolute correlation value of $\rho = 0.56$. Moreover, the distribution of cross-omics correlations showed a rather narrow shape, indicating generally lower correlations when compared to the intra-omics correlations (mRNA-mRNA, metabolite-metabolite). The metabolite-metabolite

Variable	
N (Male/Female)	712 (354/358)
	Mean (sd)
Age (years)	68.82 (4.31)
BMI (kg/m^2)	28.87 (4.56)
HDL cholesterol (mg/dl)	55.80 (13.95)
LDL cholesterol (mg/dl)	140.60 (35.97)
Triacylglycerides (mg/dl)	132.64 (75.70)

Table 3.1: Characteristics of the KORA F4 study population. N, number of individuals; BMI, body mass index; HDL, high density lipoprotein; LDL, low density lipoprotein; sd, standard deviation.

distribution was strongly skewed for positive correlation values, which is in accordance with our previous findings on a different metabolomics panel [69]. In contrast, the mRNA-mRNA distribution displayed a broad and symmetric distribution of correlation values (Figure 3.6A).

We then generated a weighted bipartite network of metabolites and transcripts by constructing an edge between a metabolite and transcript pair if the respective correlation was significant with a false discovery rate (FDR) of 0.01. This corresponded to an absolute correlation cutoff of ~ 0.181 and a p-value threshold at 1.07×10^{-6}. Obviously, the number of edges in a correlation network heavily depends on the chosen threshold. It has been shown in previous studies that a biologically reasonable threshold can be found by investigating network density as a function of the correlation cutoff value [203]. According to that study, a cutoff value slightly above the minimal density combined with a decreasing number of nodes and edges leads to biologically meaningful results. As indicated in Figure 3.6B, a clear decline in the number of included nodes and edges can be observed for increasing correlation threshold levels beginning between correlation values of 0.15 and 0.25. Minimal network density was reached for a correlation threshold value between 0.13 and 0.18 (Figure 3.7). Notably, applying the above-mentioned conventional statistical significance threshold to our data set precisely coincides with the network density-based threshold described by [203].

The resulting network, subsequently called the blood metabolome-transcriptome interface (BMTI), consisted of 636 nodes (114 metabolites, 522 transcripts) and 1109 edges, corresponding to a total network connectivity of $\sim 0.015\%$ (Figure 3.6B+D). Out of

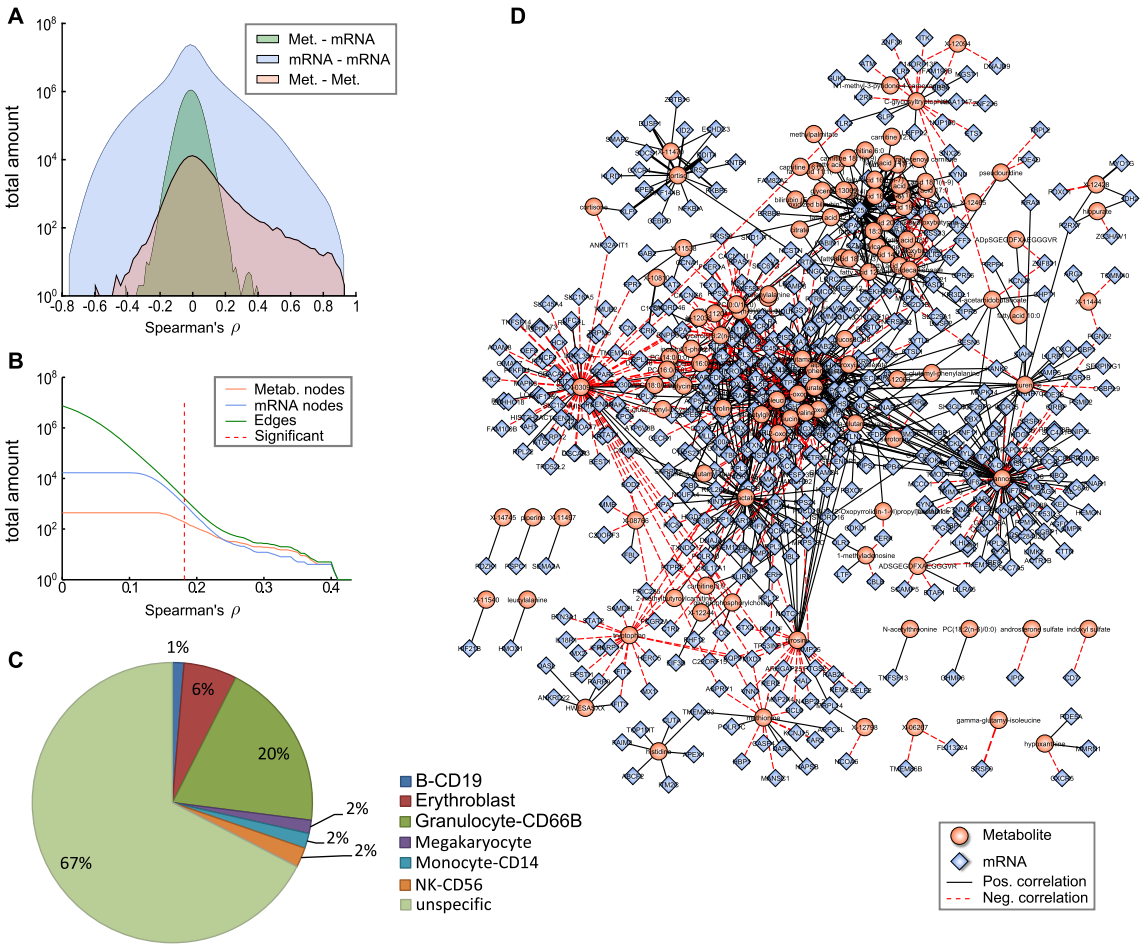


Figure 3.6: The blood metabolome-transcriptome interface. **A:** Distributions of correlation coefficients for metabolite-mRNA, mRNA-mRNA and metabolite-metabolite associations. **B:** Number of nodes and edges as a function of the absolute correlation coefficient. Red dotted line represents the correlation cutoff used in this study (0.01 FDR). **C:** Percentages of blood cell type specific transcript markers contained in the BMTI. **D:** Visualization of the blood metabolome-transcriptome interface. The correlation network consists of 114 metabolites and 522 transcripts connected by 1,109 edges. Edge widths represent correlation strengths.

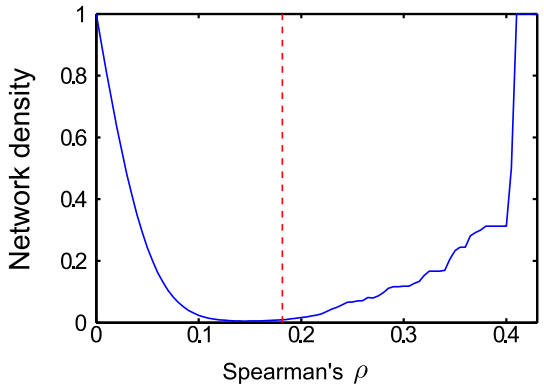


Figure 3.7: Network density. The fraction of edges above cutoff against all possible edges, plotted as a function of the absolute correlation coefficient. Red dotted line represents the correlation cutoff used in this study (0.01 FDR).

the total number of edges, 63% (699) were positive correlations and 37% (410) were negative correlations. The metabolite showing the highest degree was mannose, with significant correlations to 98 different transcripts. In contrast, the mRNA with the highest connectivity was SLC25A20 with 37 metabolites attached. We used data from the DILGOM study, which included NMR metabolomics as well as transcriptomics data for 518 individuals, for independent replication of our correlations. In total, 17 metabolites (11 amino acids, 3 lipids, 2 carbohydrates and 1 belonging to the energy metabolism) overlapped between the KORA F4 dataset and the DILGOM study, which allowed us to investigate the replication of 211 edges ($\sim$ 19% of the BMTI). 61 out of the 211 edges ($\sim$ 29%) reached a nominal significance (p-value < 0.05) in the DILGOM study of which 38 ($\sim$ 18%) remained significant after multiple testing correction (FDR < 0.05, see Table 3.2).

To investigate the possible origins of the metabolite-transcript correlations, we compared all genes represented in the BMTI with 1) two a priori defined blood cell type-specific marker gene lists, and 2) a database of more general tissue gene expression signatures (see section 3.2). For the first part, we used a list of genes derived from Palmer et al. [204] comprising 907 specifically expressed genes for 5 different blood cell types (leukocytes only) and a second list derived from the *HaemAtlas* as generated by Watkins et al. [205] comprising 1,716 genes characterizing 9 different blood cell types. For the second part, we used the *HECS database* from Shoemaker et al. [206] containing information for more than 6,000 genes and 84 tissues. Both comparisons in 1) showed that most of the BMTI genes (85% and 67%, respectively) were not specifically attributable to

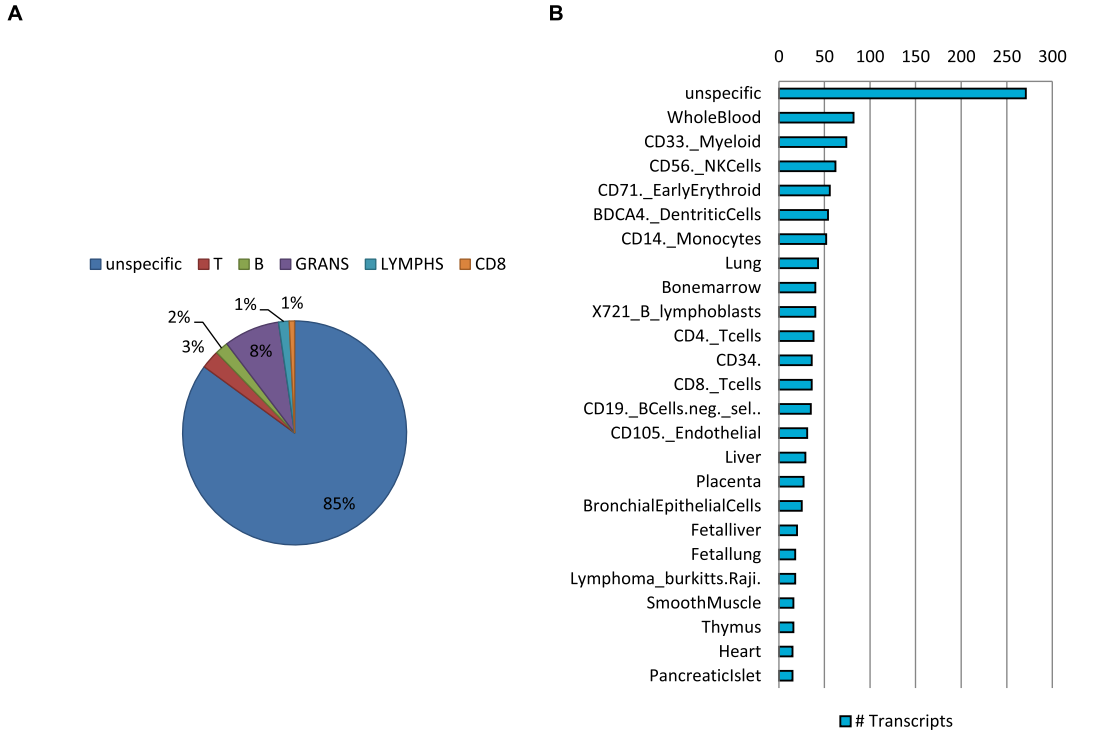


Figure 3.8: Cell type specific origins of BMTI transcripts. **A:** Comparison of BMTI transcripts with blood cell-type specific markers as identified by Palmer et al. [204] and **B:** to tissue/cell-type specific markers contained in the HECS db [206].

any blood cell type (see Figure 3.6C and 3.8A). The remaining genes could be assigned to the respective blood cell types contained in the reference lists, with granulocytes making up the largest blood cell faction in both cases (8% and 20%, respectively) and only minor signals for the other blood cell types. A similar result was observed when comparing the BMTI genes to the HECS database. 52% of the BMTI genes showed no tissue specificity, while 12 out of the 15 strongest tissue signatures were either blood cells or blood-related tissues (Figure 3.8B).

Metabolite	mRNA	KORA F4 (n=712)		DILGOM (n=518)	
		Spearman's ρ	P-value	Spearman's ρ	P-value
3-hydroxybutyrate	SLC25A20	0.35	$1.7 \cdot 10^{-22}$	0.37	$1.4 \cdot 10^{-17}$
leucine	ABCG1	-0.28	$6.7 \cdot 10^{-14}$	-0.15	$5.6 \cdot 10^{-04}$
valine	ABCG1	-0.27	$9.8 \cdot 10^{-14}$	-0.13	$4.9 \cdot 10^{-03}$
isoleucine	SLC45A3	-0.24	$6.0 \cdot 10^{-11}$	-0.31	$1.4 \cdot 10^{-12}$
isoleucine	ABCA1	-0.24	$7.5 \cdot 10^{-11}$	-0.21	$2.0 \cdot 10^{-06}$
isoleucine	AVPI1	-0.24	$1.6 \cdot 10^{-10}$	-0.12	$7.4 \cdot 10^{-03}$
lactate	S100A8	0.23	$2.4 \cdot 10^{-10}$	0.22	$6.8 \cdot 10^{-07}$
citrate	SLC25A20	0.23	$4.3 \cdot 10^{-10}$	0.32	$3.3 \cdot 10^{-13}$
tyrosine	BCL6	-0.22	$1.2 \cdot 10^{-09}$	-0.15	$6.4 \cdot 10^{-04}$
lactate	MMP8	0.22	$2.7 \cdot 10^{-09}$	0.14	$2.1 \cdot 10^{-03}$
histidine	CUTA	0.22	$4.9 \cdot 10^{-09}$	0.13	$4.8 \cdot 10^{-03}$
isoleucine	HSH2D	-0.21	$8.5 \cdot 10^{-09}$	-0.16	$5.1 \cdot 10^{-04}$
tyrosine	RERE	-0.21	$9.8 \cdot 10^{-09}$	-0.13	$2.7 \cdot 10^{-03}$
tyrosine	AQP9	-0.21	$1.0 \cdot 10^{-08}$	-0.12	$8.1 \cdot 10^{-03}$
isoleucine	NEU1	-0.21	$1.7 \cdot 10^{-08}$	-0.12	$7.4 \cdot 10^{-03}$
3-hydroxybutyrate	CPT1A	0.21	$2.4 \cdot 10^{-08}$	0.25	$8.6 \cdot 10^{-09}$
leucine	SLC45A3	-0.21	$2.7 \cdot 10^{-08}$	-0.21	$1.6 \cdot 10^{-06}$
isoleucine	HDC	-0.21	$3.0 \cdot 10^{-08}$	-0.34	$7.4 \cdot 10^{-15}$
isoleucine	TOMM7	0.21	$3.2 \cdot 10^{-08}$	0.15	$6.0 \cdot 10^{-04}$
isoleucine	C1ORF186	-0.2	$3.5 \cdot 10^{-08}$	-0.29	$2.3 \cdot 10^{-11}$
lactate	ABCG1	-0.2	$3.9 \cdot 10^{-08}$	0.13	$3.9 \cdot 10^{-03}$
leucine	AVPI1	-0.2	$4.4 \cdot 10^{-08}$	-0.13	$3.8 \cdot 10^{-03}$
3-hydroxybutyrate	PRSS33	-0.2	$4.9 \cdot 10^{-08}$	-0.14	$2.4 \cdot 10^{-03}$
isoleucine	RAB11FIP1	-0.2	$5.1 \cdot 10^{-08}$	0.14	$1.6 \cdot 10^{-03}$
leucine	ABCA1	-0.2	$5.7 \cdot 10^{-08}$	-0.17	$1.6 \cdot 10^{-04}$
3-hydroxybutyrate	PDK4	0.2	$1.2 \cdot 10^{-07}$	0.29	$7.2 \cdot 10^{-11}$
tyrosine	REM2	-0.19	$1.8 \cdot 10^{-07}$	-0.13	$3.7 \cdot 10^{-03}$
lactate	SAR1B	0.19	$1.9 \cdot 10^{-07}$	0.15	$8.4 \cdot 10^{-04}$
3-hydroxybutyrate	ABCA1	0.19	$2.1 \cdot 10^{-07}$	0.25	$8.5 \cdot 10^{-09}$
tyrosine	VNN3	-0.19	$3.1 \cdot 10^{-07}$	-0.13	$4.2 \cdot 10^{-03}$
lactate	PDCD10	0.19	$3.8 \cdot 10^{-07}$	0.12	$8.3 \cdot 10^{-03}$
histidine	APEX1	0.19	$3.9 \cdot 10^{-07}$	0.18	$8.7 \cdot 10^{-05}$
leucine	HSH2D	-0.19	$5.3 \cdot 10^{-07}$	-0.16	$3.5 \cdot 10^{-04}$
lactate	HIGD1A	0.19	$6.2 \cdot 10^{-07}$	0.12	$8.5 \cdot 10^{-03}$
histidine	FAIM3	0.18	$8.3 \cdot 10^{-07}$	0.16	$2.9 \cdot 10^{-04}$
isoleucine	TYROBP	-0.18	$8.4 \cdot 10^{-07}$	-0.15	$9.8 \cdot 10^{-04}$
leucine	HDC	-0.18	$8.9 \cdot 10^{-07}$	-0.23	$2.4 \cdot 10^{-07}$
tyrosine	RAB24	-0.18	$9.6 \cdot 10^{-07}$	-0.15	$8.0 \cdot 10^{-04}$

Table 3.2: Replication of metabolite-mRNA correlations. Shown are only significantly reproducible metabolite-mRNA associations (FDR < 0.05).

Metabolite	mRNA	Spearman's ρ	p-value
cortisol	DDIT4	0.56	$7.71 \cdot 10^{-59}$
1-oleoylglycerol (1-monoolein) | Glycerol(18:1(n-9)/0:0/0:0)	HDC	-0.41	$2.01 \cdot 10^{-29}$
oleate181n9 | oleate (18:1n9)	SLC25A20	0.40	$2.30 \cdot 10^{-29}$
palmitate (16:0) | fatty acid 16:0	SLC25A20	0.40	$3.96 \cdot 10^{-29}$
1-oleoylglycerol (1-monoolein) | Glycerol(18:1(n-9)/0:0/0:0)	SLC45A3	-0.40	$3.59 \cdot 10^{-28}$
dihomolinoleate202n6 | dihomo-linoleate (20:2n6)	SLC25A20	0.38	$8.63 \cdot 10^{-26}$
linoleate (18:2n6) | fatty acid 18:2(n-6)	SLC25A20	0.37	$1.55 \cdot 10^{-24}$
stearate (18:0) | fatty acid 18:0	SLC25A20	0.37	$2.02 \cdot 10^{-24}$
eicosenoate (20:1n9 or 11) | fatty acid 20:1(n-9/n-11)	SLC25A20	0.37	$2.67 \cdot 10^{-24}$
10-nonadecenoate (19:1n9) | fatty acid 19:1(n-9)	SLC25A20	0.36	$9.07 \cdot 10^{-24}$
cortisol	SOCS1	0.36	$2.19 \cdot 10^{-23}$
3-hydroxybutyrate (BHBA)	SLC25A20	0.35	$1.77 \cdot 10^{-22}$
5,8-tetradecadienoate	SLC25A20	0.35	$3.85 \cdot 10^{-22}$
palmitoleate (16:1n7) | fatty acid 16:1(n-7)	SLC25A20	0.35	$4.60 \cdot 10^{-22}$
10-heptadecenoate (17:1n7) | fatty acid 17:1(n-7)	SLC25A20	0.35	$7.87 \cdot 10^{-22}$
1-oleoylglycerol (1-monoolein) | Glycerol(18:1(n-9)/0:0/0:0)	GATA2	-0.35	$2.63 \cdot 10^{-21}$
cortisol	KLF9	0.34	$5.21 \cdot 10^{-21}$
linolenate [α or γ; (18:3n3 or 6)] | fatty acid 18:3(n-3/n-6)	SLC25A20	0.34	$5.13 \cdot 10^{-21}$
cortisol	DUSP1	0.34	$7.59 \cdot 10^{-21}$
margarate (17:0) | fatty acid 17:0	SLC25A20	0.34	$3.50 \cdot 10^{-20}$
myristate (14:0) | fatty acid 14:0	SLC25A20	0.32	$8.05 \cdot 10^{-19}$
1-oleoylglycerol (1-monoolein) | Glycerol(18:1(n-9)/0:0/0:0)	C1ORF186	-0.32	$4.72 \cdot 10^{-18}$
5-dodecenoate (12:1n7) | fatty acid 12:0(n-7)	SLC25A20	0.32	$3.74 \cdot 10^{-18}$
isoleucine	ABCG1	-0.32	$3.39 \cdot 10^{-18}$

Table 3.3: Top 25 network edges (i.e. strongest correlation coefficients). We observed particularly strong effects for lipid metabolism, especially around the mitochondrial transporter SLC25A20. Pipe symbols |separate alternative metabolite names.

3.4 Strong network edges in the BMTI represent known pathway mechanisms

As a first step to characterizing the BMTI, we performed a manual literature look up for the strongest absolute correlations in the network (Figure 3.1B). In the following, we provide a detailed discussion of the 25 strongest edges (Table 3.3). Notably, most of the top 25 identified associations reflect biochemically reasonable interactions like transport processes of lipids, but also regulatory signatures between signaling metabolites and transcription factors.

The strongest association in the dataset was observed between cortisol and *DNA-Damage-Inducible Transcript 4* (DDIT4, $\rho = 0.56$, p-value $= 7.70 \times 10^{-59}$), which are known to play a role in stress response [213]. Cortisol is a glucocorticoid whose release is mainly induced by exogenous stress. Via binding to the *glucocorticoid nuclear receptor*

(GR, official gene symbol NR3C1), it regulates various cellular processes like carbohydrate metabolism and the immune system by direct activation of target genes [214]. Remarkably, DDIT4 was identified as a GR target gene in mouse hepatocytes [215], rat hippocampus [216] and also in human peripheral blood lymphocytes [217] delivering a potential explanation of an indirect association for the observed correlation. Another GR target gene associated to cortisol is *Suppressor Of Cytokine Signaling 1* (SOCS1, $\rho = 0.36$, p-value $= 2.19 \times 10^{-23}$), a major constituent of the cytokine signaling pathway and inflammatory response [218]. We observed further top 25 correlations involving cortisol for *Kruppel-Like Factor 9* (KLF9) and *Dual Specificity Phosphatase 1* (DUSP1) ($\rho = 0.34$, p-value $= 5.20 \times 10^{-21}$; $\rho = 0.34$, p-value $= 7.58 \times 10^{-21}$, respectively). KLF9 is a ubiquitously expressed transcription factor involved in the regulation of diverse biological processes like cell development and differentiation in adipogenesis [219]. DUSP1 is an enzyme involved in the response to environmental stress [220]. Interestingly, for both transcripts, a cortisol-dependent regulation was already observed in epidermal cells [221] and peripheral blood mononuclear cells [222].

Another metabolite showing several strong associations to blood transcripts was 1-monoolein, which belongs to the class of monoacylgylcerols. This particular class of metabolites are bioactive compounds recently identified to be involved in various signaling processes of the immune system [223, 224]. The source of 1-monolein in humans is not fully understood. Experiments in rodents suggest that dietary 1,3-diacylglycerols are preferentially digested to 1-monoacylglycerols and free fatty acid in the small intestine, making dietary 1,3-diacylglycerols containing an oleoyl moiety at position sn-1 or sn-3 a plausible source of 1-monoolein [225]. In our analysis, 1-monoolein showed a strong negative correlation to four transcripts *GATA Binding Protein 2* (GATA2), *Histidine Decarboxylase* (HDC), *Solute Carrier Family 45, Member 3* (SLC45A3) and *Chromosome 1 Open Reading Frame 186* (C1ORF186) (ρ between -0.41 and -0.32, p-values between 2.01×10^{-29} and 4.72×10^{-18}). HDC is a cytosolic enzyme that catalyzes the conversion of histidine to histamine and thus represents an important immune system trigger molecule [226]. In addition, GATA2, a key regulator of gene expression in hematopoietic cells [227], C1ORF186 and SLC45A3, two membrane-bound proteins, were all identified to play a role in the immune response [188].

Carnitine-acylcarnitine translocase (SLC25A20) occurred in 15 of the 25 top ranked correlations. This gene encodes an enzyme which transports acylcarnitines, i.e. the transport variant of fatty acids, into the mitochondria for subsequent β-oxidation. Interestingly, the majority of SLC25A20-associated metabolites among our top 25 cor-

relations belonged to the class of long chain fatty acids (11 long chain fatty acids, 2 essential fatty acids, 1 medium chain fatty acid, 1 ketone body), which is in accordance with its function as a lipid transporter. Of note, among the metabolites associated with SLC25A20 beyond the top 25 correlations were also 5 acylcarnitines, although at lower correlation values.

We observed a significant, negative correlation between isoleucine and *ATP-Binding Cassette Sub-Family G Member 1* (ABCG1, $\rho = -0.32$, p-value $= 3.39 \times 10^{-18}$). It has been shown previously that circulating levels of branched-chain amino acids (BCAAs) affect a variety of metabolic processes such as glucose and lipid metabolism [228]. ABCG1 is a major player of lipid metabolism, controlling the transfer of cholesterol from peripheral macrophages to exogenous HDL [229]. Interestingly, an association between circulating BCAA levels and plasma HDL-C levels was also observed in a recent population study [230] and in a previous paper on the same cohort used in the present study [63].

3.5 Causality analysis of BMTI edges

To assess whether metabolite-transcript links in the BMTI contain causal effects, we performed a Mendelian randomization analysis [17]. For each metabolite-mRNA edge, we tested both the causal directions metabolite $\rightarrow$ mRNA and mRNA $\rightarrow$ metabolite given that adequate instrumental variables were available. As instruments we used SNP lists from previously published GWAS studies (see section 3.2 for details). After filtering for strong instrumental variables, we were left with 15 SNPs identified by a metabolomics GWAS study [113] associated to 16 metabolites in the BMTI. Moreover, for 157 mRNAs in the network, we selected 192 SNPs from Schramm et al. [200]. In total, we tested the causal relationship of 440 BMTI edges ($\sim$ 40%) of which 60 could be tested bi-directionally. In the BMTI, 42 metabolite-mRNA pairs (19 mRNA $\rightarrow$ metabolite; 23 metabolite $\rightarrow$ mRNA) showed a nominally significant causal effect (p-value < 0.05). At an FDR of 0.05, none of the tested pairs remained significant (Table 3.4).

SNP	Metabolite	Causal dir.	mRNA	β Met.	P-value Met.	β mRNA	P-value mRNA	$\hat{\beta}$	P-value $\hat{\beta}$	95% CI	q-value
rs1126639	dihomolinoleate202n6	$\leftarrow$	GZMB	-0.050	$2.71 \cdot 10^{-2}$	-0.430	$1.17 \cdot 10^{-31}$	0.118	$3.34 \cdot 10^{-2}$	(0.01, 0.23)	0.622
rs6465469	hexanoylcarnitine	$\leftarrow$	PDK4	-0.044	$2.80 \cdot 10^{-2}$	-0.114	$3.31 \cdot 10^{-6}$	0.3876	$2.60 \cdot 10^{-2}$	(0.05, 0.83)	0.613
rs6465469	laurylcarnitine	$\leftarrow$	PDK4	-0.069	$1.44 \cdot 10^{-2}$	-0.118	$2.82 \cdot 10^{-6}$	0.588	$7.20 \cdot 10^{-3}$	(0.16, 1.18)	0.613
rs482704	glycerophosphorylcholine	$\leftarrow$	POLR1D	0.043	$5.01 \cdot 10^{-2}$	-0.188	$2.94 \cdot 10^{-27}$	-0.232	$3.62 \cdot 10^{-2}$	(-0.46, -0.02)	0.622
rs227634	X-08766	$\leftarrow$	C20ORF3	0.079	$4.36 \cdot 10^{-2}$	-0.168	$4.13 \cdot 10^{-6}$	-0.471	$3.94 \cdot 10^{-2}$	(-0.99, -0.03)	0.622
rs17685810	4-hydroxyphenyllactate	$\leftarrow$	FECH	0.054	$1.26 \cdot 10^{-2}$	0.319	$4.18 \cdot 10^{-10}$	0.171	$5.40 \cdot 10^{-3}$	(0.05, 0.33)	0.613
rs12039959	tyrosine	$\leftarrow$	ITLN1	-0.028	$2.08 \cdot 10^{-2}$	-0.312	$5.85 \cdot 10^{-14}$	0.091	$2.90 \cdot 10^{-2}$	(0.01, 0.18)	0.613
rs2171585	3-methyl-2-oxopentanoate	$\leftarrow$	KRT1	0.028	$1.73 \cdot 10^{-2}$	0.780	$1.45 \cdot 10^{-41}$	0.036	$1.48 \cdot 10^{-2}$	(0.01, 0.07)	0.613
rs17134635	laurate120	$\leftarrow$	AKR1C3	-0.029	$6.35 \cdot 10^{-2}$	-0.142	$2.74 \cdot 10^{-5}$	0.206	$4.54 \cdot 10^{-2}$	(0.01, 0.47)	0.622
rs2695282	4-acetamidobutanoate	$\leftarrow$	CHPT1	-0.029	$6.42 \cdot 10^{-2}$	-0.229	$1.32 \cdot 10^{-16}$	0.127	$4.94 \cdot 10^{-2}$	(0.00, 0.26)	0.622
rs2545984	Cglycosyltryptophan	$\leftarrow$	ZNF30	0.028	$5.98 \cdot 10^{-2}$	-0.069	$3.33 \cdot 10^{-5}$	-0.403	$4.74 \cdot 10^{-2}$	(-0.92, -0.00)	0.622
rs1809049	X-03094	$\leftarrow$	ATG7	-0.049	$1.35 \cdot 10^{-2}$	-0.068	$1.81 \cdot 10^{-3}$	0.719	$2.58 \cdot 10^{-2}$	(0.09, 2.50)	0.613
rs3802542	4-hydroxyphenyllactate	$\leftarrow$	MARCH8	0.060	$2.15 \cdot 10^{-2}$	0.300	$1.22 \cdot 10^{-7}$	0.201	$1.76 \cdot 10^{-2}$	(0.04, 0.41)	0.613
rs1063603	mannose	$\leftarrow$	CGGBP1	0.055	$1.32 \cdot 10^{-2}$	0.063	$1.52 \cdot 10^{-4}$	0.870	$1.38 \cdot 10^{-2}$	(0.21, 1.86)	0.613
rs4802828	mannose	$\leftarrow$	SIGLEC5	-0.059	$1.81 \cdot 10^{-2}$	-0.351	$7.70 \cdot 10^{-18}$	0.169	$1.38 \cdot 10^{-2}$	(0.04, 0.31)	0.613
rs2274611	alphahydroxyisovalerate	$\leftarrow$	CTSL1	-0.045	$6.08 \cdot 10^{-2}$	-0.176	$4.98 \cdot 10^{-15}$	0.256	$4.72 \cdot 10^{-2}$	(0.00, 0.52)	0.622
rs5747027	proline	$\leftarrow$	CECR1	-0.033	$1.07 \cdot 10^{-2}$	0.064	$1.70 \cdot 10^{-4}$	-0.526	$1.00 \cdot 10^{-2}$	(-1.19, -0.14)	0.613
rs2695284	@4acetamidobutanoate	$\leftarrow$	CHPT1	-0.030	$5.16 \cdot 10^{-2}$	-0.238	$7.34 \cdot 10^{-18}$	0.128	$3.86 \cdot 10^{-2}$	(0.01, 0.25)	0.622
rs636049	@1linoleoylglycerol1monolinolein	$\leftarrow$	MRPL21	-0.061	$2.88 \cdot 10^{-2}$	-0.180	$1.01 \cdot 10^{-22}$	0.341	$2.60 \cdot 10^{-2}$	(0.04, 0.64)	0.613
rs780094	mannose	$\rightarrow$	ACTR1B	-0.102	$1.95 \cdot 10^{-9}$	0.029	$2.92 \cdot 10^{-2}$	-0.285	$2.42 \cdot 10^{-2}$	(-0.55, -0.04)	0.613
rs780094	mannose	$\rightarrow$	AGFG1	-0.102	$1.95 \cdot 10^{-9}$	-0.041	$1.42 \cdot 10^{-2}$	0.408	$1.38 \cdot 10^{-2}$	(0.09, 0.77)	0.613
rs780094	mannose	$\rightarrow$	CCR2	-0.102	$1.95 \cdot 10^{-9}$	-0.050	$2.26 \cdot 10^{-2}$	0.489	$2.64 \cdot 10^{-2}$	(0.06, 0.97)	0.613
rs780094	mannose	$\rightarrow$	CDC42	-0.102	$1.95 \cdot 10^{-9}$	-0.036	$3.89 \cdot 10^{-2}$	0.351	$3.48 \cdot 10^{-2}$	(0.03, 0.70)	0.622
rs780094	mannose	$\rightarrow$	MAX	-0.102	$1.95 \cdot 10^{-9}$	-0.034	$1.99 \cdot 10^{-2}$	0.339	$1.90 \cdot 10^{-2}$	(0.06, 0.67)	0.613
rs780094	mannose	$\rightarrow$	NR2C1	-0.102	$1.95 \cdot 10^{-9}$	0.038	$1.57 \cdot 10^{-3}$	-0.374	$2.20 \cdot 10^{-3}$	(-0.68, -0.14)	0.613
rs780094	mannose	$\rightarrow$	PPAPDC2	-0.102	$1.95 \cdot 10^{-9}$	0.035	$2.65 \cdot 10^{-2}$	-0.346	$2.62 \cdot 10^{-2}$	(-0.67, -0.04)	0.613
rs780094	mannose	$\rightarrow$	PPP1R3B	-0.102	$1.95 \cdot 10^{-9}$	-0.037	$4.52 \cdot 10^{-2}$	0.362	$4.32 \cdot 10^{-2}$	(0.01, 0.74)	0.622
rs2023634	proline	$\rightarrow$	C11ORF1	0.152	$1.51 \cdot 10^{-12}$	0.085	$7.21 \cdot 10^{-3}$	0.565	$9.20 \cdot 10^{-3}$	(0.14, 1.09)	0.613
rs2023634	proline	$\rightarrow$	COX7C	0.152	$1.51 \cdot 10^{-12}$	0.151	$1.95 \cdot 10^{-2}$	0.999	$2.14 \cdot 10^{-2}$	(0.17, 1.98)	0.613
rs2023634	proline	$\rightarrow$	CTDSP2	0.152	$1.51 \cdot 10^{-12}$	-0.081	$2.44 \cdot 10^{-3}$	-0.538	$2.60 \cdot 10^{-3}$	(-0.96, -0.19)	0.613
rs2023634	proline	$\rightarrow$	NDUFA4	0.152	$1.51 \cdot 10^{-12}$	0.117	$3.16 \cdot 10^{-2}$	0.772	$1.56 \cdot 10^{-2}$	(0.15, 1.52)	0.613
rs2023634	proline	$\rightarrow$	NDUFB3	0.152	$1.51 \cdot 10^{-12}$	0.121	$3.03 \cdot 10^{-2}$	0.799	$2.86 \cdot 10^{-2}$	(0.09, 1.67)	0.613
rs2023634	proline	$\rightarrow$	PFDN5	0.152	$1.51 \cdot 10^{-12}$	0.133	$3.31 \cdot 10^{-2}$	0.878	$4.60 \cdot 10^{-2}$	(0.02, 1.82)	0.622
rs2023634	proline	$\rightarrow$	RPL21	0.152	$1.51 \cdot 10^{-12}$	0.137	$3.00 \cdot 10^{-2}$	0.906	$2.40 \cdot 10^{-2}$	(0.12, 1.88)	0.613
rs2023634	proline	$\rightarrow$	RPL26	0.152	$1.51 \cdot 10^{-12}$	0.199	$2.49 \cdot 10^{-2}$	1.315	$2.72 \cdot 10^{-2}$	(0.16, 2.73)	0.613
rs2023634	proline	$\rightarrow$	RPL41	0.152	$1.51 \cdot 10^{-12}$	0.063	$6.26 \cdot 10^{-2}$	0.417	$4.54 \cdot 10^{-2}$	(0.01, 0.87)	0.622
rs2023634	proline	$\rightarrow$	RPS27	0.152	$1.51 \cdot 10^{-12}$	0.146	$3.73 \cdot 10^{-2}$	0.964	$3.94 \cdot 10^{-2}$	(0.04, 2.03)	0.622
rs2023634	proline	$\rightarrow$	TRAK2	0.152	$1.51 \cdot 10^{-12}$	0.070	$1.56 \cdot 10^{-2}$	0.464	$2.02 \cdot 10^{-2}$	(0.08, 0.91)	0.613
rs2023634	proline	$\rightarrow$	UQCRQ	0.152	$1.51 \cdot 10^{-12}$	0.121	$1.90 \cdot 10^{-2}$	0.802	$2.42 \cdot 10^{-2}$	(0.10, 1.63)	0.613
rs2403254	alphahydroxyisovalerate	$\rightarrow$	AHSP	-0.116	$1.97 \cdot 10^{-6}$	-0.095	$5.72 \cdot 10^{-2}$	0.821	$4.64 \cdot 10^{-2}$	(0.01, 1.78)	0.622
rs2403254	alphahydroxyisovalerate	$\rightarrow$	PITHD1	-0.116	$1.97 \cdot 10^{-6}$	-0.069	$5.25 \cdot 10^{-2}$	0.597	$4.56 \cdot 10^{-2}$	(0.01, 1.30)	0.622
rs17277546	androsteronesulfate	$\rightarrow$	LIPG	-0.618	$7.65 \cdot 10^{-11}$	0.026	$5.04 \cdot 10^{-2}$	-0.043	$3.94 \cdot 10^{-2}$	(-0.09, -0.00)	0.622

Table 3.4: Top causal metabolite-mRNA pairs. Causal direction represents the respective tested direction, metabolite $\rightarrow$ mRNA or mRNA $\rightarrow$ metabolite.

3.6 Model-based evaluation reveals systematic signatures of metabolic reactions

In order to further reveal the underlying mechanisms that determine the observed associations, we systematically analyzed whether correlating pairs of metabolites and transcripts (i.e. enzymes) correspond to the structure of the underlying metabolic network. Specifically, we investigated if strong metabolite-transcript edges of the BMTI tend to be in close proximity within biochemical pathways. All pairwise associations between metabolites and transcripts were mapped to their corresponding network nodes in the Human Recon 2 metabolic network reconstruction [82]. As a measure of metabolic network proximity, the length of the shortest path connecting each metabolite-enzyme pair was determined (Figure 3.9A). This measure is based on the common assumption that the shortest connection between two network entities corresponds to the biologically reasonable one [69, 123]. To avoid potential biologically meaningless shortcuts, we removed co-factors and currency metabolites prior to the analysis (see section 3.2 for details).

We could map 121 metabolites and 1,467 enzymes out of the 254 metabolites with known identity and 16,780 transcripts onto the metabolic network, respectively. While most pairwise correlation coefficients were closely distributed around zero for all investigated network distances, a distinct pattern was observable for statistically significant correlations. The majority of significantly correlating pairs accumulated at short distances and was dominated by positive correlations (Figure 3.9B). To determine the significance of this observation, one-tailed Fisher's exact tests were performed by either considering each distance individually or by aggregating all pairs up to a particular distance. The latter aggregation analysis combines all transcript-metabolite pairs which are reachable up to a certain number of steps (biochemical reactions) in the metabolic network. For both cases, we observed a substantial overrepresentation of significantly correlating pairs at short distances (Figure 3.9C). The strongest signals were observed for pairs that take part either directly in the same reaction ($d = 0$) or for those which are just one reaction apart ($d = 1$). For the cumulative distances we also observed significant enrichment up to a distance of $d = 2$ reactions. Proportions of significant and non-significant pairs per distance are given in Figure 3.10 and a detailed view on an exemplary path of length 2 is depicted in Figure 3.11.

To further characterize the underlying biochemical pathways, we calculated frequencies of functional annotations from Recon 2 among the significant associations for pathway distances 0 to 2 (Figure 3.9D and Figure 3.12). At a distance of 0, we identified mainly

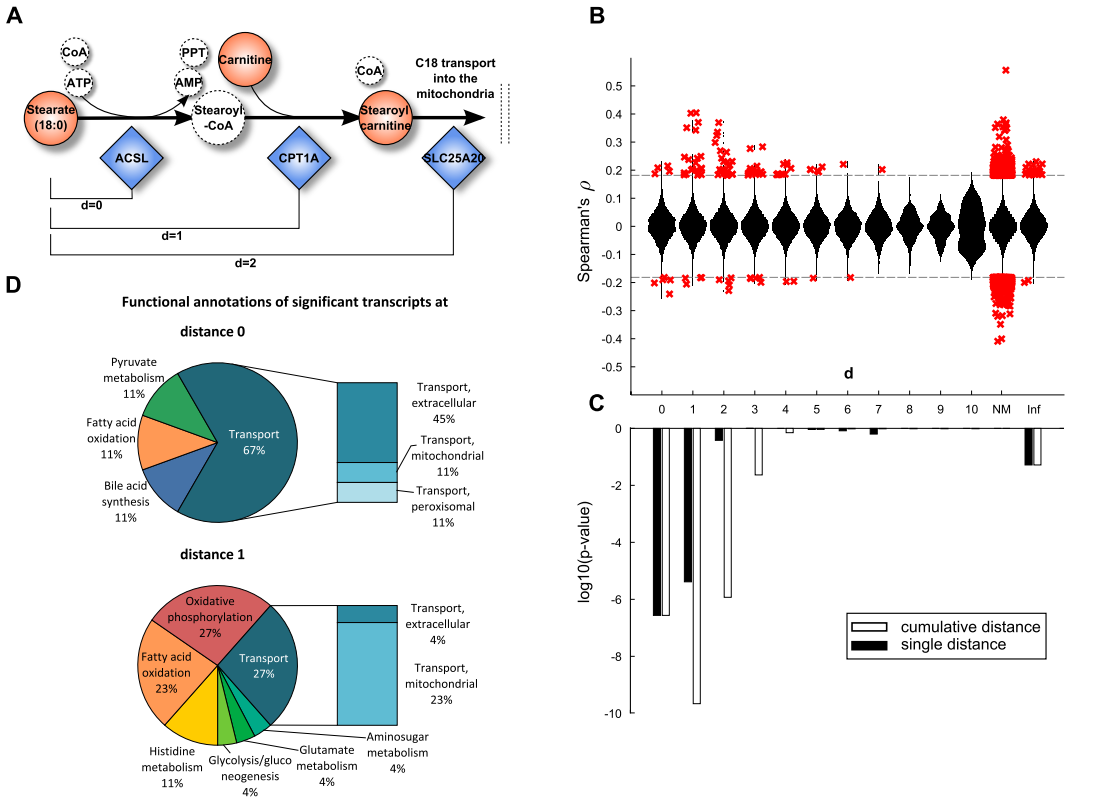


Figure 3.9: Model-based evaluation of metabolite-mRNA correlations. **A:** Schematic representation of the mitochondrial carnitine-shuttle with an explanation of network distance calculation. Note that co-factors are only illustrated for completeness, but are not considered for the calculation of the shortest path between two compounds. Dashed circles indicate unmeasured metabolites. **B:** Spearman correlation coefficient plotted against the number of pathway steps in human Recon 2. Significant correlations, i.e. those present in the BMTI are displayed as red crosses, whereas all non-significant correlations are plotted as a distribution. NM: no mapping. A distance of infinity (Inf) was assigned if there was no connection in Recon 2. **C:** Enrichment of significant correlations as determined by Fisher's exact tests. Black bars indicate $\log_{10}$ p-values assessing whether we observe more significant correlations for that particular distance than expected by chance. White bars represent the same test, only for a cumulative distance (i.e. "up to a distance of x"). **D:** Functional annotations of significantly associated transcripts at distances 0 and 1. At both distances, mainly transcripts belonging to the transport, energy, lipid and amino acid subsystems were observed.

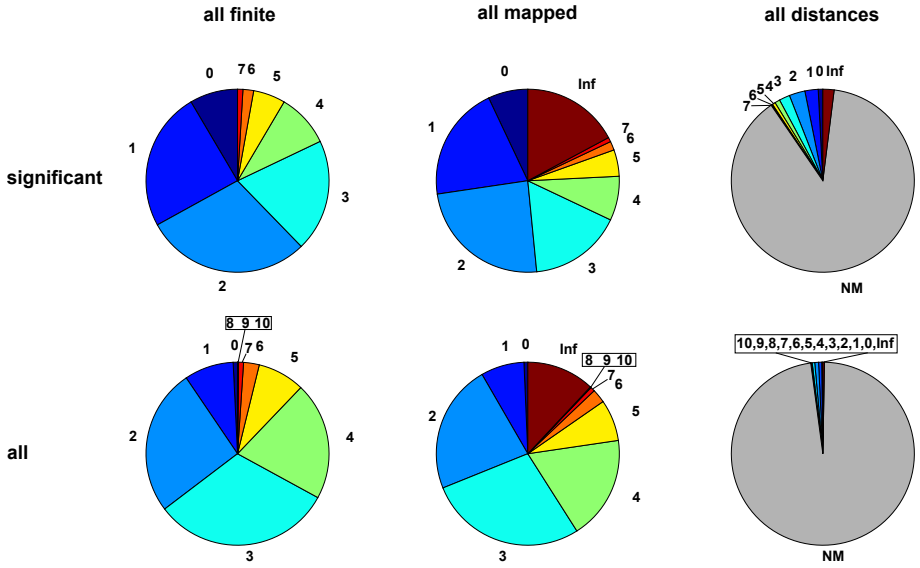


Figure 3.10: Fractions of metabolite-transcript pairs with a given pathway distance. The first column ('all finite') contains mapped pairs with non-infinite distances. The second column ('all mapped') contains all pairs of metabolites and transcripts that could be mapped to the pathway network. The third column ('all distances') shows all pairs, indicating that the larger fraction of pairs cannot be mapped to the pathway network. The first row ('significant') only contains significantly correlating pairs of metabolites and transcripts, whereas the second row ('all') shows all pairs. The increase in fraction size from 'all' to 'significant' is assessed by a Fisher's exact test and shown in Figure 3.9C.

transport reactions (67%) accompanied by reactions belonging to lipid metabolism (bile acid synthesis 11%, fatty acid oxidation 11%) and carbohydrate metabolism (pyruvate metabolism 11%). The transport reactions can be further subdivided into extracellular transport (45%), or mitochondrial transport (11%) and peroxisomal transport (11%). Similar signals can be found at distances of 1 and 2, where we additionally identified reactions belonging to energy metabolism (oxidative phosphorylation 27%) and amino acid metabolism (histidine metabolism 11%, glutamate metabolism 4%).

Taken together, the BMTI captured a systematic signal of metabolite-enzyme associations to be in close proximity when mapped onto a global metabolic network. Moreover, the strongest signals found for pathway distances of 0, 1 or 2 reflect distinct metabolic reactions mainly belonging to lipid, energy and amino acid metabolism, and transport mechanisms.

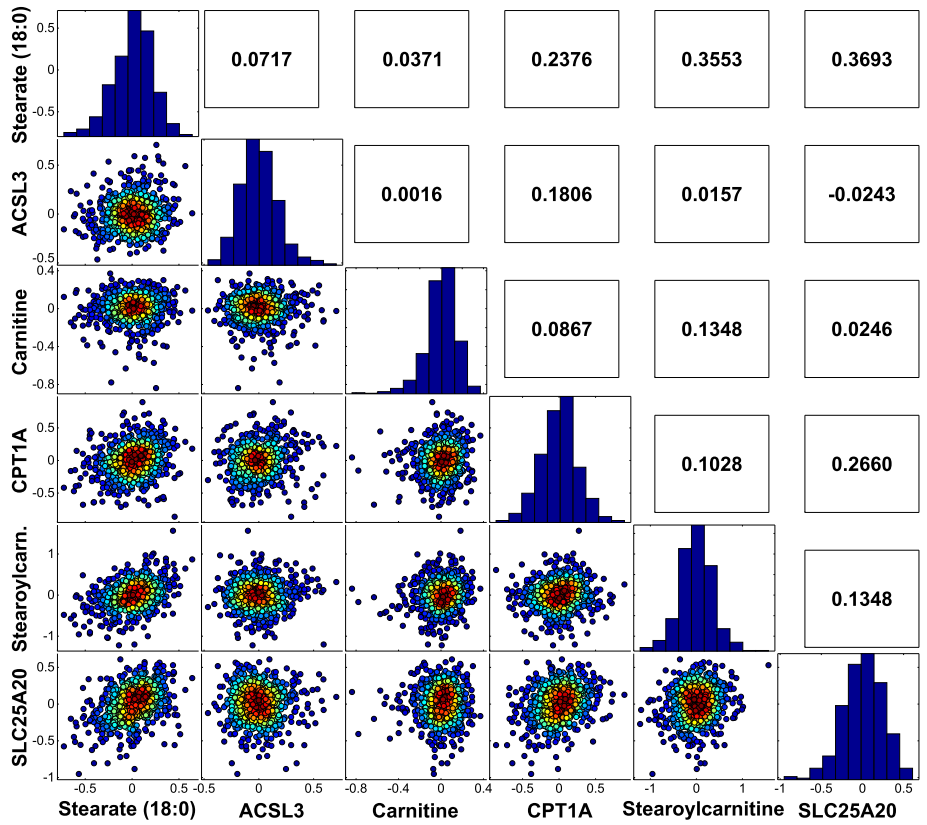


Figure 3.11: Scatter plots for an exemplary path of length 2. Note that this path corresponds to the one depicted in Figure 3.9A. The path describes the ACSL-catalyzed activation of stearate (18:0), followed by mitochondrial transport via the carnitine shuttle. The transport process is a two-step reaction including the attachment of carnitine by *Carnitine Palmitoyltransferase 1A* (CPT1A) at the outer mitochondrial membrane and subsequent internalization by *carnitine-acylcarnitine translocase* (SLC25A20). Upper triangle matrix indicates Spearman correlation coefficients. For direct substrate/product–enzyme pairs of this path we observed weak, insignificant associations ranging from $\rho = 0.0016$ to $\rho = 0.13$. In contrast, the strongest correlation was observed between stearate (18:0) and SLC25A20 ($\rho = 0.36$, p-value $= 2.02 \times 10^{-24}$), which are two reaction steps apart.

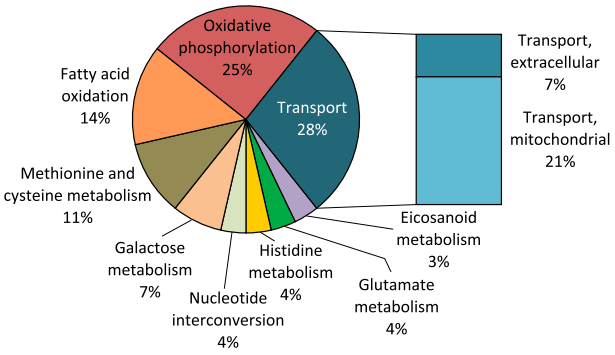


Figure 3.12: Functional annotation of transcripts at a distance of 2. Similar to shorter distances, the strongest associations between metabolites and transcripts at a distance of two mainly resemble paths describing transport processes (28%) or those belonging to energy metabolism (oxidative phosphorylation 25%), amino acid metabolism (methionine and cysteine metabolism 11%, histidine metabolism 4%, glutamate metabolism 4%) and lipid metabolism (fatty acid oxidation 14%) besides others.

3.7 Functional annotation-based aggregation of the BMTI reveals cross-talk between pathways

Up to this point, our analysis was a reaction-centered approach limited to single edges only, thereby neglecting the global network structure and cross-talk between pathways captured in the BMTI. To derive a comprehensive functional description of the biological modules included in the BMTI, we developed a novel approach based on functional annotations, which provides an integrated view on cellular processes. Briefly, the approach consists of three steps: First, we used pathway annotations to define groups of functionally related metabolites and transcripts. For metabolites, we used metabolic pathway annotations provided with the metabolomics dataset, and for transcripts we employed the Gene Ontology (GO) slim annotations. Second, an aggregated z-score ($aggZ$-Score) was calculated for each functional category. Third, we calculated correlations between $aggZ$-Scores of all functional categories. A schematic overview of this multi-step approach is provided in Figure 3.5 and described in more detail in section 3.7.

We again constructed a network (the pathway interaction network, PIN) by drawing edges between significantly correlated categories. Interestingly, even when applying a stringent Bonferroni-corrected threshold ($\alpha = 0.01$, p-value $\leq 2.2 \times 10^{-6}$) this resulted in an overly dense connected network of 166 nodes and 1220 edges. To generate a vi-

sually interpretable version of this network, an ad-hoc stringent threshold of p-value $\leq 1.0 \times 10^{-11}$ was applied to draw the network. This resulted in a PIN consisting of 113 nodes (93 GO terms, 20 metabolic pathways) connected by 244 edges (196 positive correlations, 48 negative; Figure 3.13A). Remarkably, we observed a high conformity between linked metabolic pathways and gene annotations. For example, the metabolic pathway "carnitine metabolism" was connected to the biological processes "lipid metabolic process" and "transmembrane transport". Moreover, it was linked to the cellular component "mitochondrion", indicating transport processes of fatty acids into the mitochondrion for subsequent β-oxidation. Further biologically reasonable pairs were "Valine, Leucine and Isoleucine metabolism" and "Glutamate metabolism" attached to "cellular nitrogen compound metabolic process". As a last example, "Steroid/Sterol" was connected to "response to stress" and "signal transducer activity", pointing to an interaction between hormones and regulation of gene expression. In the following, we examine two selected category-category relationships in detail, including the individual metabolites and gene transcripts that gave rise to the association.

Scenario one: Fatty acid metabolism

The first scenario contained the metabolic pathway long chain fatty acid and the gene ontology annotation lipid metabolic process (Figure 3.13B). The subnetwork that induced this association consisted of 22 individual constituents (7 mRNAs, 15 metabolites) connected by 38 edges. Notably, this subnetwork coincides well with the above-mentioned fatty acid carnitine-shuttle, i.e. the transport of long chain fatty acids into the mitochondrion for subsequent degradation. Within this subnetwork, 8 long chain fatty acids were jointly associated to CPT1A and SLC25A20, while 7 additional fatty acids were associated to SLC25A20 alone. Moreover, *acyl-CoA dehydrogenase very long-chain* (ACADVL) and *Perilipin 2* (PLIN2), both involved in β-oxidation and long chain fatty acid transport, were associated to 5 and 7 out of the 15 long chain fatty acids, respectively. Three further transcripts, *Tumor Necrosis Factor Receptor Superfamily, Member 21* (TNFRSF21), *Aldo-Keto Reductase Family 1, Member C3* (AKR1C3) and *1-Acylglycerol-3-Phosphate O-Acyltransferase 4* (AGPAT4) were correlated with 5,8-tetradecadienoate, a side product of oleate β-oxidation. While AKR1C3 and AGPAT4 are enzymes mainly related to arachidonic acid metabolism and phospholipid metabolism, potentially indicating a branching point to other pathways of the lipid metabolism, TNFRSF21 is involved in T-cell activation and immune regulation [231].

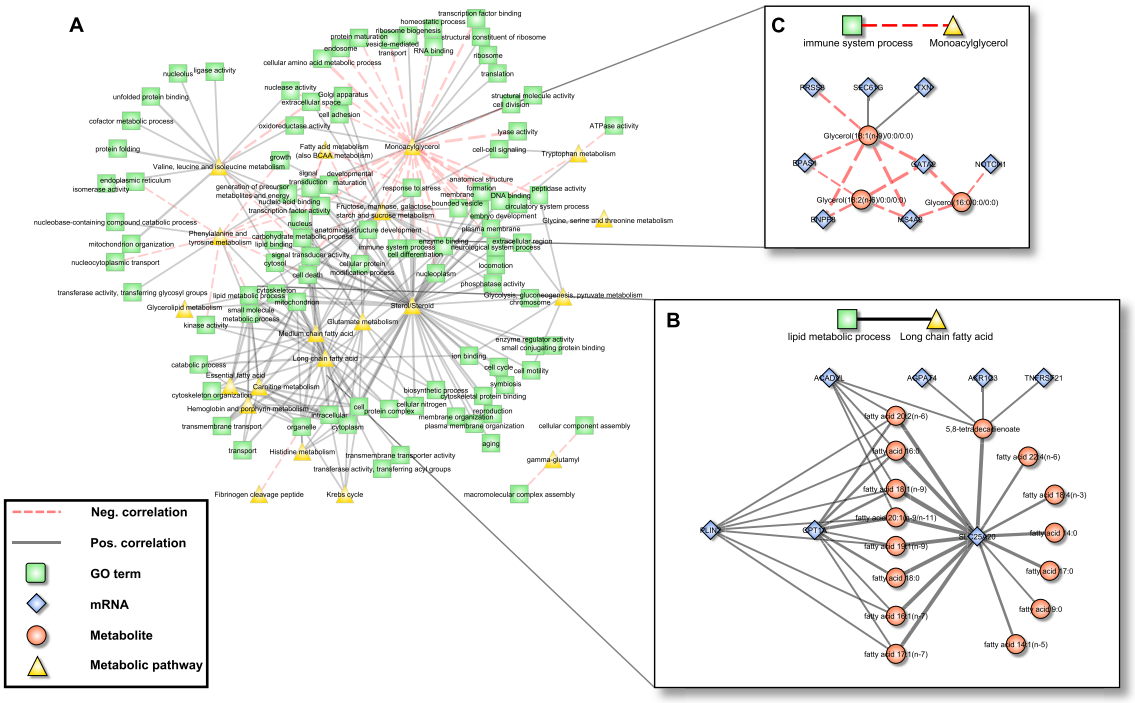


Figure 3.13: Pathway interaction network (PIN). **A:** Bipartite correlation network, where each node represents either a metabolic pathway or a gene set summarized in a GO term, while edges between them represent the correlation of the respective $aggZ$-Scores. **B+C:** Expanded view on exemplary pathway interactions. Note that zoomed parts correspond to subnetworks from Figure 3.6D that are explained by one single link of the PIN. Edge widths represent correlation strengths.

Scenario two: Monoacylglycerols and immune-related transcripts

Since whole blood transcriptomics measurements should mainly reflect the immune system, we have chosen the association between the metabolic class of "monoacylglycerols" and the GO-term "immune system process" as a second scenario (Figure 3.13C). The corresponding subnetwork contained 11 nodes (3 metabolites, 8 transcripts) and 14 edges (2 positive, 12 negative). The monoacylglycerols included 1-oleoylglycerol, 1-linoleoylglycerol and 1-palmitoylglycerol, which only differ in the attached fatty acid residues. It has been shown that monoacylglycerols are not merely intermediate lipid substances, but may also act as signaling molecules. For example, 2-arachidonoylglycerol is a known ligand of cannabinoid receptors, which are involved in the regulation of several biological processes including inhibition of pro-inflammatory and other immune system related processes [232, 233]. Among the genes summarized by the term "immune sys-

tem process", GATA2, *Endothelial PAS Domain Protein 1* (EPAS1) and Notch1 are key regulators of hematopoiesis and as such are involved in the differentiation process of immune cells [227, 234, 235]. Moreover, EPAS1 and *thioredoxin* (TXN) are associated to the response to oxidative stress [234], whereas *Membrane-Spanning 4-Domains, Subfamily A, Member 2* (MS4A2) and *Ectonucleotide Pyrophosphatase/Phosphodiesterase 3* (ENPP3) are involved in allergic responses mediated via the IgE receptor [236]. The two remaining genes, SEC61G and PRSS8, are involved in the immune processes of antigen presentation and inflammation [237].

3.8 Regulatory signatures captured by the integrated network

The BMTI contains a prominent "flower-like" network topology, i.e. many transcripts associated to a single metabolite. We therefore asked whether these coordinated changes around a metabolite and also the network topology can be explained by common transcriptional regulatory processes through transcription factors (TFs). For the following analysis, we only considered metabolites linked to at least 3 transcripts. We analyzed the promoter regions of all connected genes for an enrichment of known transcription factor binding sites (TFBS) derived from the Jaspar database [210]. This resulted in significantly enriched transcription factor binding motifs for 46 single metabolites, 24 subpathways and 7 superpathways. Section 3.2 provides a detailed explanation of the process. A summary of all enriched TFBS can be found in the online supplementary of the original publication (`http://dx.doi.org/10.1371/journal.pgen.1005274`).

In total, out of the 205 binding motif matrices used in the analysis, 189 reached a significant enrichment in at least one of the metabolite-derived gene sets, indicating a generally prevalent common regulation. Across all lists of enriched TFBS identified from our network, the motifs that occurred most frequently were *Sterol Regulatory Element Binding Transcription Factor 2* (SREBF2), *Peroxisome Proliferator-Activated Receptor Gamma* (PPARG; Jaspar motifs PPARG and PPARG::RXRA) and *Nuclear Factor, Interleukin 3 Regulated* (NFIL3). SREBF2 is a major regulator of cholesterol metabolism [238] while PPARG is known to be activated by fatty acid ligands, thereby regulating fatty acid β-oxidation and other processes [239]. NFIL3 is a regulator specifically found in activated T cells, natural killer (NK) cells, and mast cells, involved in the regulation of the immune response and the circadian rhythm [240].

Branched-chain amino acids were among the metabolites most strongly connected to SREBF2 targets. Specifically, the transcripts correlating with isoleucine and valine showed high enrichment of SREBF2 binding sites (p-value = 5.83×10^{-8} and p-value = 2.36×10^{-10}, respectively). Moreover, considering all 172 genes associated to at least one metabolite from the entire branched-chain amino acid pathway ("Valine, leucine and isoleucine metabolism") yielded significantly enriched binding sites for SREBF1 and SREBF2 (p-values 6.78×10^{-10} and 9.11×10^{-10}, respectively). Both SREBs are important regulators in lipid homeostasis, including cholesterol and fatty acid biosynthesis, further indicating a regulatory cross-link between HDL-C, TG and BCAA metabolism.

The highly interlinked network topologies of both the blood metabolome-transcriptome interface and the pathway interaction network suggest a strong coregulation between the different metabolites, processes, and pathways. As a second step of coregulation analysis, we inferred the number of pairwise shared significant TFBS to determine the extent of coregulation between single metabolites and metabolic pathways. At the single metabolite level, we found a maximum number of 27 shared TFs between histidine and X-03094 (Figure 3.14). Moreover, this highly connected unknown metabolite shared 14 TFs with another unknown metabolite (X-12442) and with a peptide (HWESASXX). For the metabolic subpathways, we observed an overlap between "histidine metabolism" and the group of "long chain fatty acids" and between "glycolysis, gluconeogenesis, and pyruvate metabolism" and the group of "fibrinogen cleavage peptides" (11 shared TFs each; Figure 3.15A). At the level of superpathways, the highest number of shared TFBS was 4, identified between "carbohydrate" and "peptide metabolism" (Figure 3.15B). Overall, we found that TF binding sites are shared to a large extent, indicating a complex coregulation not only within but also between different processes and pathways.

To gain further insight into this coregulation, we determined transcription factors which also occur as transcripts in the BMTI. 165 out of the 189 transcription factors with available binding motif were contained in the filtered data set. Only 12 of these transcription factors displayed a significant correlation to any metabolite and are thus included in the BMTI. This observation is not completely unexpected given that TFs are regulated to a large extend at a post-transcriptional level [241]. Interestingly, for two out of these 12 TFs, we also observed enriched binding sites in the promoter region of the other genes connected to the same metabolite, i.e. a "triad" network motif consisting of a metabolite, a transcription factor and its target genes (Figure 3.16A+B).

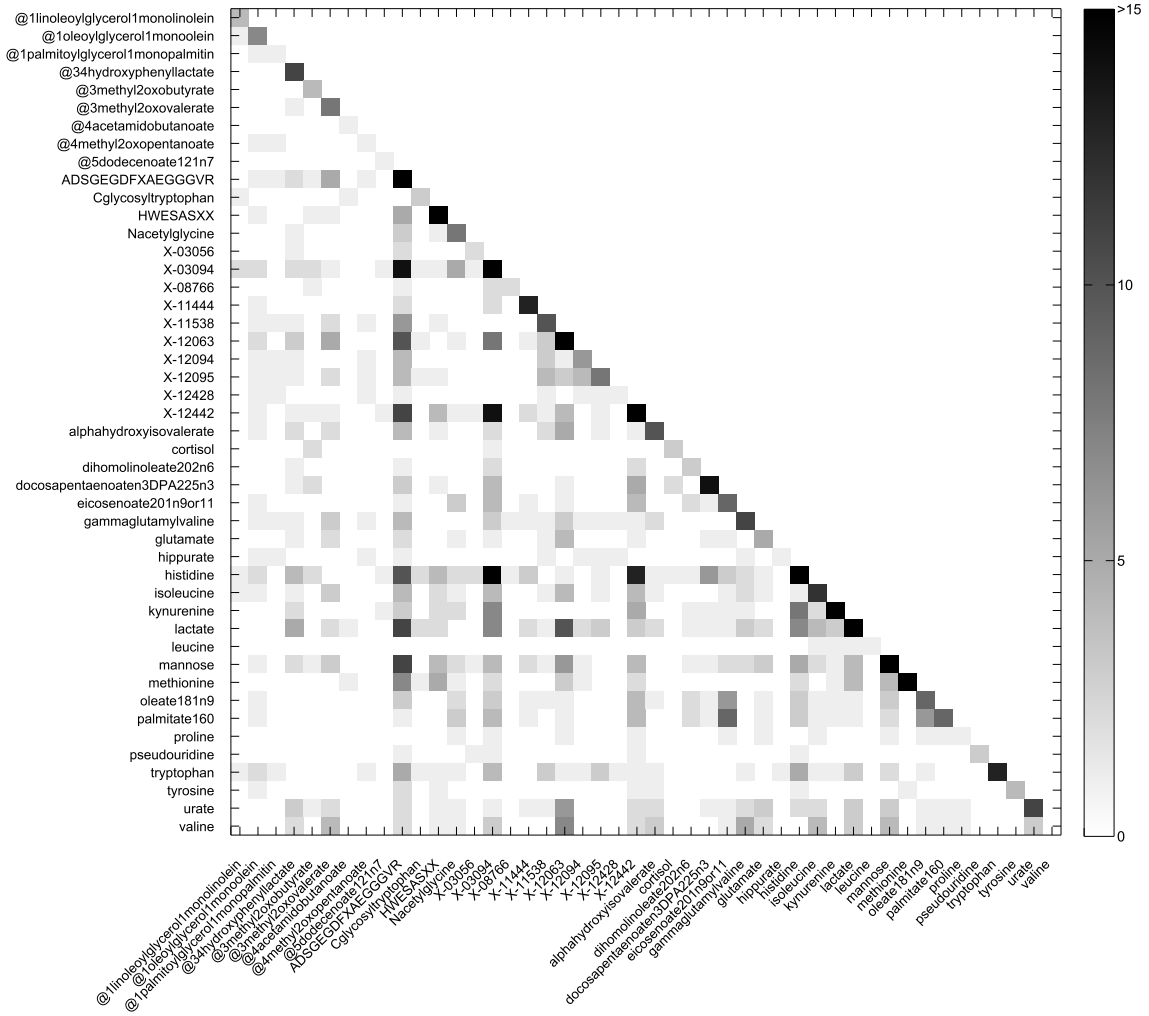


Figure 3.14: Overlap of enriched transcription-factor binding sites for all metabolites that are included in the BMTI and have at least one enriched site. At this fine grained level of single metabolites, several signatures of a shared regulation are observable, e.g. between amino acids and various other metabolites. Note that the color-scale is capped at 15, the maximum number of shared binding sites was 27 between histidine and X-03094.

The first transcription factor is *B-cell CLL/Lymphoma 6* (BCL6), a transcriptional repressor involved in the STAT-dependent interleukin 4 response of B-cells [242]. BCL6 is negatively correlated with methionine and tyrosine in our network. The TFBS enrichment analysis using all 15 genes connected to methionine within the BMTI resulted in a significant overrepresentation of the BCL6 binding motif (p-value=5.71×10^{-09}, 82% of the 15 promoter sequences contained at least one occurrence of the motif; Figure 3.16A), while no significant enrichment was observable for the genes connected to tyrosine. The

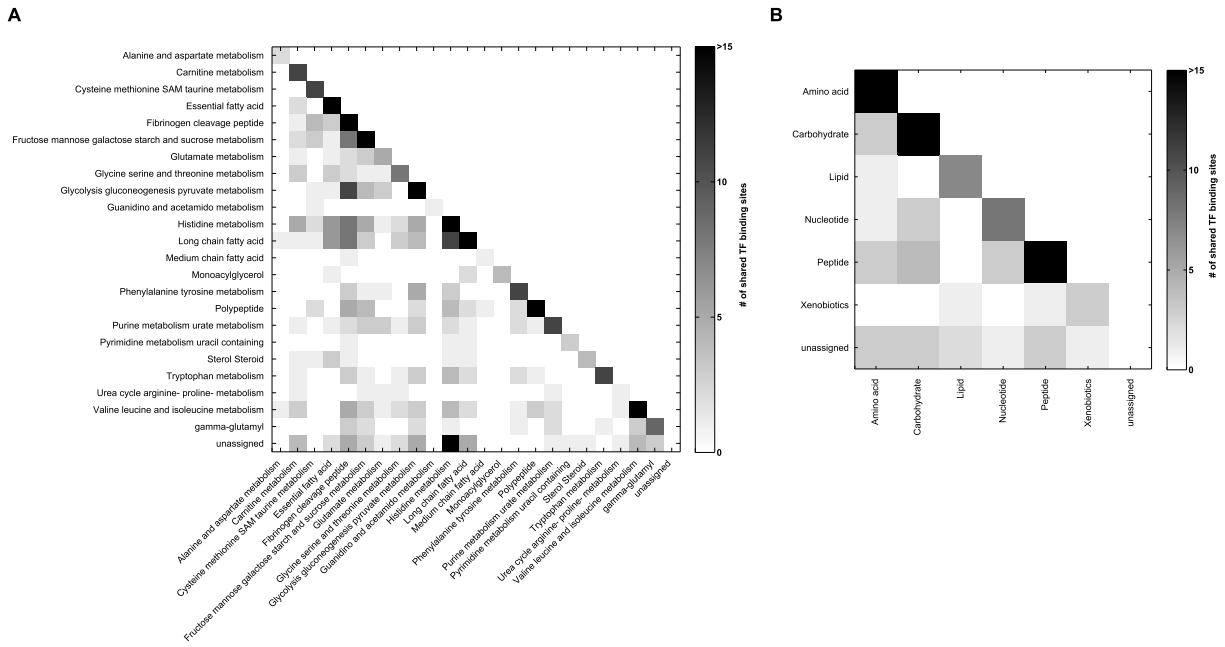


Figure 3.15: Transcription factor binding site analysis. **A:** Heatmap of shared significantly overrepresented TFBS between metabolic subpathways. Upper right triangle matrix was left out. Darker colors indicate a higher number of shared TF binding sites. The category "unassigned" includes all unknown metabolites. **B:** Overlap of enriched transcription-factor binding sites for all super-pathways. On this coarse level, only few shared regulatory signatures are observable. Note that in both comparisons, the color-scale is capped at 15.

second motif was identified around *Nuclear Receptor Subfamily 4, Group A, Member 2* (NR4A2), which was associated to 7 metabolites in our network. The 22 neighboring genes of one of those metabolites, kynurenine, showed significantly enriched binding sites for this transcription factor (p-value= 3.79×10^{-09}, 73% of the 22 promoter sequences contained at least one occurrence of the motif; Figure 3.16B).

3.9 Integration of clinical phenotypes identifies active modules

As a final analysis step, we sought to use the BMTI and the PIN to infer novel insights into the molecular mechanisms and pathways underlying complex traits. To this end, we associated the nodes of both networks with the intermediate clinical phenotypes HDL-C and LDL-C, as well as concentrations of blood triglycerides (TG), which are known risk

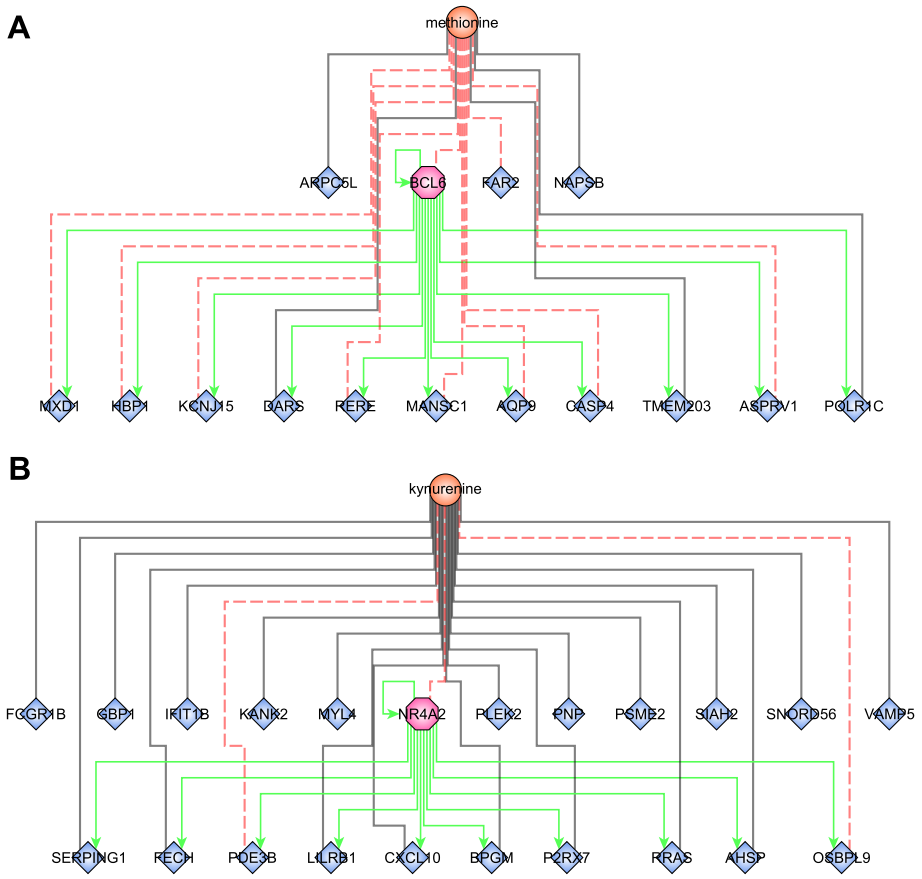


Figure 3.16: **A+B:** Identified network motifs of TFs and their respective target genes associated to the same metabolite. Green arrows indicate regulation. Black lines indicate positive correlation; red dotted lines indicate negative correlation.

factors for a variety of diseases. We performed multiple linear regression analyses with HDL-C, LDL-C and TG blood parameters as response variables and all 440 metabolites and 16780 transcripts as explanatory variables. All models were corrected for sex and age. Statistical significance was defined by a Bonferroni adjusted p-value cutoff at 2.9×10^{-6} ($\alpha = 0.05$). We then projected the $-\log_{10}$ transformed p-values from this regression as colors onto the corresponding nodes in the BMTI network. Similarly, the analysis was performed using $aggZ$-Scores of pathways/GO terms as explanatory variables and mapped to the PIN (Figure 3.17). Note that we presented similar approaches in the past for metabolomics-only networks [64, 243].

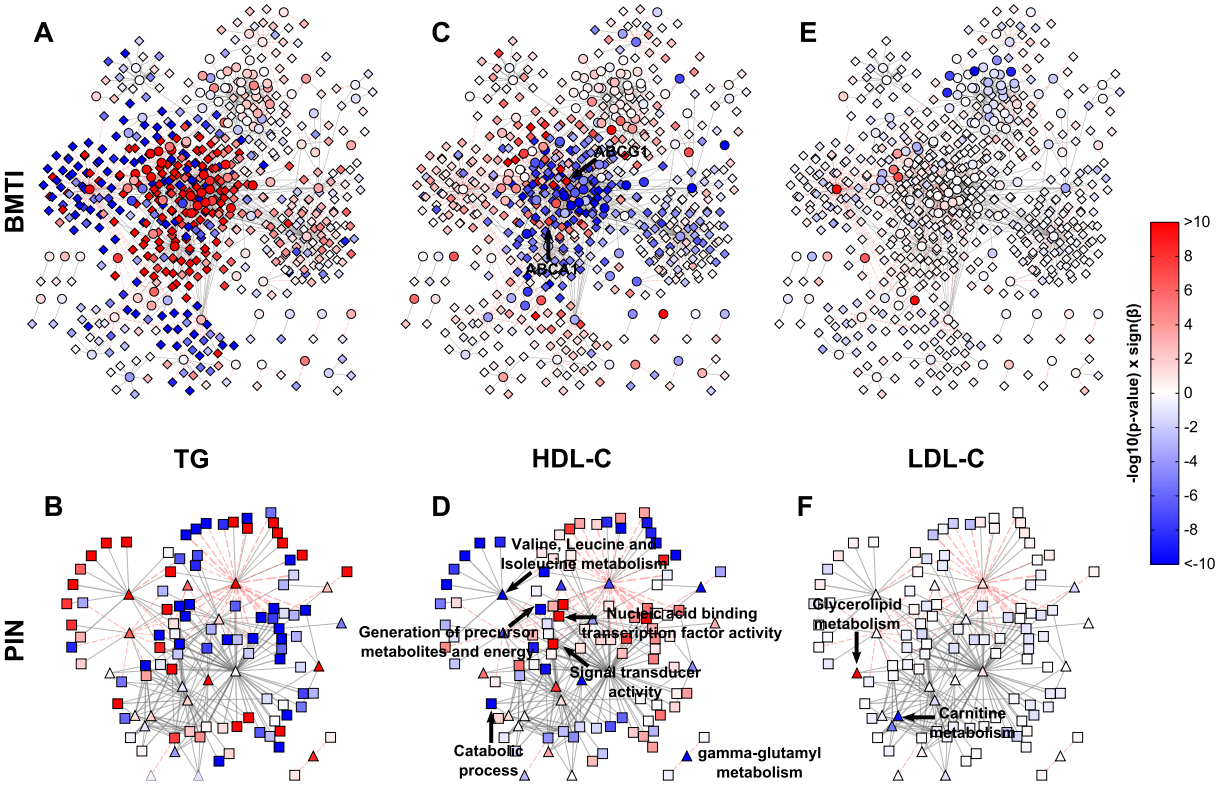


Figure 3.17: Intermediate clinical phenotype associations. Linear regression results of TG, HDL and LDL for each metabolite and transcript (**A, C, E**), or pathway and GO term (**B, D, F**) were projected onto the respective networks. Blue colors indicate negative associations, while red colors represent positive associations to the respective phenotype. Color strength of the nodes encodes the $-\log_{10}$ p-value of the respective association. β denotes the regression coefficient and its sign represents the direction of the associations (positive or negative correlation).

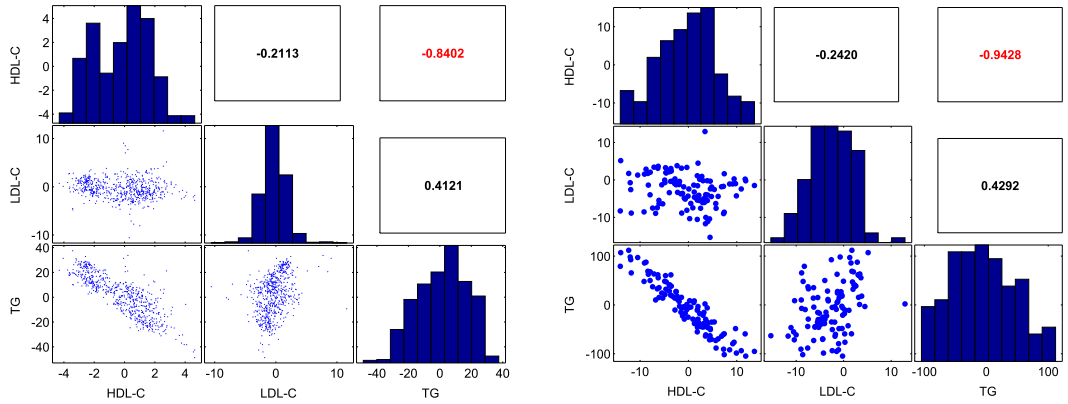
In total, regression analyses yielded 233 (54 metabolites, 179 mRNAs), 28 (28 metabolites, 0 mRNAs) and 1124 (49 metabolites, 1075 mRNAs) statistically significant

associations for HDL-C, LDL-C and TG, respectively. Of those associations, 64%, 28% and 25%, were contained in the BMTI, respectively (see online supplementary of the original publication for a complete list of associations (`http://dx.doi.org/10.1371/journal.pgen.1005274`)). We only observed few associations between LDL-C and metabolites, which can be mainly summarized in the "Glycerolipid metabolism" and "Carnitine metabolism", while none were observable for the transcripts (Figure 3.17E+F). Thus, we will focus on network associations for HDL-C and TG in the following.

Examination of the networks for HDL-C and TG revealed localized regions of similar associations, which reflect potentially co-regulated modules (Figure 3.17A+C). Interestingly, when compared to each other, there appeared to be an antagonistic pattern of associations for HDL-C and TG, which is in accordance with an overall negative correlation of the two traits ($\rho = -0.53$) . This opposing pattern also holds for the categorical networks (Figure 3.17A-D). To confirm this observation statistically, we used an approach to compare the different networks suggested by Floegel et al. [243]. Basically, we calculated the Spearman correlation of the association measures between the different clinical traits. This resulted in a strong negative correlation between the BMTI-HDL-C and the BMTI-TG network ($\rho = -0.84$) which was even more profound between the PIN-HDL-C and PIN-TG networks ($\rho = -0.94$, Figure 3.18A+B). A similar pattern of opposing associations between HDL-C and TG phenotypic traits was already described in previous studies, which suggested a pleiotropic, heritable relation between the two lipid and lipoprotein measures, potentially regulated by a common, complementary mechanism [188, 244].

In the following, we will discuss exemplary pathway mechanisms identified in the phenotype networks. ABCG1 and ABCA1, known constituents of the reverse cholesterol transport necessary for the proper formation of plasma HDL-C [229], were positively correlated with HDL-C (p-value $= 4.37 \times 10^{-12}$ and p-value $= 2.92 \times 10^{-8}$, respectively). At the pathway level, processes like "generation of precursor metabolites and energy" or "catabolic process" are negatively associated with HDL-C, while "nucleic acid binding transcription factor activity" and "signal transducer activity" are positively associated (Figure 3.17D). An inverse pattern can be seen for TG, where positive associations predominate and processes like "generation of precursor metabolites and energy" or "catabolic process" are strongly positively associated (Figure 3.17A+B).

Given the known association between HDL/TGs and branched-chain amino acids [63, 245], we investigated the phenotypic networks to further examine this metabolic class.



(A) Single transcript/metabolite level. (B) Pathway level.

Figure 3.18: Spearman's correlation of phenotype associations. Comparison of beta values from linear regression analysis. **(A)** Each dot represents either a transcript or a metabolite in the BMTI. We observe a strong anti-correlation between HDL and TG associations, modest positive correlation between TG and LDL and a very weak anti-correlation for HDL and LDL. **(B)** Dots represent either a GO term or a metabolic pathway in the PIN. Even more profound than for single transcripts and metabolites (Figure 3.18A), there is a strong anti-correlation between HDL and TG associations on the pathway level. The correlations for HDL and LDL as well as LDL and TG associations are comparable to the results from Figure 3.18A.

First, we examined the edge between isoleucine and ABCG1 within the BMTI-HDL-C network. As already mentioned, ABCG1 was strongly positively associated to HDL-C levels, while we found that isoleucine was significantly negatively associated to the concentration of HDL-C ($\beta = -4.30$, p-value $= 5.80 \times 10^{-19}$). Moreover, gamma-glutamyl variants of BCAAs belonging to "gamma-glutamyl metabolism" ($\beta = -4.84$, p-value $= 3.15 \times 10^{-14}$) and "Valine, leucine and isoleucine metabolism" ($\beta = -4.66$, p-value $= 9.17 \times 10^{-11}$) displayed profound negative associations to HDL-C (Figure 3.17D), further validating a connection between HDL-C and BCAA metabolism. For triglycerides, we observed an inverse relationship with BCAAs and BCAA-related pathways (Figure 3.17B).

3.10 Discussion

We constructed a global network model across two levels of biological information by integrating cross-sectional omics data from a large-scale population cohort. The dataset was based on circulating metabolites from plasma and transcriptional variation derived from whole blood. This analysis exploited the naturally-occurring variation caused by

genetic variation, environmental and behavioral influences from a natural population over multiple layers of organization. Such an approach was recently referred to as systems genetics, enabling the systematic exploration of information flow between the different biological scales [129].

As mentioned in the introduction, blood is a heterogeneous tissue containing a series of distinct cell-types. In this study, we utilized whole blood transcriptomics data from unsorted cells, leading to a complex mixture of transcriptional signals in the transcriptome dataset [204]. Similarly, the levels of circulating metabolites are strongly influenced by metabolically active organs [195], but also by metabolites from blood cells and those taken up from the environment. The comparison to known cell-type specific markers further suggested that a substantial amount of the signals are derived from specific blood cells. However, the analysis also showed that the majority of the BMTI contained transcripts are not assigned to any cell-type. Thus, we assume that the metabolite-mRNA associations captured in the BMTI mainly reflect cell-type unspecific processes involved in the fundamental maintenance of cellular function, besides some processes specifically related to immune functions.

Independent replication of the BMTI edges was investigated using data from the DILGOM study. Out of 211 possible associations, we were able to replicate 29% at a nominal significance and 18% after multiple testing correction (FDR < 0.05). This relatively low number of replicated associations might have various reasons. For example, (1) The DILGOM study used an NMR-based metabolomics platform in contrast to the mass spectrometry-based methodology used in KORA. (2) The smaller sample size of the DILGOM study might limit the power to detect existing associations between metabolites and transcripts. (3) Differences in laboratory procedures and protocols or the population structure can affect replication across cohorts. Future high-powered studies with more similar measurement platforms can further address the stability of metabolite-transcript correlations across studies.

A comprehensive analysis of the strongest associations between transcripts and metabolites clearly revealed biologically reasonable relationships, such as signaling and transport mechanisms. Many identified associations, e.g. between cortisol and DDIT4 or between SLC25A20 and multiple long chain fatty acids, were in consent with known signaling or metabolic pathways. Others support and extend results from previous studies. As one example, nearly all transcripts included in the lipid-leukocyte (LL) module identified by Inouye et al. [193] were among the top scoring association pairs. For instance,

we were able to confirm associations between CP3A, FCER1A, GATA2, HDC, MS4A2, and SLC45A3, core genes of the LL module, and leucine, isoleucine, and several lipids. In addition, we found associations which, to the best of our knowledge, have not been described before. These include associations between 1-monoolein and GATA2, a key regulator of hematopoiesis, or SLC45A3, a known diagnostic marker for prostate cancer [246]. The identified associations extend the current knowledge about the connection between system-wide metabolism and immunity-related pathways.

Causal inference of the metabolite-mRNA associations using Mendelian randomization yielded no statistically significant results. There are various possible reasons for this negative outcome. First, there might be no causal effect in either direction between the investigated transcripts and metabolites. Besides that, the lack of significant findings could also be caused by the limitations of Mendelian randomization. For instance, MR is known to require large numbers of samples to detect true causal relationships, and the power in our study (n=712) might have been too low [17]. We therefore decided to leave a more detailed discussion and analysis of causal effects to future, high-powered studies.

Comparison of the blood metabolome-transcriptome interface with the most recent human genome-scale metabolic reconstruction [82] allowed to assess whether transcript-metabolite correlations also recapitulate known biochemical reactions at a systematic level. We were able to show that strong associations between enzymes (represented by their respective transcripts) and metabolites are significantly accumulated at shorter pathway distances (Figure 3.9B+C), which is consistent with previous studies [95, 123, 247]. Further functional characterization identified transport, energy, lipid and amino acid subsystems to be predominately present at short pathway distances (Figure 3.9D + Figure 3.12). This observation may reflect metabolic proximity through the uptake of metabolic nutrients by metabolically active blood cells. For instance, in our analysis we found signatures for all three major sources for energy production: lipids, proteins (in terms of amino acids) and carbohydrates indicating an active use of fuel molecules for energy generation by the blood cells.

Our model-based analysis has several limitations. As mentioned above, any such analysis is heavily dependent on the quality of the underlying metabolic reconstructions, which are still far from being complete [158]. This incompleteness, together with a prevalent inconsistent nomenclature of metabolites allowed us to map only 121 out of 254 measured metabolites onto the metabolic network model. Another limitation is the

incomplete coverage of the metabolome, which is owed to the capabilities of currently available technologies. In this study we used measurements of 440 metabolites, which corresponds to just $\sim$ 10% of the estimated human serum metabolome [248]. Nevertheless, we believe that despite incomplete pathway mappings, our observations further indicate that combined metabolomics and transcriptomics data from human blood reflect reaction signatures of system-wide biological processes.

To further functionally characterize the blood metabolome-transcriptome interface at a global level, we developed a network approach based on functional annotations. To this end, we aggregated z-score transformed measurements of metabolites and transcripts into their corresponding metabolic pathways and gene ontology categories, respectively. This approach allowed us to calculate correlation values between different functional categories, rather than between single metabolites and transcripts only. From these associations, we generated a pathway interaction network (PIN) of associated metabolic pathways and Gene Ontology terms, substantially reducing the complexity of the original network and thus facilitating functional interpretations. Detailed inspection of the PIN revealed that correlating nodes resembled not only signatures of well-known biological processes, like the carnitine shuttle, but also suggested novel interactions such as a crosstalk between monoacylglycerols and immune system processes. Taken together, the pathway interaction network enabled us to verify and elevate observations from the single reaction level (see section 3.6) onto a pathway level. Moreover, we are now able to explore associations between biological processes/pathways across different biological scales including those that are not necessarily covered by metabolism, such as signaling or transcriptional processes.

Given the high interconnectivity of the BMTI and the PIN, we asked whether these associations contain information about regulatory interactions across the different metabolite classes and pathways. Enrichment analysis of transcription factor binding sites in the promoter regions of the genes contained in our network identified regulatory signatures for transcripts associated to the same metabolite, which are additionally largely shared between metabolites belonging to different metabolic pathways (Figure 3.14 and Figure 3.15). There is a series of possible explanations for this observation. On the one hand, our findings could indicate that single metabolites/transcripts are fulfilling multiple roles, thus sharing several biochemical pathways. On the other hand, it might reflect regulatory interactions operationally linking different metabolic pathways. In depth investigation of 12 transcription factors included in the BMTI additionally revealed two triad network motifs between transcription factors BCL6 and NR4A2, their target genes

and the metabolites methionine and kynurenine, respectively. Remarkably, in a study conducted on mice fed a methionine and choline deficient diet, a significant reduction in the expression of BCL6 was observed [249]. It is widely known that metabolites can act as intermediates in cellular signaling, thereby also regulating gene expression, and together with our findings we suggest that characteristics of metabolic regulation are captured in the BMTI. However, from a correlation network, the detection of an association between a metabolite and a transcript does not necessarily imply a regulatory relationship nor can a conclusion be drawn about the directionality of the relationship. Yet, a combined analysis might offer the opportunity to identify novel molecular mechanisms behind cellular regulation that need to be validated further by experimental evidence.

Besides transcriptional regulation mediated by TFs, a substantial fraction of transcripts are expected to be regulated by epigenetic processes [250]. Comparing 1350 reported methylation site-metabolite associations from a recent epigenome-wide association study [195] with our results surprisingly revealed only a single overlapping hit: X-03094 and the MAN2A2 transcript correlated in our study and also displayed a comparable methylation-metabolite association in the EWAS study. This sparse overlap could be explained by a phenomenon termed "phenotypic buffering" [129], where effects in one organizational layer (e.g. epigenetics) are not detectable anymore on the next layer (e.g. transcriptomics). A detailed explanation of this observation is beyond the scope of the present paper and needs further investigation.

Further following the scheme of a systems genetics approach, we integrated the two identified networks with intermediate clinical trait data. To this end, we investigated the relationships between changing levels of HDL-C, LDL-C and TG and all measured metabolites and transcripts, metabolic pathways and GO terms (Figure 3.17). A similar study already described an association between a gene-module derived from whole blood transcriptomics data and circulating lipid parameters including apolipoprotein B (APOB), HDL-C and triglycerides (TG) from a Finnish population cohort [193]. Our systematic analysis identified a large number of metabolites, transcripts, metabolic pathways, and functional GO categories that are all associated with the levels of circulating lipids. These findings further strengthen the assumption of a close link between system-wide metabolism, reflected by circulating metabolites and clinical lipid markers, and intracellular gene regulatory processes of blood cells. In addition, an opposite pattern between HDL-C and TG associations (Figure 3.17A-F) was observed from the phenotype networks which supports a previously suggested antagonistic regulation of both clinical

traits [244, 251]. However, the precise molecular mechanism behind this regulation is not entirely known, and our results might provide a basis for future studies to gain novel insights into the regulatory mechanisms of intermediate physiological phenotypes.

Combining results from all analysis steps allows for novel hypothesis generation. For example, for the well-known interactions between HDL-C, TG and BCAAs [63, 245], we discovered a potential regulatory pattern on different biological scales. In our first analysis step, we identified a strong negative association between the branched-chain amino acid isoleucine and ABCG1, a major constituent of lipid homeostasis and cholesterol metabolism [229, 252]. Second, at a more global level, the phenotype networks revealed an inverse association between HDL-C and TG, and also between HDL-C, TG and BCAAs (BCAAs are positively associated to TG, negatively to HDL-C, see Table S9). Third, in the TFBS enrichment analysis we were able to identify a clear regulatory signature of SREBPs in the vicinity of BCAAs, which are known to regulate cholesterol metabolism, indicating a potential coregulation between BCAAs and cholesterol metabolism at the transcriptional level. Interestingly, a combined study using cultured hepatocytes in a branched-chain amino acid-rich medium and obese mice showed that BCAAs directly induce the expression of SREBP1C which leads to hypertriglyceridemia, further supporting the suggested regulatory cross-link between HDL-C, TG and BCAAs [245]. This link is of particular interest since all three molecular traits have been associated to various diseases such as coronary artery disease, obesity and diabetes type II [253–255] and our observations might contribute to further decipher their underlying mechanisms.

In summary, our study highlights the potential of a systems genetics approach for understanding interactions across multiple biological scales in this case circulating metabolites and blood cellular gene expression - and how those insights can be used to generate novel hypothesis on mechanisms underlying common diseases.

Chapter 4

Plasma metabolites as proxy markers for inter-organ processes in diabetic mice

4.1 Background

By only monitoring single tissues, important regulatory interactions between tissues as well as tissue-specific processes relevant to understand the disease etiology might be missed. However, especially in humans, invasive measurements from relevant tissues/organs are often either infeasible, such as biopsies of brain tissue, or in case of other organs pose an aggravated risk for the patients of a normal population study. Such analysis are, however, possible in appropriate model organisms closely recapitulating human disease states and progression. For instance, using transcriptomics data measured in seven tissues of rat strains modeling hypertension, metabolic syndrome and cardiovascular disease, Xiao et al. [256] identified both ubiquitously present and tissue-specific gene co-expression cluster with disease relevance. In another study, Gao et al. [257] integrated transcriptomics data measured in six different tissues of a diabetes mouse model with clinical trait data to construct a multi-tissue trait-pathway network. Subsequent functional analysis revealed many trait relevant pathways across the different tissues. As a last example, Dobrin et al. [258] constructed bipartite co-expression networks between hypothalamus, liver and adipose tissue of around 300 outbred M16 (obese) and ICR (control) mice. Investigation of the tissue-to-tissue interactions revealed

several obesity related processes such as energy balance, stress response and immune response.

As already discussed in the introduction (Section 1.4), instead of using invasive procedures, another promising option to study disease-induced effects on relevant organs might be the usage of blood as non-invasive "surrogate tissue". In its physiological function, blood can be seen as road system between all tissues and organs of the body, not only supplying them with nutrients, but also removing metabolic waste products arising from the different tissues and organs [248]. As a result, the metabolite profiles in blood may be interpreted as a mixture of process signals, each of which can be attributed to a specific tissue or organ. Clinicians take advantage of this unique characteristic of blood, i.e. they determine the levels of molecules (markers) in blood tests for prognosis and diagnosis of diseases or for the assessment of organ function. For example, in glucose tolerance tests, blood glucose levels are used to test for diabetes, insulin resistance and impaired β-cell function [259]. Moreover, liver function tests, which measure the blood levels of proteins such as albumin, aspartate transaminase (AST) and alanine transaminase (ALT) are used to assess the liver function or injury state [260]. Using blood metabolomics data, several studies systematically linked altered metabolite levels with various physiological conditions including insulin resistance [113], type 2 diabetes [261] and cardiovascular disease [262].

In a previous study, Liew et al. [134] investigated the transcriptome of human peripheral blood cells using both expressed sequence tags (ESTs) and microarray technology. By systematically comparing the blood transcriptome with expression data from nine different human tissues, the authors detected a shared expression of 80% of all genes between blood and any tissue. Moreover, they experimentally validated the expression of 'tissue-specific' genes in blood cells and demonstrated that the expression levels of these transcripts in human blood cells are indicative of changes in their physiological environment (cf. Chapter 1.4). Motivated by these observations, we ask whether blood metabolite concentrations analogously to the expression of blood cell transcripts, reflect biological processes in organs under various physiological conditions, which could be utilized for diagnostic/prognostic purposes. More specifically, we will investigate whether easily accessible plasma metabolite concentrations can be used as proxy markers of organ processes. To this end, we used adipose tissue, kidney, liver and muscle samples, all simultaneously obtained from a diabetic mouse model and healthy controls (db/db and wild-type) and studied the association between organ and plasma metabolomes in an integrative approach. Moreover, a global gene expression profiling on kidney and liver

samples was carried out to further investigate whether signatures of transcriptional processes are also reflected by plasma metabolites. The only studies which, to the best of our knowledge, investigated the associations between different organs/fluids and plasma are (1) Do et al. [263], who calculated Gaussian graphical models (GGMs) of metabolite concentrations within and between three biological fluids, plasma, urin and saliva and (2) Torell et al. [61], who used hierarchical modeling to investigate how the metabolomes of different organs contribute to the plasma metabolic profile of mice.

In the following, we first systematically compare the global metabolomes of plasma, adipose tissue (adipose), kidney, liver and muscle samples, which are all central organs in the regulation of systemic metabolism, and investigate the distribution of detectable metabolites across the various organs (Figure 4.1). To examine the associations between plasma and organ metabolites, we compute pairwise bipartite correlation networks between each organ and plasma at both a pathway and a molecular level. In contrast to the previous Chapter 3, where we could exploit the large biological variation between individuals of a population study to asses the associations between molecules, this time, we use molecular data derived from a clonal mice population where a much smaller variation has to be expected. To account for this, we do not correct the data for the different treatments, instead we simply removed xenobiotics and those metabolites which where not detected in all treatments from further analysis. Moreover, the tissue samples from clonal mice are biological replicates which still contain a substantial amount of variation, for instance caused by variabilities in enzyme activities/ concentrations or nutritional states between the animals. Note that, despite the fact that GGMs have been proven to successfully reconstruct associations between metabolites [69], we again decided not to use them to construct the organ-plasma networks for the following reasons: (1) Many metabolites are shared between the various organs, thus conditioning on them might obscure the inference of associations. (2) In this study we are situated in a $n << p$ scenario. (3) We are not directly interested in reconstructing metabolic pathways between organs and plasma, but in identifying all possible organ processes which are associated to and thus reflected by plasma metabolites.

Using the constructed networks, we compare and analyze the extent to which the organ metabolomes are reflected by plasma metabolites starting at a pathway level through to single metabolites. We then extend the plasma-organ metabolite correlation network with an additional functional layer by integrating gene expression data measured in organ samples from kidney and liver. Functional enrichment analysis on the transcripts correlating with plasma metabolites reveals distinct signals for the two organs. More-

over, based on the observations from the network analysis, we infer different types of plasma-organ metabolite pairs that likely reflect different biological processes between organs and blood. Furthermore, we identify potential plasma proxy markers which either carry information on all investigated organs, or are specifically associated with molecules measured in one particular organ. Finally, we check whether we can shed further light onto type 2 diabetes as exemplary complex phenotype using our multi-tissue network and if concentrations of plasma metabolites can be used as proxy markers to detect disease-specific changes of metabolite and gene expression in the various organs. To this end, we determine differential gene expression/metabolite concentrations between db/db and wild-type animals and map significantly changed plasma-organ molecule pairs onto the multi-tissue metabolomics/transcriptomics network. Furthermore, we discuss the relevance of some of the diabetes-related pairs across the investigated organs with a particular focus on the earlier identified plasma proxy candidates.

The work presented in this chapter has been performed in collaboration with the group of Susanne Neschen at the Institute of Experimental Genetics. A manuscript is currently prepared for publication:

⋆ **Bartel, J.**, Neschen, S., Fridrich, B., Scheerer, M., Zukunft, S., Irmler, M., Kastenmüller, G., de Angelis, M.H., Adamski, J., Beckers, J., Theis, F.J., and Krumsiek, J. Plasma metabolite profiles as proxy markers for inter-organ processes in diabetic mice. *in preparation.*

4.2 Methods

All experimental procedures described in this section were performed by our collaboration partners Barbara Fridrich, Markus Scheerer, Sven Zukunft and Susanne Neschen at the Institute of Experimental Genetics from the Helmholtz Center Munich.

Animals and sample acquisition

We used data from 80 male leptin receptor deficient db/db mice and 20 male Dock7m+/+ (wt) littermates used in a previously published study on the effect of antidiabetic drugs in diabetic mouse models. Details on the sample acquisition as well as on the experimental procedures can be found in [264]. Briefly, db/db animals served as model for obesity induced type 2 diabetes and wild-type animals were used as lean, non-diabetic

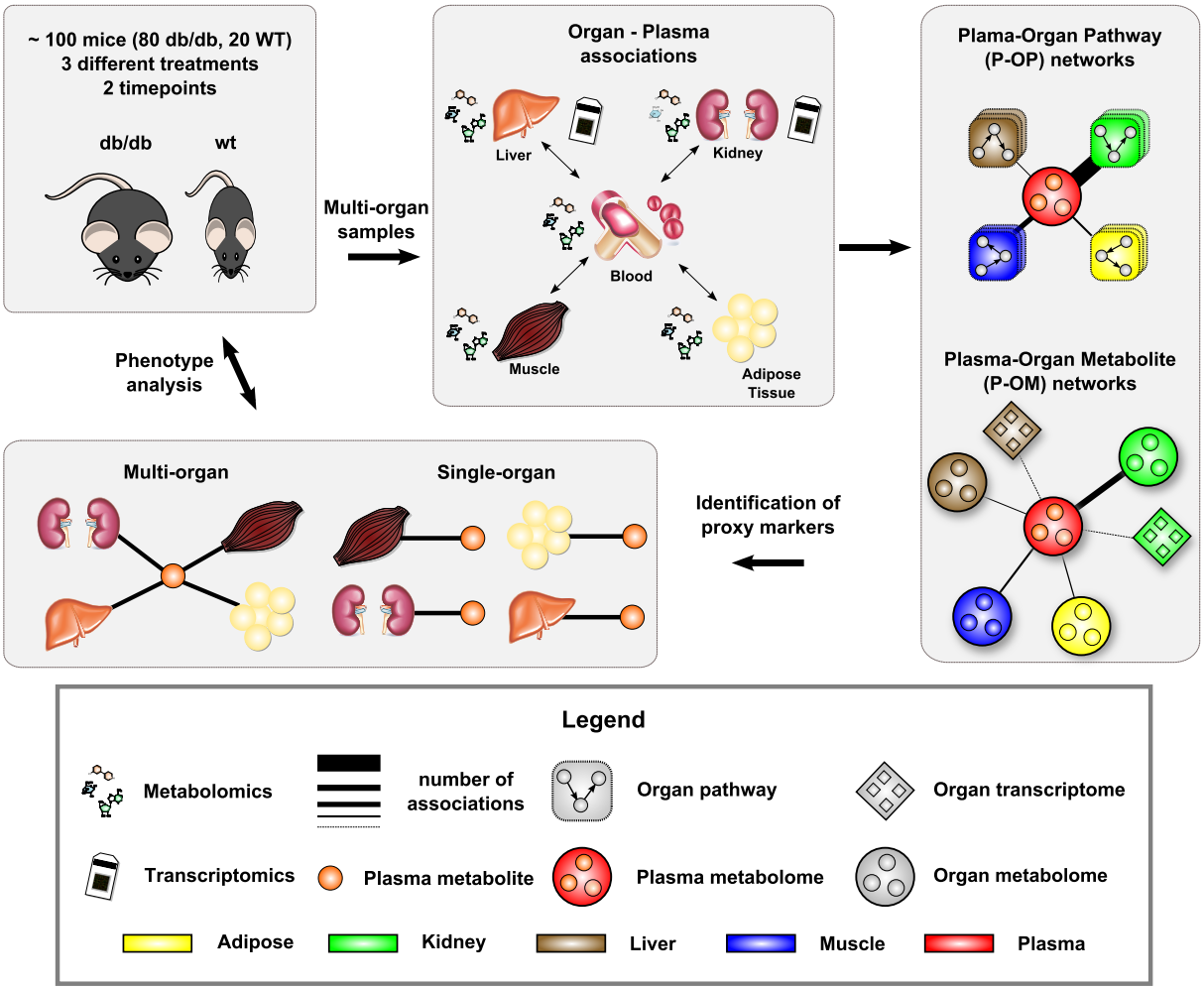


Figure 4.1: Data integration workflow for the systematic analysis of inter-organ associations. We combine high-throughput metabolomics data from five different organs (adipose tissue, kidney, liver, muscle and plasma) and transcriptomics data from two organs (kidney and liver) in inter-organ association networks to study how organ pathways (P-OP) and organ metabolite concentrations (P-OM) are reflected in plasma metabolites. Based on these associations, we then identify multi-organ and single-organ proxy markers. Finally, we investigate the usability of these identified proxy markers in the context of type 2 diabetes.

controls (Figure 4.1). In its original purpose, the study was designed to investigate the antidiabetic effect of two pharmaceuticals, via single treatment and in combination. After three weeks of age, all 100 mice were fed a commercially available, standardized high-fat diet. After eight weeks, animals received treatment either with vehicle solution (controls), metformin, SLGT-2 inhibitor or a combination of both compounds by oral gavage. At this point, mice were evenly split into an acute treatment group (10 wt, 40 db/db), which was sampled right after the first treatment, and a sub-chronic treatment group, which received further daily treatment for another two weeks until they were sacrificed. Before sacrifice, mice of both groups fasted for 4h, then plasma and organ samples were taken and stored at $-80\,^{\circ}\mathrm{C}$ for subsequent analysis. Animal experiments were approved by the Upper-Bavarian district government (Regierung von Oberbayern, Gz.55.2-1-54-2532-4-11).

Untargeted metabolomic profiling and preprocessing

Samples from a total of 100 mice were used for metabolomic profiling. Untargeted metabolic profiling was performed on plasma, adipose, kidney, liver and muscle samples from each mouse using ultrahigh-performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry (UPLC-MS/MS, GC-MS/MS) by Metabolon, Inc. as described previously [265, 266]. In total, 334 metabolites (123 known, 111 unknowns) were quantified in adipose tissue, 504 (300 knowns, 204 unknowns) in kidney, 508 (307 knowns, 201 unknown metabolites) in liver, 520 (312 known, 208 unknown metabolites) in muscle and 441 in plasma (276 known, 165 unknown metabolites). Note that numbers reported in the results might deviate from the total number of quantified metabolites due to sub-selection of mice samples or subsequent processing. Since the analysis focus was on endogenous metabolites, all molecules annotated as 'Xenobiotics' were removed from further analysis (3 in adipose tissue, 7 in kidney, 5 in liver, 6 in muscle, 8 in plasma). After the comparison of detectable metabolites between organs, metabolites with more than 20% missing values across samples were excluded from further analysis (104 in adipose tissue, 43 in kidney, 114 in liver, 66 in muscle, 88 in plasma). All metabolite concentrations were logarithmized and remaining missing values were imputed using the 'mice' R package [267]. A heatmap representation of the metabolomics data is shown in Figure 4.2.

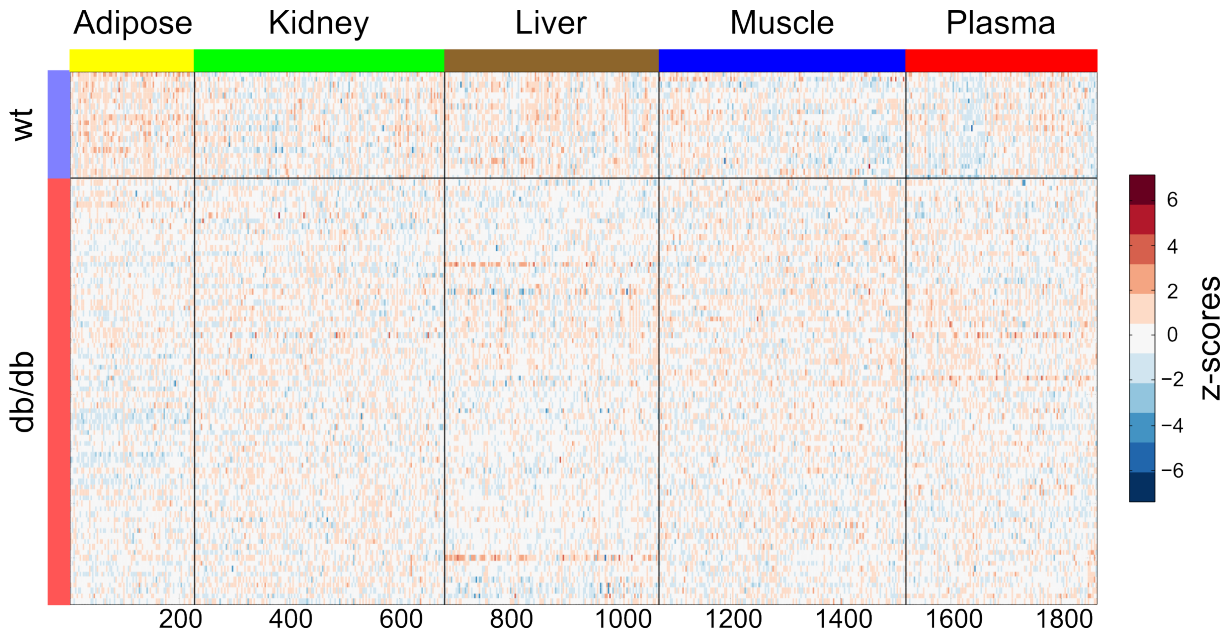


Figure 4.2: Heatmap representation of the metabolomics data across organs. The metabolite concentrations were z-score normalized in column direction. Metabolites are depicted on the x-axis across all organs, animals are shown on the y-axis.

Transcriptional profiling, preprocessing and quality control

For RNA profiling, kidney and liver samples of all 100 mice were collected. Total RNA was extracted from liver and kidney tissue employing the miRNeasy Mini kit (Qiagen) including on-column digestion of remaining DNA. The RNA quality was monitored by RNA 6000 Pico Assay (Agilent) and samples with RNA integrity number (RIN) values higher than 7 were used for microarray analysis. Total RNA was amplified using the Ambion Total Prep RNA Amplification Kit and all samples were hybridized on Illumina BeadChip MouseRef-8 v2 arrays. The arrays were scanned (Illumina HiScan) and the background subtraction was performed using the Illumina GenomeStudio V2009.1 software. An offset was added to all values in order to ensure non-negativity of all raw, non-transformed expression values. The data was then log-transformed and quantile normalized (CARMAweb). Genes with an average expression < 50 (before transformation) in at least one group were excluded from further analysis. For liver samples, 25,649 probes passed this criteria while all 25,697 probes passed the filtering for kidney samples.

Data adjustment

In order to avoid spurious false positive associations generated by genetic or treatment duration effects, a linear model with genotype and time as covariates was used to adjust the metabolomics and transcriptomics data. All subsequent calculations were performed on the respective residuals. Note that for correlation network generation, the data was not corrected for treatment effects, since it was shown that strong non-technical variation e.g. due to varying conditions is beneficial for the reconstruction of correlation networks [69] . For the phenotype analysis (Section 4.8), only the vehicle-treated mice were used (n=20 wt, n=20 db/db) and both data sets were corrected against duration of treatment (time) effects.

Bipartite network generation

All bipartite correlation networks were calculated based on Spearman's rank correlation coefficients. For the plasma-organ pathway network, aggregated z-scores ($aggZ$-Scores) were calculated for each organ super- and subpathway in a similar way as already described previously (see Chapter 3.2). Subpathways with less than two members were discarded, leaving 47 subpathways in adipose tissue, 54 in muscle, and 62 in liver and kidney for further analysis. Moreover, unknown metabolites were excluded from the pathway analysis. Subsequently, correlation coefficients were calculated between the concentrations of all possible plasma metabolite-organ pathway pairs ($345 * 7$ in case of superpathways, and $345 * X$; X= number of tissue subpathways). Taking the broad level of some super- and subpathway annotations into account, a more loose significance threshold for metabolite-pathway correlations was set at $\alpha = 0.05$ after Bonferroni correction. The plasma-organ metabolite network was constructed analogously by calculating all possible pairwise correlations between plasma metabolites and organ metabolites for each tissue separately. Note that this time unknown metabolites were included into the association analysis. Statistical significance was determined at $\alpha = 0.01$ after Bonferroni correction and the four pairwise correlation networks were merged into one network based on the associated plasma metabolites. In a subsequent step, the network was extended by an additional functional level, i.e. we calculated correlations between the concentrations of all plasma metabolites and liver/kidney transcript pairs. We have seen in the previous chapter that correlations between metabolites and transcripts are generally lower than between two metabolites or two transcripts (recall Figure 3.6A). To account for this, again a more loose statistical significance threshold was set for as-

sociations between metabolites and transcripts at a false discovery rate (FDR) of 0.05. Significant correlations were visualized as networks using yEd graph editor (yWorks GmbH, Tuebingen; `http://www.yworks.com`).

Enrichment analysis

Functional enrichment of gene sets contained in the extended plasma-organ metabolite (P-OM) network was assessed using pathway enrichment analysis on: (1) all terms belonging to the 'biological process' (BP) domain of the full Gene Ontology (GO) resource and (2) terms of the BP and 'molecular function' (MF) domains contained in the generic GO Slim subset collection. Statistical significance of gene sets in the list of network contained genes was evaluated using hypergeometric tests. To account for multiple testing, a significance threshold for the tested terms was set at a FDR ≤ 0.05.

Statistical data analysis

Principal component analysis (PCA) was performed on 85 of the 104 metabolites shared between all investigated organs and surpassing filtering criteria (i.e. missing value exclusion) to identify major groupings in the metabolomics data. To assess the variation between all investigated organs, each tissue sample was considered separately, that is 5 samples per mouse and treatment condition (data matrix 498×85). Prior to the analysis, each metabolite variable was scaled to zero mean and standard deviation one. In order to determine an adequate number of principal components, the explained variance per principal component was calculated according to $\lambda_i / \sum_{i=1}^{n} \lambda_i$, with λ_i being the eigenvalue of the i-th original random variable. Visual inspection reveals clear breaks in the explained variance between the first and second, and forth and fifth principal components (Figure 4.3). Since the first two principal components already explained roughly two-thirds of the total variability, a model with two principal components was chosen. Differential analysis of both transcript expression and metabolite concentrations between all organ samples of wild-type and diabetic (db/db) animals was carried out using Student's t-tests. Significantly changed metabolites/transcripts were determined at $\alpha = 0.05$ after multiple testing correction by the FDR procedure. All statistical analyses and data visualizations were performed using Matlab (version 7.12.0) statistical software.

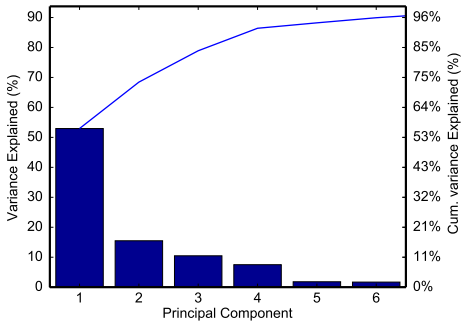


Figure 4.3: Scree plot of PCA on metabolites shared between all organs. Left y-axis shows the variance explained by each principal component (x-axis). Right y-axis and blue line indicate the cumulative variance explained. The first two principal components already explain $\sim$ 70% of the total variance.

4.3 Distribution of detectable metabolites across different organs

In the first step of our analysis, we compared the sets of metabolites detectable in the different organ samples. To this end, we counted all metabolites (with known and unknown structural identity) detectable in at least one organ sample. To avoid spurious counts in the metabolomes caused by the different treatments, for instance drug metabolites and their degradation products, only samples of mice treated with a vehicle solution (wt = 20, db/db= 20) were used in this analysis step.

In total, 1048 unique metabolites were profiled within the 'control' condition across all tissue comparisons. Most metabolites were detected in muscle (514), followed by liver (499), kidney (495), plasma (432) and adipose tissue (330). Out of the 1048 metabolites, around 46% were observable in a single tissue whereas only 11% are shared across all tissues (Figure 4.4A). The largest number of uniquely detected metabolites was found in plasma (129), followed by muscle (111), liver (107), kidney (99) and adipose tissue (39). Partitioning all metabolites detected in a single tissue into their metabolite classes revealed an equal coverage for uniquely detected metabolites across five metabolites classes (amino acids, carbohydrates, cofactors and vitamins, lipids, unknowns), irrespective of the underlying tissue (Figure 4.4B). Peptides are most frequently detected in muscle and plasma, nucleotides in kidney, and energy metabolites in liver. Out of the metabolites detected in all tissues, the largest fractions belong to energy, lipid, amino acids and nucleotides (Figure 4.4B).

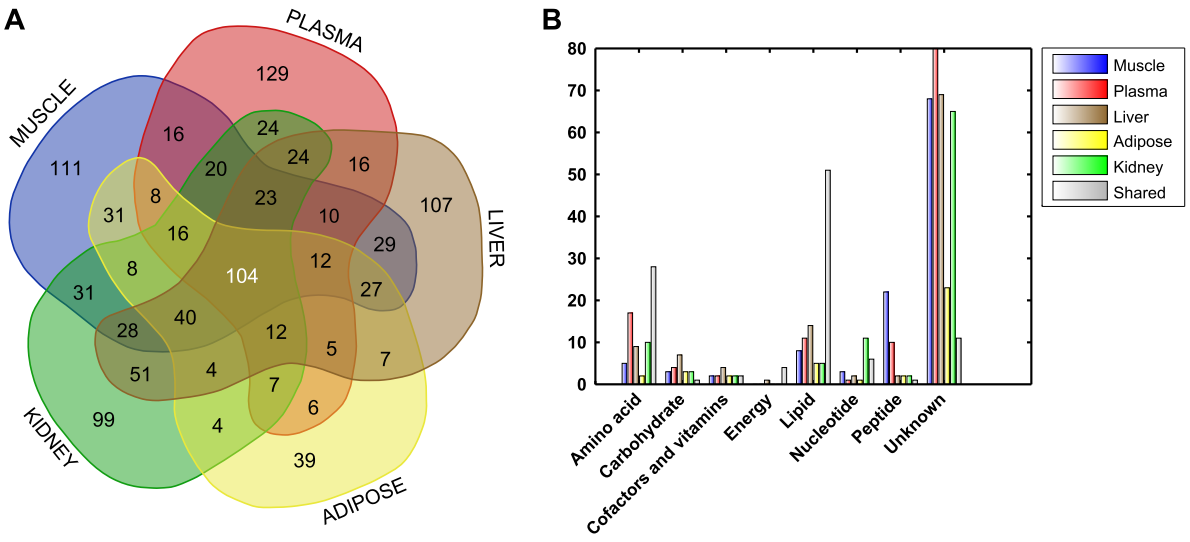


Figure 4.4: Comparison of metabolomes across organs. **A:** Venn diagram showing the overlap of detected metabolites between the measured tissues. Plasma yields the highest number of metabolites exclusively detected in a single tissue, while the fewest were found in fat. In a pairwise comparison with the plasma metabolome, most metabolites were shared between plasma and kidney, followed by muscle, liver and adipose tissue. Around 11% of the total detected metabolites were shared between all tissues. **B:** Metabolite classes of exclusively detected and ubiquitously shared metabolites. The bars indicate the fraction of metabolites belonging to the respective metabolite class. For uniquely detected metabolites, an even distribution between organs can be observed for most metabolite classes. Exceptions are the energy, nucleotide and peptide classes. For metabolites commonly detected in all organs, the largest fractions belong to the energy, lipid, amino acid and nucleotide classes.

Class	Plasma	Adipose	Kidney	Liver	Muscle
Amino Acids	71 (84)	44 (63)	76 (83)	68 (76)	69 (77)
Carbohydrate	12 (15)	8 (14)	23 (28)	30 (31)	28 (30)
Cofactors and vitamins	10 (13)	6 (11)	15 (15)	17 (21)	10 (11)
Energy	7 (7)	6 (6)	8 (8)	7 (8)	7 (7)
Lipid	112 (119)	73 (89)	107 (113)	100 (126)	103 (120)
Nucleotide	10 (14)	15 (23)	30 (30)	18 (21)	20 (22)
Peptide	10 (16)	2 (14)	15 (16)	6 (17)	31 (39)
Unknown	113 (164)	73 (110)	180 (202)	143 (199)	180 (208)
Sum	345 (432)	227 (330)	454 (495)	389 (499)	448 (514)

Table 4.1: Metabolite class distributions in the final data set. All quantified metabolites were classified according to their primary superpathway. Note that only metabolites with $< 20\%$ missing values were kept for further analysis. Numbers in brackets indicate detected metabolites per metabolite class prior to the removal of missing values.

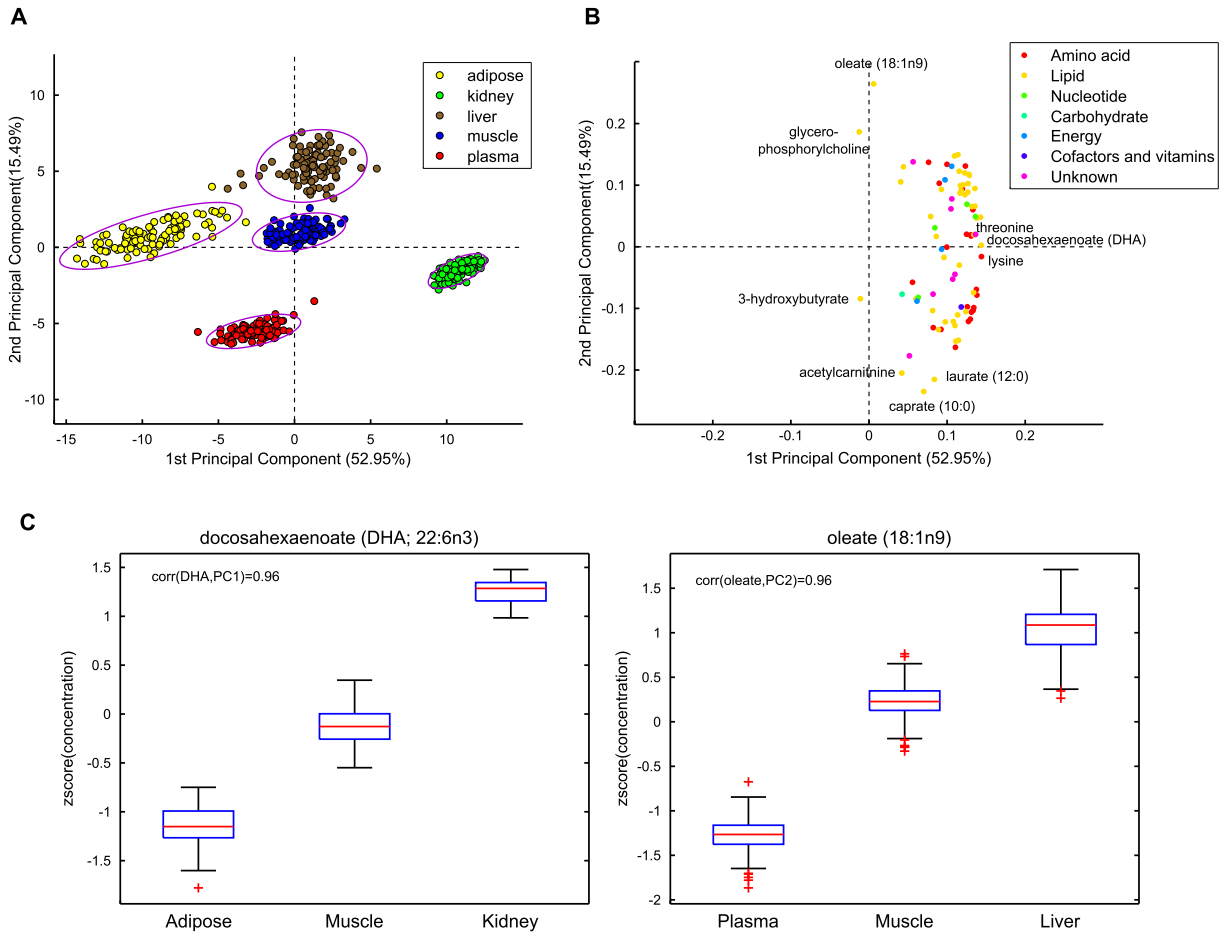


Figure 4.5: PCA score and respective loadings plot of all organ samples. The PCA was performed using the metabolite overlap (85) between all organs. **A**: Shown are the scores of the first and second principle components, which together account for 68.44% of the total variation in the data. Purple ellipses illustrate the 95% confidence interval for each group of tissue samples. A clear separation between samples of the different organs was observable; numbers in brackets depict the variance explained by the respective principal component.**B**: Loadings plot for the first two principle components. Loadings are colored according to their superpathway annotations. Selected loadings are labeled by the metabolite name. **C**: Boxplots showing the concentrations of docosahexaenoate (DHA) and oleate across three organs, respectively. The concentrations of DHA show a strong correlation with the first principal component, which defines a common (linear) axis of variation between adipose tissue, muscle and kidney. Similarly, oleate defines a common axis of variation between plasma, muscle and liver.

Prior to all subsequent analysis steps, metabolomics measurements were further pre-processed by removing metabolites with $> 20\%$ missing values, resulting in a slightly different metabolite panel (Table 4.1; see Section 4.2 for details). Among the identified metabolites, the lipid class represented the largest fraction across all investigated

tissues, followed by amino acids. A larger difference between the organs can be seen for the classes of carbohydrates, cofactors and vitamins, nucleotides and peptides. For example, 30 carbohydrates were quantified in liver, but only 8 in adipose tissue, or 31 peptides in muscle, but only 2 in adipose. Lastly, a large fraction of metabolites with unknown structure were identified within all investigated organs.

Using the metabolites shared across all five tissues and all samples (data matrix 498×85), we performed a principal component analysis (PCA) to investigate the relationships of the metabolite profiles between the different organs. Based on the metabolic signatures, a clear discrimination between the different organs was observable, whereas samples taken from the same organ clustered together closely (Figure 4.5A). In comparison, metabolite profiles of samples derived from adipose and liver showed a higher variability in the first two principal components than those of the remaining organs (as indicated by the size of the 95% confidence intervals). Inspecting the principal components in detail, we observed that the first principal component mainly separated kidney and adipose tissue samples from the rest of the data, thereby forming a common axis between adipose tissue, muscle and kidney samples. Similarly, the second principal component mainly separated liver from plasma samples, again forming a common axis with muscle. These observations are also reflected in the corresponding loadings (Figure 4.5B). Almost all loadings pointed in the direction of kidney samples indicating that these metabolites have a higher level in kidney when compared to the remaining organs. Only a few loadings pointed into another direction further emphasizing that for almost all loadings the lowest levels can be found in adipose tissue. Moreover, we identified two metabolites, docosahexaenoate (DHA) and oleate, which most strongly correlated with the first and second principal components, respectively (Figure 4.5B). Interestingly, the concentrations of these two metabolites, showed a linear increase between adipose tissue, muscle and kidney (Figure 4.5C, left panel) and plasma, muscle and liver (Figure 4.5C, right panel) further indicating the above-mentioned common linear axes of variation.

Taken together, these results show that the overall metabolic profiles are distinct between the examined organs pointing to a strong metabolic compartmentalization and - at least with the analytical technique applied here - many metabolites are only detectable in specific tissues. In the following, we will investigate to which extent the footprint of systemic metabolism is reflected in circulating plasma metabolites. To this end, we analyze the correlation structure between plasma and organ metabolites at both a pathway level and individual molecule level.

4.4 Specific organ metabolic processes are reflected in circulating plasma metabolites

To assess the associations between plasma metabolites and metabolic pathways occurring in the distinct organs, we again carried out an aggregated z-score analysis ($aggZ$-Score; cf. Section 3.2) on each metabolic sub- and superpathway to construct bipartite plasma-organ pathway (P-OP) networks. Briefly, for each metabolic pathway, we calculated the mean z-score of all associated metabolites in the respective tissue. We then calculated the Spearman's rank correlation between each plasma metabolite and pathway $aggZ$-Score. An edge was included in the network model if it is significantly different from zero with α=0.05 after Bonferroni correction.

At the superpathway level, the resulting P-OP network consisted of a total of 291 edges between 120 unique plasma metabolites and 18 super pathway nodes across the different organs (Figure 4.6A). The largest fraction of the 291 edges was observed between plasma metabolites and pathways located in kidney ($\sim$ 81%; 6 superpathways, 100 plasma metabolites), followed by muscle ($\sim$ 14%; 6 superpathways, 30 plasma metabolites), liver ($\sim$ 4%; 3 superpathways, 12 plasma metabolites) and adipose ($\sim$ 1%; 3 superpathways, 2 plasma metabolites). Of note, 216 out of the 236 edges ($\sim$ 92%) between plasma metabolites and kidney superpathways represent negative correlations, whereas a more even distribution between negative and positive associations can be seen for the remaining organs. Focusing on the 120 plasma metabolites that display significant associations with organ superpathways, 79 plasma metabolites were exclusively connected to kidney pathways, 15 exclusively to muscle, 3 exclusively to liver and 1 exclusively to adipose tissue. The remaining 22 plasma metabolites were mainly shared between kidney and muscle (12) or kidney and liver (8). Figure 4.6B shows only the strongest signals observed for each super pathway in the respective tissue. Notably, not a single tissue exhibited significant correlations for all seven super pathways. For kidney and muscle, significant correlations between plasma metabolites and six super pathways were observable, 3 for liver and adipose tissue. The overall strongest signal was observed between an unknown plasma metabolite, namely X-12649 and 'carbohydrate metabolism' located in kidney (ρ = -0.776; Figure 4.6C).

In the next step of our analysis, we calculated the bipartite P-OP correlation network at a subpathway level, which allows a more fine-grained functional analysis. Applying a significance threshold of α=0.05 after Bonferroni correction resulted in an overall network of 980 edges connecting 172 plasma metabolites with 82 organ pathways (Figure

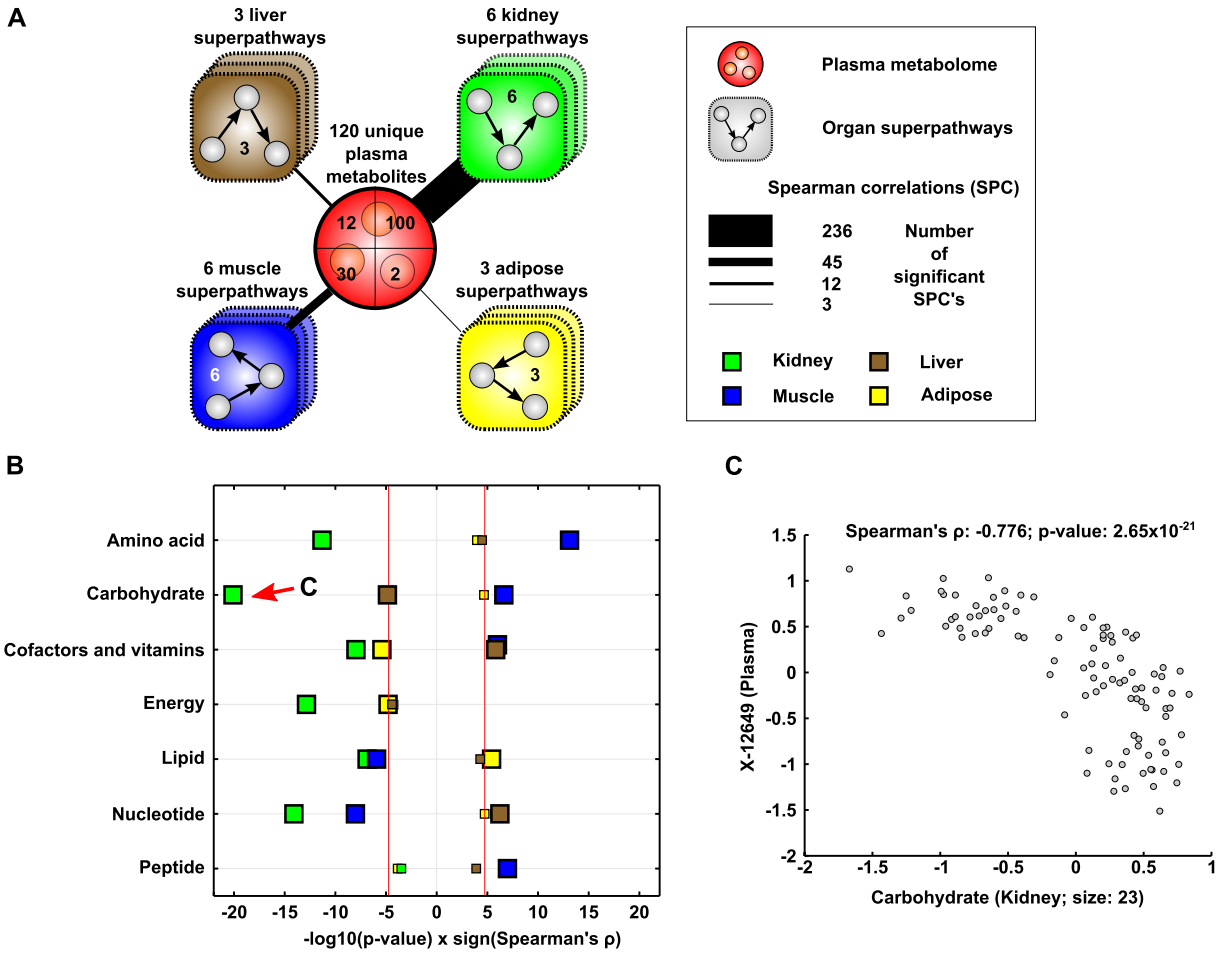


Figure 4.6: Plasma-Organ superpathway analysis. **A:** Schematic network representation between plasma metabolites and organ superpathways. Numbers within the circle indicate plasma metabolites connected to the respective number of organ superpathways **B:** Strongest associations observed between plasma metabolites and each superpathway of the respective organs. Red vertical lines denote the significance threshold with $\alpha = 0.05$ after Bonferroni correction and a corresponding p-value of $1.81 \times 10^{-}5$. Insignificant associations are shown as smaller rectangles. Overall, the strongest signals were observed for kidney (green rectangles) and muscle (blue rectangles) superpathways, while less strongly signals are present for liver (brown rectangles) and adipose tissue (yellow rectangles). **C**: Scatterplot of the strongest superpathway association. The *aggZ*-Score of kidney 'carbohydrate metabolism' comprises the concentrations of 23 metabolites. For this pathway, we observe a strong negative association with an unknown plasma metabolite. Scales are based on the linear regression residuals from data adjustments (Section 4.2).

4.7A). Again, most edges ($\sim$ 72%) were observable between 30 kidney subpathways and 130 plasma metabolites, followed by muscle ($\sim$ 19%; 30 subpathways, 84 plasma me-

tabolites), liver ($\sim$ 7%; 15 subpathways, 60 plasma metabolites) and adipose ($\sim$ 2%; 7 subpathways, 12 plasma metabolites). Also consistent with the observations made at a superpathway level, a strong bias towards negative associations was observable between plasma metabolites and kidney subpathways ($\sim$ 89%), whereas an even distribution between positive and negative associations was observable for the remaining organs. The largest amount of the 172 plasma metabolites included in the network are exclusively connected to kidney pathways, followed by liver and muscle, while no metabolite was uniquely connected to adipose tissue (Figure 4.7B). We found representatives of all metabolite classes among the plasma metabolites connected to organ pathways, the largest of which are unknowns (33%), lipids (33%) and amino acids (23%; Figure 4.7C). The strongest, significant associations observed between plasma metabolites and organ subpathways are given in Table 4.2. Similar to the observations from the superpathway analysis above, the unknown plasma metabolite X-12649 was involved in the overall strongest signal associating with 'glycine, serine and threonine metabolism' in kidney (ρ = -0.815). The refined assignments into subpathways allowed more precise interpretation of the observed associations. For instance, focusing on adipose tissue, we identified a variety of significant associations for 'carnitine metabolism', 'medium-' and 'long chain fatty acid metabolism' and 'sphingolipid metabolism', all subpathways belonging to the lipid superpathway (Table 4.2). But we also observed signals of adipose tissue pathways which did not reach significance in the superpathway analysis, such as glutamate metabolism, polyamine metabolism and urea cycle, arginine-, proline metabolism, which belong to the amino acid metabolism.

Taken together, these results confirm that levels of single plasma metabolites globally reflect signatures of specific metabolic pathways proceeding in the different organs to a varying extent, both at a coarse superpathway level and at the more fine-grained subpathway level.

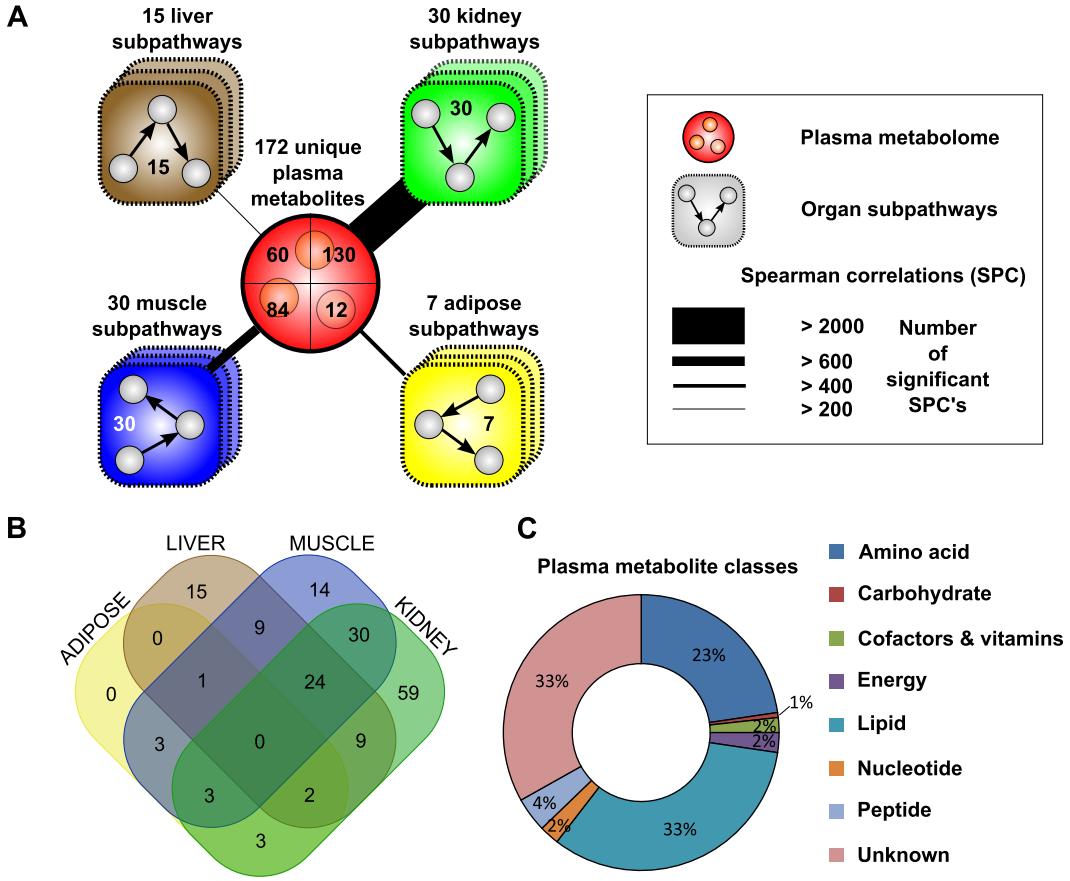


Figure 4.7: Plasma-reflected organ subpathways. **A:** Schematic representation of the P-OP network at a subpathway level. Indicated are the numbers of plasma metabolites and organ subpathways connected to each other. Edge widths represent the relation of observed edges in the respective tissue against the smallest number of observed edges (liver). **B:** Overlap of plasma metabolites associated with at least one pathway of the respective organ. Many plasma metabolites are uniquely connected to a single organ. **C:** Distribution of metabolite classes across the P-OP plasma metabolites.

Tissue	Pathway	Size	ρ	P-value	Plasma Metabolite
ADIPOSE	Glutamate metabolism	2	0.511	$5.69E-08$	dimethylglycine
	Polyamine metabolism	2	0.493	$1.83E-07$	glycerol
	Long chain fatty acid	17	0.483	$3.50E-07$	X-03002
	Urea cycle; arginine-, proline-, metabolism	6	0.477	$5.14E-07$	X-11787
	Sphingolipid	2	-0.475	$5.94E-07$	X-17160
	Carnitine metabolism	4	0.473	$6.80E-07$	dimethylglycine
	Medium chain fatty acid	6	-0.45	$2.70E-06$	X-16982
KIDNEY	Glycine, serine and threonine metabolism	8	-0.815	$6.35E-25$	X-12649
	Pyrimidine metabolism, uracil containing	6	-0.731	$5.59E-18$	phenylalanine
	Sterol/Steroid	7	-0.727	$1.16E-17$	2-palmitoylglycerophosphocholine*
	Krebs cycle	5	-0.723	$1.97E-17$	X-12649
	Fructose, mannose, galactose, starch, and sucrose metabolism	3	-0.718	$4.09E-17$	X-12649
	Urea cycle; arginine-, proline-, metabolism	6	-0.715	$6.39E-17$	2-palmitoylglycerophosphocholine*
	Alanine and aspartate metabolism	6	0.7	$5.20E-16$	X-12776
	Nucleotide sugars, pentose metabolism	8	-0.689	$2.41E-15$	tryptophan
	Glycolysis, gluconeogenesis, pyruvate metabolism	9	-0.686	$3.60E-15$	X-12649
	Cysteine, methionine, SAM, taurine metabolism	9	-0.685	$3.69E-15$	X-12029
	Monoacylglycerol	4	-0.683	$4.72E-15$	2-palmitoylglycerophosphocholine*
	Sphingolipid	4	-0.679	$8.34E-15$	X-12649
	Inositol metabolism	3	-0.665	$4.31E-14$	2-palmitoylglycerophosphocholine*
	Aminosugars metabolism	3	-0.644	$4.72E-13$	X-12649
	Pyrimidine metabolism, thymine containing	2	-0.634	$1.39E-12$	methionine
	Diacylglycerol	2	0.631	$1.90E-12$	X-09045
	Glycerolipid metabolism	6	-0.625	$3.83E-12$	2-aminoadipate
	Lysine metabolism	4	-0.622	$5.16E-12$	pantothenate
	Essential fatty acid	7	0.603	$3.30E-11$	3-hydroxybutyrate (BHBA)
	Ascorbate and aldarate metabolism	4	-0.587	$1.34E-10$	methionine
	Purine metabolism, urate metabolism	2	-0.578	$2.93E-10$	methionine
	Butanoate metabolism	2	-0.577	$3.31E-10$	gamma-glutamylleucine
	Oxidative phosphorylation	3	0.576	$3.50E-10$	X-09045
	Purine metabolism, guanine containing	5	0.569	$6.81E-10$	X-12776
	Creatine metabolism	2	-0.559	$1.46E-09$	2-palmitoylglycerophosphocholine*
	Purine metabolism, (hypo)xanthine/inosine containing	5	-0.536	$8.88E-09$	methionine
	Phenylalanine & tyrosine metabolism	4	-0.466	$1.02E-06$	1-stearoylglycerophosphocholine
	Eicosanoid	2	-0.466	$1.05E-06$	X-12385
	Fatty acid metabolism (also BCAA metabolism)	2	0.461	$1.36E-06$	X-12465
	Valine, leucine and isoleucine metabolism	17	-0.518	$3.41E-08$	2-docosapentaenoylglycerophosphocholine*
LIVER	Fatty acid metabolism (also BCAA metabolism)	2	0.763	$2.69E-20$	N-acetylglycine
	Valine, leucine and isoleucine metabolism	5	0.583	$1.97E-10$	N-acetylglycine
	Medium chain fatty acid	4	-0.571	$5.44E-10$	levulinate (4-oxovalerate)
	Glutamate metabolism	6	0.563	$1.05E-09$	2-aminobutyrate
	Nicotinate and nicotinamide metabolism	2	0.557	$1.81E-09$	X-11533
	Fructose, mannose, galactose, starch, and sucrose metabolism	9	-0.55	$2.98E-09$	spermidine
	Glutathione metabolism	6	0.528	$1.70E-08$	2-aminobutyrate
	Urea cycle; arginine-, proline-, metabolism	5	0.526	$1.97E-08$	X-11875
	Purine metabolism, (hypo)xanthine/inosine containing	4	0.509	$6.31E-08$	N-acetylglycine
	Polyamine metabolism	3	-0.504	$9.08E-08$	tryptophan
	Glycolysis, gluconeogenesis, pyruvate metabolism	9	-0.5	$1.18E-07$	X-11787
	Inositol metabolism	2	-0.489	$2.40E-07$	X-12786
	Sphingolipid	2	-0.482	$3.82E-07$	lysine
	Nucleotide sugars, pentose metabolism	8	-0.456	$1.81E-06$	leucine
	Glycerolipid metabolism	8	0.454	$2.13E-06$	allantoin
	Pantothenate and CoA metabolism	4	0.453	$2.28E-06$	histidine
MUSCLE	Ketone bodies	2	0.701	$9.86E-16$	3-hydroxybutyrate (BHBA)
	Dipeptide derivative	2	0.658	$1.72E-13$	methyl palmitate (15 or 2)
	Glycine, serine and threonine metabolism	9	-0.641	$1.16E-12$	X-16209
	Medium chain fatty acid	5	0.638	$1.64E-12$	methyl palmitate (15 or 2)
	Cysteine, methionine, SAM, taurine metabolism	9	0.623	$7.59E-12$	X-12776
	Pyrimidine metabolism, cytidine containing	3	0.608	$3.18E-11$	methyl palmitate (15 or 2)
	Tryptophan metabolism	3	0.605	$4.34E-11$	X-12776
	Oxidative phosphorylation	2	-0.604	$4.54E-11$	X-13581
	Creatine metabolism	2	0.574	$6.72E-10$	methyl palmitate (15 or 2)
	Pyrimidine metabolism, uracil containing	4	-0.57	$8.84E-10$	X-16588
	Sphingolipid	2	0.568	$1.02E-09$	glycerol
	Urea cycle; arginine-, proline-, metabolism	7	-0.556	$2.71E-09$	2-aminobutyrate

Continued on next page

Table 4.2 – continued from previous page

Tissue	Pathway	Size	ρ	P-value	Plasma Metabolite
	Sterol/Steroid	5	0.551	$4.02E-09$	glycerol
	Glutamate metabolism	4	-0.549	$4.96E-09$	X-16588
	Histidine metabolism	2	0.544	$6.82E-09$	methyl palmitate (15 or 2)
	Glutathione metabolism	6	0.539	$1.04E-08$	azelate (nonanedioate)
	Carnitine metabolism	7	-0.536	$1.31E-08$	X-16588
	Valine, leucine and isoleucine metabolism	10	0.528	$2.25E-08$	X-04766
	Phenylalanine & tyrosine metabolism	3	0.527	$2.54E-08$	X-12505
	Nucleotide sugars, pentose metabolism	7	-0.512	$7.04E-08$	methyl palmitate (15 or 2)
	Purine metabolism, (hypo)xanthine/inosine containing	4	0.509	$8.49E-08$	X-12505
	Pantothenate and CoA metabolism	5	0.503	$1.26E-07$	dimethylglycine
	Monoacylglycerol	4	0.491	$2.90E-07$	acetylcarnitine
	Fatty acid, branched	2	-0.483	$4.87E-07$	X-16588
	Alanine and aspartate metabolism	5	-0.474	$8.02E-07$	X-12649
	Glycerolipid metabolism	7	-0.469	$1.14E-06$	proline
	Lysine metabolism	4	0.466	$1.35E-06$	azelate (nonanedioate)
	Fructose, mannose, galactose, starch, and sucrose metabolism	9	0.462	$1.68E-06$	glycerate
	Long chain fatty acid	17	-0.462	$1.73E-06$	X-16588
	Dipeptide	28	0.459	$1.97E-06$	methyl palmitate (15 or 2)

Table 4.2: Strongest significant associations observed between single plasma metabolites and all possible organ subpathways. Associations with p-values $< 3.08 \cdot 10^{-6}$ meet global significance, which corresponds to a significance level of $\alpha = 0.05$ after Bonferroni correction. Note that only the strongest hit per subpathway is shown. For kidney and muscle, significant associations with plasma metabolites were observed for 30 subpathways, 15 for liver and 7 for adipose tissue. ρ, Spearman's rank correlation coefficient.

4.5 Bipartite plasma-organ metabolic network reveals to which extent the organ metabolomes are mirrored in plasma at a molecular level.

Having established an observable association between plasma metabolites and activity signatures of metabolic pathways in organs, we next sought to systematically investigate to which extent concentrations of organ metabolites are reflected by plasma metabolites. More specifically, we evaluated the associations between plasma and organ metabolite concentrations at a molecular level, which might not only provide mechanistic insights into the interplay between organs, but also enables the integration of other molecular levels, such as transcriptional regulation, and the identification of potential proxy markers. Analogous to the previous section, we again calculated Spearman's rank correlation coefficients between the different molecules. Note that throughout this analysis, we are only focusing on associations between plasma metabolites and organ molecules (Figure 4.8A), since our major aim is to identify plasma proxy markers for organ processes. Nevertheless, to get a global overview of the observable associations between different organs, we will provide a short comparison of cross-organ associations in the

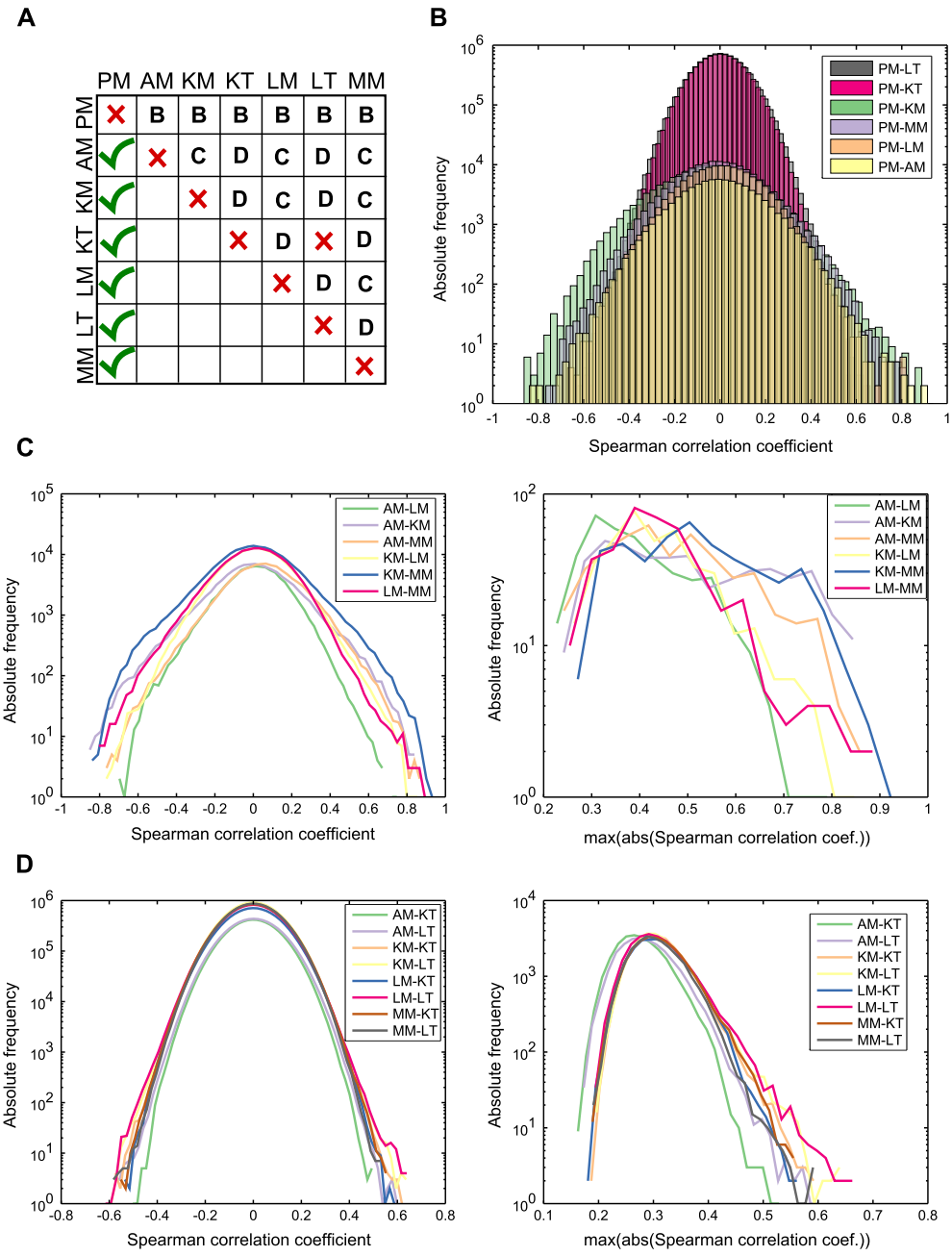
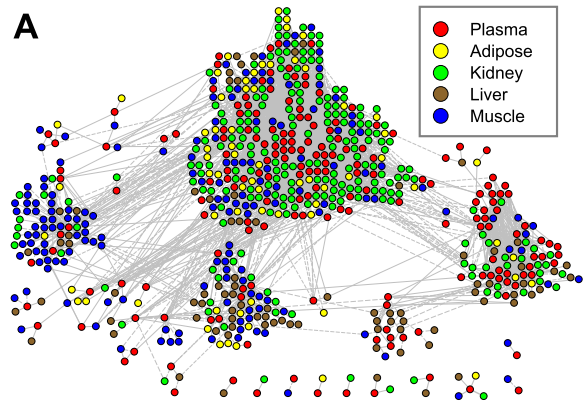


Figure 4.8: Analysis matrix and cross-organ correlations. **A:** Overview matrix showing the cross-organ associations considered in this study. Green checkmarks highlight cross-organ associations which are thoroughly investigated in this analysis. Red crosses indicate associations which were deliberately omitted from our analysis. Letters indicate in which subplot the respective cross-organ association is shown. Abbreviations: PM=plasma metabolome, AM=adipose tissue metabolome, KM=kidney metabolome, LM=liver metabolome, MM=muscle metabolome, KT=kidney transcriptome, LT=liver transcriptome **B:** Distributions of correlation coefficients for all relevant plasma-organ associations. **C:** Distributions of correlation coefficients between metabolomes of all remaining cross-organ associations. The left panel shows all pairwise correlations, the right panel shows only the maximal absolute correlation coefficient per metabolite. **D:** Distributions of correlation coefficients between metabolomes of all organs and liver and kidney transcriptomes. The left panel shows all pairwise correlations, the right panel shows only the maximal absolute correlation coefficient per metabolite.



B

Class	Plasma	Adipose	Kidney	Liver	Muscle
Amino Acids	61%	25%	49%	27%	61%
Carbohydrate	33%	25%	78%	23%	21%
Cofactors and vitamins	40%	0%	40%	24%	20%
Energy	71%	33%	88%	29%	29%
Lipid	47%	21%	38%	15%	21%
Nucleotide	50%	13%	47%	33%	55%
Peptide	40%	0%	27%	0%	26%
Unknown	58%	48%	47%	23%	39%
Total	53%	30%	47%	22%	37%

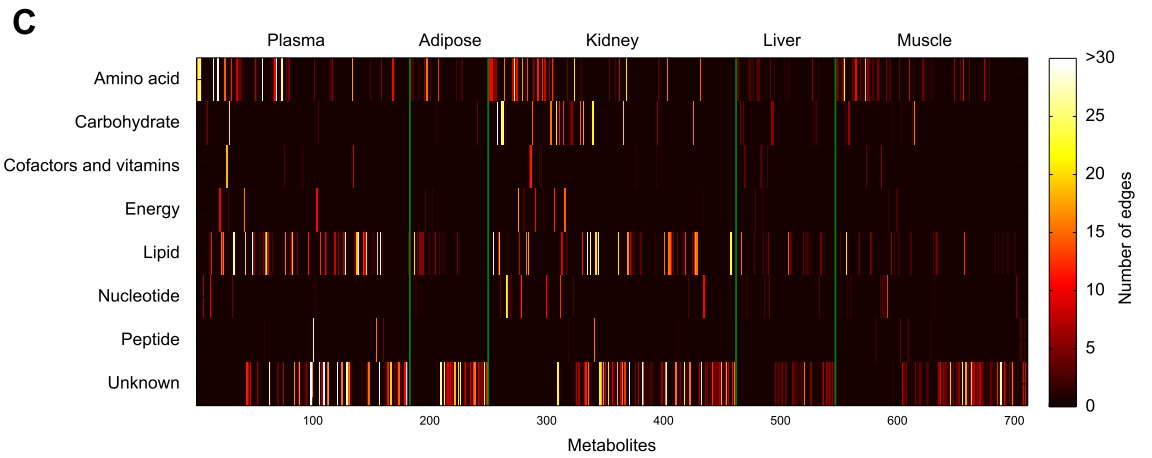


Figure 4.9: Bipartite plasma-organ metabolite network. **A:** Fused network representation of the significant Spearman correlations ($\alpha = 0.01$ after Bonferroni correction) between plasma metabolites and metabolites from the four investigated organs. Many plasma metabolites exhibit a high connectivity observable from the central positions they occupy in the P-OM network. **B:** Fractions of total measured metabolites per organ occurring in the P-OM network, partitioned according to their metabolite class. **C:** Heatmap illustrating the number of edges for each metabolite in the network grouped by metabolic classes. Overall, a weak connectivity is observable for metabolites quantified in liver and adipose. The metabolite exhibiting the strongest connectivity in the network is the plasma metabolite 2-palmitoylglycerophosphocholine, which is connected to 132 organ metabolites.

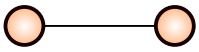
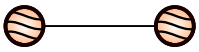
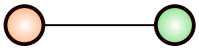
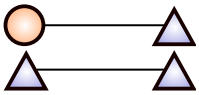
following. First, focusing on the associations between plasma metabolites and metabolites/transcripts of organs, we observed very similar distributions of correlations between plasma metabolites and both transcriptomes and only minor differences between the distributions of correlations between plasma metabolites and metabolites from the four organs (Figure 4.8B). Similarly to our previous observations from Chapter 3, cross-omics correlations between metabolites and transcripts are generally lower than correlations between metabolites, even across organs. Comparing the distributions of correlations between metabolites of the remaining organs, a similar pattern of almost equal distributions was observed (Figure 4.8C, left panel). Notably, the associations between kidney and muscle metabolites are generally higher then the remaining cross-organ metabolite associations, which is also observable when only focusing on the maximal absolute correlations (Figure 4.8C, right panel). For associations between organ metabolites and liver and kidney transcripts we again observed very similar correlation coefficient distributions for all organ pairs (Figure 4.8D, left panel). Interestingly, correlations involving metabolites from adipose tissue appeared generally lower than the others, which becomes even more emphasized for the distributions of the maximum absolute correlations (Figure 4.8D, right panel). Surprisingly, no generally stronger intra-organ correlations between liver metabolites and transcripts, and between kidney metabolites and transcripts were observed in comparison to the cross-organ associations, and even from the maximum correlations only slightly higher correlations between liver metabolites and transcripts were noticeable.

In the next step of our analysis, we constructed a bipartite correlation network between plasma and organ metabolites, the plasma-organ metabolite (P-OM) network. Correlations were included into the network model if they exceed a significance threshold of $\alpha = 0.01$ after Bonferroni correction. This resulted in an overall network consisting of 711 nodes connected by 3385 edges significantly different from zero (Figure 4.9A). Similar to the analysis at a pathway level, the strongest connectivity was observable between plasma and kidney metabolites, comprising $\sim$ 61% of all observed edges between 144 plasma and 212 kidney metabolites. Remarkably, more than three quarters of these edges represent negative associations, which is consistent with the results at pathway level. The second largest fraction of edges ($\sim$ 20%) was observed between 129 plasma and 164 muscle metabolites, which were almost evenly distributed between positive and negative associations. In contrast to the pathway approach, we observed less edges between liver and plasma metabolites ($\sim$ 7% of all observed edges; 85 liver, 79 plasma metabolites) than between adipose tissue and plasma metabolites ($\sim$ 12% of all observed edges; 67 adipose, 111 plasma metabolites).

Next, we studied the distribution of metabolites included in the P-OM network. In total, 53% of all quantified plasma metabolites are contained in the P-OM network (Figure 4.9B). These metabolites reflect organ metabolic signals to a varying extent. For example, $\sim$ 47% of the kidney metabolites display a significant association to plasma metabolites, but only $\sim$ 22% of the metabolites quantified in liver. Further subdividing all P-OM metabolites into their metabolic classes provides additional information about the nature of metabolites reflected in blood. Focusing on plasma, the largest fractions of measured metabolite classes connected to the different organs were 'energy', 'amino acid', 'nucleotide' and 'lipid', but also a large fraction of 'unknowns' (Figure 4.9B). Besides 'energy', 'amino acid' and 'nucleotide', a large fraction of the kidney metabolites connected to plasma belonged to 'carbohydrates'. From liver and muscle, the largest fractions were 'amino acid', 'nucleotide' and 'energy'. Generally, except for adipose and liver, organ metabolites from all classes are associated to plasma metabolites to a varying degree. In the P-OM network, a heterogeneous pattern of connectivity across organs and metabolite classes was observable. For instance, in plasma, many metabolites belonging to the Amino acid and the Lipid classes show a high connectivity (Figure 4.9C). In kidney, besides the Amino acid and Lipid class, also members of the Carbohydrate, Energy and Nucleotide classes exhibit higher degrees. In comparison, metabolites measured in liver and adipose tissue show only a weak connectivity across all measured metabolite classes.

It is known that the metabolic steady state of an organism is caused by a vast interplay among organs and many metabolites are signaling molecules in the regulation of these processes [268]. In order to investigate whether there is a tendency of plasma and organ metabolites to correlate with each other, i.e. if edges in the P-OM network represent biological processes, such as an active blood-tissue exchange of metabolites or an inter-organ association of pathway intermediates, we systematically analyzed all 3385 correlating pairs. For each organ, we therefore categorize the pairwise metabolite combinations into four manually defined groups: (1) The plasma metabolite and the associated organ metabolite are identical. For such a combination, the most likely explanation is an active exchange between plasma and the respective organ, i.e. a transport reaction taking up from or releasing the metabolite into plasma. (2) Both metabolites are not the same compound, but are members of the same pathway/metabolite class. This case might indicate a dependence on the same substrate pool, i.e. pathway-wide correlations between intermediates caused by mass flow. (3) Both metabolites are neither the same compound, nor do they share the same pathway. These pairs might reflect branching points between different metabolic pathways where a pathway cross-talk occurs, or a

common upstream regulation of different metabolic pathways. (4) One or both metabolites of the connected pair are unidentified metabolites. For such correlated pairs, no functional interpretation is possible.

Type	Adipose	Kidney	Liver	Muscle
	5	13	9	9
	49	244	23	60
	30	676	59	108
	330	1142	147	481

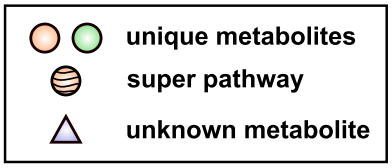


Table 4.3: Different types of pairwise associations. All observed edges in the P-OM network are assigned to one out of four groups: **(1):** Plasma and organ metabolite are equal (edge with two orange nodes). **(2):** Plasma and organ metabolite share the same pathway/ metabolite class (edge with two hatched orange nodes). **(3):** Plasma and organ metabolite are different on metabolite and pathway level (edge with orange and green node). **(4):** One or both metabolites are unknown compounds (edges with one or two purple triangles).

The smallest amount of edges belonged to group (1) whereas most edges include one or two unknowns across all investigated organs and thus, were categorized into group (4) (Table 4.3). Following our interpretation from above, all investigated organs might share the same source pool for metabolites in group (1) with an active inter-organ exchange taking place. Except for adipose tissue, the second largest proportion of edges was observed for group (3), i.e. between different metabolites across all tissues. This might indicate an extensive cross-talk between the different plasma and organ metabolic pathways in terms of substrate availability, signal flow or even receptor function and gene expression.

Inspecting the metabolites contained in group (1) in detail, we mainly observe amino acids or their derivatives, besides lipids, all related to energy homeostatic processes. Interestingly, three plasma metabolites, 3-dehydrocarnitine, 3-hydroxybutyrate (BHBA) and X-12465, can be traced across all organs, i.e. they are strongly associated with their respective organ proxies across all four investigated organs pointing to a common source for these metabolites in systemic metabolism (Figure 4.10).

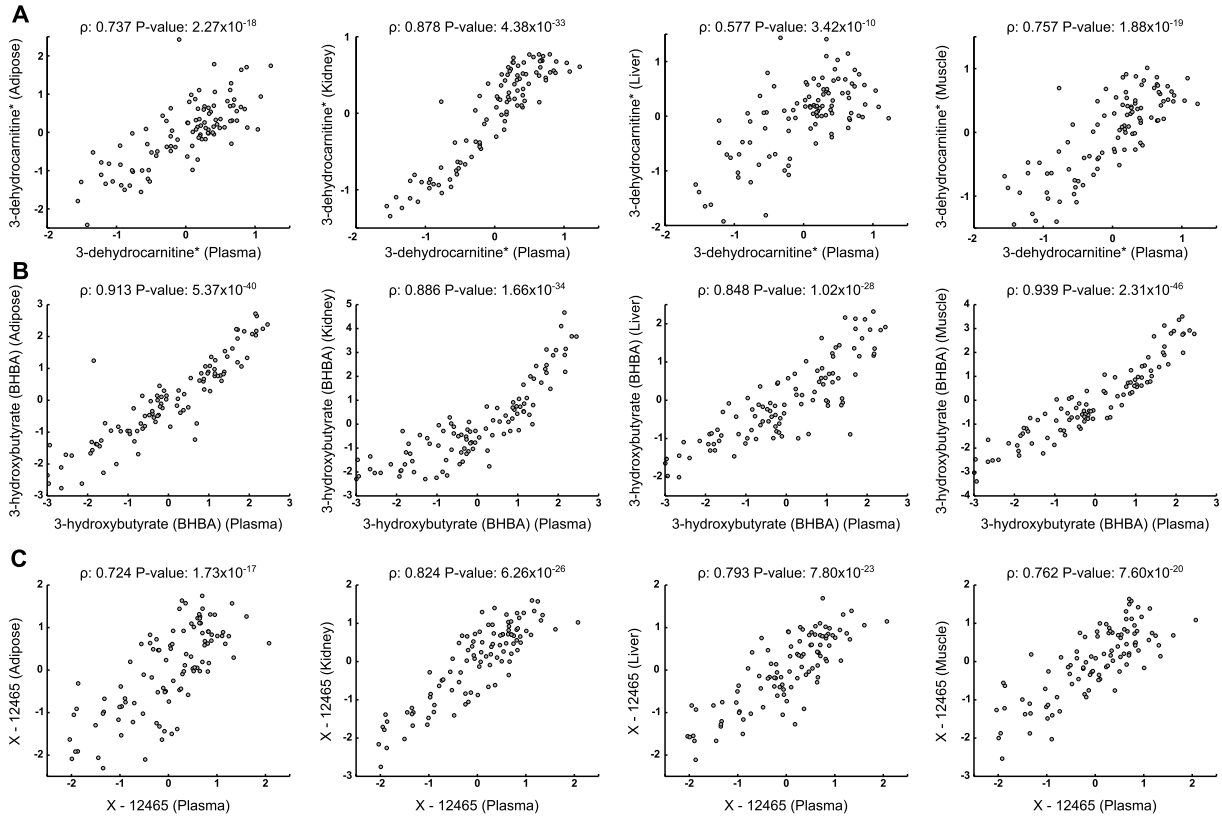


Figure 4.10: Metabolites sharing a common source across organs. The plasma metabolites 3-dehydrocarnitine (**A**), 3-hydroxybutyrate (BHBA) (**B**) and X-12465 (**C**) are equally strongly associated with their respective organ proxies across all tissues. Note that the scatter plots are based on the linear regression residuals from data adjustments (Section 4.2).

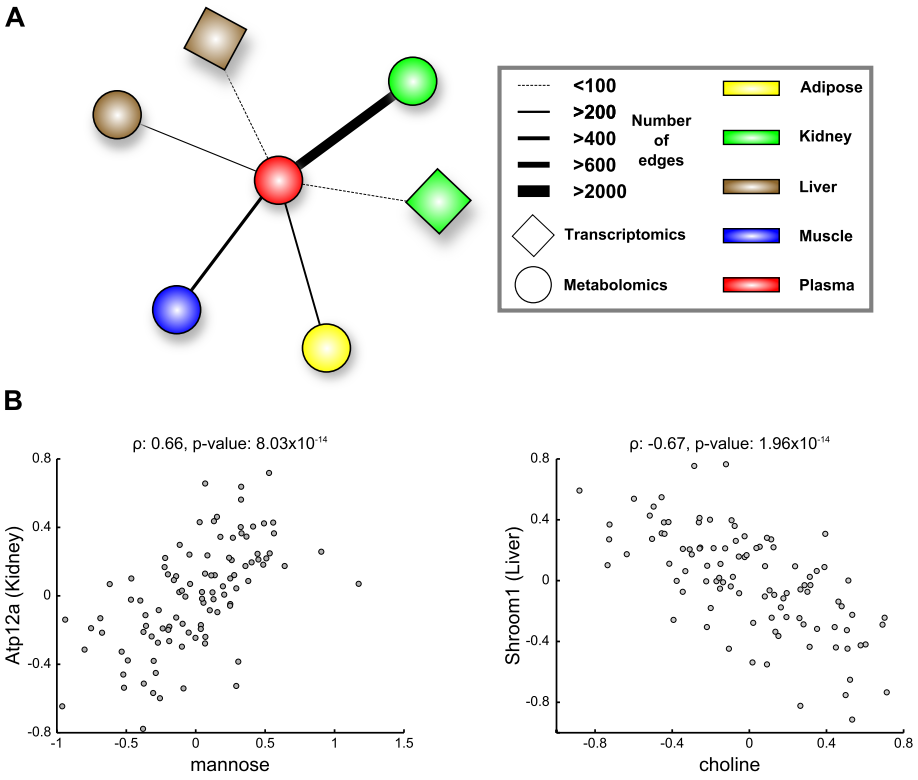


Figure 4.11: Extension of the P-OM network by organ transcriptomics data. **A:** Schematic representation of the P-OM network extended by kidney and liver transcripts. Node colors illustrate the different organs, edge widths indicate the relative number of edges found between two organs with respect to the smallest number of edges observed. Only a small number of edges was observed between plasma metabolites and kidney or liver transcripts. **B:** The strongest associations between plasma metabolites and organ transcripts were observed between Atp12a (kidney) and mannose and between Shroom1 (liver) and choline. ρ, Spearman's rank correlation coefficient.

4.6 Organ gene regulatory signatures are reflected by plasma metabolites

To investigate whether plasma metabolites also carry signatures of organ transcript levels, we used gene expression profiling data from kidney and liver samples and extended the P-OM network by a transcriptional layer. To this end, we calculated the pairwise Spearman correlations between all plasma metabolite concentrations and both kidney and liver transcript profiles. Statistical significance of the correlation coefficients was assessed by a false discovery rate of $q = 0.05$. In total, we observed 97 significant associations (44 positive, 53 negative) between 58 liver transcripts and 26 plasma metabolites,

and 55 (32 positive, 23 negative) between 50 kidney transcripts and 12 plasma metabolites (Figure 4.11A). As described above and reported previously, the correlations between metabolites and transcripts are generally lower than between metabolites, even within organs (see Figure 4.8B and Chapter 3). Thus, the low numbers of significant associations in comparison to the metabolite-metabolite correlations between plasma and organs were not unexpected, given the large molecular and functional distance between metabolites and transcripts, and also the compartmentalization between organs and blood. Out of the 50 kidney transcripts, 43 are associated with only two plasma metabolites, namely hexanoylcarnitine and X-11407, both of which are disconnected in the network (i.e. not connected to any other metabolite or transcript). In comparison, the liver transcripts are tightly integrated and dispersed all over the P-OM network. The strongest association with a kidney transcript was observed between *Atp12a*, a membrane ion channel, and mannose ($\rho = 0.66$, p-value $= 8.03 \cdot 10^{-14}$), which are both related to energy metabolism (oxidative phosphorylation and carbohydrate metabolism). For liver transcripts, the strongest association was found between *Shroom1* and choline, both of which are involved in the development of cell membranes ($\rho = -0.67$, p-value $= 1.96 \cdot 10^{-14}$; Figure 4.11B) [269, 270].

In order to characterize the observed relationships between organ transcripts and plasma metabolites functionally, we performed an enrichment analysis using GO terms on the plasma-associated transcripts of both kidney and liver. Enrichment was calculated for all terms belonging to the 'biological process' and 'molecular function' domain that are contained in the GO slim collection (see also Chapter 3.2 for information about GO slim). In total, 114 terms were tested and significance was determined at $\alpha = 0.05$ after Bonferroni correction (p-values $< 4.38 \cdot 10^{-4}$). Note that the aggregated z-score approach is not applicable here due to the high number of variables (transcripts) and the coarse-grained level of GO slim terms.

For the liver transcripts, we found significant enrichment of 60 GO terms. These included 'signal transduction' (p-value $= 1.29 \cdot 10^{-5}$), 'cell-cell signaling' (p-value $= 2.01 \cdot 10^{-4}$), 'cell adhesion' and 'cell morphogenesis' (p-values $= 1.07 \cdot 10^{-4}$ and $= 1.99 \cdot 10^{-4}$), but also several terms associated to metabolism, such as 'cellular nitrogen compound metabolic process' (p-value $= 9.93 \cdot 10^{-5}$), 'generation of precursor metabolites and energy' (p-value $= 2.98 \cdot 10^{-4}$) and 'small molecule metabolic process' (p-value $= 2.98 \cdot 10^{-4}$; see Table 4.4 for top 10 identified terms). Interestingly, also terms like 'response to stress' (p-value $= 2.75 \cdot 10^{-4}$) and 'circulatory system process' (p-value $= 3.19 \cdot 10^{-4}$) were among the significant categories. The first term describes processes occurring

GO Term	P-value	Counts	Name	Ontology
GO:0007165	$1.29 \cdot 10^{-5}$	7 (5193)	signal transduction	biological process
GO:0006397	$9.61 \cdot 10^{-5}$	5 (4128)	mRNA processing	biological process
GO:0034641	$9.93 \cdot 10^{-5}$	6 (4443)	cellular nitrogen compound metabolic process	biological process
GO:0007155	$1.07 \cdot 10^{-4}$	6 (4425)	cell adhesion	biological process
GO:0007010	$1.75 \cdot 10^{-4}$	5 (3989)	cytoskeleton organization	biological process
GO:0000902	$1.99 \cdot 10^{-4}$	5 (3959)	cell morphogenesis	biological process
GO:0030198	$2.00 \cdot 10^{-4}$	5 (3958)	extracellular matrix organization	biological process
GO:0007267	$2.01 \cdot 10^{-4}$	5 (3956)	cell-cell signaling	biological process
GO:0007059	$2.20 \cdot 10^{-4}$	5 (3935)	chromosome segregation	biological process
GO:0007009	$2.64 \cdot 10^{-4}$	5 (3891)	plasma membrane organization	biological process

Table 4.4: Top 10 enriched GO terms identified for liver transcripts. Counts in brackets indicate the total number of transcripts within the respective GO term. Significance of the GO terms was assessed at $\alpha = 0.05$ after Bonferroni correction. A total number of 25,649 transcripts was used as background set in the hypergeometric tests.

after a disturbance in organismal homeostasis and the second relates to any process involved in the exchange of extracellular fluids to and from tissue, both of which are in accordance with the physiological environment (e.g. treatment and fasting conditions; blood circulatory system) under investigation. Taken together, the most prominently identified GO terms for plasma-associated liver transcripts primarily represent signaling and/or metabolic processes, besides cellular development.

In comparison, we observe a completely different pattern of enriched GO terms for the kidney transcripts. Only seven terms are above the significance level which are almost all related to transport processes. The most significant term is 'protein transporter activity' (p-value = $3.58 \cdot 10^{-7}$), but also 'transmembrane transport' (p-value = $5.22 \cdot 10^{-7}$) and 'vesicle-mediated transport' (p-value = $5.53 \cdot 10^{-7}$) were among the top hits (Table 4.5). Since only a few GO terms of the coarse-grained GO Slim catalog reached significance, we additionally performed an enrichment analysis using GO terms of the entire biological process ontology. This results in a more fine-grained view on the plasma-reflected processes and our findings further confirmed the observations already made on the broad level of GO slim terms. For instance, among the most significant hits were terms such as 'carnitine transport' (p-value = $1.43 \cdot 10^{-12}$), 'spermine transport' (p-value = $2.85 \cdot 10^{-9}$), 'single-organism transport' (p-value = $3.33 \cdot 10^{-9}$) and 'acetylcholine transport' (p-value = $1.62 \cdot 10^{-8}$) to mention just a few examples (analysis details not shown).

GO Term	P-value	Counts	Name	Ontology
GO:0008565	$3.58 \cdot 10^{-7}$	23 (2216)	protein transporter activity	molecular function
GO:0055085	$5.22 \cdot 10^{-7}$	23 (2262)	transmembrane transport	biological process
GO:0016192	$5.53 \cdot 10^{-7}$	23 (2269)	vesicle-mediated transport	biological process
GO:0022857	$1.29 \cdot 10^{-5}$	10 (519)	transmembrane transporter activity	molecular function
GO:0006605	$8.32 \cdot 10^{-5}$	34 (5909)	protein targeting	biological process
GO:0006810	$1.40 \cdot 10^{-4}$	34 (6032)	transport	biological process
GO:0007165	$1.87 \cdot 10^{-4}$	8 (5194)	signal transduction	biological process
GO:0007034	$2.01 \cdot 10^{-3}$	31 (5865)	vacuolar transport	biological process
GO:0030705	$2.11 \cdot 10^{-3}$	31 (5880)	cytoskeleton-dependent intracellular transport	biological process
GO:0006913	$2.19 \cdot 10^{-3}$	31 (5892)	nucleocytoplasmic transport	biological process

Table 4.5: Top 10 enriched GO terms identified for kidney transcripts. Counts in brackets indicate the total number of transcripts contained in the respective GO term. Note that the last three terms did not reach the significance threshold of p-values $< 4.38 \cdot 10^{-4}$ and should just be merely considered as suggestive. A total number of 25,697 transcripts was used as background set in the hypergeometric tests.

4.7 Network-based identification of plasma proxy markers of organ processes

Up to this point, we systematically analyzed the relationship between metabolites quantified in plasma and those measured in four distinct organs. The results obtained from this analysis clearly showed that organ-specific metabolic signals, both at pathway and at single metabolite level, are mirrored in plasma metabolite concentrations. In addition, we could show that plasma metabolites carry organ-specific transcriptional signals. In the following, we will use the P-OP networks at super- (P-OSup) and subpathway (P-OSub) level and P-OM networks to identify plasma metabolites that might serve as two different types of proxy markers: (1) As single-organ informative proxy markers, potentially reflecting organ-specific biochemical changes induced by phenotypic traits. Thus, these proxy markers provide specific information about the (patho)physiological state of a particular organ. (2) As multi-organ informative, system-wide proxy markers, carrying information on all investigated organs, thus potentially reflecting the inter-organ state of systemic metabolism. This is particularly based on the observation that plasma metabolites associated with the investigated organs to a varying degree across all three constructed networks, i.e. each network contained plasma metabolites either connected to molecules of all four organs, or exclusively connected to molecules of a single organ.

Tissues	P-OSup network	P-OSub network	P-OM network
4	0	0	46
3	1	30	49
2	21	54	44
1	98 (1 A, 79 K, 3 L, 15 M)	88 (0 A, 59 K, 15 L, 14 M)	44 (6 A, 18K, 10 L, 10 M)
Sum	120	172	183

Table 4.6: Number of associated tissues per plasma metabolite and network (P-OM metabolome only). For all three constructed networks, the contained plasma metabolites are grouped according to the number of connected tissues. A = Adipose, K = Kidney, L = Liver, M = Muscle.

To identify potential proxy markers of both types among the plasma metabolites, we performed two analysis steps. First, we assigned the plasma metabolites into different groups according to the number of associated organs. This step was performed separately for each network including the P-OM (metabolomics only) and extended P-OM (metabolomics & transcriptomics). Second, to identify organ proxy markers of high confidence associated at all biological levels, we focused on those plasma metabolites associated with one or four tissues and subsequently looked for an overlap between all potential proxy markers identified from the different networks.

When examining all three constructed networks, only the plasma-organ metabolite (P-OM) network contained 46 plasma metabolites associating with metabolites of all four other organs (Table 4.6). This observation might be explained by stronger heterogeneity between organs at pathway level in comparison to single metabolite concentrations. Out of these 46 potential multi-organ proxy markers, carnitine and 3-hydroxybutyrate (BHBA) are among the top six correlations across all investigated organs (Table 4.7). One possible explanation for the observation of these two metabolites as multi-organ proxy markers might be the globally prevalent energy deprivation due to the fasting conditions of the mice in our study. Carnitine and its esters (i.e. acetyl-carnitines) are tightly linked in energy metabolism, particularly involved in the transfer of long-chain fatty acids into mitochondria for β-oxidation (c.f. Sections 3.4 and 3.7). The total pool of free and acetylated carnitine is largely dependent on extrinsic conditions (e.g. nutritional status, physical exercise) and is mainly stored in muscle, liver and kidney [271]. In addition, the major ketone body 3-hydroxybutyrate (BHBA) has been described as a central signaling molecule in the regulation of food intake and, under fasting conditions, it is synthesized from acetyl-CoA in liver and can be used as an alternative energy source [272]. In an earlier study investigating the relationship between long-chain fatty acid oxidation and ketogenesis in fed and fasting rats, a carnitine dependent stimulation of ketogenesis was observed, suggesting a regulatory role for carnitine in the synthesis of

ketone bodies [273]. Moreover, in a study conducted on the carnitine metabolism in fasting rats, a decrease in the levels of carnitine was observed, whereas the levels of 3-hydroxybutyrate (BHBA) were elevated during the fast indicating a shift in metabolism towards fatty acids as primary energy source [274]. This finding fits well to the strong negative association observed between carnitine and 3-hydroxybutyrate (BHBA) in our study (Table 4.7) and suggests that these multi-tissue proxy markers globally reflect the nutritional state of the system under investigation.

All three networks (P-OSup, P-OSub and P-OM) contained plasma metabolites exclusively connected to a single organ. Interestingly, more metabolites connected to a single organ are observable in the P-OP networks (superpathway 98, subpathway 88) compared to the plasma-organ metabolite network (44). To derive a set of high confidence single-organ proxy markers, we systematically compared the lists of metabolites exclusively connected to one tissue between all three networks (Figure 4.12). For adipose tissue, no overlap was found between 7 identified single-organ metabolites (1 from the P-OP network at super pathway level and 6 identified in the P-OM network). Moreover, no liver-specific single-organ plasma metabolite was contained in all three networks. However, ethanolamine, spermidine, X - 08889 and butyrylglycine were identified as liver-specific by different combinations of two networks (Figure 4.12).

For kidney and muscle, we identified 7 and 2 potential single-organ proxy markers shared between all three networks and solely connected to the respective organs (Table 4.8 for detailed information on proxy marker candidates). Notably, seven out of the nine candidates are known markers for organ function or specific diseases (see last column of Table 4.8). For instance, creatinine is a constantly produced metabolic product of creatine phosphate breakdown in muscle, which is subsequently cleared from the blood stream by the kidneys. Abnormal blood creatinine levels thus provide information about renal function [275]. As another example, blood levels of branched-chain amino acids, such as valine, have been associated to a variety of diseases including phenylketonuria [276], Alzheimer's disease [277] and also chronic renal failure [278].

In the next step, we additionally incorporated the transcripts into the proxy marker analysis. Again, we were mainly interested in plasma metabolites solely connected to one organ and those providing information on all tissues. As already mentioned above, 86% of the kidney transcripts that correlate with any plasma metabolite are associated with just two metabolites, namely hexanoylcarnitine and X-11407, which are disconnected components in the network. Of the remaining seven kidney transcripts, five

Organ	Plasma metabolite	Organ metabolite	ρ	P-value
Adipose	carnitine	3-hydroxybutyrate (BHBA)	-0.8393	$1.09 \cdot 10^{-27}$
Kidney	carnitine	3-hydroxybutyrate (BHBA)	-0.8518	$2.83 \cdot 10^{-29}$
Liver	carnitine	3-hydroxybutyrate (BHBA)	-0.8088	$2.51 \cdot 10^{-24}$
Muscle	carnitine	3-hydroxybutyrate (BHBA)	-0.8565	$2.51 \cdot 10^{-29}$
Adipose	3-hydroxybutyrate (BHBA)	3-hydroxybutyrate (BHBA)	0.9132	$5.37 \cdot 10^{-40}$
Kidney	3-hydroxybutyrate (BHBA)	X - 14839	0.8881	$7.55 \cdot 10^{-35}$
Liver	3-hydroxybutyrate (BHBA)	3-hydroxybutyrate (BHBA)	0.8476	$1.02 \cdot 10^{-28}$
Muscle	3-hydroxybutyrate (BHBA)	3-hydroxybutyrate (BHBA)	0.9392	$2.31 \cdot 10^{-46}$

Table 4.7: Top scoring multi-organ proxy markers. Among the 46 plasma metabolites shared between all 4 investigated organs, carnitine and 3-hydroxybutyrate show the strongest associations with metabolites quantified in organs.

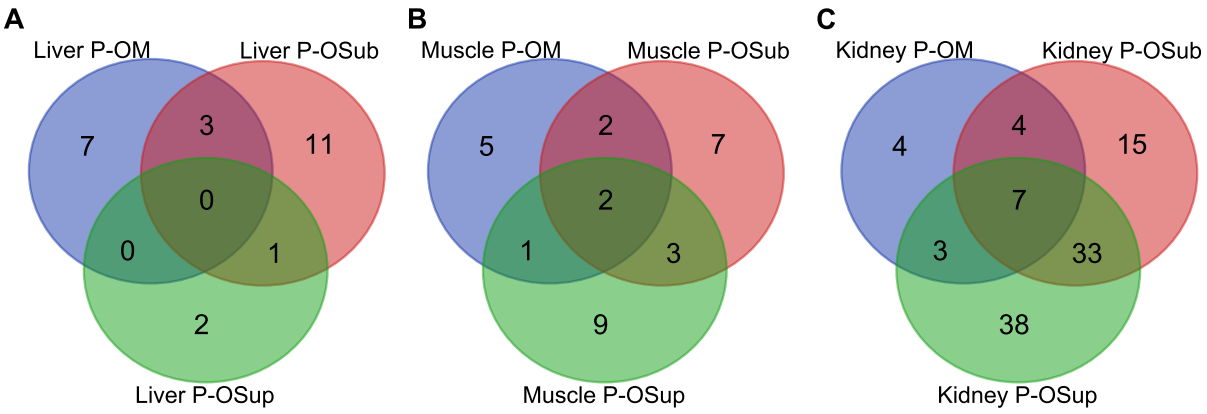


Figure 4.12: Comparison of all marker candidates identified by the three different networks. Shown is the overlap between plasma metabolites which are exclusively associated with metabolites and pathways of a single organ across all three constructed networks (see Table 4.6). For adipose tissue, no overlap was observable (not shown). P-OSub = plasma-organ pathway network at a subpathway level, P-OSup = plasma-organ pathway network at super pathway level.

Organ	Candidate	P-OP (Super)	P-OP (Sub)	P-OM	Published associations
Kidney	pantothenate	Amino acid	Sterol/Steroid; Creatine metabolism; Nucleotide sugars, pentose metabolism; Lysine metabolism; Urea cycle, arginine-, proline-, metabolism; Inositol metabolism	histidine, uracil, threonine, lysine, gamma-aminobutyrate (GABA), ethanolamine, glycerate, taurine, xylitol, glucuronate, 3-methyl-2-oxovalerate, aspartate, 4-methyl-2-oxopentanoate, scyllo-inositol, choline phosphate, X-15509, X-15520, X-15527, X-17115	heart disease [279, 280], atherosclerosis [281], rheumatoid arthritis [282, 283]
	valine	Carbohydrate, Energy	Nucleotide sugars, pentose metabolism; Pyrimidine metabolism, uracil containing; Butanoate metabolism; Aminosugars metabolism; Diacylglycerol	X - 11081	Phenylketonuria [276], maple syrup urine disease [276], Alzheimer's disease [277], lung cancer [284], epilepsy [285], chronic renal failure [278]
	butyrylcarnitine	Cofactors and vitamins, Lipid, Nucleotide	Monoacylglycerol; Diacylglycerol; Pyrimidine metabolism, thymine containing; Ascorbate and aldarate metabolism	oleate (18:1n9), 1-palmitoylglycerol	short-chain acyl-CoA dehydrogenase (SCAD) / isobutyryl-CoA dehydrogenase deficiency [286], celiac disease [287]
	pseudouridine	Energy	Creatine metabolism	glucuronate	canavan disease [288], uremia [289], malignant proliferative diseases [290], kidney disease [291], heart failure [292]
	2-oleoylglycero-phosphocholine	Carbohydrate, Cofactors and vitamins	Glycolysis, gluconeogenesis, pyruvate metabolism; Butanoate metabolism; Ascorbate and aldarate metabolism; Monoacylglycerol	gluconate	
	1-palmitoylglycero-phosphoethanolamine	Carbohydrate	Glycine, serine and threonine metabolism	lactate, alanine, glycine, dihydrocholesterol, X-08169, X-10500, X-11533, palmitoyl sphingomyeline	metabolic syndrome [293]
	X-16569	Carbohydrate	Glycine, serine and threonine metabolism	glucuronate, stearoyl sphingomyelin, xylonate, X - 11081	
Muscle	creatinine	Amino acid, Nucleotide	Valine, leucine and isoleucine metabolism; Purine metabolism, (hypo)xanthine/inosine containing	methionine	canavan disease [288], chronic renal failure [275], hyperoxalemia [294]
	azelate (nonanedioate)	Amino acid, Nucleotide, Peptide	Pyrimidine metabolism, cytidine containing; Lysine metabolism; Urea cycle, arginine-, proline-, metabolism; Glutathione metabolism; Dipeptide derivative	tryptophan, proline, anserine, histidylleucine	diabetes [295], kidney cancer [296]

Table 4.8: Potential single-organ proxy markers confirmed by all three networks. Plasma metabolites exclusively connected to metabolites, superpathways and subpathways of the respective organ. The last column contains information on known disease associations for the identified plasma metabolites.

Proxy marker type	Kidney	Liver
Single-organ	hexanoylcarnitine, beta-hydroxypyruvate, heptanoate (7:0), X-11407, X-17383, X-11372	corticosterone, gamma-glutamylisoleucine*, X-16589, spermidine, X-08889
Multi-organ	3-hydroxybutyrate (BHBA), carnitine, N-acetylglycine, X-12775, X-14839	3-hydroxybutyrate (BHBA), carnitine, N-acetylglycine, hexanoyl-glycine, indolelactate, isovalerylglycine, pro-line, X-12101, X-14839, X-12465

Table 4.9: Plasma proxy markers for transcriptional processes in kidney and liver. Shown are plasma metabolites contained in the extended plasma-organ metabolite network, which are either connected exclusively with molecules (metabolites + transcripts) measured in kidney or liver, or associated with molecules of all investigated organs.

are associated to four additional metabolites that are exclusively connected to kidney transcripts, whereas only a single transcript, *Pdk4*, is integrated in largest connected component. Therein it is associated with five central plasma metabolites connected to compounds of all 4 investigated tissues (see Table 4.9). Among these multi-organ proxy metabolites, we again observe carnitine and 3-hydroxybutyrate. Similarly, when investigating the multi-organ proxy metabolites associated with liver transcripts, these two metabolites are among the 10 identified potential proxy markers, further indicating their global information content (Table 4.9). Moreover, we identify five plasma metabolites solely connected to liver, out of which three were only associated with transcripts. The remaining two, spermidine and X-08889, additionally associated with liver metabolites. Remarkably, these two metabolites were also among the potential liver proxy markers shared between the P-OM and the P-OSub network (Figure 4.12) adding another piece of confirmation for these two metabolites as single-organ proxy markers.

4.8 Determining plasma proxy markers for diabetes-induced alterations in organs

As mentioned above, type 2 diabetes is a complex multi-organ disease, involving dysregulations of various processes occurring in several distinct organs. Studies focusing

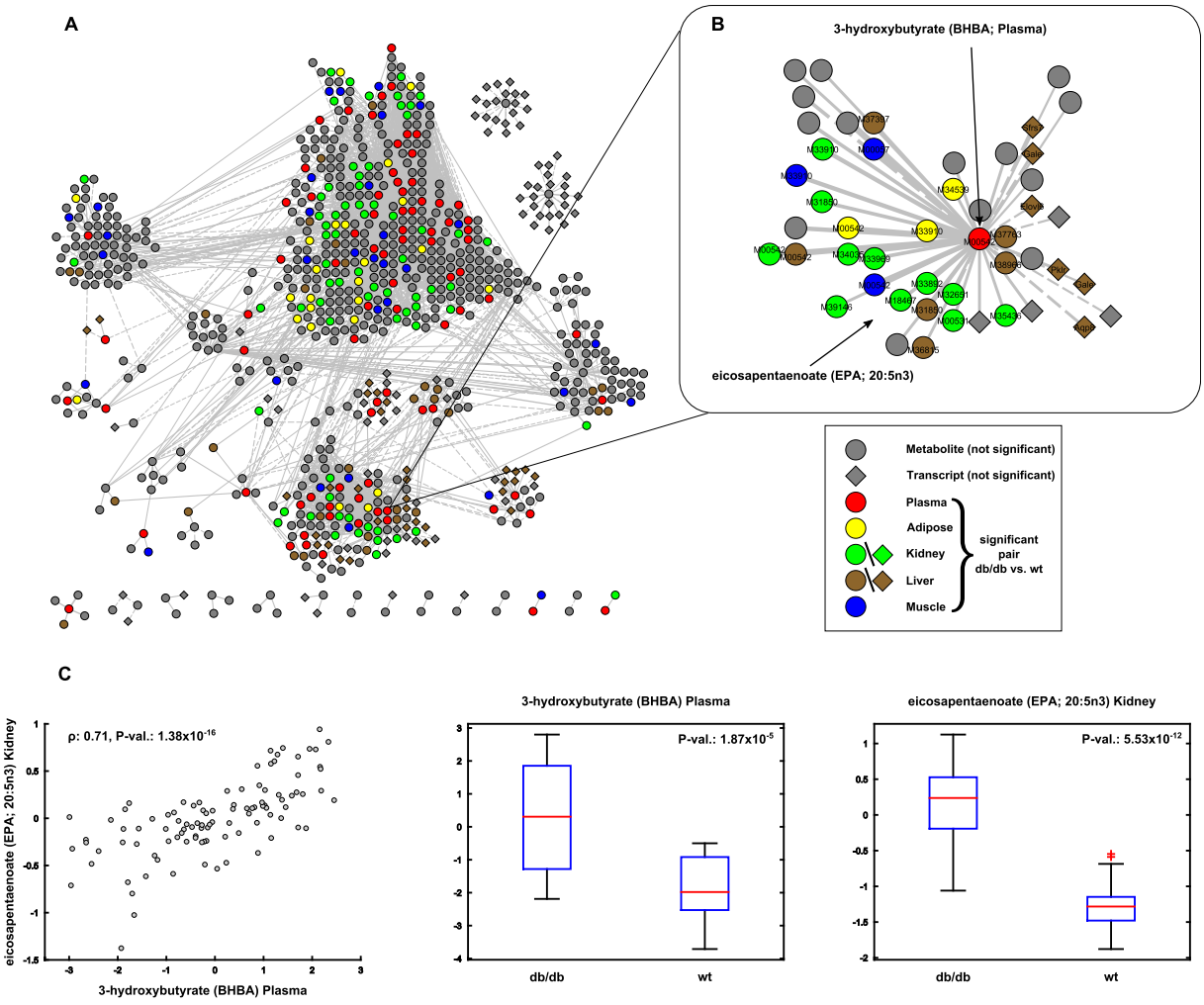


Figure 4.13: Diabetes-related organ changes are reflected in the plasma metabolome. **A:** Illustration of diabetes related changes in the extended plasma-organ metabolite network. Nodes are only colored if both the plasma and the organ metabolite directly connected by an edge are changed significantly under the investigated condition. The remaining nodes are colored gray. **B:** Subnetwork showing 3-hydroxybutyrate (BHBA) and its direct neighbors. 3-hydroxybutyrate (BHBA) is a potential multi-organ proxy marker and several of its neighbors from the investigated organs are significantly changed under diabetic conditions. **C:** Example of a significantly associated network pair between plasma and kidney (left). Both the concentrations of 3-hydroxybutyrate (BHBA; middle) measured in plasma and eicosapentaenoate (EPA; right) measured in kidney are significantly elevated under diabetic conditions. An overview providing the number of such examples in given in Table 4.11.

on metabolic alterations in plasma already discovered many biomarkers associated with type 2 diabetes [261, 297]. However, most of these studies were restricted to a single tissue, allowing only a limited scope on the complex nature of the underlying molecular changes in the pathogenesis of type 2 diabetes. In particular, only little is known about the origin of these plasma-observable metabolic changes in disease, i.e. how metabolic and transcriptional changes occurring in the involved tissues are mirrored in blood metabolic profiles. To systematically examine the potential of plasma metabolites to serve as proxy markers for diabetes-induced molecular changes in organs, we carried out a differential analysis of plasma and organ metabolite profiles, and transcriptional expression signatures of kidney and liver. For this analysis, only data from the mice treated with vehicle solution was used, ensuring that results are not obscured by any treatment effects. Thus, for each organ and plasma, we identified differentially expressed molecules between db/db (n=20) and wt (n=20) animals using Student's t statistics. The corresponding p-values were adjusted for multiple testing by controlling the false discovery rate (FDR) at 0.05. To identify plasma proxy markers, we mapped significantly changed metabolites/transcripts onto the extended P-OM network whenever an edge exists between a significantly changed plasma metabolite and organ molecule (metabolite or transcript; Figure 4.13A).

Diabetes induced organ metabolome changes were most pronounced in liver, followed by kidney, plasma , adipose tissue and muscle (Table 4.10). At the transcript level, again most prominent differences were observable in liver (5292) followed by kidney (1677). After mapping significant organ-plasma pairs onto the extended P-OM network, we identify a total of 597 edges ($\sim$ 17%) connecting 82 significantly altered plasma metabolites (48%) with 24 metabolites (16%) in adipose tissue, 69 metabolites (34%) in kidney, 31 metabolites in liver (13%) and 32 metabolites (22%) in muscle (Figure 4.13A). At the transcript level, 32 differential liver transcripts were connected to significantly changed plasma metabolites, while no significant pairs were found for kidney. Again, we categorized the significantly changed pairs into groups according to the observable molecule combinations between plasma and organs (cf. Section 4.5). We observe significantly changed pairs of each type distributed relatively even between the different organs, further supporting the hypothesis that disease affected organ processes are reflected in plasma metabolite profiles (Table 4.11).

In the following, we discuss some of the above-mentioned potential multi-organ and single-organ proxy markers for each tissue (see Section 4.7) in the context of type 2

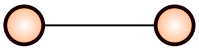
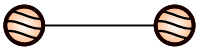
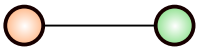
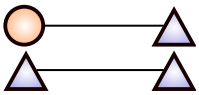
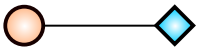
Organ	db/db vs wt, significant	In network	# associations (plasma-organ pairs)
Adipose	149	24	73
Kidney M.	202	69	338
Kidney T.	1677	0	0
Liver M.	235	31	70
Liver T.	5292	32	40
Muscle	143	32	76
Plasma	172	82	597

Table 4.10: Significantly changed molecules in diabetes type 2. Shown are the numbers of significantly changed metabolites/transcripts between db/db and wild-type mice. Abbreviations: Liver M.= liver metabolites, Liver T.= liver transcripts, Kidney M.= kidney metabolites, Kidney T.= kidney transcripts.

diabetes. For those organs where we did not find any network-based plasma proxy markers, we discuss other potentially disease-relevant metabolites.

Multi-organ proxy markers in diabetes

Out of all 46 multi-organ proxy metabolites identified in the P-OM network (Table 4.6), 24 are significantly changed under diabetic conditions and have at least one significantly changed molecule attached. Among these are also the two above-mentioned top correlated multi-organ proxy metabolites, carnitine and 3-hydroxybutyrate (BHBA) (Table 4.7). 3-hydroxybutyrate (BHBA) was associated to 23 significantly altered organ metabolites (3 adipose, 11 kidney, 6 liver and 3 muscle) and 6 differentially expressed liver transcripts (Figure 4.13B). For instance, eicosapentaenoate (EPA) concentrations in kidney are strongly associated with plasma BHBA (Figure 4.13C (left)), and both the levels of 3-hydroxybutyrate (BHBA) and EPA are markedly elevated in diabetic mice when compared to controls (Figure 4.13C (middle, right)). Increased concentrations of 3-hydroxybutyrate (BHBA) in diabetes were already observed in another study using db/db mice [298] and also in diabetic humans [266], indicating a mild ketotic condition. Eicosapentaenoate is a ω-3 fatty acid and precursor of important bioactive molecules, such as eicosanoids and prostaglandins, which have been linked to variety of diseases including hepatic insulin resistance [299] and diabetes type 2 [300]. As another example, plasma carnitine was significantly reduced in db/db mice compared to wild-type animals (p-value = $2.33 \cdot 10^{-7}$). Plasma carnitine levels have been shown to be reduced in diabetes type 2, especially under ketoacidotic conditions, which might result from an increased utilization for the generation of acylcarnitines in tissues and a subsequent

Type	Adipose	Kidney	Liver	Muscle
	2 (5)	6 (13)	3 (9)	4 (9)
	18 (49)	49 (244)	7 (23)	6 (60)
	10 (30)	132 (676)	15 (59)	14 (108)
	43 (330)	151 (1142)	45 (147)	52 (481)
	-	0 (55)	40 (97)	-
Σ	73 (414)	338 (2130)	110 (335)	76 (658)

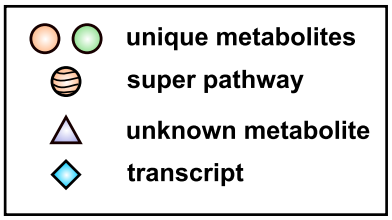


Table 4.11: Types of significantly changed plasma-organ molecule pairs in diabetes. Edges connecting significantly changed plasma metabolites and organ molecules categorized as previously explained (Section 4.5). Numbers indicate all significantly changed pairs of the respective type; numbers in brackets denote all edges contained in the extended P-OM network of the respective type. Note that transcripts were only measured in kidney and liver.

redistribution between organs according to the needs [271, 301, 302]. This could also be an explanation for the identification of carnitine as multi-organ proxy marker in our P-OM network, where it was connected to 3 significantly changed metabolites in adipose tissue, 14 in kidney, 4 in liver and 2 in muscle. Moreover, it was associated to 7 significantly altered liver transcripts.

Kidney proxy markers

Of the seven identified proxy markers for kidney, three are significantly changed and additionally connected to significantly different kidney metabolites. Both plasma pseudouridine and X-16569 are associated to only one significantly changed kidney metabolite, namely glucuronate. Increased plasma levels of pseudouridine have been associated with a variety of diseases including several types of cancer [303], heart failure [292] and renal dysfunction [291]. In our analysis, plasma pseudouridine was negatively associated with the concentration of glucuronate in kidney, a carbohydrate derived from glucose, which is involved in the detoxification of a variety of substances via glucuronidation

[304]. As a consequence, these toxic substances can be excreted much easier by the liver, intestine and kidney. Interestingly, both the concentrations of pseudouridine in plasma and glucuronate in kidney are decreased under diabetic conditions possibly due to an increased renal activity and subsequent clearance of pseudouridine from plasma.

The third kidney proxy marker, pantothenate, was associated to nine kidney metabolites (4 amino acid metabolism, 2 unknowns, 1 carbohydrate and 1 lipid) which were significantly changed under diabetic conditions. Pantothenate is an essential nutrient and important constituent of energy metabolism, required for the synthesis of coenzyme-A (coA) [305]. Diabetes is characterized by high glucose and decreased insulin levels, conditions which are known to cause an increased production of coA [306]. In our P-OM network, pantothenate was negatively associated with eight out of the nine differential metabolites. While the concentrations of plasma pantothenate are strongly increased in diabetic mice, the levels of all eight negatively associated metabolites are decreased in kidney. A study on the pantothenate metabolism in diabetic rats showed a shift in the pantothenate uptake of several tissues and a decreased urinary excretion [307]. This finding fits well to the elevated plasma levels of pantothenate along with the decreased levels of kidney metabolites observed in our study. One thus might speculate a decreased uptake of pantothenate into kidney and an active redistribution to organs of elevated demand such as liver for energy (glucose) production [308].

Liver proxy markers

Of the two proxy markers exclusively associated with liver molecules (Table 4.9), only an unidentified metabolite, X-08889, was significantly changed under diabetic conditions and was associated with the significantly changed liver transcript Gprc5a (*G protein-coupled receptor, family C, group 5, member A*). Gprc5a is an orphan transmembrane receptor whose expression is induced by retinoic acid and plays important roles in a couple of signaling pathways including cAMP signaling, NF-κB and STAT3 signaling pathways [309]. In our study, we observe a positive association between X-08889 and Gprc5a, but while the concentrations of the plasma metabolite are decreased in diabetic mice, the expression of Gprc5a was induced. Dysregulation of Gprc5a has recently been related to various types of cancer, chronic obstructive pulmonary disease and type 2 diabetes to name just a few diseases [309, 310]. However, very little is known about the precise function of orphan receptors and also the associated metabolite is not identified yet, this identified association might of interest for future investigations in the diabetes context.

Adipose tissue & Muscle proxy markers

No adipose tissue proxy marker was found by our network-based approach (Figure 4.12). However, we were able to map 73 edges connecting 24 adipose tissue metabolites with 33 plasma metabolites in the P-OM network, which were significantly altered under diabetic conditions. Interestingly, the plasma metabolite 2-docosapentaenoylglycerophosphocholine, a lysolipid and multi-organ proxy marker, was associated with a third of the 24 significantly changed adipose tissue metabolites (6 lipids, 2 unknowns), suggesting a potential marker for lipid composition of adipose tissue.

For the two identified muscle candidates, no significantly altered pair was found under diabetic conditions. However, we identified 76 significantly altered metabolite pairs connecting 32 muscle with 36 plasma metabolites. Two plasma metabolites, X-12649 and tryptophan are associated with 9 respectively 8 changed muscle metabolites. Notably, all these muscle metabolites are changed in concordant direction with the two plasma metabolites implying further plasma proxies for diabetes induced changes in muscle.

4.9 Discussion

The use of metabolites measured in biofluids to assess organ function and also as prognostic/diagnostic markers for a variety of diseases is well established in clinical applications [259, 260]. The underlying paradigm is, that changes occurring in the respective organs are reflected in a easily accessible biofluid such as blood. In order to systematically investigate the potential of plasma metabolites as proxy markers for metabolic as well as transcriptomic processes in organs, we performed a comprehensive analysis integrating data from multiple organs and molecular levels. Thereby, we combined metabolomics data measured in adipose tissue, kidney, liver, muscle and plasma, and transcriptomics data from kidney and liver of db/db and wild-type mice.

First, we compared the metabolomes of the investigated organs and showed that not only the detectable metabolites differ between organs, but also the metabolite profiles form organ-specific clusters which are perfectly separated from each other. This data-driven observation demonstrates that each organ has a unique metabolic profile which is also consistent with existing knowledge [311]. Next, we constructed bipartite correlation networks connecting plasma metabolites and the four investigated organs at pathway and single molecule level. In total, more than 50% of the quantified plasma metabolites

carried signatures of any organ metabolite/pathway already highlighting their potential as organ proxy markers. Subsequent functional investigation provided clear evidence that plasma metabolites generally reflect metabolic processes and pathways of the investigated organs to a varying extent. For instance, kidney processes are most strongly reflected by plasma metabolites, whereas liver processes are among the least reflected. This finding is in line with previous observations of a less strong liver contribution to the plasma metabolic profile [61] and might be explained by differences in metabolite turn-over times between the investigated organs [298]. In contrast to the remaining organs, most correlations observed between plasma and kidney were negative, which might reflect the physiological function of kidney as blood filter. Furthermore, we systematically examined all pairwise associations between plasma metabolites and organ molecules and inferred four different biologically meaningful categories. For instance, the first category was between identical metabolites quantified in plasma and organs. Within this category, we identified associations between 3 identical metabolites which can be traced through all five investigated organs, namely 3-hydroxybutyrate (BHBA), 3-dehydrocarnitine and X-12465, indicating a common organism wide source for these metabolites. Further pairwise associations between different metabolites (other categories) might indicate intermediates of the same pathway or branching points between distinct pathways indicating potential inter-organ pathway cross-talk.

By further integrating transcriptomics data from kidney and liver, we demonstrated that plasma metabolites additionally carry specific signatures of transcriptional regulation occurring in both organs. For example, between plasma and liver, we predominantly identified signatures of signaling pathways and metabolic processes, whereas mainly transport processes were reflected between plasma and kidney. Based on the observations from all constructed networks, we prioritized some plasma metabolites as proxy marker candidates for (1) multi-organ processes occurring in all organs and (2) single-organ specific processes. Finally, we demonstrated the usefulness of plasma metabolites as proxies for diabetes-related changes in the investigated organs and discussed several of our newly identified inter-organ proxy markers in the context of type 2 diabetes. We thereby confirm observations of a perturbed energy metabolism with a shift from glucose utilization as energy source to an increased use of fatty acids. This process affects all investigated organs and is widely reflected in the profiles of plasma metabolites.

Our approach could be extended in several directions: (1) The study was conducted on an animal model and despite studies showing that the db/db mouse model is well-suited to study the pathophysiology of type 2 diabetes [312, 313], it is still questionable

if results from mouse models are transferable to humans. (2) While transcriptomics measurements are considered to be comprehensive (i.e. genome-wide), metabolomics measurement techniques are lacking behind and a similar analysis with more advanced measurement technologies might identify further hits. (3) The integration of measurements from further organs might influence the identification of organ proxy markers, especially of those currently associated with a single organ.

In conclusion, our analysis has shown that plasma metabolites reflect metabolic and transcriptional processes occurring in the analyzed organs and changes in plasma are valid proxies for phenotype-related changes in the investigated tissues. Beyond the potential usage in clinical applications as prognostic and/or diagnostic markers, our study may also help to identify candidate processes for organ-specific investigations that are relevant for the pathophenotype under investigation. Lastly, our methodology can be readily transferred to other studies, possibly investigating different phenotypes, different organs, or even different omics.

Chapter 5

Summary & Outlook

Higher order biological systems are extremely complex, at a cellular level built up from different types of macromolecules such as DNA, RNA or proteins, but also metabolites, which are organized into multiple molecular layers. Cells themselves are hierarchically structured into different biological scales, ranging from tissues, over organs to ultimately whole organisms. The traditional reductionist approach in cell biology, i.e. disassembling biological systems into their individual components and subsequently studying their function in isolation has provided many valuable insights into cellular mechanisms [314]. For instance, gene knock-out studies have been successfully applied to elucidate the function of individual genes or proteins. However, together with the progress in molecular research, it became apparent that individual components of biological systems never work alone. Instead, they operate as interacting elements of hierarchically structured and tightly regulated molecular networks [75]. These different molecular networks or layers again are densely interconnected and operate in an orchestrated fashion, which is of central importance for cellular function and to sustain life as such. Modern high-throughput 'omics' technologies nowadays provide the tools to systematically analyze the relationships between the components of biological systems. For example, metabolic pathways have been successfully reconstructed from metabolomics data [69] or gene regulatory networks using transcriptomics data [182]. Moreover, many studies now also started to produce data from multiple omics technologies in parallel. Analyzing and understanding the relationships between entities of different molecular levels, however, remains a challenging task. To gain a better understanding of health and disease, the embedding of metabolism into the omics landscape is of particular interest, since

metabolic profiles are commonly seen as endpoints of biological processes - thus most closely linked to the observable phenotype [42].

In this thesis, a particular focus was laid on the integration of metabolomics data with varying other omics measurements derived from different biological scales. At a cellular level, we integrated metabolomics data measured on two different platforms with proteomics data to investigate the effect of an environmental pollutant on the metabolism of T cells (Chapter 2). At a tissue level, we integrated metabolomic, transcriptomic and genetic data derived from a cross-sectional human population study to elucidate general signaling, regulatory and metabolic processes observable in blood (Chapter 3). Furthermore, at a multi-organ level, we integrate metabolomics and transcriptomics data from various metabolically active organs with metabolomics data from plasma to investigate cross-organ associations and to identify plasma markers for organ-specific processes (Chapter 4).

Integration of metabolomics and proteomics at a cellular level

The advancing industrialization leads to an ever-increasing emission of environmental pollutants, which are known to adversely influence the immune system. However, the precise mechanisms behind this impact remains largely unknown. For a better understanding of the effect of an environmental pollutant, benzo[a]pyrene, on the metabolism of T cells, both in activated and resting state, we systematically analyzed metabolomics data from two different platforms combined with proteomics data measured in independent (repeated) experiments (Chapter 2). Due to this study design and the relative low number of replicate samples, no direct statistical integration (e.g. correlation-based) was possible. Instead a bottom-up approach was required, which utilizes *a priori* existing knowledge about the associations between two omics levels of interest in order to integrate them. Classical approaches from this field include enrichment analysis methods that integrate data with known biological pathways to identify affected processes. However, most of these methods were designed for single omics levels with only a few, often proprietary exceptions. Another interesting alternative to integrate omics datasets are model-based approaches, which map, in an automated manner, measured molecules onto pathways or networks and subsequently make use of mathematical algorithms to find those paths/regions where significant changes between two conditions accumulate. Random walks on graphs are one example of such an algorithm, which has been shown

to extract relevant subgraphs in relation to a given set of seeds (nodes of interest) from a single omics level [162].

We thus suggested the use of random walks for the integrated analysis of metabolomics and proteomics data in the context of a genome-scale metabolic network model to systematically identify affected metabolic processes. At the time this analysis was performed, there was only a low level of agreement between the existing pathway databases, but their content was at least in part complementary to each other [158]. Thus, we first constructed a consensus metabolic network model from the three major metabolic pathway resources [81, 155, 157]. The goal of this step was to generate a model that provides a maximal coverage of known biochemical associations. In metabolic networks, so-called currency metabolites are unspecifically taking part in a plethora of biochemical reactions, thereby introducing biologically meaningless shortcuts to the network model. To account for this and to force the transition probabilities of random walks towards biologically meaningful reaction paths, a topology-based weighting strategy was applied. Significantly changed proteins and metabolites between the different conditions were mapped onto the metabolic network model and subsequently used as seed nodes for the random walks approach. This allowed us to extract condition-specific metabolic subnetworks containing the most relevant metabolic pathways in relation to the changed proteins and metabolites. As a result, we identified enriched signatures of several pathways in T cell metabolism affected by B[a]P exposure, both in activated and resting state. Importantly, we observed signatures of leukotriene metabolism in resting T cells exposed to B[a]P, and IL-7 and phosphatidylinositol signaling in activated T cells, all of which are known to play prominent roles in T cell development and function. The identification of the phosphatidylinositol signaling system is of special interest, since it is known to act upstream of another important regulator of T cell development and metabolism, the mTOR signaling pathway [180], which was also identified by IPA. A B[a]P mediated dysregulation of these pathways thus might lead to alterations in the cell cycle and activity of T cells [315]. Taken together, these observations might help to further understand the the mechanisms behind the toxicological effects of benzo[a]pyrene.

From this project, we propose random walks on a metabolic model as useful tool to identify enriched pathways from multi-omics data. Random walks have several advantages when compared to classic enrichment approaches. For instance, the metabolic pathways are treated as joint entities, which allows for a more intuitive interpretation of the results by means of a continuous propagation through the metabolic network. Moreover, spurious results caused by the redundancy of metabolites and enzymes can

be avoided by this approach. Lastly, the random walks approach can be easily adapted to any kind of biological network further emphasizing the flexibility of this approach. To the best of our knowledge, we were the first to show that a joint network-based random walks approach on metabolomics and proteomics data can be used to extract condition-specific biologically meaningful subnetworks.

The human blood metabolome-transcriptome interface

Blood represents a heterogeneous mixture of blood cells and circulating metabolites derived from a variety of processes such as food intake, transport processes and excretory mechanisms of degradation products from and between various organs. The close proximity of metabolic compounds and metabolically active white blood cells, together with its easy non-invasive accessibility, renders blood an ideal tissue to systematically investigate their interactions for fundamental research. A particularly convenient study type to systematically integrate and analyze omics measurements from multiple molecular layers in humans are large-scale population studies, where natural inter-individual variation can be exploited to investigate how dysregulations and interactions in- and between these molecular levels contribute to phenotypic traits. Such an approach constitutes a new direction in the field of systems biology, which has recently been referred to as *Systems Genetics* [129]. However, to date, only very few well-powered human studies have made use of such data, especially when combining high-throughput metabolic and transcriptomic measurements [316]. More precisely, to the best of our knowledge, there is only one pioneer study investigating the interplay between circulating metabolites in blood and whole blood transcriptomics in a population-based environment [193].

In this study, we aimed to systematically investigate the interface between genome-wide transcriptomics and metabolomics data, i.e. to examine the complex interplay between circulating metabolites and blood cell transcripts on a systems scale (Chapter 3). We addressed this goal by integrating cross-sectional genome-wide transcriptional profiles and untargeted metabolomics data collected from blood of 712 individuals of the KORA F4 population study. From this dataset, we generated a global network of metabolite-transcript associations, reflecting the interplay between these organizational layers in human blood. In contrast to the study design from the previous Chapter 2, the parallel measurements of both omics in the same samples allowed us to directly infer these associations from the data by using a correlation-based approach. By applying both manual investigation and a systematic comparison of our data-derived network with an

existing metabolic pathway database [82], we were able to show that pairwise correlations between serum metabolites and whole blood transcripts carry a systematic signature of the underlying biochemical processes, especially around lipid transport, branched-chain amino acids and, more generally, immune system processes.

Moreover, despite a very limited overlap of 17 metabolites, we replicated 69 out of 211 possible associations ($\sim$ 19% of all BMTI edges) in an independent population cohort. Investigation of the possible origin of transcripts included in the blood metabolome-transcriptome interface with blood cell-type and tissue-specific gene expression databases revealed an unspecific expression of $\sim$ 70% of the transcripts included in the BMTI. This observation suggests a general validity of most of the reported associations, irrespective of the underlying tissue, or more precisely, that these associations represent signatures of fundamental molecular processes performed by all types of (blood) cells.

Another question that always arises when dealing with associations is, whether an observed correlation represents a direct causal effect. In order to assess whether any of our identified associations represents a causal effect, we performed a Mendelian randomization approach using genetic markers (SNPs) as instrumental variables. We used both SNPs associated with metabolites (mQTLs) and transcripts (eQTLs) to test for both causal directions metabolite $\rightarrow$ mRNA and mRNA $\rightarrow$ metabolite. We observed nominally significant causal effects for both directions, however none of them remained significant after multiple testing correction. There can be different reasons why we were not able to detect strong evidences for causal effects, the most probable being the low statistical power owed to the sample size.

When comparing the metabolite-transcript associations with a metabolic pathway database, we observed that most of the correlating mRNAs do not encode enzymes and thus are not contained in the database. Moreover, these databases are typically incomplete and also the mapping capabilities of metabolites onto their respective metabolic pathway nodes due to, for instance ambiguous nomenclatures, are often rather limited. To account for this, we developed a novel bi-directional enrichment approach based on aggregated z-scores of functional annotations. Constructing a pathway interaction network by this approach allowed us to embed metabolites with missing pathway evidence and non enzymatic transcripts into our functional analysis, and also to investigate the inherent cross-talk between biological pathways in a data-driven fashion. Furthermore, in many cases these aggregated pathway signatures displayed stronger correlations with

each other than any single molecule alone, suggesting a higher robustness due to lower individual-specific variations in the metabolite and transcript levels.

Inspecting the BMTI network topology, a prominent 'flower-like' network motif, i.e. an accumulation of one-to-many associations between metabolites and transcripts, respectively, was observable. This suggested a systematic coregulation between transcripts associated with the same metabolite or metabolic pathway. By systematically screening the promoter sequences of all transcripts associated to a certain metabolite/metabolic pathway for enriched transcription factor binding sites, we found concrete evidences for this hypothesis further suggesting that correlations between transcripts and metabolites not only reflect actual metabolic pathway reactions, but are also of regulatory nature.

In addition, we provided a showcase how our data-derived networks can be used to investigate clinical phenotypes and to generate novel hypotheses on the underlying mechanisms. To this end, we projected metabolite and transcript associations with clinical risk factors (HDL-C, LDL-C, and triglycerides) onto the metabolite-transcript and the pathway interaction network and identified phenotype associated cellular molecules as well as whole pathways which have never been reported before.

In conclusion, our study clearly highlighted the potential of a systems genetics approach for understanding interactions across multiple biological scales and further advances the insights into the nature of interactions between circulating metabolites and blood cellular gene expression.

Plasma proxy markers for inter-organ processes

Because of its close connection with all organs and tissues, blood harbors systemic information about whole organism processes. Clinicians exploit this unique characteristic of blood for prognosis and diagnosis of diseases or for the assessment of organ function by monitoring the levels of specific molecules (markers) in blood tests. More recently, metabolomics-based discovery studies started to systematically link levels of blood metabolites with various pathophysiological conditions [113, 261] promising a great potential for clinical research and diagnostic applications. However, only little is known about the precise origins of blood metabolites, i.e. which organ activities give rise to the detectable metabolite signals, hampering the possibility to draw conclusions on the mechanism of disease and limiting the potential of an improved diagnosis. Moreover, knowing which organ processes contribute to the metabolite levels opens new possibil-

ities for the use of blood metabolites as proxy markers that, for instance, measure the physiological status of less easily accessible organs.

In this project, we therefore asked whether blood metabolites can be used as proxy markers of organ processes. To this end, we used adipose tissue, kidney, liver and muscle samples, all simultaneously obtained from 100 mice (80 db/db and 20 wild-type) and investigated the associations between organ and plasma metabolomes. Moreover, we integrated global gene expression profiles from kidney and liver samples and analyzed whether plasma metabolites additionally reflect transcriptional processes in organs.

A systematic comparison of the organ and plasma metabolomes revealed that metabolic profiles are representative of their source organ, which might be explained by the compartmentalization between them. In addition, a large fraction of the measured metabolites are only detectable in specific tissues. Calculation of pairwise correlation networks between each organ and plasma both at a pathway and single molecule level showed that metabolites and metabolic processes in kidney are most strongly reflected by plasma metabolites, followed by muscle, liver and adipose tissue. Notably, more than three quarters of the associations observed between plasma and kidney were negative, suggesting a reflection of the physiological function of kidney as major blood filter.

Based on the constructed pairwise correlation networks between each organ and plasma, we inferred three biologically meaningful association categories: (1) Associations between identical metabolites across organs. These associations most likely reflect an active exchange between plasma and the respective organ. (2) Associations between metabolites of the same metabolite class. These cases might be explained by a dependence on the same substrate pool, i.e. pathway-wide correlations between intermediates caused by mass flow. (3) Associations between metabolites from different pathways across organs. These pairs might reflect a pathway cross-talk, or a common upstream regulation of different metabolic pathways. Notably, within the first category, we identified associations between 3 identical metabolites which were traceable through all five investigated organs, namely 3-hydroxybutyrate (BHBA), 3-dehydrocarnitine and X-12465, further supporting our hypothesis of a common organism wide source for these metabolites.

Integration of transcriptomics data measured in kidney and liver samples further demonstrated that plasma metabolites carry a distinct signal for both organs. For instance, from liver, we predominantly identified signatures of signaling pathways and metabolic processes, whereas mainly transport processes were reflected between plasma and kidney.

Combining the information obtained from all constructed networks, we prioritized some plasma metabolites as proxy marker candidates for (1) multi-organ processes occurring in all organs and (2) single-organ specific processes. In a final step, we investigated diabetes-related changes in all organs and identified more than 500 altered organ-plasma association pairs demonstrating the potential of plasma metabolites as proxy markers for organ processes. For instance, blood levels of 3-hydroxybutyrate (BHBA) were indicative of changes in all four investigated organs suggesting a global shift from glucose utilization as energy use to an increased use of fatty acids.

In conclusion, our analysis shed more light onto the origins of plasma metabolites in terms of reflected processes and pathways of the examined organs. Moreover, we showed that plasma metabolites can be used as proxies for phenotype-related changes in organs which might further help to understand the underlying disease mechanisms and to identify potential leverage points for therapeutic interventions.

Extensions and future directions

There are various possible extensions to the omics integration approaches presented in this thesis which we will discuss in the following.

Random walks

- In weighted networks, the visitation probabilities for nodes or edges are strongly influenced by their respective initial weights, the weights of their direct neighbors, as well as the local network connectivity. Thus, the weighting strategy is of great importance for the outcome of the random walks approach. Our applied weighting strategy could be extended in several directions: (1) Integration of data values derived from statistical tests. For instance, p-values from t-tests could be mapped as weights all nodes which would not only consider dichotomously determined seeds for the subnetwork extraction, but also accounts for more subtle changes. (2) Integration of the direction of changes, i.e. up- or down-regulation, like for instance applied in a similar approach by Zur et al. [93]. This would direct the random walks towards active (up-regulated) reaction paths which might lead to biologically more relevant subnetworks. (3) Integration of reaction stoichiometry and directions as utilized in flux balance analysis [89]. Again this would direct the random walks towards metabolically feasible reaction paths.

- Another factor that strongly influences the results of a random walks approach is the quality and completeness of the underlying network model. Our analysis could be repeated with the recently published most comprehensive human specific Recon 2 metabolic model [82]. Furthermore, data-derived network models such as correlation networks might be an interesting alternative, since mapping issues would disappear and correlation strengths between two nodes could be directly used as weights for random walks directing them towards edges with a higher confidence.

- The threshold for subgraph extraction could be determined in different ways. For instance, inverse edge relevances could be used to define edge costs and an algorithm could be used to find the subgraph with minimal total costs connecting the seed nodes [162]. Or the thresholds could be determined empirically be randomly shuffling seed nodes on the network and calculating p-values for the edges from the distributions of node/ edge relevances.

Human blood metabolome-transcriptome interface

- A major issue when dealing with whole blood transcriptomics data is, that the data represents a readout from a mixture of blood cells. Especially in correlation analysis this might introduce false positives in terms of spurious correlations caused by differences in cell-types between individuals rather than in the levels of the actual molecules. Moreover, cell-type specific correlations between molecules might be masked by the contributions of other cell types and thus can only hardly be identified using whole blood data. The effect of the cellular composition of blood on our identified correlations should be investigated, for instance, by applying the Houseman method, which has been shown to accurately estimate the actual blood cellular composition of each individual based on differentially methylated DNA regions [317]. Moreover, the estimated cellular distributions across individuals could subsequently be used for deconvolution of the whole blood transcriptomics data [318] and identification of cell-type specific associations.

- Another biological process that might influence the associations between metabolites and transcripts is the circadian rhythm [319]. Thus, strong differences in the state of circadian rhythm across individuals, e.g. caused by different sleep-wake cycles, might introduce strong variations into the levels of associated transcripts and metabolites also influencing the results from a correlation analysis. A possibility

to account for this and align the circadian rhythm between individuals might be the usage of circulating cortisol levels to adjust the data, which have been shown to be stable markers for the individual circadian rhythm [320].

- Applying Mendelian randomization to our data resulted in the identification of only nominally significant causal associations between metabolites and transcripts. This might just be a power issue due to low sample sizes or weak instruments. A possibility to overcome this limitation might be the use of multiple variants combined into a weighted allele score [17].

- We developed a novel bi-enrichment method that aggregates gene expression levels and metabolite concentrations based on respective functional annotations which subsequently can be associated with each other. This novel method allowed us to functionally interpret associations between molecules without any or only weak database support. Currently, our group is working on a extension to this approach using principal components instead of aggregated z-scores to integrate the expression values of molecules. In addition we could make the method publicly available as part of a package for R.

- Moreover, in Section 3.9 we have shown how the BMTI can be complemented with phenotypic information by coloring the nodes according to the direction and strength of association with the respective phenotype. This information could be directly used as input for our random walks approach in Chapter 2 to identify the BMTI subnetwork most relevant for the phenotype of interest. This could facilitate the functional interpretation and provide novel insights into the underlying mechanisms.

- As data becomes available, the human metabolome-transcriptome interface could be extended by further molecular levels which could provide even more insights into the interplay between the different layers.

Plasma proxy markers for inter-organ processes

- As a first step, we particularly focused on the pairwise associations between plasma metabolites and organ metabolites and transcripts, respectively. However, the analysis could be extended by the remaining possible intra- and inter-organ associations to generate a comprehensive picture of the systemic metabolism.

- Furthermore, a systematic comparison of the intra-organ associations between metabolites and transcripts in kidney and liver with the blood metabolite-transcript associations in humans might enable us to separate generally valid associations from those that are blood specific.
- The utilized study was conducted on mice, which is the most widely used model organism to study human diseases. However, care must be taken when transferring observations between organisms. Ideally, the analysis should be repeated in humans to validate the identified organ proxy markers, which yet might be unfeasible due to the necessary invasive procedures.
- Moreover, the analysis was conducted with a relatively low sample size when compared to the population study used in Chapter 3. Thus, the robustness of the identified associations should be checked, for instance by cross validation.

Conclusion

Complex biological systems operate on multiple, intertwined molecular levels that can nowadays be quantitatively assessed by high-throughput measurement methods, the so-called omics technologies. A major aim in the field of systems biology is to understand the flow of biological information between the different layers at a systems level in both health and disease. In this thesis, we integrated metabolomics with proteomics and transcriptomics data from a cellular to an organ level and systematically investigated the relationships between these molecular levels. We thereby generated novel insights into the structure and regulation of systemic metabolism.

Bibliography

[1] Watson, J.D. and Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953. ISSN 0028-0836.

[2] Watson, J.D. and Crick, F.H. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, 1953. ISSN 0028-0836.

[3] Crick, F.H. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958. ISSN 0081-1386.

[4] Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356, 1961. ISSN 0022-2836.

[5] Beadle, G.W. and Tatum, E.L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499–506, 1941. ISSN 0027-8424.

[6] Waddington, C.H. The epigenotype. 1942. *International Journal of Epidemiology*, 41(1):10–13, 2012. ISSN 1464-3685.

[7] Holliday, R. DNA Methylation and Epigenetic Inheritance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 326(1235):329–338, 1990. ISSN 0962-8436, 1471-2970.

[8] Berget, S.M., Moore, C., and Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175, 1977. ISSN 0027-8424, 1091-6490.

[9] Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, 1977. ISSN 0092-8674.

[10] Noble, D. *The Music of Life: Biology Beyond the Genome*. OUP Oxford, 2006. ISBN 978-0-19-929573-9.

[11] Kitano, H. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[12] Lockhart, D.J. and Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–836, 2000. ISSN 0028-0836.

[13] Sanger, F., Nicklen, S., and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1977. ISSN 0027-8424.

[14] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512, 1995. ISSN 0036-8075.

[15] Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girn, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Khri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., and Flicek, P. Ensembl 2015. *Nucleic Acids Research*, page gku1010, 2014. ISSN 0305-1048, 1362-4962.

[16] Karczewski, K.J., Daneshjou, R., and Altman, R.B. Chapter 7: Pharmacogenomics. *PLoS Computational Biology*, 8(12):e1002817, 2012. ISSN 1553-7358.

[17] Davey Smith, G. and Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98, 2014. ISSN 0964-6906, 1460-2083.

[18] Hirschhorn, J.N. and Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005. ISSN 1471-0056, 1471-0064.

[19] Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D., Meisinger, C., Wichmann, H.E., Hoffmann, W., Vlzke, H., Vlker, U., Teumer, A., Biffar, R., Kocher, T., Felix, S.B., Illig, T., Kroemer, H.K., Gieger, C., Rmisch-Margl, W., and Nauck, M. A genome-wide association study of metabolic traits in human urine. *Nature genetics*, 43(6):565–569, 2011. ISSN 1546-1718.

[20] Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.P., Walter, K., Menni, C., Chen, L., Vasquez, L., Valdes, A.M., Hyde, C.L., Wang, V., Ziemek, D., Roberts, P., Xi, L., Grundberg, E., The Multiple Tissue Human Expression Resource (MuTHER) Consortium, Waldenberger, M., Richards, J.B., Mohney, R.P., Milburn, M.V., John, S.L., Trimmer, J., Theis, F.J., Overington, J.P., Suhre, K., Brosnan, M.J., Gieger, C., Kastenmller, G., Spector, T.D., and Soranzo, N. An

atlas of genetic influences on human blood metabolites. *Nature Genetics*, advance online publication, 2014. ISSN 1061-4036.

[21] Miller, M.B. and Tang, Y.W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4):611–633, 2009. ISSN 1098-6618.

[22] Drghici, S. *Statistics and Data Analysis for Microarrays Using R and Bioconductor, Second Edition*. CRC Press, 2011. ISBN 978-1-4398-0975-4.

[23] Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. ISSN 1471-0056.

[24] Zelezniak, A., Pers, T.H., Soares, S., Patti, M.E., and Patil, K.R. Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. *PLoS Comput Biol*, 6(4):e1000729, 2010.

[25] Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1), 2007. ISSN false.

[26] Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.

[27] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[28] Jiang, D., Tang, C., and Zhang, A. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[29] Huang, D.W., Sherman, B.T., and Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. ISSN 1362-4962.

[30] Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., and Williams, K.L. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology & Genetic Engineering Reviews*, 13:19–50, 1996. ISSN 0264-8725.

[31] Cox, J. and Mann, M. Is proteomics the new genomics? *Cell*, 130(3):395–398, 2007. ISSN 0092-8674.

[32] Klein, J.B. and Thongboonkerd, V. Overview of proteomics. *Proteomics Nephrol*, 141:1–10, 2004.

[33] Cha, S., Imielinski, M.B., Rejtar, T., Richardson, E.A., Thakur, D., Sgroi, D.C., and Karger, B.L. In Situ Proteomic Analysis of Human Breast Cancer Epithelial Cells Using Laser Capture Microdissection: Annotation by Protein Set Enrichment Analysis and Gene Ontology. *Molecular & Cellular Proteomics*, 9(11):2529–2544, 2010. ISSN 1535-9476, 1535-9484.

[34] Isabel Padro, A., Ferreira, R., Vitorino, R., and Amado, F. Proteome-base biomarkers in diabetes mellitus: progress on biofluids' protein profiling using mass spectrometry. *Proteomics. Clinical Applications*, 6(9-10):447–466, 2012. ISSN 1862-8354.

[35] Rao, P.V., Reddy, A.P., Lu, X., Dasari, S., Krishnaprasad, A., Biggs, E., Roberts, C.T., and Nagalla, S.R. Proteomic Identification of Salivary Biomarkers of Type-2 Diabetes. *Journal of Proteome Research*, 8(1):239–245, 2009. ISSN 1535-3893.

[36] Wienkoop, S., Morgenthal, K., Wolschin, F., Scholz, M., Selbig, J., and Weckwerth, W. Integration of Metabolomic and Proteomic Phenotypes. *Molecular & Cellular Proteomics : MCP*, 7(9):1725–1736, 2008. ISSN 1535-9476.

[37] Gibbs, D.L. and McWeeney, S.K. Multi-omic network signatures of disease. *Frontiers in Systems Biology*, 4:309, 2014.

[38] Seliger, B., Dressler, S.P., Wang, E., Kellner, R., Recktenwald, C.V., Lottspeich, F., Marincola, F.M., Baumgrtner, M., Atkins, D., and Lichtenfels, R. Combined analysis of transcriptome and proteome data as a tool for the identification of candidate biomarkers in renal cell carcinoma. *PROTEOMICS*, 9(6):1567–1581, 2009. ISSN 16159853, 16159861.

[39] MCATEER, J.G.P., SKERRETT, S.J., LIGGITT, D., and FREVERT, C.W. A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections. In *Pacific Symposium on Biocomputing 2009: Kohala Coast, Hawaii, USA, 5-9 January 2009*, page 451. 2009.

[40] Oliver, S.G., Winson, M.K., Kell, D.B., and Baganz, F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16(9):373–378, 1998. ISSN 0167-7799.

[41] Nicholson, J.K. and Lindon, J.C. Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056, 2008. ISSN 0028-0836.

[42] Fiehn, O. Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2):155–171, 2002. ISSN 0167-4412.

[43] Ludwig, C. and Viant, M.R. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical analysis: PCA*, 21(1):22–32, 2010. ISSN 1099-1565.

[44] Roux, A., Lison, D., Junot, C., and Heilier, J.F. Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clinical biochemistry*, 44(1):119–135, 2011. ISSN 1873-2933.

[45] Patti, G.J., Yanes, O., and Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, 2012. ISSN 1471-0072.

[46] Lindon, J.C., Holmes, E., and Nicholson, J.K. Metabonomics techniques and applications to pharmaceutical research & development. *Pharmaceutical research*, 23(6):1075–1088, 2006.

[47] Wishart, D.S. Advances in metabolite identification. *Bioanalysis*, 3(15):1769–1782, 2011.

[48] Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wgele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., Hammond, C.J., de Angelis, M.H., Kastenmller, G., Kttgen, A., Kronenberg, F., Mangino, M., Meisinger, C., Meitinger, T., Mewes, H.W., Milburn, M.V., Prehn, C., Raffler, J., Ried, J.S., Rmisch-Margl, W., Samani, N.J., Small, K.S., Wichmann, H.E., Zhai, G., Illig, T., Spector, T.D., Adamski, J., Soranzo, N., and Gieger, C. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, 2011. ISSN 1476-4687.

[49] Gieger, C., Geistlinger, L., Altmaier, E., Hrab de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T., and Suhre, K. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, 4(11):e1000282, 2008. ISSN 1553-7404.

[50] Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W., Clish, C.B., and Mootha, V.K. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science (New York, N.Y.)*, 336(6084):1040–1044, 2012. ISSN 1095-9203.

[51] Fav, G., Beckmann, M.E., Draper, J.H., and Mathers, J.C. Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics? *Genes & nutrition*, 4(2):135–141, 2009. ISSN 1555-8932.

[52] Bondia-Pons, I., Nordlund, E., Mattila, I., Katina, K., Aura, A.M., Kolehmainen, M., Orei, M., Mykknen, H., and Poutanen, K. Postprandial differences in the plasma metabolome of healthy Finnish subjects after intake of a sourdough fermented endosperm rye bread versus white wheat bread. *Nutrition journal*, 10:116, 2011. ISSN 1475-2891.

[53] Fendt, S.M., Buescher, J.M., Rudroff, F., Picotti, P., Zamboni, N., and Sauer, U. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Molecular systems biology*, 6:356, 2010. ISSN 1744-4292.

[54] Vander Heiden, M.G. Targeting cancer metabolism: a therapeutic window opens. *Nature reviews. Drug discovery*, 10(9):671–684, 2011. ISSN 1474-1784.

[55] Bartel, J., Krumsiek, J., and J. Theis, F. Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal*, 4(5), 2013. ISSN 20010370.

[56] Saltiel, A.R. and Kahn, C.R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, 2001. ISSN 0028-0836.

[57] DeSouza, C. and Fonseca, V. Therapeutic targets to reduce cardiovascular disease in type 2 diabetes. *Nature Reviews Drug Discovery*, 8(5):361–367, 2009. ISSN 1474-1776.

[58] Forbes, J.M. and Cooper, M.E. Mechanisms of Diabetic Complications. *Physiological Reviews*, 93(1):137–188, 2013. ISSN 0031-9333, 1522-1210.

[59] Zammitt, N.N. and Frier, B.M. Hypoglycemia in Type 2 Diabetes Pathophysiology, frequency, and effects of different treatment modalities. *Diabetes Care*, 28(12):2948–2961, 2005. ISSN 0149-5992, 1935-5548.

[60] Hotamisligil, G.S., Peraldi, P., Budavari, A., Ellis, R., White, M.F., and Spiegelman, B.M. IRS-1-mediated inhibition of insulin receptor tyrosine kinase activity in TNF-alpha- and obesity-induced insulin resistance. *Science (New York, N.Y.)*, 271(5249):665–668, 1996. ISSN 0036-8075.

[61] Torell, F., Bennett, K., Cereghini, S., Rnnar, S., Lundstedt-Enkel, K., Moritz, T., Haumaitre, C., Trygg, J., and Lundstedt, T. Multi-Organ Contribution to the Metabolic Plasma Profile Using Hierarchical Modelling. *PLoS ONE*, 10(6):e0129260, 2015.

[62] Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, 101(27):10205–10210, 2004. ISSN 0027-8424, 1091-6490.

[63] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Bayesian Independent Component Analysis Recovers Pathway Signatures from Blood Metabolomics Data. *Journal of Proteome Research*, 11(8):4120–4131, 2012. ISSN 1535-3893.

[64] Mittelstrass, K., Ried, J.S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., Meitinger, T., Kronenberg, F., Weidinger, S., Wichmann, H.E., Suhre, K., Wang-Sattler, R., Adamski, J., and Illig, T. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genet*, 7(8):e1002215, 2011.

[65] Hori, S., Nishiumi, S., Kobayashi, K., Shinohara, M., Hatakeyama, Y., Kotani, Y., Hatano, N., Maniwa, Y., Nishio, W., Bamba, T., Fukusaki, E., Azuma, T., Takenawa, T., Nishimura, Y., and Yoshida, M. A metabolomic approach to lung cancer. *Lung Cancer*, 74(2):284–292, 2011. ISSN 0169-5002.

[66] Quaranta, M., Knapp, B., Garzorz, N., Mattii, M., Pullabhatla, V., Pennino, D., Andres, C., Traidl-Hoffmann, C., Cavani, A., Theis, F.J., Ring, J., Schmidt-Weber, C.B., Eyerich,

S., and Eyerich, K. Intraindividual genome expression analysis reveals a specific molecular signature of psoriasis and eczema. *Science Translational Medicine*, 6(244):244ra90–244ra90, 2014. ISSN 1946-6234, 1946-6242.

[67] Xia, J. and Wishart, D.S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(Web Server issue):W71–77, 2010. ISSN 1362-4962.

[68] Chagoyen, M. and Pazos, F. MBRole: enrichment analysis of metabolomic data. *Bioinformatics (Oxford, England)*, 27(5):730–731, 2011. ISSN 1367-4811.

[69] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011.

[70] Langfelder, P. and Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9:559, 2008. ISSN 1471-2105.

[71] Cvijovic, M., Olivares-Hernndez, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K.R., and Nielsen, J. BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acids Research*, page gkq404, 2010. ISSN 0305-1048, 1362-4962.

[72] Backes, C., Rurainski, A., Klau, G.W., Mller, O., Stckel, D., Gerasch, A., Kntzer, J., Maisel, D., Ludwig, N., Hein, M., Keller, A., Burtscher, H., Kaufmann, M., Meese, E., and Lenhof, H.P. An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Research*, page gkr1227, 2011. ISSN 0305-1048, 1362-4962.

[73] Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, 2013. ISSN 1548-7091.

[74] Joyce, A.R. and Palsson, B.. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006. ISSN 1471-0072.

[75] Barabsi, A.L. and Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. ISSN 1471-0056.

[76] Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A.P., Cayting, P., Charos, A., Chen, D.Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J.J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E.C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T.E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K.Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P.J., Myers, R.M., Weissman, S.M., and Snyder,

M. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012. ISSN 0028-0836.

[77] Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muiz-Rascado, L., Garca-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martnez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernndez, S., Alquicira-Hernndez, K., Lpez-Fuentes, A., Porrn-Sotelo, L., Huerta, A.M., Bonavides-Martnez, C., Balderas-Martnez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jimnez-Jacinto, V., Vega-Alvarado, L., Del Moral-Chvez, V., Hernndez-Alvarez, A., Morett, E., and Collado-Vides, J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(Database issue):D203–213, 2013. ISSN 1362-4962.

[78] Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stmpflen, V., Mewes, H.W., Ruepp, A., and Frishman, D. The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, 21(6):832–834, 2005. ISSN 1367-4803.

[79] Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–772, 2009. ISSN 1362-4962.

[80] Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Weerasinghe, D., Zhang, P., and Karp, P.D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):D459–D471, 2014. ISSN 0305-1048, 1362-4962.

[81] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2011. ISSN 0305-1048, 1362-4962.

[82] Thiele, I., Swainston, N., Fleming, R.M.T., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., Thorleifsson, S.G., Agren, R., Blling, C., Bordel, S., Chavali, A.K., Dobson, P., Dunn, W.B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J.J., Juty, N., Keating, S., Nookaew, I., Le Novre, N., Malys, N., Mazein, A., Papin, J.A., Price, N.D., Selkov Sr, E., Sigurdsson, M.I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J.H.G.M., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H.V., Kell, D.B., Mendes, P., and Palsson, B..

A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31(5):419–425, 2013. ISSN 1087-0156.

[83] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003. ISSN 1088-9051.

[84] Hu, Z., Chang, Y.C., Wang, Y., Huang, C.L., Liu, Y., Tian, F., Granger, B., and DeLisi, C. VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic Acids Research*, 41(W1):W225–W231, 2013. ISSN 0305-1048, 1362-4962.

[85] Meierhofer, D., Weidner, C., and Sauer, S. Integrative Analysis of Transcriptomics, Proteomics, and Metabolomics Data of White Adipose and Liver Tissue of High-Fat Diet and Rosiglitazone-Treated Insulin-Resistant Mice Identified Pathway Alterations and Molecular Hubs. *Journal of Proteome Research*, 13(12):5592–5602, 2014. ISSN 1535-3893.

[86] Shahzad, K. and Loor, J.J. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Current Genomics*, 13(5):379–394, 2012. ISSN 1389-2029.

[87] Çakir, T. and Khatibipour, M.J. Metabolic Network Discovery by Top-Down and Bottom-Up Approaches and Paths for Reconciliation. *Frontiers in Bioengineering and Biotechnology*, 2, 2014. ISSN 2296-4185.

[88] Edwards, J.S. and Palsson, B.O. Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype. *Journal of Biological Chemistry*, 274(25):17410–17416, 1999. ISSN 0021-9258, 1083-351X.

[89] Orth, J.D., Thiele, I., and Palsson, B.. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010. ISSN 1087-0156.

[90] Papin, J.A., Price, N.D., and Palsson, B.. Extreme Pathway Lengths and Reaction Participation in Genome-Scale Metabolic Networks. *Genome Research*, 12(12):1889–1900, 2002. ISSN 1088-9051.

[91] Famili, I., Frster, J., Nielsen, J., and Palsson, B.O. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences*, 100(23):13134–13139, 2003.

[92] Ibarra, R.U., Edwards, J.S., and Palsson, B.O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–189, 2002. ISSN 0028-0836.

[93] Zur, H., Ruppin, E., and Shlomi, T. iMAT: an integrative metabolic analysis tool. *Bioinformatics (Oxford, England)*, 26(24):3140–3142, 2010. ISSN 1367-4811.

[94] Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. *PLoS Comput Biol*, 8(5):e1002518, 2012.

[95] Zelezniak, A., Sheridan, S., and Patil, K.R. Contribution of Network Connectivity in Determining the Relationship between Gene Expression and Metabolite Concentration Changes. *PLoS Comput Biol*, 10(4):e1003572, 2014.

[96] Patil, K.R. and Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685–2689, 2005.

[97] Oliveira, A.P., Patil, K.R., and Nielsen, J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Systems Biology*, 2(1):17, 2008. ISSN 1752-0509.

[98] Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergstrm, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bckhed, F. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013. ISSN 0028-0836.

[99] Çakir, T., Patil, K.R., Önsan, Z.I., Ülgen, K.O., Kirdar, B., and Nielsen, J. Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular Systems Biology*, 2, 2006. ISSN 1744-4292.

[100] Mitra, K., Carvunis, A.R., Ramesh, S.K., and Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013. ISSN 1471-0056.

[101] Yeung, K.Y., Dombek, K.M., Lo, K., Mittler, J.E., Zhu, J., Schadt, E.E., Bumgarner, R.E., and Raftery, A.E. Construction of regulatory networks using expression time-series data of a genotyped population. *Proceedings of the National Academy of Sciences*, 108(48):19436–19441, 2011. ISSN 0027-8424, 1091-6490.

[102] Astola, L., Groenenboom, M., Gomez Roldan, V., van Eeuwijk, F., Hall, R., Bovy, A., and Molenaar, J. Metabolic pathway inference from time series data: a non iterative approach. *Pattern Recognition in Bioinformatics*, pages 97–108, 2011.

[103] Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, 2003. ISSN 1367-4803, 1460-2059.

[104] Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., and Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory

Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006. ISSN 1471-2105.

[105] Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79–82, 2013. ISSN 0028-0836.

[106] Pearson, K. Contributions to the Mathematical Theory of Evolution. III. Regression, Heredity, and Panmixia. *Proceedings of the Royal Society of London*, 59(353-358):69–71, 1895.

[107] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003. ISSN 1061-4036.

[108] Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, 302(5643):249–255, 2003. ISSN 1095-9203.

[109] Bergmann, S., Ihmels, J., and Barkai, N. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biol*, 2(1):e9, 2003.

[110] Mabbott, N.A., Baillie, J.K., Brown, H., Freeman, T.C., and Hume, D.A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, 14(1):632, 2013. ISSN 1471-2164.

[111] Camacho, D., de la Fuente, A., and Mendes, P. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, 2005. ISSN 1573-3882.

[112] Morgenthal, K., Weckwerth, W., and Steuer, R. Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Bio Systems*, 83(2-3):108–117, 2006. ISSN 0303-2647.

[113] Krumsiek, J., Suhre, K., Evans, A.M., Mitchell, M.W., Mohney, R.P., Milburn, M.V., Wgele, B., Rmisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F.J., and Kastenmller, G. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genet*, 8(10):e1003005, 2012.

[114] Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics*, 16(6):283–293, 2002. ISSN 1099-128X.

[115] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620, 2000. ISSN 1066-5277, 1557-8666.

[116] Freudenberg, J., Wang, M., Yang, Y., and Li, W. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. *BMC Bioinformatics*, 10(Suppl 1):S66, 2009. ISSN 1471-2105.

[117] Opgen-Rhein, R. and Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(1):37, 2007. ISSN 17520509.

[118] Steuer, R., Kurths, J., Daub, C.O., Weise, J., and Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, 2002. ISSN 1367-4803, 1460-2059.

[119] Spearman, C. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[120] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.

[121] Greenbaum, D., Jansen, R., and Gerstein, M. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics (Oxford, England)*, 18(4):585–596, 2002. ISSN 1367-4803.

[122] Carrari, F., Baxter, C., Usadel, B., Urbanczyk-Wochniak, E., Zanor, M.I., Nunes-Nesi, A., Nikiforova, V., Centero, D., Ratzka, A., Pauly, M., Sweetlove, L.J., and Fernie, A.R. Integrated Analysis of Metabolite and Transcript Levels Reveals the Metabolic Shifts That Underlie Tomato Fruit Development and Highlight Regulatory Aspects of Metabolic Network Behavior. *Plant Physiology*, 142(4):1380–1396, 2006. ISSN 0032-0889.

[123] Walther, D., Strassburg, K., Durek, P., and Kopka, J. Metabolic Pathway Relationships Revealed by an Integrative Analysis of the Transcriptional and Metabolic Temperature Stress-Response Dynamics in Yeast. *OMICS: A Journal of Integrative Biology*, 14(3):261–274, 2010. ISSN 1536-2310, 1557-8100.

[124] Su, G., Burant, C.F., Beecher, C.W., Athey, B.D., and Meng, F. Integrated metabolome and transcriptome analysis of the NCI60 dataset. *BMC bioinformatics*, 12(Suppl 1):S36, 2011.

[125] Clish, C.B., Davidov, E., Oresic, M., Plasterer, T.N., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., Adourian, A., et al. Integrative biological analysis of the APOE* 3-leiden transgenic mouse. *Omics: a journal of integrative biology*, 8(1):3–13, 2004.

[126] Zhu, J., Sova, P., Xu, Q., Dombek, K.M., Xu, E.Y., Vu, H., Tu, Z., Brem, R.B., Bumgarner, R.E., and Schadt, E.E. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLoS Biol*, 10(4):e1001301, 2012.

[127] Stifanelli, P.F., Creanza, T.M., Anglani, R., Liuzzi, V.C., Mukherjee, S., and Ancona, N. A comparative study of Gaussian Graphical Model approaches for genomic data. *arXiv preprint arXiv:1107.0261*, 2011.

[128] Markowetz, F. and Spang, R. Inferring cellular networks a review. *BMC Bioinformatics*, 8(Suppl 6):S5, 2007. ISSN 1471-2105.

[129] Civelek, M. and Lusis, A.J. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, 2013. ISSN 1471-0056, 1471-0064.

[130] Ghazalpour, A., Bennett, B.J., Shih, D., Che, N., Orozco, L., Pan, C., Hagopian, R., He, A., Kayne, P., Yang, W.p., Kirchgessner, T., and Lusis, A.J. Genetic regulation of mouse liver metabolite levels. *Molecular Systems Biology*, 10(5), 2014. ISSN 1744-4292, 1744-4292.

[131] Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., Clark, M.J., Pagnon, J., Yang, W.P., He, A., Truong, A., Patel, S., Nelson, S.F., Horvath, S., Berliner, J.A., Kirchgessner, T.G., and Lusis, A.J. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences of the United States of America*, 103(34):12741–12746, 2006. ISSN 0027-8424.

[132] Holle, R., Happich, M., Lwel, H., Wichmann, H.E., and MONICA/KORA Study Group. KORA–a research platform for population based health research. *Gesundheitswesen (Bundesverband Der rzte Des ffentlichen Gesundheitsdienstes (Germany))*, 67 Suppl 1:S19–25, 2005. ISSN 0941-3790.

[133] Orkin, S.H. and Zon, L.I. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell*, 132(4):631–644, 2008. ISSN 0092-8674.

[134] Liew, C.C., Ma, J., Tang, H.C., Zheng, R., and Dempsey, A.A. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *The Journal of laboratory and clinical medicine*, 147(3):126–132, 2006. ISSN 0022-2143.

[135] Cai, C., Langfelder, P., Fuller, T.F., Oldham, M.C., Luo, R., Berg, L.H.v.d., Ophoff, R.A., and Horvath, S. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*, 11(1):589, 2010. ISSN 1471-2164.

[136] Bomprezzi, R., Ringnr, M., Kim, S., Bittner, M.L., Khan, J., Chen, Y., Elkahloun, A., Yu, A., Bielekova, B., Meltzer, P.S., Martin, R., McFarland, H.F., and Trent, J.M. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Human Molecular Genetics*, 12(17):2191–2199, 2003. ISSN 0964-6906.

[137] Tang, Y., Nee, A.C., Lu, A., Ran, R., and Sharp, F.R. Blood genomic expression profile for neuronal injury. *Journal of Cerebral Blood Flow & Metabolism*, 23(3):310–319, 2003.

[138] Lampe, J.W., Stepaniants, S.B., Mao, M., Radich, J.P., Dai, H., Linsley, P.S., Friend, S.H., and Potter, J.D. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 13(3):445–453, 2004. ISSN 1055-9965.

[139] Connolly, P.H., Caiozzo, V.J., Zaldivar, F., Nemet, D., Larson, J., Hung, S.P., Heck, J.D., Hatfield, G.W., and Cooper, D.M. Effects of exercise on gene expression in human peripheral blood mononuclear cells. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, 97(4):1461–1469, 2004. ISSN 8750-7587.

[140] Herder, C., Karakas, M., and Koenig, W. Biomarkers for the Prediction of Type 2 Diabetes and Cardiovascular Disease. *Clinical Pharmacology & Therapeutics*, 90(1):52–66, 2011. ISSN 0009-9236.

[141] Hardin, J.A., Hinoshita, F., and Sherr, D.H. Mechanisms by which benzo[a]pyrene, an environmental carcinogen, suppresses B cell lymphopoiesis. *Toxicology and Applied Pharmacology*, 117(2):155–164, 1992. ISSN 0041-008X.

[142] Krieger, J.A., Davila, D.R., Lytton, J., Born, J.L., and Burchiel, S.W. Inhibition of sarcoplasmic/endoplasmic reticulum calcium ATPases (SERCA) by polycyclic aromatic hydrocarbons in HPB-ALL human T cells and other tissues. *Toxicology and Applied Pharmacology*, 133(1):102–108, 1995. ISSN 0041-008X.

[143] Murugaiyan, J., Rockstroh, M., Wagner, J., Baumann, S., Schorsch, K., Trump, S., Lehmann, I., Bergen, M.v., and Tomm, J.M. Benzo[a]pyrene affects Jurkat T cells in the activated state via the antioxidant response element dependent Nrf2 pathway leading to decreased IL-2 secretion and redirecting glutamine metabolism. *Toxicology and Applied Pharmacology*, 269(3):307–316, 2013. ISSN 1096-0333.

[144] Yang, K. and Chi, H. mTOR and metabolic pathways in T cell quiescence and functional activation. *Seminars in Immunology*, 24(6):421–428, 2012. ISSN 1044-5323.

[145] Uno, S., Dalton, T.P., Shertzer, H.G., Genter, M.B., Warshawsky, D., Talaska, G., and Nebert, D.W. Benzo[a]pyrene-induced toxicity: paradoxical protection in Cyp1a1(-/-) knockout mice having increased hepatic BaP-DNA adduct levels. *Biochemical and Biophysical Research Communications*, 289(5):1049–1056, 2001. ISSN 0006-291X.

[146] Chuang, H.Y., Rassenti, L., Salcedo, M., Licon, K., Kohlmann, A., Haferlach, T., Foa, R., Ideker, T., and Kipps, T.J. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood*, 120(13):2639–2649, 2012. ISSN 0006-4971, 1528-0020.

[147] Morine, M.J., Tierney, A.C., van Ommen, B., Daniel, H., Toomey, S., Gjelstad, I.M.F., Gormley, I.C., Prez-Martinez, P., Drevon, C.A., Lpez-Miranda, J., and Roche, H.M. Transcriptomic Coordination in the Human Metabolic Network Reveals Links between n-3 Fat Intake, Adipose Tissue Gene Expression and Metabolic Health. *PLoS Computational Biology*, 7(11):e1002223, 2011. ISSN 1553-7358.

[148] Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D., and Wishart, D.S. MetaboAnalyst 2.0a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 2012. ISSN 0305-1048, 1362-4962.

[149] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research*, 39(Database issue):D712–717, 2011. ISSN 1362-4962.

[150] Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917–2918, 2011. ISSN 1367-4803, 1460-2059.

[151] Thiele, I., Price, N.D., Vo, T.D., and Palsson, B.. Candidate Metabolic Network States in Human Mitochondria IMPACT OF DIABETES, ISCHEMIA, AND DIET. *Journal of Biological Chemistry*, 280(12):11683–11695, 2005. ISSN 0021-9258, 1083-351X.

[152] Faust, K., Dupont, P., Callut, J., and van Helden, J. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, 26(9):1211–1218, 2010. ISSN 1367-4803, 1460-2059.

[153] Komurov, K., White, M.A., and Ram, P.T. Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data. *PLoS Computational Biology*, 6(8):e1000889, 2010. ISSN 1553-7358.

[154] Schneider, U., Schwenk, H.U., and Bornkamm, G. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *International Journal of Cancer. Journal International Du Cancer*, 19(5):621–626, 1977. ISSN 0020-7136.

[155] Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007. ISSN 0027-8424, 1091-6490.

[156] Schellenberger, J., Park, J.O., Conrad, T.M., and Palsson, B.. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, 2010. ISSN 1471-2105.

[157] Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, 3(1), 2007. ISSN false.

[158] Stobbe, M.D., Houten, S.M., Jansen, G.A., van Kampen, A.H.C., and Moerland, P.D. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology*, 5(1):165, 2011.

[159] Rockstroh, M., Mller, S., Jende, C., Kerzhner, A., Von Bergen, M., and Tomm, J.M. Cell fractionation - an important tool for compartment proteomics. *Journal of Integrated OMICS*, 1(1), 2011. ISSN 2182-0287.

[160] Baumann, S., Rockstroh, M., Bartel, J., Krumsiek, J., Otto, W., Jungnickel, H., Potratz, S., Luch, A., Willscher, E., Theis, F., Von Bergen, M., and Tomm, J. Subtoxic concentrations of benzo[a]pyrene induce metabolic changes and oxidative stress in non-activated and affect the mTOR pathway in activated Jurkat T cells. *Journal of Integrated OMICS*, 4(1), 2014. ISSN 2182-0287.

[161] Oberbach, A., Blher, M., Wirth, H., Till, H., Kovacs, P., Kullnick, Y., Schlichting, N., Tomm, J.M., Rolle-Kampczyk, U., Murugaiyan, J., Binder, H., Dietrich, A., and Bergen, M.v. Combined Proteomic and Metabolomic Profiling of Serum Reveals Association of the Complement System with Obesity and Identifies Novel Markers of Body Fat Mass Changes. *Journal of Proteome Research*, 10(10):4769–4788, 2011. ISSN 1535-3893, 1535-3907.

[162] Dupont, P., Callut, J., Dooms, G., Monette, J.N., and Deville, Y. Relevant subgraph extraction from random walks in a graph. *Universite catholique de Louvain, UCL/INGI, Number RR*, 7, 2006.

[163] Kemeny, J.G. and Snell, J.L. *Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix"*. Springer New York, 1983. ISBN 978-0-387-90192-3.

[164] Callut, J., Fanoisse, K., Saerens, M., and Dupont, P. Semi-supervised classification in graphs using bounded random walks. In *Proceedings of the 17th Annual Machine Learning Conference of Belgium and the Netherlands (Benelearn)*, pages 67–68. 2008.

[165] Croes, D., Couche, F., Wodak, S.J., and van Helden, J. Inferring Meaningful Pathways in Weighted Metabolic Networks. *Journal of Molecular Biology*, 356(1):222–236, 2006. ISSN 00222836.

[166] Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. ISSN 1742-5468.

[167] Tuosto, L. NF-B family of transcription factors: biochemical players of CD28 costimulation. *Immunology Letters*, 135(1-2):1–9, 2011. ISSN 1879-0542.

[168] Gebauer, F. and Hentze, M.W. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology*, 5(10):827–835, 2004. ISSN 1471-0072.

[169] Chapman, N.M. and Chi, H. mTOR links environmental signals to T cell fate decisions. *T Cell Biology*, 5:686, 2015.

[170] Calvano, S.E., Xiao, W., Richards, D.R., Felciano, R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein, B.H., Cobb, J.P., Tschoeke, S.K., Miller-Graziano, C., Moldawer, L.L., Mindrinos, M.N., Davis, R.W., Tompkins, R.G., Lowry, S.F., and Large Scale Collab. Res. Program, I.a.H.R.t.I. A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061):1032–1037, 2005. ISSN 0028-0836, 1476-4679.

[171] Macropol, K., Can, T., and Singh, A.K. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10(1):283, 2009. ISSN 1471-2105.

[172] Chen, H., Qin, J., Wei, P., Zhang, J., Li, Q., Fu, L., Li, S., Ma, C., and Cong, B. Effects of leukotriene B4 and prostaglandin E2 on the differentiation of murine Foxp3+ T regulatory cells and Th17 cells. *Prostaglandins, Leukotrienes, and Essential Fatty Acids*, 80(4):195–200, 2009. ISSN 1532-2823.

[173] Niu, N. and Qin, X. New insights into IL-7 signaling pathways during early and late T cell development. *Cellular & Molecular Immunology*, 10(3):187–189, 2013. ISSN 1672-7681.

[174] Huang, Y.H. and Sauer, K. Lipid Signaling in T-Cell Development and Function. *Cold Spring Harbor Perspectives in Biology*, 2(11):a002428, 2010. ISSN , 1943-0264.

[175] Wolterbeek, A.P.M., Roggeband, R., van Moorsel, C.J.A., Baan, R.A., Koeman, J.H., Feron, V.J., and Rutten, A.A.J.J.L. Vitamin A and -carotene influence the level of benzo[a]pyrene-induced DNA adducts and DNA-repair activities in hamster tracheal epithelium in organ culture. *Cancer Letters*, 91(2):205–214, 1995. ISSN 0304-3835.

[176] Israels, L.G., Walls, G.A., Ollmann, D.J., Friesen, E., and Israels, E.D. Vitamin K as a regulator of benzo(a)pyrene metabolism, mutagenesis, and carcinogenesis. Studies with rat microsomes and tumorigenesis in mice. *The Journal of Clinical Investigation*, 71(5):1130–1140, 1983. ISSN 0021-9738.

[177] Charalabopoulos, K., Karkabounas, S., Charalabopoulos, A.K., Papalimneou, V., Ioachim, E., and Giannakopoulos, X. Inhibition of benzo(a)pyrene-induced carcinogenesis by vitamin C alone and by vitamin C/vitamin E and selenium/glutathione. *Biological Trace Element Research*, 93(1-3):201–212, 2003. ISSN 0163-4984.

[178] Matsunawa, M., Akagi, D., Uno, S., Endo-Umeda, K., Yamada, S., Ikeda, K., and Makishima, M. Vitamin D Receptor Activation Enhances Benzo[a]pyrene Metabolism via CYP1A1 Expression in Macrophages. *Drug Metabolism and Disposition*, 40(11):2059–2066, 2012. ISSN , 1521-009X (Online).

[179] Berg, J.M., Tymoczko, J.L., and Stryer, L. *Biochemistry*. W. H. Freeman, 2010. ISBN 978-1-4292-2936-4.

[180] Shimobayashi, M. and Hall, M.N. Making new contacts: the mTOR network in metabolism and signalling crosstalk. *Nature Reviews Molecular Cell Biology*, 15(3):155–162, 2014. ISSN 1471-0072.

[181] Lisse, T.S. and Hewison, M. Vitamin D. *Cell Cycle*, 10(12):1888–1889, 2011. ISSN 1538-4101.

[182] Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 2005. ISSN 1061-4036.

[183] Lefebvre, C., Rajbhandari, P., Alvarez, M.J., Bandaru, P., Lim, W.K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B.C., Basso, K., Beltrao, P., Krogan, N., Gautier, J., Dalla-Favera, R., and Califano, A. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:377, 2010. ISSN 1744-4292.

[184] Nayak, R.R., Kearns, M., Spielman, R.S., and Cheung, V.G. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Research*, 19(11):1953–1962, 2009. ISSN 1088-9051.

[185] Doering, T., Crawford, A., Angelosanto, J., Paley, M., Ziegler, C., and Wherry, E.J. Network Analysis Reveals Centrally Connected Genes and Pathways Involved in CD8+ T Cell Exhaustion versus Memory. *Immunity*, 37(6):1130–1144, 2012. ISSN 1074-7613.

[186] He, F., Chen, H., ProbstKepper, M., Geffers, R., Eifes, S., Sol, A.d., Schughart, K., Zeng, A.P., and Balling, R. PLAU inferred from a correlation network is critical for suppressor function of regulatory T cells. *Molecular Systems Biology*, 8(1), 2012. ISSN 1744-4292, 1744-4292.

[187] Saris, C.G., Horvath, S., Vught, P.W.v., Es, M.A.v., Blauw, H.M., Fuller, T.F., Langfelder, P., DeYoung, J., Wokke, J.H., Veldink, J.H., Berg, L.H.v.d., and Ophoff, R.A. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics*, 10(1):405, 2009. ISSN 1471-2164.

[188] Inouye, M., Silander, K., Hamalainen, E., Salomaa, V., Harald, K., Jousilahti, P., Mnnist, S., Eriksson, J.G., Saarela, J., Ripatti, S., Perola, M., van Ommen, G.J.B., Taskinen, M.R., Palotie, A., Dermitzakis, E.T., and Peltonen, L. An Immune Response Network Associated with Blood Lipid Levels. *PLoS Genet*, 6(9):e1001113, 2010.

[189] Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., Kasturi, S., Carlone, G.M., Quinn,

C., Chaussabel, D., Palucka, A.K., Mulligan, M.J., Ahmed, R., Stephens, D.S., Nakaya, H.I., and Pulendran, B. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology*, 15(2):195–204, 2014. ISSN 1529-2908.

[190] Orei, M., Tang, J., Seppnen-Laakso, T., Mattila, I., Saarni, S.E., Saarni, S.I., Lnnqvist, J., Sysi-Aho, M., Hytylinen, T., Perl, J., and Suvisaari, J. Metabolome in schizophrenia and other psychotic disorders: a general population-based study. *Genome Medicine*, 3(3):19, 2011. ISSN 1756-994X.

[191] Kujala, U.M., Mkinen, V.P., Heinonen, I., Soininen, P., Kangas, A.J., Leskinen, T.H., Rahkila, P., Wrtz, P., Kovanen, V., Cheng, S., Sipil, S., Hirvensalo, M., Telama, R., Tammelin, T., Savolainen, M.J., Pouta, A., OReilly, P.F., Mntyselk, P., Viikari, J., Khnen, M., Lehtimki, T., Elliott, P., Vanhala, M.J., Raitakari, O.T., Jrvelin, M.R., Kaprio, J., Kainulainen, H., and Ala-Korpela, M. Long-term Leisure-time Physical Activity and Serum Metabolome. *Circulation*, 127(3):340–348, 2013. ISSN 0009-7322, 1524-4539.

[192] Valcarcel, B., Ebbels, T.M.D., Kangas, A.J., Soininen, P., Elliot, P., Ala-Korpela, M., Jarvelin, M.R., and de Iorio, M. Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity. *Journal of the Royal Society Interface*, 11(94), 2014. ISSN 1742-5689.

[193] Inouye, M., Kettunen, J., Soininen, P., Silander, K., Ripatti, S., Kumpula, L.S., Hmlinen, E., Jousilahti, P., Kangas, A.J., Mnnist, S., Savolainen, M.J., Jula, A., Leivisk, J., Palotie, A., Salomaa, V., Perola, M., Ala-Korpela, M., and Peltonen, L. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Molecular Systems Biology*, 6, 2010. ISSN 1744-4292.

[194] Homuth, G., Teumer, A., Vlker, U., and Nauck, M. A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling. *The Journal of endocrinology*, 215(1):17–28, 2012. ISSN 1479-6805.

[195] Petersen, A.K., Zeilinger, S., Kastenmller, G., Rmisch-Margl, W., Brugger, M., Peters, A., Meisinger, C., Strauch, K., Hengstenberg, C., and Pagel, P. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Human molecular genetics*, 23(2):534–545, 2014.

[196] Arsenault, B.J., Boekholdt, S.M., and Kastelein, J.J.P. Lipid parameters for measuring risk of cardiovascular disease. *Nature Reviews Cardiology*, 8(4):197–206, 2011. ISSN 1759-5002.

[197] Wichmann, H.E., Gieger, C., and Illig, T. KORA-gen - Resource for Population Genetics, Controls and a Broad Spectrum of Disease Phenotypes. *Das Gesundheitswesen*, 67(S 01):26–30, 2005. ISSN 0941-3790, 1439-4421.

[198] Rathmann, W., Strassburger, K., Heier, M., Holle, R., Thorand, B., Giani, G., and Meisinger, C. Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. *Diabetic Medicine*, 26(12):1212–1219, 2009. ISSN 1464-5491.

[199] Mehta, D., Heim, K., Herder, C., Carstensen, M., Eckstein, G., Schurmann, C., Homuth, G., Nauck, M., Vlker, U., and Roden, M. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *European Journal of Human Genetics*, 2012.

[200] Schramm, K., Marzi, C., Schurmann, C., Carstensen, M., Reinmaa, E., Biffar, R., Eckstein, G., Gieger, C., Grabe, H.J., Homuth, G., Kastenmller, G., Mgi, R., Metspalu, A., Mihailov, E., Peters, A., Petersmann, A., Roden, M., Strauch, K., Suhre, K., Teumer, A., Vlker, U., Vlzke, H., Wang-Sattler, R., Waldenberger, M., Meitinger, T., Illig, T., Herder, C., Grallert, H., and Prokisch, H. Mapping the Genetic Architecture of Gene Regulation in Whole Blood. *PLoS ONE*, 9(4):e93844, 2014.

[201] Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Drr, M., Endlich, K., Felix, S.B., Gieger, C., Grallert, H., Herder, C., Hoffmann, W., Homuth, G., Illig, T., Kruppa, J., Meitinger, T., Mller, C., Nauck, M., Peters, A., Rettig, R., Roden, M., Strauch, K., Vlker, U., Vlzke, H., Wahl, S., Wallaschofski, H., Wild, P.S., Zeller, T., Teumer, A., Prokisch, H., and Ziegler, A. Analyzing Illumina Gene Expression Microarray Data from Different Tissues: Methodological Aspects of Data Analysis in the MetaXpress Consortium. *PLoS ONE*, 7(12):e50938, 2012. ISSN 1932-6203.

[202] Du, P., Kibbe, W.A., and Lin, S.M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548, 2008. ISSN 1367-4803, 1460-2059.

[203] Aoki, K., Ogata, Y., and Shibata, D. Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology. *Plant and Cell Physiology*, 48(3):381–390, 2007. ISSN 0032-0781, 1471-9053.

[204] Palmer, C., Diehn, M., Alizadeh, A.A., and Brown, P.O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, 7:115, 2006. ISSN 1471-2164.

[205] Watkins, N.A., Gusnanto, A., de Bono, B., De, S., Miranda-Saavedra, D., Hardie, D.L., Angenent, W.G.J., Attwood, A.P., Ellis, P.D., Erber, W., Foad, N.S., Garner, S.F., Isacke, C.M., Jolley, J., Koch, K., Macaulay, I.C., Morley, S.L., Rendon, A., Rice, K.M., Taylor, N., Thijssen-Timmer, D.C., Tijssen, M.R., van der Schoot, C.E., Wernisch, L., Winzer, T., Dudbridge, F., Buckley, C.D., Langford, C.F., Teichmann, S., Gottgens, B., Ouwehand, W.H., and on behalf of the Bloodomics Consortium. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, 113(19):e1–e9, 2009. ISSN 0006-4971, 1528-0020.

[206] Shoemaker, J.E., Lopes, T.J., Ghosh, S., Matsuoka, Y., Kawaoka, Y., and Kitano, H. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC genomics*, 13(1):460, 2012.

[207] Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008. ISSN 02776715, 10970258.

[208] Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., and Kastenmuller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, 2014. ISSN 1367-4803, 1460-2059.

[209] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, 2000. ISSN 1061-4036.

[210] Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W.W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkt997, 2013. ISSN 0305-1048, 1362-4962.

[211] Tonon, L., Touzet, H., and Varr, J.S. TFM-Explorer: mining cis-regulatory regions in genomes. *Nucleic Acids Research*, 38(suppl 2):W286–W292, 2010. ISSN 0305-1048, 1362-4962.

[212] Defrance, M. and Touzet, H. Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics*, 7(1):396, 2006. ISSN 1471-2105.

[213] Sinclair, D., Fillman, S.G., Webster, M.J., and Weickert, C.S. Dysregulation of glucocorticoid receptor co-factors FKBP5, BAG1 and PTGES3 in prefrontal cortex in psychotic illness. *Scientific Reports*, 3, 2013.

[214] Schoneveld, O.J.L.M., Gaemers, I.C., and Lamers, W.H. Mechanisms of glucocorticoid signalling. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1680(2):114–128, 2004. ISSN 0167-4781.

[215] Wong, S., Tan, K., Carey, K.T., Fukushima, A., Tiganis, T., and Cole, T.J. Glucocorticoids stimulate hepatic and renal catecholamine inactivation by direct rapid induction of the dopamine sulfotransferase Sult1d1. *Endocrinology*, 151(1):185–194, 2010. ISSN 1945-7170.

[216] Polman, J.A.E., Hunter, R.G., Speksnijder, N., van den Oever, J.M.E., Korobko, O.B., McEwen, B.S., de Kloet, E.R., and Datson, N.A. Glucocorticoids modulate the mTOR pathway in the hippocampus: differential effects depending on stress history. *Endocrinology*, 153(9):4317–4327, 2012. ISSN 1945-7170.

[217] Schmidt, S. Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. *Blood*, 107(5):2061–2069, 2006. ISSN 0006-4971, 1528-0020.

[218] Philip, A.M., Daniel Kim, S., and Vijayan, M.M. Cortisol modulates the expression of cytokines and suppressors of cytokine signaling (SOCS) in rainbow trout hepatocytes. *Developmental and comparative immunology*, 38(2):360–367, 2012. ISSN 1879-0089.

[219] Pei, H., Yao, Y., Yang, Y., Liao, K., and Wu, J.R. Krppel-like factor KLF9 regulates PPAR transactivation at the middle stage of adipogenesis. *Cell Death & Differentiation*, 18(2):315–327, 2011. ISSN 1350-9047.

[220] Liu, Y.X., Wang, J., Guo, J., Wu, J., Lieberman, H.B., and Yin, Y. DUSP1 Is Controlled by p53 during the Cellular Response to Oxidative Stress. *Molecular Cancer Research*, 6(4):624–633, 2008. ISSN 1541-7786, 1557-3125.

[221] Sprl, F., Korge, S., Jrchott, K., Wunderskirchner, M., Schellenberg, K., Heins, S., Specht, A., Stoll, C., Klemz, R., Maier, B., Wenck, H., Schrader, A., Kunz, D., Blatt, T., and Kramer, A. Krppel-like factor 9 is a circadian transcription factor in human epidermis that controls proliferation of keratinocytes. *Proceedings of the National Academy of Sciences*, 109(27):10903–10908, 2012. ISSN 0027-8424, 1091-6490.

[222] Charmandari, E., Chrousos, G.P., Lambrou, G.I., Pavlaki, A., Koide, H., Ng, S.S.M., and Kino, T. Peripheral CLOCK Regulates Target-Tissue Glucocorticoid Receptor Transcriptional Activity in a Circadian Fashion in Man. *PLoS ONE*, 6(9):e25612, 2011.

[223] Zechner, R., Zimmermann, R., Eichmann, T.O., Kohlwein, S.D., Haemmerle, G., Lass, A., and Madeo, F. FAT SIGNALS - Lipases and Lipolysis in Lipid Metabolism and Signaling. *Cell Metabolism*, 15(3):279–291, 2012. ISSN 1550-4131.

[224] Fonseca, B.M., Costa, M.A., Almada, M., Correia-da Silva, G., and Teixeira, N.A. Endogenous cannabinoids revisited: A biochemistry perspective. *Prostaglandins & Other Lipid Mediators*, 102103:13–30, 2013. ISSN 1098-8823.

[225] Kondo, H., Hase, T., Murase, T., and Tokimitsu, I. Digestion and assimilation features of dietary DAG in the rat small intestine. *Lipids*, 38(1):25–30, 2003.

[226] Schneider, E., Leite-de Moraes, M., and Dy, M. Histamine, Immune Cells and Autoimmunity. In R.L. Thurmond, editor, *Histamine in Inflammation*, number 709 in Advances in Experimental Medicine and Biology, pages 81–94. Springer US, 2010. ISBN 978-1-4419-8055-7 978-1-4419-8056-4.

[227] Bresnick, E.H., Katsumura, K.R., Lee, H.Y., Johnson, K.D., and Perkins, A.S. Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Research*, page gks281, 2012. ISSN 0305-1048, 1362-4962.

[228] Zhang, Y., Guo, K., LeBlanc, R.E., Loh, D., Schwartz, G.J., and Yu, Y.H. Increasing Dietary Leucine Intake Reduces Diet-Induced Obesity and Improves Glucose and Cholesterol Metabolism in Mice via Multimechanisms. *Diabetes*, 56(6):1647–1654, 2007. ISSN 0012-1797, 1939-327X.

[229] Kennedy, M.A., Barrera, G.C., Nakamura, K., Baldn, n., Tarr, P., Fishbein, M.C., Frank, J., Francone, O.L., and Edwards, P.A. ABCG1 has a critical role in mediating cholesterol efflux to HDL and preventing cellular lipid accumulation. *Cell Metabolism*, 1(2):121–131, 2005. ISSN 1550-4131.

[230] Cheng, S., Rhee, E.P., Larson, M.G., Lewis, G.D., McCabe, E.L., Shen, D., Palma, M.J., Roberts, L.D., Dejam, A., Souza, A.L., Deik, A.A., Magnusson, M., Fox, C.S., O'Donnell, C.J., Vasan, R.S., Melander, O., Clish, C.B., Gerszten, R.E., and Wang, T.J. Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation*, 125(18):2222–2231, 2012. ISSN 1524-4539.

[231] Klma, M., Broukov, A., Koc, M., and Andra, L. T-cell activation triggers death receptor-6 expression in a NF-B and NF-AT dependent manner. *Molecular Immunology*, 48(1213):1439–1447, 2011. ISSN 0161-5890.

[232] Kendall, A.C. and Nicolaou, A. Bioactive lipid mediators in skin inflammation and immunity. *Progress in Lipid Research*, 52(1):141–164, 2013. ISSN 0163-7827.

[233] Alhouayek, M., Masquelier, J., and Muccioli, G.G. Controlling 2-arachidonoylglycerol metabolism as an anti-inflammatory strategy. *Drug Discovery Today*, 19(3):295–304, 2014. ISSN 1359-6446.

[234] Scortegagna, M. The HIF family member EPAS1/HIF-2 is required for normal hematopoiesis in mice. *Blood*, 102(5):1634–1640, 2003. ISSN 0006-4971, 1528-0020.

[235] Radtke, F., Fasnacht, N., and MacDonald, H.R. Notch Signaling in the Immune System. *Immunity*, 32(1):14–27, 2010. ISSN 1074-7613.

[236] Bhring, H.J., Streble, A., and Valent, P. The Basophil-Specific Ectoenzyme E-NPP3 (CD203c) as a Marker for Cell Activation and Allergy Diagnosis. *International Archives of Allergy and Immunology*, 133(4):317–329, 2004. ISSN 1423-0097, 1018-2438.

[237] Frateschi, S., Camerer, E., Crisante, G., Rieser, S., Membrez, M., Charles, R.P., Beermann, F., Stehle, J.C., Breiden, B., Sandhoff, K., Rotman, S., Haftek, M., Wilson, A., Ryser, S., Steinhoff, M., Coughlin, S.R., and Hummler, E. PAR2 absence completely rescues

inflammation and ichthyosis caused by altered CAP1/Prss8 expression in mouse skin. *Nature Communications*, page 161, 2011.

[238] Zeng, L., Liao, H., Liu, Y., Lee, T.S., Zhu, M., Wang, X., Stemerman, M.B., Zhu, Y., and Shyy, J.Y.J. Sterol-responsive element-binding protein (SREBP) 2 down-regulates ATP-binding cassette transporter A1 in vascular endothelial cells: a novel role of SREBP in regulating cholesterol metabolism. *The Journal of Biological Chemistry*, 279(47):48801–48807, 2004. ISSN 0021-9258.

[239] Tontonoz, P., Nagy, L., Alvarez, J.G., Thomazy, V.A., and Evans, R.M. PPARgamma promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell*, 93(2):241–252, 1998. ISSN 0092-8674.

[240] Cowell, I.G. E4BP4/NFIL3, a PAR-related bZIP factor with many roles. *BioEssays*, 24(11):1023–1029, 2002. ISSN 1521-1878.

[241] Everett, L., Hansen, M., and Hannenhalli, S. Regulating the regulators: modulators of transcription factor activity. *Methods in Molecular Biology (Clifton, N.J.)*, 674:297–312, 2010. ISSN 1940-6029.

[242] Tsuruoka, N., Arima, M., Arguni, E., Saito, T., Kitayama, D., Sakamoto, A., Hatano, M., and Tokuhisa, T. Bcl6 is required for the IL-4-mediated rescue of the B cells from apoptosis induced by IL-21. *Immunology Letters*, 110(2):145–151, 2007. ISSN 0165-2478.

[243] Floegel, A., Wientzek, A., Bachlechner, U., Jacobs, S., Drogan, D., Prehn, C., Adamski, J., Krumsiek, J., Schulze, M.B., Pischon, T., and Boeing, H. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *International journal of obesity (2005)*, 2014. ISSN 1476-5497.

[244] Mahaney, M.C., Blangero, J., Comuzzie, A.G., VandeBerg, J.L., Stern, M.P., and MacCluer, J.W. Plasma HDL Cholesterol, Triglycerides, and Adiposity A Quantitative Genetic Test of the Conjoint Trait Hypothesis in the San Antonio Family Heart Study. *Circulation*, 92(11):3240–3248, 1995. ISSN 0009-7322, 1524-4539.

[245] Li, S., Ogawa, W., Emi, A., Hayashi, K., Senga, Y., Nomura, K., Hara, K., Yu, D., and Kasuga, M. Role of S6K1 in regulation of SREBP1c expression in the liver. *Biochemical and biophysical research communications*, 412(2):197–202, 2011. ISSN 1090-2104.

[246] Perner, S., Rupp, N.J., Braun, M., Rubin, M.A., Moch, H., Dietel, M., Wernert, N., Jung, K., Stephan, C., and Kristiansen, G. Loss of SLC45A3 protein (prostein) expression in prostate cancer is associated with SLC45A3-ERG gene rearrangement and an unfavorable clinical course. *International Journal of Cancer. Journal International Du Cancer*, 132(4):807–812, 2013. ISSN 1097-0215.

[247] Hancock, T., Wicker, N., Takigawa, I., and Mamitsuka, H. Identifying Neighborhoods of Coordinated Gene Expression and Metabolite Profiles. *PLoS ONE*, 7(2):e31345, 2012. ISSN 1932-6203.

[248] Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B., Young, N., Xia, J., Knox, C., Dong, E., Huang, P., Hollander, Z., Pedersen, T.L., Smith, S.R., Bamforth, F., Greiner, R., McManus, B., Newman, J.W., Goodfriend, T., and Wishart, D.S. The Human Serum Metabolome. *PLoS ONE*, 6(2):e16957, 2011.

[249] Greene, M.W., Burrington, C.M., Lynch, D.T., Davenport, S.K., Johnson, A.K., Horsman, M.J., Chowdhry, S., Zhang, J., Sparks, J.D., and Tirrell, P.C. Lipid Metabolism, Oxidative Stress and Cell Death Are Regulated by PKC Delta in a Dietary Model of Nonalcoholic Steatohepatitis. *PLoS ONE*, 9(1):e85848, 2014. ISSN 1932-6203.

[250] The ENCODE Project Consortium, T.E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 0028-0836.

[251] Calkin, A.C. and Tontonoz, P. Transcriptional integration of metabolism by the nuclear sterol-activated receptors LXR and FXR. *Nature Reviews Molecular Cell Biology*, 2012. ISSN 1471-0072, 1471-0080.

[252] Yvan-Charvet, L., Wang, N., and Tall, A.R. The role of HDL, ABCA1 and ABCG1 transporters in cholesterol efflux and immune responses. *Arteriosclerosis, thrombosis, and vascular biology*, 30(2):139–143, 2010. ISSN 1079-5642.

[253] Newgard, C.B., An, J., Bain, J.R., Muehlbauer, M.J., Stevens, R.D., Lien, L.F., Haqq, A.M., Shah, S.H., Arlotto, M., Slentz, C.A., Rochon, J., Gallup, D., Ilkayeva, O., Wenner, B.R., Yancy Jr., W.S., Eisenson, H., Musante, G., Surwit, R.S., Millington, D.S., Butler, M.D., and Svetkey, L.P. A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cell Metabolism*, 9(4):311–326, 2009. ISSN 1550-4131.

[254] Wang, T.J., Larson, M.G., Vasan, R.S., Cheng, S., Rhee, E.P., McCabe, E., Lewis, G.D., Fox, C.S., Jacques, P.F., Fernandez, C., O'Donnell, C.J., Carr, S.A., Mootha, V.K., Florez, J.C., Souza, A., Melander, O., Clish, C.B., and Gerszten, R.E. Metabolite profiles and the risk of developing diabetes. *Nature medicine*, 17(4):448–453, 2011. ISSN 1546-170X.

[255] O'Connell, T.M. The Complex Role of Branched Chain Amino Acids in Diabetes and Cancer. *Metabolites*, 3(4):931–945, 2013.

[256] Xiao, X., Moreno-Moral, A., Rotival, M., Bottolo, L., and Petretto, E. Multi-tissue Analysis of Co-expression Networks by Higher-Order Generalized Singular Value Decomposition Identifies Functionally Coherent Transcriptional Modules. *PLoS Genet*, 10(1):e1004006, 2014.

[257] Gao, S., Roberts, H.K., and Wang, X. Cross Tissue Trait-Pathway Network Reveals the Importance of Oxidative Stress and Inflammation Pathways in Obesity-Induced Diabetes in Mouse. *PLoS ONE*, 7(9):e44544, 2012.

[258] Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M.L., Carlson, S., Allan, M.F., Pomp, D., and Schadt, E.E. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*, 10(5):R55, 2009.

[259] DeFronzo, R.A. and Abdul-Ghani, M. Assessment and treatment of cardiovascular risk in prediabetes: impaired glucose tolerance and impaired fasting glucose. *The American Journal of Cardiology*, 108(3 Suppl):3B–24B, 2011. ISSN 1879-1913.

[260] Johnston, D.E. Special considerations in interpreting liver function tests. *American Family Physician*, 59(8):2223–2230, 1999. ISSN 0002-838X.

[261] Floegel, A., Stefan, N., Yu, Z., Muhlenbruch, K., Drogan, D., Joost, H.G., Fritsche, A., Haring, H.U., Hrabe de Angelis, M., Peters, A., Roden, M., Prehn, C., Wang-Sattler, R., Illig, T., Schulze, M.B., Adamski, J., Boeing, H., and Pischon, T. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. *Diabetes*, 62(2):639–648, 2012. ISSN 0012-1797, 1939-327X.

[262] Shah, S.H., Kraus, W.E., and Newgard, C.B. Metabolomic Profiling for the Identification of Novel Biomarkers and Mechanisms Related to Common Cardiovascular Diseases Form and Function. *Circulation*, 126(9):1110–1120, 2012. ISSN 0009-7322, 1524-4539.

[263] Do, K.T., Kastenmller, G., Mook-Kanamori, D.O., Yousri, N.A., Theis, F.J., Suhre, K., and Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *Journal of Proteome Research*, 14(2):1183–1194, 2015. ISSN 1535-3907.

[264] Neschen, S., Scheerer, M., Seelig, A., Huypens, P., Schultheiss, J., Wu, M., Wurst, W., Rathkolb, B., Suhre, K., Wolf, E., Beckers, J., and Angelis, M.H.d. Metformin Supports the Antidiabetic Effect of a Sodium Glucose Cotransporter 2 Inhibitor by Suppressing Endogenous Glucose Production in Diabetic Mice. *Diabetes*, 64(1):284–290, 2015. ISSN 0012-1797, 1939-327X.

[265] Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M., and Milgram, E. 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. *Analytical Chemistry*, 81(16):6656–6667, 2009. ISSN 0003-2700, 1520-6882.

[266] Suhre, K., Meisinger, C., Dring, A., Altmaier, E., Belcredi, P., Gieger, C., Chang, D., Milburn, M.V., Gall, W.E., Weinberger, K.M., Mewes, H.W., Hrab de Angelis, M., Wichmann,

H.E., Kronenberg, F., Adamski, J., and Illig, T. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS ONE*, 5(11):e13953, 2010.

[267] Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3), 2011.

[268] Leverve, X.M. Inter-organ substrate exchanges in the critically ill. *Current Opinion in Clinical Nutrition and Metabolic Care*, 4(2):137–142, 2001. ISSN 1363-1950.

[269] Dietz, M.L., Bernaciak, T.M., Vendetti, F., Kielec, J.M., and Hildebrand, J.D. Differential actin-dependent localization modulates the evolutionarily conserved activity of Shroom family proteins. *The Journal of Biological Chemistry*, 281(29):20542–20554, 2006. ISSN 0021-9258.

[270] Zeisel, S.H. and da Costa, K.A. Choline: An Essential Nutrient for Public Health. *Nutrition reviews*, 67(11):615–623, 2009. ISSN 0029-6643.

[271] Gatti, R., Palo, C.B.D., Spinella, P., and Palo, P.E.F.D. Free carnitine and acetyl carnitine plasma levels and their relationship with body muscular mass in athletes. *Amino Acids*, 14(4):361–369, 1998. ISSN 0939-4451, 1438-2199.

[272] Laeger, T., Metges, C.C., and Kuhla, B. Role of beta-hydroxybutyric acid in the central regulation of energy balance. *Appetite*, 54(3):450–455, 2010. ISSN 1095-8304.

[273] McGarry, J.D., Robles-Valdes, C., and Foster, D.W. Role of carnitine in hepatic ketogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 72(11):4385–4388, 1975. ISSN 0027-8424.

[274] Brass, E.P. and Hoppel, C.L. Carnitine metabolism in the fasting rat. *The Journal of Biological Chemistry*, 253(8):2688–2693, 1978. ISSN 0021-9258.

[275] Kikuchi, T., Orita, Y., Ando, A., Mikami, H., Fujii, M., Okada, A., and Abe, H. Liquid-chromatographic determination of guanidino compounds in plasma and erythrocyte of normal persons and uremic patients. *Clinical Chemistry*, 27(11):1899–1902, 1981. ISSN 0009-9147.

[276] Deng, C., Shang, C., Hu, Y., and Zhang, X. Rapid diagnosis of phenylketonuria and other aminoacidemias by quantitative analysis of amino acids in neonatal blood spots by gas chromatography-mass spectrometry. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, 775(1):115–120, 2002. ISSN 1570-0232.

[277] Fonteh, A.N., Harrington, R.J., Tsai, A., Liao, P., and Harrington, M.G. Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. *Amino Acids*, 32(2):213–224, 2007. ISSN 0939-4451.

[278] Cano, N.J.M., Fouque, D., and Leverve, X.M. Application of Branched-Chain Amino Acids in Human Pathological States: Renal Failure. *The Journal of Nutrition*, 136(1):299S–307S, 2006. ISSN 0022-3166, 1541-6100.

[279] Bertolini, S., Donati, C., Elicio, N., Daga, A., Cuzzolaro, S., Marcenaro, A., Saturnino, M., and Balestreri, R. Lipoprotein changes induced by pantethine in hyperlipoproteinemic patients: adults and children. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, 24(11):630–637, 1986. ISSN 0174-4879.

[280] Gaddi, A., Descovich, G.C., Noseda, G., Fragiacomo, C., Colombo, L., Craveri, A., Montanari, G., and Sirtori, C.R. Controlled evaluation of pantethine, a natural hypolipidemic compound, in patients with different forms of hyperlipoproteinemia. *Atherosclerosis*, 50(1):73–83, 1984. ISSN 0021-9150.

[281] Carrara, P., Matturri, L., Galbussera, M., Lovati, M.R., Franceschini, G., and Sirtori, C.R. Pantethine reduces plasma cholesterol and the severity of arterial lesions in experimental hypercholesterolemic rabbits. *Atherosclerosis*, 53(3):255–264, 1984.

[282] Barton-Wright, E.C. and Elliott, W.A. The pantothenic acid metabolism of rheumatoid arthritis. *The Lancet*, 282(7313):862–863, 1963.

[283] Calcium pantothenate in arthritic conditions. A report from the General Practitioner Research Group. *The Practitioner*, 224(1340):208–211, 1980. ISSN 0032-6518.

[284] Stretch, C., Eastman, T., Mandal, R., Eisner, R., Wishart, D.S., Mourtzakis, M., Prado, C.M.M., Damaraju, S., Ball, R.O., Greiner, R., and Baracos, V.E. Prediction of skeletal muscle and fat mass in patients with advanced cancer using a metabolomic approach. *The Journal of Nutrition*, 142(1):14–21, 2012. ISSN 1541-6100.

[285] Rainesalo, S., Kernen, T., Palmio, J., Peltola, J., Oja, S.S., and Saransaari, P. Plasma and cerebrospinal fluid amino acids in epileptic patients. *Neurochemical Research*, 29(1):319–324, 2004. ISSN 0364-3190.

[286] Koeberl, D.D., Young, S.P., Gregersen, N.S., Vockley, J., Smith, W.E., Benjamin, D.K., An, Y., Weavil, S.D., Chaing, S.H., Bali, D., McDonald, M.T., Kishnani, P.S., Chen, Y.T., and Millington, D.S. Rare disorders of metabolism with elevated butyryl- and isobutyryl-carnitine detected by tandem mass spectrometry newborn screening. *Pediatric Research*, 54(2):219–223, 2003. ISSN 0031-3998.

[287] Bene, J., Komlsi, K., Gasztonyi, B., Juhsz, M., Tulassay, Z., and Melegh, B. Plasma carnitine ester profile in adult celiac disease patients maintained on long-term gluten free diet. *World journal of gastroenterology: WJG*, 11(42):6671–6675, 2005. ISSN 1007-9327.

[288] Tavazzi, B., Lazzarino, G., Leone, P., Amorini, A.M., Bellia, F., Janson, C.G., Di Pietro, V., Ceccarelli, L., Donzelli, S., Francis, J.S., and Giardina, B. Simultaneous high performance liquid chromatographic separation of purines, pyrimidines, N-acetylated amino

acids, and dicarboxylic acids for the chemical diagnosis of inborn errors of metabolism. *Clinical Biochemistry*, 38(11):997–1008, 2005. ISSN 0009-9120.

[289] Niwa, T., Takeda, N., and Yoshizumi, H. RNA metabolism in uremic patients: Accumulation of modified ribonucleosides in uremic serum. *Kidney International*, 53(6):1801–1806, 1998. ISSN 0085-2538.

[290] Motyl, T., Traczyk, Z., Cieluk, S., Daniewska-Michalska, D., Kukulska, W., Kauzny, Z., Podgurniak, M., Orzechowski, A., and Debski, B. Blood plasma pseudouridine in patients with malignant proliferative diseases. *European Journal of Clinical Chemistry and Clinical Biochemistry: Journal of the Forum of European Clinical Chemistry Societies*, 31(11):765–771, 1993. ISSN 0939-4974.

[291] Dzrik, R., Spustov, V., and Janekov, K. The prevalence of insulin resistance in kidney disease patients before the development of renal failure. *Nephron*, 69(3):281–285, 1995. ISSN 1660-8151.

[292] Dunn, W.B., Broadhurst, D.I., Deepak, S.M., Buch, M.H., McDowell, G., Spasic, I., Ellis, D.I., Brooks, N., Kell, D.B., and Neyses, L. Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. *Metabolomics*, 3(4):413–426, 2007. ISSN 1573-3882, 1573-3890.

[293] Rhee, E.P., Ho, J.E., Chen, M.H., Shen, D., Cheng, S., Larson, M.G., Ghorbani, A., Shi, X., Helenius, I.T., ODonnell, C.J., Souza, A.L., Deik, A., Pierce, K.A., Bullock, K., Walford, G.A., Vasan, R.S., Florez, J.C., Clish, C., Yeh, J.R.J., Wang, T.J., and Gerszten, R.E. A Genome-Wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell metabolism*, 18(1):130–143, 2013. ISSN 1550-4131.

[294] Ogawa, Y., Machida, N., Jahana, M., Gakiya, M., Chinen, Y., Oda, M., Morozumi, M., and Sugaya, K. Major factors modulating the serum oxalic acid level in hemodialysis patients. *Frontiers in Bioscience: A Journal and Virtual Library*, 9:2901–2908, 2004. ISSN 1093-9946.

[295] Januszewski, A.S., Jenkins, A.J., Baynes, J.W., and Thorpe, S.R. Lipid-derived modifications of plasma proteins in experimental and human diabetes. *Annals of the New York Academy of Sciences*, 1043:404–412, 2005. ISSN 0077-8923.

[296] Ganti, S., Taylor, S.L., Abu Aboud, O., Yang, J., Evans, C., Osier, M.V., Alexander, D.C., Kim, K., and Weiss, R.H. Kidney Tumor Biomarkers Revealed by Simultaneous Multiple Matrix Metabolomics Analysis. *Cancer Research*, 72(14):3471–3479, 2012. ISSN 0008-5472, 1538-7445.

[297] Menni, C., Fauman, E., Erte, I., Perry, J.R., Kastenmller, G., Shin, S.Y., Petersen, A.K., Hyde, C., Psatha, M., Ward, K.J., Yuan, W., Milburn, M., Palmer, C.N., Frayling, T.M., Trimmer, J., Bell, J.T., Gieger, C., Mohney, R.P., Brosnan, M.J., Suhre, K., Soranzo, N.,

and Spector, T.D. Biomarkers for Type 2 Diabetes and Impaired Fasting Glucose Using a Nontargeted Metabolomics Approach. *Diabetes*, 62(12):4270–4276, 2013. ISSN 0012-1797.

[298] Giesbertz, P., Padberg, I., Rein, D., Ecker, J., Hfle, A.S., Spanier, B., and Daniel, H. Metabolite profiling in plasma and tissues of ob/ob and db/db mice identifies novel markers of obesity and type 2 diabetes. *Diabetologia*, pages 1–11, 2015. ISSN 0012-186X, 1432-0428.

[299] Schaefer, A., Neschen, S., Kahle, M., Sarioglu, H., Gaisbauer, T., Imhof, A., Adamski, J., Hauck, S.M., and Ueffing, M. The Epoxyeicosatrienoic Acid Pathway Enhances Hepatic Insulin Signaling and Is Repressed in Insulin-Resistant Mouse Liver. *Molecular & cellular proteomics: MCP*, 2015. ISSN 1535-9484.

[300] Figueras, M., Olivan, M., Busquets, S., Lpez-Soriano, F.J., and Argils, J.M. Effects of Eicosapentaenoic Acid (EPA) Treatment on Insulin Sensitivity in an Animal Model of Diabetes: Improvement of the Inflammatory Status. *Obesity*, 19(2):362–369, 2011. ISSN 1930-739X.

[301] Genuth, S.M. and Hoppel, C.L. Plasma and Urine Carnitine in Diabetic Ketosis. *Diabetes*, 28(12):1083–1087, 1979. ISSN 0012-1797, 1939-327X.

[302] Palo, E.D., Gatti, R., Sicolo, N., Padovan, D., Vettor, R., and Federspil, G. Plasma and urine free L-Carnitine in human diabetes mellitus. *Acta diabetologia latina*, 18(1):91–95, 1981. ISSN 0001-5563, 1432-5233.

[303] Seidel, A., Brunner, S., Seidel, P., Fritz, G.I., and Herbarth, O. Modified nucleosides: an accurate tumour marker for clinical diagnosis of cancer, early detection and therapy control. *British Journal of Cancer*, 94(11):1726–1733, 2006. ISSN 0007-0920.

[304] King, C.D., Rios, G.R., Green, M.D., and Tephly, T.R. UDP-glucuronosyltransferases. *Current Drug Metabolism*, 1(2):143–161, 2000. ISSN 1389-2002.

[305] Slyshenkov, V.S., Dymkowska, D., and Wojtczak, L. Pantothenic acid and pantothenol increase biosynthesis of glutathione by boosting cell energetics. *FEBS letters*, 569(1-3):169–172, 2004. ISSN 0014-5793.

[306] Li, Y., Chang, Y., Zhang, L., Feng, Q., Liu, Z., Zhang, Y., Zuo, J., Meng, Y., and Fang, F. High glucose upregulates pantothenate kinase 4 (PanK4) and thus affects M2-type pyruvate kinase (Pkm2). *Molecular and Cellular Biochemistry*, 277(1-2):117–125, 2005. ISSN 0300-8177, 1573-4919.

[307] Reibel, D.K., Wyse, B.W., Berkich, D.A., Palko, W.M., and Neely, J.R. Effects of diabetes and fasting on pantothenic acid metabolism in rats. *The American Journal of Physiology*, 240(6):E597–601, 1981. ISSN 0002-9513.

[308] Stipanuk, M.H. and Caudill, M.A. *Biochemical, Physiological, and Molecular Aspects of Human Nutrition*. Elsevier Health Sciences, 2013. ISBN 978-0-323-26695-6.

[309] Zhou, H. and Rigoutsos, I. The emerging roles of GPRC5A in diseases. *Oncoscience*, 1(12):765–776, 2014. ISSN 2331-4737.

[310] Soni, A., Amisten, S., Rorsman, P., and Salehi, A. GPRC5B a putative glutamate-receptor candidate is negative modulator of insulin secretion. *Biochemical and Biophysical Research Communications*, 2013. ISSN 1090-2104.

[311] Berg, J.M., Stryer, L., and Tymoczko, J.L. *Stryer Biochemie*. Springer-Verlag, 2015. ISBN 978-3-8274-2989-6.

[312] Kobayashi, K., Forte, T.M., Taniguchi, S., Ishida, B.Y., Oka, K., and Chan, L. The db/db mouse, a model for diabetic dyslipidemia: molecular characterization and effects of Western diet feeding. *Metabolism: Clinical and Experimental*, 49(1):22–31, 2000. ISSN 0026-0495.

[313] Belke, D.D. and Severson, D.L. Diabetes in mice with monogenic obesity: the db/db mouse and its use in the study of cardiac consequences. *Methods in Molecular Biology (Clifton, N.J.)*, 933:47–57, 2012. ISSN 1940-6029.

[314] Regenmortel, M.H.V. Reductionism and complexity in molecular biology. *EMBO Reports*, 5(11):1016–1020, 2004. ISSN 1469-221X.

[315] Gao, A., Liu, B., Shi, X., Jia, X., Ye, M., Jiao, S., You, B., and Huang, C. Phosphatidylinositol-3 kinase/Akt/p70S6K/AP-1 signaling pathway mediated benzo(a)pyrene-induced cell cycle alternation via cell cycle regulatory proteins in human embryo lung fibroblasts. *Toxicology Letters*, 170(1):30–41, 2007. ISSN 0378-4274.

[316] Suhre, K. and Gieger, C. Genetic variation in metabolic phenotypes: study designs and applications. *Nature reviews. Genetics*, 2012. ISSN 1471-0064.

[317] Houseman, E., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012. ISSN 1471-2105.

[318] Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H.F. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS ONE*, 4(7):e6098, 2009. ISSN 1932-6203.

[319] Eckel-Mahan, K.L., Patel, V.R., Mohney, R.P., Vignola, K.S., Baldi, P., and Sassone-Corsi, P. Coordination of the transcriptome and metabolome by the circadian clock. *Proceedings of the National Academy of Sciences*, 109(14):5541–5546, 2012. ISSN 0027-8424, 1091-6490.

[320] Selmaoui, B. and Touitou, Y. Reproducibility of the circadian rhythms of serum cortisol and melatonin in healthy subjects: a study of three different 24-h cycles over six weeks. *Life Sciences*, 73(26):3339–3349, 2003. ISSN 0024-3205.