HELMHOLTZ
ASSOCIATION

Alliance on Systems Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Network-based analysis of multi-fluid metabolomics data

Kieu Trinh Do

June 2018

TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Network-based analysis of multi-fluid metabolomics data

Kieu Trinh Do

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:

Prof. Dr. D. Frishman

Prüfer der Dissertation:

1. Prof. Dr. Dr. F. J. Theis
2. Prof. Dr. J. Baumbach

Die Dissertation wurde am 08.06.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 12.06.2019 angenommen.

Danksagung

Ich möchte mich bei einigen Menschen bedanken, ohne die diese Arbeit kaum möglich gewesen wäre.

Allen voran bedanke ich mich bei meinem Doktorvater, Prof. Dr. Dr. Fabian Theis, dass ich bei ihm zunächst meine Masterarbeit, und anschließend meine Doktorarbeit machen durfte. Ich danke Fabian dafür, dass ich Teil eines tollen Instituts mit einer wunderbaren Arbeitsatmosphäre sein durfte. Ich bedanke mich auch für seine Unterstützung sowohl in fachlichen, als auch persönlichen Belangen. Ich weiß, dass das nicht selbstverständlich ist.

Auch bedanken möchte ich mich bei Prof. Dr. Jan Baumbach für die Zweitbegutachtung meiner Dissertation, sowie Prof. Dr. Frishman für den Vorsitz der Prüfungskommission.

Ich möchte mich von ganzem Herzen bei Dr. Jan Krumsiek für seine fantastische Betreuung bedanken. Ich habe nicht nur unendlich viel fachliches Wissen von ihm gelernt, sondern durfte durch ihn auch erleben wie es ist, eine vertrauensvolle und entspannte Arbeitsatmosphäre zu haben, in der man sich sowohl persönlich entfalten kann, als auch jederzeit benötigte Hilfestellung bekommt.

Des Weiteren möchte ich mich bei Dr. Gabi Kastenmüller und Prof. Dr. Karsten Suhre bedanken. Ihr seid nicht nur großartige Kollaborationspartner, sondern auch geduldige und liebe Betreuer für mich gewesen.

Ein großes Dankeschön geht an das tolle SysDiab-Team: Mustafa, Parviz, Michl, Alida, Helena, und Elisa. Danke für eure Freundschaft und eure schlauen Köpfe. Insbesondere möchte ich Elisa danken; du bist eine Kollegin und eine Freundin ohnesgleichen, und der Held meines Arbeitsalltags!

Ganz besonders möchte ich mich bei meiner gesamten Familie bedanken. Meinen Großeltern, Tanten und Onkeln in Vietnam, die immer an mich geglaubt haben. Alles was ich bin, habe ich meinen Eltern, Kieu Hanh und Tien Toan, zu verdanken. Ihr habt mich mit Liebe zu der Person erzogen, die ich heute bin. Und natürlich danke ich euch für euer uneingeschränktes Vertrauen und Unterstützung seit 29 Jahren. Besonderer Dank gilt meiner kleinen Schwester, Kieu Mi, für ihre Liebe und ihre Unterstützung. Ich habe von dir – trotz deines Status als “kleine” Schwester – einiges gelernt und bin stolz auf die Frau, zu der du dich entwickelt hast.

Ich danke auch meiner Schwiegerfamilie, die mich in ihr Herz und ihre Familie aufgenommen haben. Das sind meine Schwiegereltern, Brigitte und Jürgen, Oma Trudi, und natürlich Geli und Lexi. Ich danke euch dafür, dass ihr mich mit eurem Stolz und eurer Zuneigung stets motiviert habt. Und für die Baby-Sitter Dienste in den letzten zwei Jahren!

Besonderer Dank gilt natürlich auch euch, meinen lieben Freunden. Ihr habt mich über Jahre hinweg motiviert, unterstützt und habt immer an mich geglaubt.

Zum Schluss möchte ich den beiden wichtigsten Menschen in meinem Leben danken, mein wundervoller Mann und mein kleiner Wurm. Ihr seid mein größtes Glück und die größte Motivation für alles, was ich je erreichen möchte. Stefan, ich bin unendlich dankbar für deine grenzenlose Liebe, dein bedingungsloses Vertrauen, dein Stolz und deine ungefragte, uneingeschränkte Unterstützung. Du hast alles gemacht und alles gegeben, damit diese Arbeit ein Erfolg werden konnte. Und Lucas, du bist das Beste in meinem Leben und meine größte Freude.

Abstract

Metabolomics, the study of metabolite profiles at a global level, is an established tool to gain insights into pathophysiological outcomes. In contrast to other *omics* data, metabolomics measurements are informative on the overall status of the organism irrespective of where they have been measured since metabolites are rather small molecules and diffuse quickly throughout the body. In particular, due to their non-invasiveness, measurements in easily accessible samples such as blood and urine have great clinical advantages.

Previous metabolomics studies were mostly limited to only blood, however, recent technical advances have been allowing for the generation of complex, multi-fluid data. Therefore, a recent trend is the simultaneous analysis of metabolomics data in multiple body fluids to unravel cross-fluid processes and their links to clinical endpoints. To exploit the full potential of this complex data, appropriate statistical approaches are needed. The aim of this cumulative doctoral thesis was to contribute to (i) improved data preprocessing, (ii) more powerful data analysis, and (iii) facilitated data interpretation of metabolomics studies. To this end, each of the aforementioned aspects was addressed in three main studies.

Untargeted MS-based metabolomics data typically contains a tremendous amount of missing values. We explored the missing values patterns in real data and exploited the gained insights to evaluate missing value handling strategies to finally provide concrete handling recommendations.

In our second project, we proposed a network-based approach for the analysis of complex multi-fluid data, aiming to statistically reconstruct biochemical processes within and between different bio fluids. In addition, we introduced an approach for embedding associations of metabolites with clinical outcomes into the reconstructed multi-fluid metabolic network. The generic character and the potential of our study were demonstrated in four follow-up applications.

Finally, we aimed to facilitate interpretation of complex associations of metabolism with clinical outcomes, which can either cover only few metabolite associations ("sparse" effects) or a large number of associations ("dense" effects). Our idea was to explore phenotype associations in form of functional modules at different layers of resolution, from single metabolites to entire pathways, to facilitate interpretation for both sparse and dense cases. To this end, we developed a greedy algorithm based on data-driven networks to systematically identify phenotype-driven modules. We demonstrated that our approach enables to extract biologically relevant insights that could not have been identified with classical association analysis. Following the current important trend to make novel generic algorithms available to the scientific community, we implemented our approach as an open-source and easy-to-use R package.

In summary, we extensively evaluated existing and developed novel methods for improved metabolomics data preprocessing, more powerful data analysis, and easier data interpretation. The work in this doctoral thesis will substantially help future metabolomics studies to unravel the potential of complex multi-fluid data.

Zusammenfassung

Metabolomics, die Erforschung metabolischer Profile auf globalem Level, hat sich als ein bewährtes Hilfsmittel etabliert, um Einblicke in die patho-physiologischen Ausprägungen eines Organismus zu erlangen. Im Gegensatz zu anderen *omics*-Daten sind metabolische Messungen von jeder Art Probe informativ, da Metaboliten klein sind und daher schnell in den gesamten Körper diffundieren. Das hat insbesondere klinische Vorteile, da Metabolomics-Messungen aus leicht zugänglichen Körperflüssigkeiten wie Blut und Urin viel Auskunft über den Organismus geben.
Die meisten veröffentlichten Humanstudien im Metabolomics-Bereich basieren auf Messungen innerhalb nur einer Körperflüssigkeit, üblicherweise Blut. Ein schnell um sich greifender Trend, der durch jüngste technische Fortschritte ermöglicht wird, ist die Messung und Analyse von sogenannten *multi-fluid*-Daten – Messungen aus unterschiedlichen Körperflüssigkeiten derselben Person. Diese Daten erlauben, nicht nur lokale, sondern auch organübergreifende metabolische Prozesse sowie deren Verknüpfungen zu klinischen Endpunkten zu untersuchen. Um das volle Potenzial dieser komplexen *multi-fluid*-Daten auszuschöpfen, ist die Entwicklung neuer, geeigneter statistischer Ansätze von Nöten. Das Ziel dieser kumulativen Dissertation war es in diesem Zusammenhang, geeignete Methoden zum Preprocessing, zur Datenanalyse und zur Dateninterpretation von Metabolomics-Studien beizusteuern. In drei Hauptprojekten wurde jeweils einer dieser Aspekte behandelt.

Untargeted MS-basierte Metabolomics-Daten beinhalten typischerweise eine enorme Anzahl an *missing values* – fehlende Einträge in der Datenmatrix. Wir analysierten das Auftreten und die Systematik von *missing values* in echten Daten, um die daraus abgeleiteten Erkenntnisse in die Evaluierung von möglichen Handhabungsstrategien einfließen zu lassen. Mit dieser Studie konnte erstmalig eine ausführliche Beschreibung von *missing values*-Eigenschaften und -Empfehlungen für die beste Umgangsstrategie geliefert werden.
Der Fokus des zweiten Hauptprojekts lag auf der netzwerkbasierten Analyse von komplexen *multi-fluid*-Daten, um metabolische Prozesse innerhalb und zwischen unterschiedlichen Körperflüssigkeiten statistisch zu rekonstruieren. Zusätzlich haben wir einen Ansatz zur Analyse phänotypischer Outcomes im Zusammenhang mit den rekonstruierten metabolischen Prozessen vorgestellt. Der generische Charakter und das Potenzial unserer Methoden wurde erfolgreich in weiteren Follow-up-Studien unter Beweis gestellt.
Schließlich zielten wir in einer dritten Studie darauf ab, die Interpretation von komplexen Assoziationen zwischen phänotypischem Outcome und hochdimensionalen Daten zu erleichtern. Metabolische Einheiten bilden typischerweise funktionale Module, die mit dem Phänotypen assoziiert sind. Diese Assoziationen können in der Netzwerkdarstellung sowohl

in zerstreuter (*sparse effects*) als auch dichter Form (*dense effects*) auftreten. Um beide Fälle angemessen interpretieren zu können, bestand unser Ansatz darin, funktionale Module auf unterschiedlichem Auflösungs-Level (von Metaboliten-Level über Pathway-Level) zu untersuchen. Während *sparse* Assoziationen auf Metaboliten-Level gut interpretierbar sind, sind *dense* Assoziationen auf Pathway-Level übersichtlicher und einfacher auszuwerten. Wir entwickelten eine netzwerkbasierte Methode, die systematisch funktionale Module auf unterschiedlichen Levels sucht. Entsprechend dem gegenwärtigen, wichtigen Trend, neue generische Algorithmen für die wissenschaftliche Gemeinschaft zugänglich zu machen, haben wir unseren Ansatz als kostenfreies R package implementiert.

Zusammenfassend ist festzuhalten, dass wir einerseits bereits existierende Methoden umfassend evaluiert und andererseits neue Ansätze entwickelt haben, die im Metabolomics-Bereich massiv dazu beitragen werden, ein adäquates Datenpreprocessing und eine statistisch starke Datenanalyse sicherzustellen. Durch anschauliche Visualisierungsmethoden und innovative Betrachtungsweisen wird die Dateninterpretation erheblich vereinfacht. Die Erkenntnisse, die im Rahmen dieser Dissertation erarbeitet wurden, werden dazu beitragen, in künftigen Metabolomics-Studien das volle Potenzial von komplexen *multi-fluid*-Daten zu erkennen und auszuschöpfen.

List of contributed articles

Publications in the context of my doctoral studies

All results of my doctoral thesis have been published in peer-reviewed journals or are currently in publication process. These publications are listed below in chronological order.

(i) **Do KT**, Kastenmüller G, Mook-Kanamori DO, Yousri NA, Theis FJ, Suhre K, J Krumsiek (2015), **Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva.** *J Proteome Res, 14: 1183-1194*

(ii) Yousri NA, Mook-Kanamori DO, Selim MME-D, Takiddin AH, Al-Homsi H, Al-Mahmoud KAS, Karoly ED, Krumsiek J, **Do KT**, Neumaier U, Mook-Kanamori MJ, Rowe J, Chidiac OM, McKeon C, Al Muftah WA, Kader SA, Kastenmüller G, Suhre K (2015), **A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control.** *Diabetologia 58: 1855-1867*

(iii) Schulte EC, Altmaier E, Berger HS, **Do KT**, Kastenmüller G, Wahl S, Adamski J, Peters A, Krumsiek J, Suhre K, Haslinger B, Ceballos-Baumann A, Gieger C, Winkelmann J (2016), **Alterations in Lipid and Inositol Metabolisms in Two Dopaminergic Disorders.** *PloS One 11: e0147129*

(iv) Knacke H, Pietzner M, **Do KT**, Römisch-Margl W, Kastenmüller G, Völker U, Völzke H, Krumsiek J, Artati A, Wallaschofski H, Nauck M, Suhre K, Adamski J, Friedrich N (2016), **Metabolic fingerprints of circulating IGF-I and the IGF-I/IGFBP-3 ratio: a multi-fluid metabolomics study.** *J. Clin. Endocrinol. Metab.: jc20162588*

(v) **Do KT**, Pietzner M, Rasp D, Friedrich N, Nauck M, Kocher T, Suhre K, Mook-Kanamori DO, Kastenmüller G, Krumsiek J (2017), **Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations.** *NPJ Syst Biol Appl. 2017;3:28*

(vi) **Do KT**, Rasp D, Kastenmüller G, Suhre K, Krumsiek J (2017), **MoDentify: a tool for phenotype-driven identification of modules in high-throughput data.** *Under review* in *Oxford Bioinformatics*

(vii) **Do KT***, Wahl S*, Raffler J, Molnos S, Laimighofer M, Adamski J, Suhre K, Strauch K, Peters A, Gieger C, Langenberg C, Stewart I, Theis FJ, Grallert H, Kastenmüller G, Krumsiek J (2017), **Characterization of missingness in untargeted MS-based metabolomics data and evaluation of missing data handling strategies.** *Under review* in *Metabolomics*

Further publications

I was also involved in further projects and collaborations during the course of my time as PhD student, resulting in the two publications listed below.

(i) Krumsiek J, Mittelstrass K, **Do KT**, Stückler F, Ried J, Adamski J, Peters A, Illig T, Kronenberg F, Friedrich N, Nauck M, Pietzner M, Mook-Kanamori DO, Suhre K, Gieger C, Grallert H, Theis FJ, Kastenmüller G (2016), **Gender-specific pathway differences in the human serum metabolome.** *Metabolomics 11: 1815–18337*

(ii) Piontek U, Wallaschofski H, Kastenmüller G, Suhre K, Völzke H, **Do KT**, Artati A, Nauck M, Adamski J, Friedrich N, Pietzner M (2017) **Sex-specific metabolic profiles of androgens and its main binding protein SHBG in a middle aged population without diabetes.** Sci. Rep. 7: 2235

Table of Contents

Chapter 1

Introduction

1.1 Metabolomics in the Central Dogma of Molecular Biology

The *Central Dogma of Molecular Biology* was introduced for the first time by Francis Crick in 1958, stating that genetic information in form of the DNA is transcribed into transportable messenger RNA, and finally translated into proteins. By now, further molecular layers and their interactions (Figure 1.1) have been identified and are investigated with rapidly evolving *omics* technologies [1]. One of the fields in the *omics* era is *metabolomics*, the identification and quantification of intermediates and products of metabolism (*metabolites*) in a given biological sample. The *metabolome* is the set of all small molecules present in cells, tissues, organs or biological fluids such as carbohydrates, amino acids, or lipids [2]. Metabolites can be produced by the host organism, but are also obtainable from microorganisms or exogeneous sources such as diet [3]. There are several differences when comparing the metabolome to the more classical "genetic *omes*" that cover compounds directly arising from the DNA sequence such as RNAs or proteins:

First, metabolites can be considered as the only molecular species that diffuses rapidly and is systematically transported through all organs and tissues [5]. Therefore, metabolomics measurements throughout the body are expected to provide a dynamic reflection of tissue and organ metabolism and their interactions. Epidemiological studies and their clinical applications are preferably performed in easily accessible fluids such as blood to ensure minimal invasiveness in sampling [6], but also saliva and urine have been proposed as promising resources for non-invasive diagnostics [7–9]. Metabolomics data can be biologically conclusive in all of these fluids, for instance, with respect to understanding cellular processes [10] or identification of disease signatures [9, 11]. In contrast, a comprehensive global picture of other molecular pools such as the transcriptome or proteome can only be obtained from specific tissue samples such as hepatic specimens, which are usually challenging to obtain from living humans.

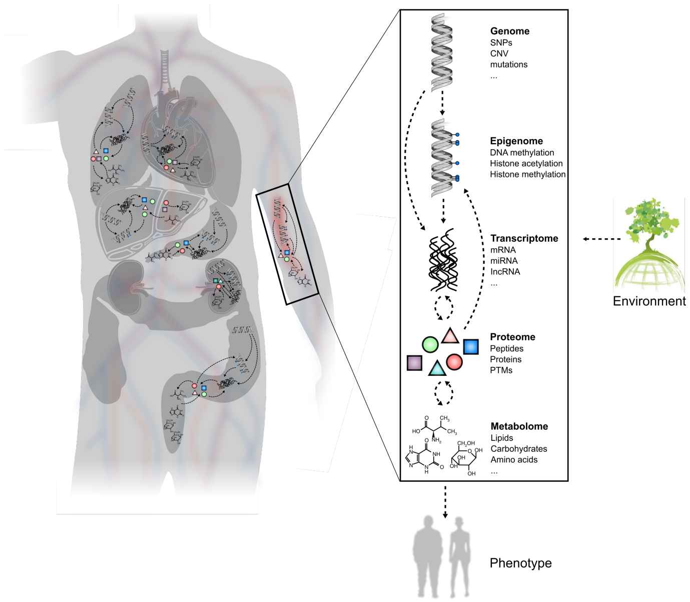


Figure 1.1: Extended *Central Dogma of Molecular Biology* (Central Dogma box adapted from [4]). The interplay between different molecular layers, which are captured by various *omics* technologies, occurs body-wide, in all tissues, organs, and fluids. Although the Central Dogma reflects the directional information flow from the genome to the metabolome, downstream layers can also influence upstream molecular species. Here, only the most obvious and dominant relationships between the *omics* layers are shown, but the information flow is much more complex and more links are possible. As depicted in this simplified scheme, the metabolome can be seen as the endpoint of preceding *omes* and therefore, serves as a convenient readout that is closely linked to physiological phenotypes.

Second, the metabolome is highly dynamic and rapidly changing. Compared to genetic *omes*, environmental influences such as diet, drugs, or physical activity can change the metabolic pool within minutes [12]. The fast response of metabolism to internal and external factors can be attributed to the fact that the metabolic pool is based on rapid chemical reactions instead of slow regulatory processes. That is, compared to other *omes*, which indicate that a change in the organism's phenotype *may* occur, changes in metabolite concentrations produce *directly observable* changes, which is needed for understanding of cellular functions in a living system [13].

Finally, it is often assumed that metabolomics data is a particularly convenient molecular readout of a clinical endpoint, since it is considered to represent the downstream biochemical end products of the preceding *omes*, their interactions as well as influences from environmental factors [2, 14–16]. Metabolites are the smallest molecular building blocks of the organism and final result of regulatory processes. They perform the actual biochemical functions, while genes or proteins provide the "blueprint" for these functions [17]. Therefore, metabolomics can be seen as the chemical phenotype of an organism and is frequently used to analyze patterns in pathophysiological states [18, 19], to identify potential biomarkers, and to predict disease incidence and progression [20].

1.2 Metabolomics workflow

A typical scheme of a metabolomics study comprises metabolomics measurement, data quality control and preprocessing, statistical analysis covering both univariate and multivariate analyses, and biological interpretation (Figure 1.2).

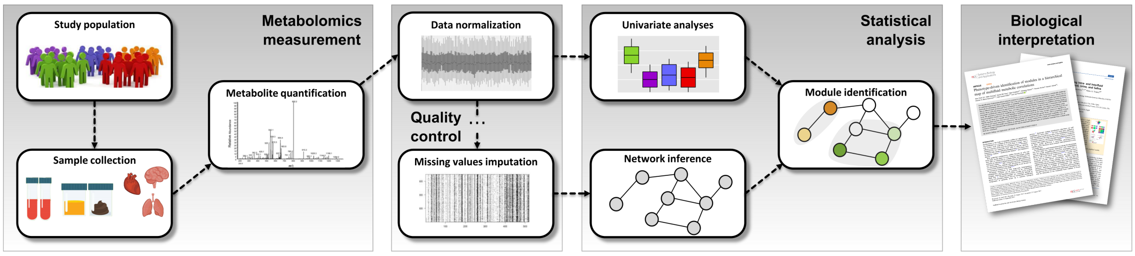


Figure 1.2: Typical workflow of a metabolomics study. After identification and quantification of metabolites from certain samples of a study population, the data quality is first surveyed. Necessary preprocessing steps, such as normalization and missing values imputation, are performed. Subsequently, depending on the research question, univariate or multivariate statistical approaches are applied. In our studies, one major multivariate approach is network inference with subsequent relation to clinical outcomes. Finally, the obtained results are biologically interpreted.

1.2.1 Metabolomics measurement technology

In the past decade, technical improvements have accelerated the identification and quantification of metabolites. One of the main approaches for metabolomics measurement is mass-spectrometry (MS), which consists of three essential steps: The molecule is first ionized by adding or removing an electron, resulting in a negatively or positively charged compound, respectively. The ions are then separated according to their mass and charge in a mass analyzer by a magnetic field. Finally, a detector captures and quantifies the separated ions, thereby producing a unique fingerprint of the original molecule as a spectrum of mass-to-charge ratios (m/z) and a relative intensity of the measured compound. Due to the substantial complexity of biological samples, gas- (GC) or liquid chromatography (LC) is often performed prior to MS. In chromatography, a flow-through column is used to separate molecules according to their different retention times (amount of time a molecule needs to pass the column), which is dependent on chemical and physical properties of each compound. Both the retention time and the m/z ratio are finally used to identify and quantify the metabolites [21].

Metabolomics measurements can be performed either in a targeted or untargeted manner [2]. In targeted approaches only a limited number of already known, structurally characterized and biochemically annotated metabolites are captured. The measurements are sensitive, robust and absolute quantifications can be obtained. In contrast, untargeted approaches offer the discovery of novel compounds as the measurement is not limited to pre-defined signals. However, the annotation and quantification of metabolites is much more challenging than for targeted measurements. For instance, in many untargeted metabolomics data sets, for nearly half of the quantifiable compounds the chemical identity is not known. Although these metabolites have been reported to associate with various diseases [22], their clinical usability remains limited.

1.2.2 Data quality control

The quality of metabolomics data can be influenced by several factors including sample type, sample preparation, batch effects or runday-based instrument sensitivity. Therefore, preprocessing steps are necessary to ensure high data quality and to prevent false positives or false negatives in hypothesis testing. Typical preprocessing steps are data normalization, data transformation, outlier handling, missing values handling, and data scaling (Figure 1.3).

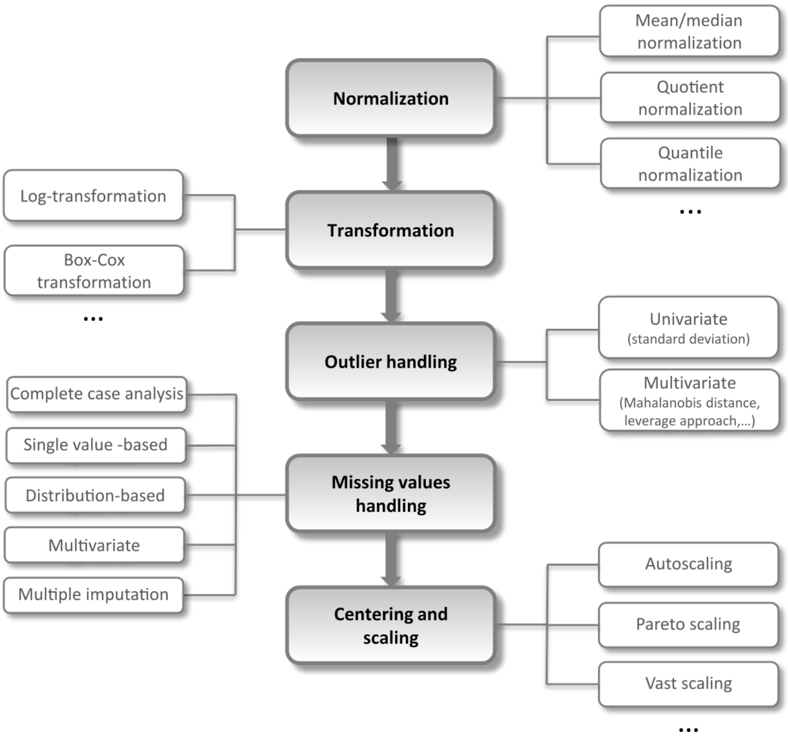


Figure 1.3: Typical preprocessing steps for metabolomics studies. The data is normalized to remove unwanted biological or technical variation, for instance, resulting from batch effects. Log-transformation is applied since metabolite concentrations are assumed to follow a log-normal distribution. Subsequently, univariate (point) and multivariate (sample) outliers as well as missing values are handled. Finally, the data is scaled to force the variables to the same unit.

Normalization

Variation in metabolomics data can occur either due to technical or biological sources. Whether biological variation is desirable depends on the research question and the study design. For instance, when analyzing metabolomics data from multiple body fluids simultaneously, it might be reasonable to normalize for osmolality, since fluids such as urine or saliva are substantially influenced by water intake and diet. Batch effects, which are technical sources of variation, are usually undesired. For instance, measurements can vary due to different sample acquisition, preparation, or storage. Moreover, in large-scale metabolomics studies, the measurements are spread across multiple days (rundays), which can be even several months apart. Differences between these rundays such as in machine performances or environmental conditions in the laboratory like temperature also introduce variation into the measurements, which have no biological cause. There are several normalization strategies to remove unwanted variation in the data, including mean or median normalization, quotient normalization, or normalization by an external factor such as urinary creatinine levels.

Transformation

Data transformation is performed to convert multiplicative to additive relations [23]. This correction for heteroscedasticity also makes the data more symmetric such that the data distribution resembles a normal distribution, a requirement for many statistical approaches. A common choice in metabolomics is log-transformation, since metabolites are often assumed to follow a log-normal distribution [24]. Another option is power transformation, in particular the one-dimensional Box-Cox transformation where an exponent (λ) is estimated, usually by maximum likelihood estimation, to transform the data Y to Y^{λ} [25]. Although Box-Cox transformations might be more flexible and, therefore, more efficient to make the data normally distributed than log-transformations, one major drawback is that each variable can be transformed by a different λ. Thus, their different scales limit result interpretation, for instance, if ratios of variables are investigated.

Outlier handling

Outliers can be single data points that lie outside of an expected value range (univariate outliers) or whole samples that show unexpected deviations from the data variability (multivariate outliers). Univariate outliers can be identified based on data location and scatter: assuming a normal distribution of the investigated variable, a value is defined as an outlier if it is more than x standard deviations from the mean. In the metabolomics field, x is commonly set to three or four according to experience. Often, univariate outliers are excluded for univariate analysis by omitting samples with an outlier for each variable. However, for multivariate analysis this would lead to substantial loss of information such that univariate outliers are preferably kept or replaced by a reasonable value such as the mean or median metabolite concentration. Multivariate outliers can be detected using metrics assessing the distance of a sample from all the other variables' distributions such as the Mahalanobis distance or the leverage approach [26]. Samples identified as multivariate outliers are usually excluded for further downstream analysis.

Scaling

Finally, in untargeted metabolomics data the variables are not on the same scale, since only the relative concentrations of the metabolites are captured. This can distort multivariate statistics relying on dimensionally homogeneous data such as principal component analysis. Therefore, scaling should be performed to force all variables to the same unit. For this purpose, most often the data is expressed in standard deviations (z-scores) with mean equal to 0 and variance equals 1.

1.2.3 Missing values in MS-based metabolomics data

Beyond the above described preprocessing steps, a crucial aspect with often underestimated impact on downstream statistical analysis is the handling of missing values. The existence of missing values is an inevitable property of high-throughput measurements due to either systematic or random loss of data. Broadly speaking, missing values are gaps in the data matrix where a value is not available as a result of technical challenges or real biological absence of the metabolite in the respective sample (Figure 1.4). Technical reasons include instrument sensitivity thresholds (limit of detection, LOD), computational challenges in spectral processing, or matrix effects through co-eluting compounds [27]. Biological absence is typically observed for exogenous compounds such as drug metabolites, which are only measured in patients taking the medication. Also compounds only obtainable from nutrition are only detected in people with the respective diet. Since the reason why a certain value is missing is challenging or even unfeasible to determine, missing values are often statistically characterized as completely at random (MCAR), at random (MAR), or not at random (MNAR). In the MCAR case, the probability of missing values does not depend on the observed nor the unobserved measurements, i.e., the probability to be missing is constant for all units. In MAR, the assumption is that the pattern of missingness depends on the observed data, i.e., the missingness would be random if all conditional variables were corrected for. For example, two chemically closely related metabolites could have very similar retention times when measured with chromatography based platforms (see section 1.2.1). This would result in their peaks being hardly separable, and therefore only one of the two metabolites would appear as measured in each sample. Missingness in these metabolites would be classified as MAR. In contrast, MNAR describes the occurrence of a missing value dependent on the unobserved measurements, e.g., a limit of detection.

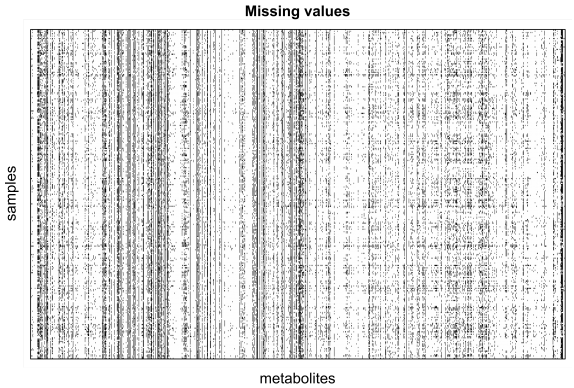


Figure 1.4: Data matrix with missing values in untargeted metabolomics. Data matrices resulting from untargeted metabolomics measurements contain many missing values (black), which can either occur systematically within a metabolite or a sample, or randomly spread across the matrix. Excluding whole samples or whole metabolites with missing values would lead to considerable loss of information. Therefore, an alternative missing values handling strategy is imputation, the replacement of missing entries by reasonable values.

In untargeted metabolomics, typically nearly one third of the data points is missing (see Figure 1.4) [28–30], which can be handled by two main strategies: in complete case analysis (CCA), all observations with missing values are excluded from the variable of interest. For multivariate approaches this procedure would cause substantial loss of information. For this reason, an alternative strategy, known as imputation, is to replace missing entries by a reasonable substitution value. The definition of such a "reasonable" value differs for each imputation approach, depending on their assumptions on the nature of missing values. For instance, some methods assume low concentrations for missing entries due to an LOD such as minimum imputation, which is widely used in the metabolomics field [31–33]. Here, missing entries are replaced by the minimal value observed for the respective metabolite. Other methods such as mean or median replacement assume random missing values [9, 12, 30, 34, 35] and replace them by the mean or median of the observed concentration values, respectively. Although more sophisticated, multivariate missing value handling strategies exist, including multiple imputation by chained equations (MICE) [36], imputation based on k-nearest neighbors [37], imputation based on random forests [30], or missing value invariant PCA [38] and diffusion maps [39], in practice most often simple substitution methods are applied. However, these simple approaches can distort statistical analysis and mislead biological interpretation if the underlying assumptions are not correct for the given data. Examples are illustrated in Figure 1.5.

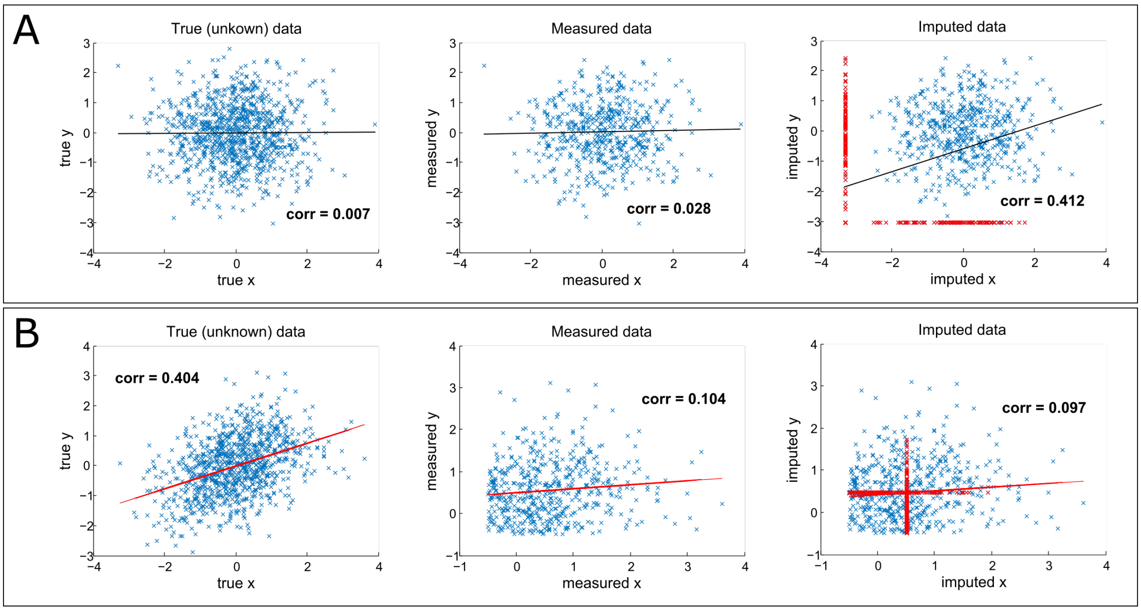


Figure 1.5: Bias induced by oversimplified (but commonly used) imputation on statistical analysis and interpretation. Missing values imputation with wrong assumptions on the missingness pattern can substantially affect the statistical analysis and biological interpretation of the data. In the simulated scenario **A**, two metabolites are not correlated and both contain random missing values. Applying minimum imputation, as widely used in the metabolomics field, would lead to a false positive correlation between the two metabolites. In scenario **B**, two metabolites are truly correlating and have missing values due to a limit of detection. Considering only the measured data, the correlation would be considerably diminished. Mean imputation would even produce a false negative result.

In the first scenario (Figure 1.5A), two originally uncorrelated metabolites with random missing values falsely become strongly correlated after minimum imputation. In a second scenario (Figure 1.5B), two truly correlated metabolites have very low concentrations that are below the sensitivity threshold of the machine and thus, can not be measured (missing values due to LOD). The correlation estimated only based on observed values is already substantially decreased, but after performing mean imputation the two metabolites are not correlated anymore. These two examples demonstrated that if an imputation approach is applied with incorrect assumptions on the real nature of missing values, false positive or false negative statistics resulting in wrong biological interpretation can be a consequence.

The effects of imputation on downstream analysis of metabolomics data have been evaluated by only few previous studies so far. All of these studies observed that overall missing values treatment had substantial effects on the outcome and interpretation of statistical analysis and that simple substitution approaches such as minimum and mean imputation consistently showed poor performances [24, 29, 30, 40, 41]. Most of the studies unanimously showed that imputation based on k-nearest neighbors was the optimal imputation approach according to their evaluation schemes [24, 29, 41, 42].

Despite these comprehensive studies, two main aspects have been insufficiently addressed. First, in all studies, evaluation of imputation methods was based on artificial data. Although simulation studies have the advantage that the true underlying values are known, and thus, evaluation can be performed in a completely controlled setting, the challenge is to generate artificial data resembling real data. In all previous studies, data simulation was based on assumptions of *only* random or *only* LOD-based missing values, without confirming the validity of these assumptions for the measured data. To create artificial data that captures the essential properties of the real data, a detailed description of missing values patterns is required. Second, previous studies only focused on the *statistical* evaluation, while the evaluation of imputation methods by *biological* validity of the results has not yet been addressed. That is, besides the reconstruction of statistical estimates between simulated variables, insights can also be gained by evaluating whether real data retains its overall biological conclusiveness after imputation (e.g., in metabolomics, if biochemical pathways can still be reconstructed).

We tackled these two challenges in a study, where for the first time, an extensive investigation of missing values patterns in real MS-based metabolomics data was performed, followed by both statistical and biological evaluation of different imputation methods (**Research question I.** in section 1.4). Briefly, we used the insights gained from the real data to create appropriate artificial data that reflects the real patterns. In this controlled setting, we evaluated the reconstruction of statistical estimates for 31 imputation approaches. We addressed biological validity of the imputed data by assessing the effects of imputation on the reconstruction of biochemical pathways, and the association of genetic variants with metabolite levels.

1.3 Metabolomics in biomedical research

Metabolomics is frequently used to identify patterns associated with pathophysiological states, since the metabolome is considered to represent an interplay of the preceding *omes* and environmental factors as depicted in the extended "Central Dogma of Molecular Biology" (Figure 1.1). It provides a readout that is closely linked to the phenotype of interest, therefore, metabolomics has a wide range of applications in many research areas, including nutrition [43], biomarker identification [5, 6], exploration of pathogenesis [5, 20, 44, 45], and drug discovery [5, 46, 47]. One of the most active areas centers around metabolomics biomarker discovery since classical diagnostic modalities for widespread diseases such as angiography for coronary artery disease (CAD) or biopsy analysis for cancer are costly and can be rather invasive [6]. In contrast, early disease diagnosis based on metabolomics data from easily accessible fluids would lower invasiveness in diagnostic testing and financial expenses.

In the past decade, many potential metabolic biomarkers for various health conditions have been identified. For instance, several metabolites have been found to be strongly associated with insulin resistance [48–51] and can even be predictive of type 2 diabetes (T2D) [52–57]. Metabolomics was also extensively researched with respect to coronary artery diseases (CAD), where some compounds were found to be associated with or have predictive power for CAD [58–62]. In cancer research, metabolomics is often used for case-control comparisons, but other fields such as patient prognosis, therapy control and tumor classification are also investigated [63]. Biomarker identification in the metabolomics area has been performed for other pathophysiological outcomes as well, including asthma [64–66], multiple sclerosis [67], or restless legs syndrom and parkinson disease [68].

Although links between metabolism and many clinical endpoints have been investigated extensively in these studies, two aspects need further attention. First, metabolomics studies have mainly been performed in blood, but metabolites can also be measured in other human fluids and tissues or even in cell cultures. But the most promosing resources for diagnostics are urine and saliva [7–9], since these fluids can be obtained in an even less invasive way than blood. Second, besides *identifying* disease markers for early diagnostics, *understanding* the functional mechanisms of these biomarkers and why they are linked to a certain phenotype is essential for disease therapy. For instance, BCAAs have been reported to be associated with insulin resistance in many studies [48, 50, 51]. Newgard *et al.* contributed to the understanding of physiologic mechanisms of insulin resistance by reporting that the excess of BCAAs, which partly accounts for impaired efficiency of fatty acid oxidation in insulin-resistant status, may derive from the gut microbiome [69]. Two bacterial species in the human gut produce BCAAs from organic precursors and partly induce dysregulation of metabolic pathways and eventually the development of T2D [69, 70]. This example illustrates that for a global and comprehensive view of disease pathogenesis, a systematic embedding of potential biomarkers into metabolic processes is required. In particular, these metabolic processes should also cover the crosstalk between different body compartments.

We aimed to address these two aspects in our studies, for which reason a more detailed description is provided below (see sections 1.3.1 and 1.3.2).

1.3.1 Multi-fluid metabolomics

Although the vast majority of metabolomics studies have been performed in only one bio fluid, due to recent technical advances enabling cheaper metabolomics measurements with higher sensitivity, a quickly spreading trend is to measure metabolite concentrations in multiple body fluids of the same individuals. This trend is reflected in the increasing number of studies covering metabolomic analysis of multiple fluids (Figure 1.6). For instance, based on plasma and urine metabolomics measurements, metabolite signatures for pubertal development [71], glycemic state [72], insulin resistance [48], type 1 diabetes [73], HIV/AIDS [74], and drug effects [75] have been identified. Plasma, urine, and saliva metabolomics samples have been used to analyze the effect of dietary components on human metabolomics profiles [76]. In another study, the effects of the polyphenol resveratrol, a dietary supplement assumed to act like an antioxidant, has been explored in metabolism of blood, urine, fat, and muscle tissue in men [77].

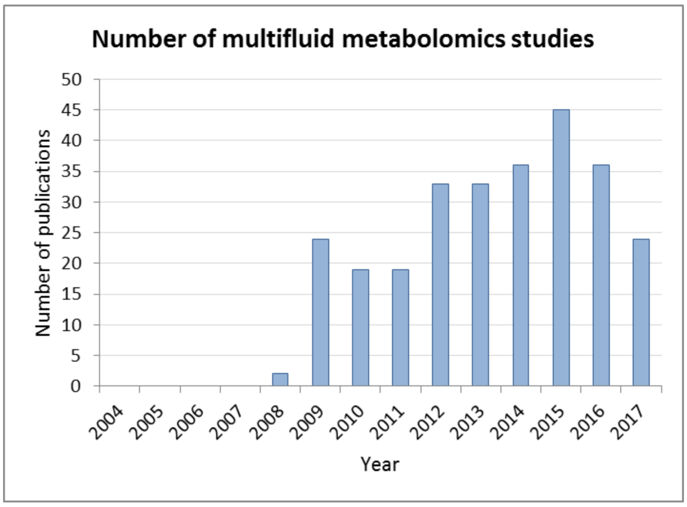


Figure 1.6: Number of multi-fluid (blood and urine) metabolomics studies (as of 12th January 2018). These numbers were obtained from a PubMed search for publications, which have the terms "blood", "urine", "human", and "metabolomics" in the title and abstract. Metabolomics studies have long been performed on only blood or urine. However, recent technical advancements have facilitated the measurement and simultaneous analysis of multiple body fluids, as illustrated by the increasing number of studies covering both plasma and urine (or more) metabolomics measurements from the same individuals.

Despite great achievements in this field, the full potential of multi-fluid data to unravel biological processes has not yet been exploited. For instance, these studies investigated the data in each fluid separately and, subsequently, compared the results. Although metabolomics data is highly informative in all of the investigated fluids or tissues, each body compartment has its own physiological function and can therefore, contain fluid- or tissue- specific information. Moreover, small molecules such as metabolites are easily exchangeable between different compartments. Thus, the metabolic crosstalk between different fluids or tissues is required for a fundamental understanding of the dynamics of whole-body metabolism. One step towards this objective was already reported in a study by Adourian *et al.*, who analyzed correlations between plasma analytes and liver molecules in rats [78]. However, a systematic investigation of metabolic interactions between different human body fluids has not yet been reported (**Research question II.** in section 1.4).

1.3.2 Phenotype-driven metabolic networks

Associations between the metabolome and phenotypes such as clinical variables or disease status are expected to pass beyond single metabolites and span an entire set of compounds. Typically, metabolites are assigned to "sets" (termed "pathways") based on their structural or biochemical properties (metabolite-centric such as "branched-chain amino acids") or their participation in metabolic processes (process-centric such as "glycolysis"). Pathway annotations can be obtained from various public databases including KEGG [79], HMDB [80], MetaCyc [81], or Recon [82]. A common strategy to explore pathways is to perform pathway enrichment analysis, which evaluates whether the metabolites associated with a certain phenotype are represented in a pathway more than expected by chance. A considerable limitation of this analysis is that phenotype associations are explored within pre-defined groups of metabolites. Thus, the results are strongly dependent on the quality and type of pathway annotation.

In an alternative unbiased strategy, the aim is to identify sets of metabolites that are co-regulated or driven by the respective phenotype, referred to as functional modules, without relying on pre-defined metabolite pathways. Module identification approaches are well established for *omics* data and are mainly based on biological networks [83–88]. In these networks, a node corresponds to a molecule (e.g., protein, metabolite, transcript, etc.) and edges between two nodes represent the correlation between the two respective compounds. Network-based module identification methods aim to detect connected parts in the underlying network that are associated with the given phenotype more than when the single components are considered alone.

Metabolomics networks are considered as a model of metabolic processes in a given tissue of a given organism, where nodes represent metabolites and edges correspond to interactions between two metabolites. Most prominent publicly available models of metabolism to date are stored in KEGG and Recon. However, public databases are not complete and the metabolite nomenclature is inconsistent. Therefore, mappings of the measured metabolites in the data at hand to database entries can be substantially challenging and the obtained network is not easy to handle. Data-driven networks can circumvent mapping difficulties since they are statistically estimated directly from the given data. Often, Pearson correlation networks are used. However, for high-dimensional data these networks are difficult to interpret due to their density induced by indirect correlations. Therefore, Krumsiek *et al.* proposed Gaussian graphical models (GGM), which are based on partial correlations [89]. Since partial correlations represent conditional dependencies in multivariate Gaussian distributions, the GGM only exhibits direct correlations and, hence, facilitates interpretation due to its sparsity. Based on cross-sectional metabolomics data, Krumsiek *et al.* showed that GGMs indeed reflect real metabolic pathways [89] and are a valuable tool to reconstruct metabolism in a data-driven manner.

Based on metabolomics networks, several studies have proposed approaches to identify phenotype-related modules [33, 64, 90–94]. However, there are two shortcomings inherent to these studies: First, modules were identified visually after the phenotype association of each single metabolite has been mapped onto the network by coloring the nodes according to the strength of the association. Thus, the selection of modules in these studies was arbitrary, mostly based on manual evaluation. A more systematic approach was performed by Krumsiek *et al.*, who used the network to group the underlying metabolomics data to highly correlated clusters, which were then explored with respect to the phenotype [95]. However, similar to pathway enrichment analysis, this approach also relies on the definition of "modules" (= clusters) before investigating phenotype associations. An *automatic* method for the *systematic* identification of modules *driven* by the phenotype has not yet been reported in the metabolomics field. Second, none of the published module identification approaches consider that phenotype associations can occur at different scales, ranging from global associations spanning entire pathways or sets of pathways (e.g., "dense" associations between metabolomics and gender or BMI), to localized associations with only a few metabolites (e.g., “sparse” associations between metabolomics and insulin-like growth factor I or asthma). In particular for dense phenotype associations, modules at metabolite level are challenging to interpret due to their overwhelming number. In contrast, modules at a hierarchically super-ordinate level such as a pathway network, where the plethora of information at the fine-grained level is condensed, would notably facilitate interpretation.

Taken together, an approach for the phenotype-driven and systematic module identification approach operating at different layers of resolution is still missing in the metabolomics field. We addressed this need with **Research question III.** (section 1.4).

1.4 Research questions

For a fundamental understanding of metabolism and its link to certain clinical endpoints, metabolic processes within and between different body compartments need to be explored. Due to increasing technical capabilities and decreasing costs, measurement of multiple fluids from the same individual has become feasible. This complex data needs appropriate quality control and preprocessing methods, and in particular, in relation with a phenotype, requires statistical approaches enabling unbiased and comprehensive data analysis and interpretation. Within the scope of this cumulative doctoral thesis the following research questions were addressed to contribute to three major steps of any typical metabolomics study (see Figure 1.2): data quality control, statistical analysis, and biological interpretation.

I. *Which missingness patterns can be observed in real metabolomics data and how can these missing values be handled appropriately?*

II. *How can metabolic processes within and between different body fluids be investigated, and how can we use these processes to analyze pathophysiological outcomes?*

III. *How can interpretable functional modules be identified in a systematic and phenotype-driven manner?*

1.5 Overview of this thesis

This cumulative doctoral thesis consists of four main chapters. **Chapter 1** is an introductory chapter on metabolomics, achievements in this field, the needs remained for appropriate handling and fundamental understanding of the data, and our research questions to address these needs. In **Chapter 2**, the available data, all computational and statistical methods applied or developed in our studies are described. **Chapter 3** gives a summary of the first-author studies published within the scope of this doctoral thesis. Finally, in **Chapter 4**, the strengths and weaknesses of our studies and possible future perspectives are discussed.

Chapter 2

Materials and methods

In this chapter, we briefly describe the data cohorts and the statistical approaches used in the first-author publications. In the first section, three large-scale metabolomics data sets are portrayed. In section two, we describe how missing values patterns in untargeted MS-based metabolomics data were investigated, and how artificial data was generated based on the gained insights. Moreover, we introduce 31 imputation approaches, which we grouped into four categories that are outlined in the second section. Next, we describe two approaches to analyze metabolic pathways with respect to a given phenotype. Network inference, visualization, and analysis are explained in section four. And finally, we developed a network-based algorithm for phenotype-driven identification of functional modules, which is illustrated in the last section.

2.1 Metabolomics cohorts

German Cooperative Health Research in the Region of Augsburg (KORA) F4

The KORA F4 study is a regional population-based cohort established in 1996 [96]. Metabolomics measurements using ultra-high performance liquid-phase chromatography (UHPLC) and gas-chromatography (GC) separation coupled with tandem mass spectrometry (MS/MS) in both positive and negative modes were performed by Metabolon, Inc. for fasting serum samples of 910 females and 858 males, aged between 25 and 74 years. Samples were divided in 53 rundays, with 34 samples on average per runday. In total, 516 metabolites were quantified, of which 303 have known chemical structures. Each metabolite was annotated with one of 68 sub-pathways (see section 2.3) representing biochemical subclasses or metabolic pathways (e.g., Branched-chain amino acid), and one of 8 super-pathways representing more global metabolite classes (e.g., Amino acid).

Qatar Metabolomics Study on Diabetes (QMDiab)

QMDiab is a cross-sectional case-control study that was conducted in 2012 at the Dermatology Department of Hamad Medical Corporation in Doha, Qatar. The cohort comprises 180 females and 189 males of Arab and Asian ethnicity aged between 17 and 81 years. 188 and 181 individuals were classified as type 2 diabetes patients and non-diabetic controls, respectively. Untargeted metabolomics measurements were performed for non-fasting plasma, urine, and saliva samples on 11 rundays using UHPLC-MS/MS and GC/MS by Metabolon, Inc. In addition, targeted metabolomics measurements were performed for plasma by Biocrates Life Sciences AG (MS-based) and for urine samples by Chenomx, Inc. (nuclear magnetic resonance based). For our studies, untargeted measurements from Metabolon, Inc. were used. A total of 2,251 metabolites, of which 1,563 represent unique compounds (Figure 2.1) were measured. 762 compounds have a known chemical structure. Each known metabolite was assigned to one out of 85 sub-pathways and to one out of 8 super-pathways.

Study of Health in Pomerania (SHIP-TREND)

SHIP-TREND is a population-based study conducted between 2008 and 2011 in West Pomerania, Germany, comprising 4,420 participants between 20 and 81 years. Untargeted metabolomics measurements for fasting plasma, urine, and saliva samples of 561 females and 439 males were performed by the Genome Analysis Center, Helmholtz Zentrum Munich, Germany on the platform UHPLC-MS/MS developed by Metabolon, Inc. A total of 1,665 metabolites across all fluids, of which 1,191 represented unique compounds were determined (Figure 2.1). Each metabolite was assigned to one out of 73 sub-pathways and to one out of 8 super-pathways.

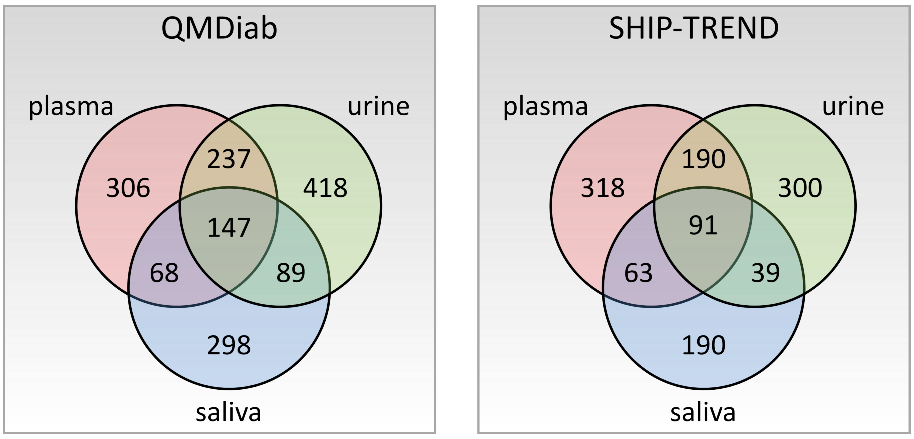


Figure 2.1: Number of metabolites measured in different fluids in the QMDiab and SHIP-TREND studies. Both cohorts comprise metabolomics measurements of plasma, urine, and saliva samples from the same individuals. Only a small fraction of the identified metabolites was measured in all three fluids, and the vast majority was only detected in one of the three fluids.

2.2 Missing values

In this section, we give a brief overview of the statistical approaches performed in our study [97], but detailed descriptions can be found in the Methods section and in the supplementary material of our publication.

2.2.1 Derivation of missingness mechanisms from real data

For this section, let $X = \{x_{i,j}\}$ be an $\mathbb{R}^{n \times p}$ matrix of metabolite concentrations with samples $i, i = 1, ..., n$ and metabolites $j, j = 1, ..., p$, where n is the total number of observations and p is the total number of metabolites. In many studies, the number of samples to be measured exceeds the number of samples that can be measured on one day by far. Hence, the measurements are spread across multiple "rundays". Each sample was measured on a runday $d, d = 1, ..., l$, and I_d is the set of samples measured on runday d (Figure 2.2). The data can contain missing values, which are gaps in the data matrix, where the measurement value is not available (denoted as NA).

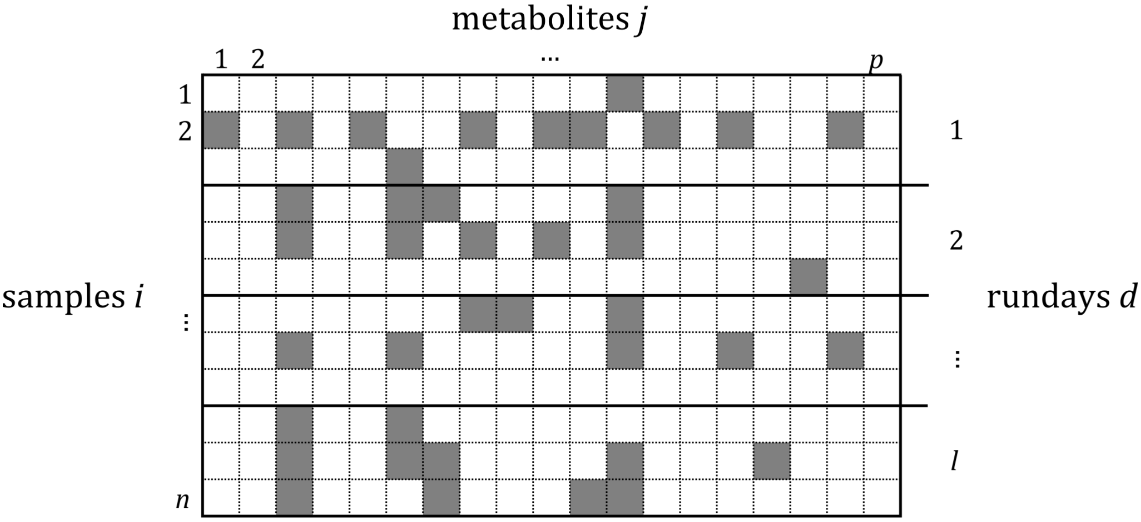


Figure 2.2: Schematic metabolomics data matrix with missing values and rundays. Typically, metabolomics data is represented as an $\mathbb{R}^{n \times p}$ matrix $X = \{x_{i,j}\}$ with the samples $i = 1, ..., n$ in rows and the metabolites $j = 1, ..., p$ in columns. In large cohorts, the measurements are divided into different rundays. Therefore, in the data matrix, each row is additionally assigned to one respective runday d. Black cells correspond to missing values, which are gaps in the data matrix where a measurement was not available (denoted as NA).

Limit of detection

It is commonly assumed that missing values in metabolomics data correspond to very low concentrations of the respective compound, which could not be measured due to a limit of detection (LOD). We investigated this assumption for each metabolite with more than 10% and less than 70% missing values in the KORA F4 data set using Pearson correlation analysis. For a given metabolite j, an *auxiliary* metabolite j_{aux} defined as the metabolite with the highest Pearson correlation coefficient above $r = 0.3$ to j was determined. A Wilcoxon-Mann-Whitney test was applied to compare the concentrations of j_{aux} in samples with missing values and samples with observed values in j. The missingness pattern of j was assumed to follow an LOD tendency, if the p-value of the Wilcoxon-Mann-Whitney test was significant with $\alpha = 0.05$ after Bonferroni correction for multiple testing.

Runday effects

To check runday effects on the occurrence of missing values, we compared the proportions of missing values across rundays within each platform. For a metabolite j and a runday d, let $q_{j,d} = |\{x_{i,j} \mid x_{i,j} \text{ is } NA \text{ and } i \in I_d\}|$ be the number of missing measurements of the given metabolite in the samples measured on the given runday. Then the normalized missingness $m_{j,d}$ was calculated as the number of missing values for j in d divided by the total number of samples measured in d, divided by the median value of missing data of j over all run days.

$$m_{j,d} = \frac{\frac{1}{|I_d|} \cdot q_{j,d}}{median(q_{j,1},\ldots,q_{j,l})}. \tag{2.1}$$

A normalized proportion of missingness $m_{j,d} = 1$ is the average runday-specific amount of missing values. Cross-platform dependencies were investigated by calculating the Pearson correlation coefficient and corresponding p-value between the median normalized runday-specific missing values across all metabolites of two platforms.

Runday-dependent limit of detection

Previous studies have already suspected that batch (runday) effects can give rise to multiple detection limits [98, 99]. To verify this assumption, we calculated the Pearson correlation between the runday-specific proportion of missing values and the runday-specific mean measurement for each metabolite. A positive correlation indicates higher amounts of missing values with higher concentrations of the respective metabolite, while a negative correlation illustrates an inverse relationship of runday mean and runday-specific missing values.

Missingness mechanisms

Based on the insights gained from the real data, we derived five mechanisms for the occurrence of missing values (Figure 2.3). In the *Fixed LOD* mechanism, missing values were assumed to be below a global limit of detection. The *Probabilistic LOD* mechanism was an attenuated form of *Fixed LOD* where probabilistic means that the probability of a value being missing gets lower with higher values. *Runday-specific fixed LOD* should mirror cases where the LOD is assumed to vary across rundays. *Runday-specific probabilistic LOD* was a mechanism where the probabilistic LOD is assumed to occur in each runday. Finally, *Unsystematic missingness* should represent missing values occurring completely randomly. Single missingness mechanisms are described in more detail below.

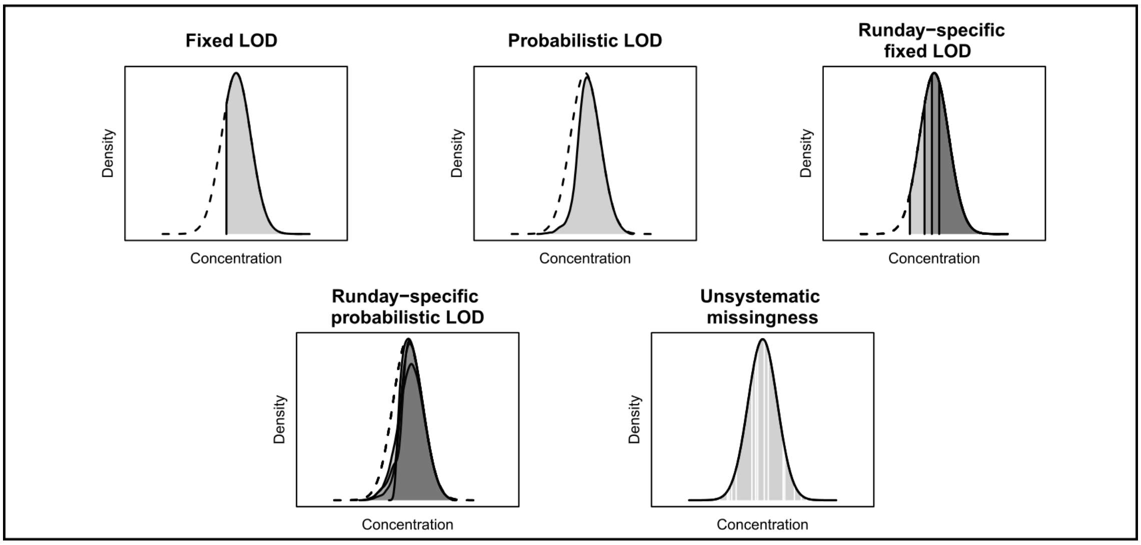


Figure 2.3: Derived missing values mechanisms (figure from [97]). Based on insights gained from real untargeted MS-based metabolomics data, five mechanisms for the occurrence of missing values were derived. *Fixed LOD* is a mechanism where missing values occur due to a fixed threshold. *Probabilistic LOD* is a mechanism where the probability of a value to be missing inversely depends on the value. *Runday-specific fixed and probabilistic LOD* describe the two aforementioned mechanisms within each runday. Finally, *Unsystematic missingness* is a mechanism where missing values occur randomly.

2.2.2 Simulation of missingness mechanisms in artificial data

For this section, let x and y be two normally distributed vectors of length n, where $x_{(i)}$ denotes entry i in the vector of ordered concentrations in x.

Fixed LOD

It was assumed that all values below a fixed limit of detection are missing. This LOD was defined as $LOD = x_{round(n*miss)}$ with $miss$ as the proportion of missing values to be simulated and $round()$ is a probabilistic rounding function.

Probabilistic LOD

It was assumed that LOD is not fixed but the probability that a value is missing inversely depends on itself. The probability of missingness for each observation i, of a variable x was modeled as a function of x as $P(x_i\ missing) = logistic(\beta_0 + \beta_1 x_i)$, with $logistic(a) = \frac{e^a}{1+e^a}$ is the logistic function and coefficient β_1 resembles the dependence of the probability of missingness from variable value. β_0 is the intercept and was estimated by solving the equation $\frac{1}{n}\sum_{i=1}^{n} P(x_i \text{ missing}) = miss$, where $miss$ was achieved by drawing $n \times miss$ times from a multinomial distribution with probability vector $(P(x_i\ missing))_i = 1, ..., n$.

Runday-specific fixed LOD

It was assumed that the runday-specific fixed LOD varies around the global fixed LOD according to a normal distribution. For variables x and y and a runday d, we used the approximation

$$\begin{pmatrix} miss_{x,d} \\ miss_{y,d} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} miss_x \\ miss_y \end{pmatrix}, \begin{pmatrix} \sigma^2_{miss,x} & r_{miss} \cdot \sigma_{miss,x} \cdot \sigma_{miss,y} \\ r_{miss} \cdot \sigma_{miss,x} \cdot \sigma_{miss,y} & \sigma^2_{miss,y} \end{pmatrix} \right). \tag{2.2}$$

$miss_x$ and $miss_y$ are the global amount of missing values for the two variables, $\sigma^2_{miss,x}$ and $\sigma^2_{miss,y}$ denote the variation of the proportion of missing values across rundays, and r_{miss} is the correlation of runday-specific missingness between the two variables. σ^2_{miss} was estimated with

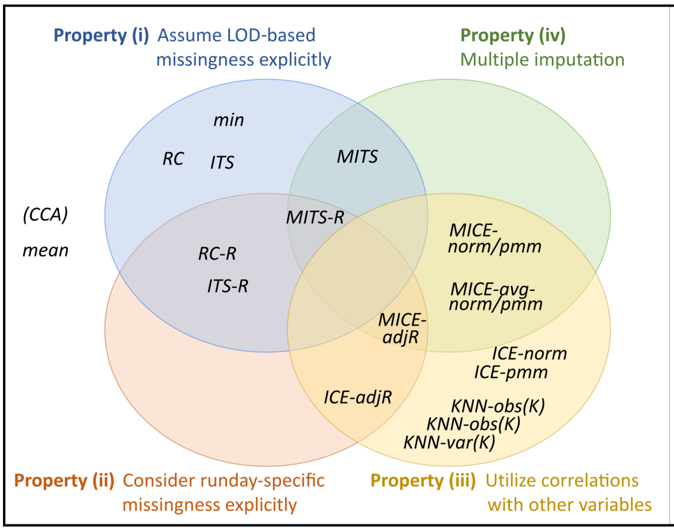
$$\sigma^2_{miss} = \begin{cases} \left(\frac{min(miss, 1-miss)}{4}\right)^2, & \text{moderate variability of missingness across rundays} \\ \left(\frac{min(miss, 1-miss)}{2}\right)^2, & \text{strong variability of missingness across rundays} \end{cases} \tag{2.3}$$

Runday-specific probabilistic LOD

This missingness mechanism was simulated similarly to the runday-specific fixed LOD. The runday-specific β_{1d} was either kept constant across rundays or was simulated to vary similarly to $miss$ with mean at the global β_1 and variance $\sigma^2_{\beta_1} = \left(\frac{\beta_{1d}}{2}\right)^2$ for strong variation of β_{1d} across rundays.

2.2.3 Imputation methods

In our study [97], we evaluated 31 imputation methods, each having at least one of the following features (Figure 2.4).


Property (i) Assume LOD-based missingness explicitly
Property (iv) Multiple imputation
min
RC
ITS
MITS
(CCA)
mean
MITS-R
RC-R
ITS-R
MICE-norm/pmm
MICE-avg-norm/pmm
MICE-adjR
ICE-norm
ICE-pmm
ICE-adjR
KNN-obs(K)
KNN-obs(K)
KNN-var(K)
Property (ii) Consider runday-specific missingness explicitly
Property (iii) Utilize correlations with other variables

CCA	Complete case analysis
min	Minimum imputation
RC	Richardson & Ciampi
ITS	Imputation by truncated sampling
MITS	Muliple ITS
RC-R	RC within rundays
ITS-R	ITS within rundays
MITS-R	MITS within rundays
mean	Mean imputation
ICE-norm	Imputation by chained equations using Bayesian regression imputation
ICE-pmm	ICE using predictive mean matching
ICE-adjR	ICE with random runday intercept
MICE-norm	Multiple ICE-norm, pooling statistics
MICE-pmm	Multiple ICE-pmm, pooling statistics
MICE-avg-norm	Multiple ICE-norm, pooling data
MICE-avg-pmm	Multiple ICE-pmm, pooling data
MICE-adjR	Multiple ICE-adjR
KNN-var(K)	*K*-nearest neighbor imputation per variable
KNN-obs(K)	KNN per observation
KNN-obs-sel(K)	KNN per observation using selected variables

Figure 2.4: Imputation approaches (figure from [97]). The venn diagram on the left panel shows four different properties inherent to 31 imputation approaches. Note that the figure contains complete case analysis (CCA), which is not an imputation method, and is noted in brackets. CCA and *mean* were placed outside of the Venn diagram, as they do not comprise any of the four characteristics. LOD: limit of detection. The right panel consists of names and short descriptions of the imputation methods.

Imputation based on an LOD

Methods that explicitly assume LOD-based missing values replace missing entries with very low values. The most commonly used approach is minimum imputation, where missing values are replaced by (half of, or squared of) the lowest number observed in the data. A related method was proposed by Richardson and Ciampi (in the following referred to as RC method), who assumed that metabolites with missing values follow a left-truncated normal distribution [100]. They suggested to estimate the truncated distribution of a vector x with maximum likelihood estimation (MLE) using the smallest observed value as truncation point and the log-likelihood function

$$\log L(x) = (n-o) \times \ln\Phi\Big(\frac{LOD-\mu_x}{\sigma_x}\Big) - o\ln\big(\sqrt{2\pi}\sigma_x\big) - \sum_{i=1}^{o}\frac{(x_i-\mu_x)^2}{2\sigma_x^2}, \tag{2.4}$$

where L is the likelihood, $i, i = 1, ..., n$ are the observations, o is the number of observed values for x, $n-o$ is the number of missing values for x, μ_x and σ_x are the mean and standard deviation of observed values in x, and $\Phi(x)$ is the cumulative Gaussian distribution function. After computing the marginal MLEs $\hat{\mu}_x$ and $\hat{\sigma}_x$, missing values were replaced by the expectation value:

$$E(x|x \leq LOD) = \hat{\mu}_x - \frac{\hat{\sigma}_x^{-\frac{(LOD-\hat{\mu}_x)^2}{2\hat{\sigma}_x^2}}}{\Phi(\frac{LOD-\hat{\mu}_x}{\hat{\sigma}_x})}. \tag{2.5}$$

Based on the idea of Richardson and Ciampi, we developed an approach for imputation by truncated sampling (ITS). Similarly to RC, a left-truncated normal distribution was estimated with MLE with the lowest observed value as truncation point. But missing values were then replaced by random draws from the originally censored part of the estimated distribution.

Imputation considering runday effects

Based on observations from real untargeted MS-based metabolomics data, we implemented runday-specific versions of RC and ITS, termed RC-R and ITS-R, respectively. To this end, the imputation methods were performed for each runday before runday normalization (see section 1.2.2). For a robust performance of MLE, we set the minimal required number of non-missing observations per runday to 17 (half of the average number of observations per runday in KORA F4). For RC-R the missing entries in the remaining rundays, i.e. rundays with less than 17 observations, were set to the mean expected values across all rundays with a sufficient number of values. For ITS-R, the remaining missing values were replaced using imputation by chained equations with Bayesian regression as primary model (see below).

Multivariate imputation

Multivariate approaches take the correlation between the variables into account for imputation. We applied multiple imputation by chained equations (MICE) and k-nearest neighbors (KNN) imputation with different parameter settings.

MICE assumes that missing values are missing at random (MAR), that is, the probability of a value being missing depends only on observed values (see section 1.2.3). Therefore, the idea is that if all available variables are controlled for, then the remaining missingness is completely random [101]. The principle of imputation by chained equations (ICE) is a repeated chain of equations through the incomplete variables - variables with at least on missing entry -, where in each imputation model the given incomplete variable is modeled as a function of the remaining variables. This procedure can be summarized in a few steps:

- First, an initial mean imputation is performed for the incomplete data set.
- Then for a given incomplete variable x, the mean imputed entries are set back to missing (NA) and a regression model (also called "primary model") with x as response and the other variables in the data set as predictors is formed.
- The missing values in x are imputed with predictions from this regression model.
- This imputation procedure is performed for each incomplete variable in the data set as response variable using the newest imputations for the predictors (one "iteration").
- At the end of an iteration, all missing values were replaced by predictions from the variable-specific regression models, which reflect the relationship of the observed data.
- A user-specific number of iterations can be performed, resulting in one final imputed data set that is assumed to contain converged (stable) imputations.
- To perform *multiple* ICE (MICE), the described ICE procedure is performed m times, generating m imputed data sets (see Multiple imputation in section 2.2.3).

In our work [97], we used predictive mean matching and Bayesian regression (for details see [101] and [36]) as primary models and performed variable pre-selection for computational speed-up by only using predictors with at least $r = 0.1$ correlation to the respective incomplete variable. Additionally, for each model age, sex, and BMI were included as covariates. (M)ICE was performed using the R package *mice*, version 2.25. A detailed description of the complete MICE algorithm is provided in [101].

KNN-based imputation replaces missing values for each variable by the weighted k nearest variables or observations, defined by a distance measure. In our projects, we used the Euclidean distance as distance measure. Weights were chosen as e^{-dist}, where $dist$ is the distance between two variables or observations. In case the nearest neighbors were defined as observations, we performed variable pre-selection by including only strong correlated variables at $r \geq 0.2$ in the calculation of the distances [37]. On the one hand, this dimension reduction decreases the computational time, and on the other hand, pre-selection of variables to only highly correlated pairs was shown to improve the imputation performance [97], most likely by reducing the noise in identification of the k nearest neighbors.

Multiple imputation

Usually single imputation does not take the uncertainty of the imputed data into account, for which reason multiple imputation (MI) is often performed. The idea of MI is to impute the given data multiple times to generate multiple imputed data sets. The differences between these data sets reflect the uncertainty of the missing values.

Multiple imputation techniques usually cover the three steps (i) imputation, (ii) analysis, and (iii) pooling. In (i), an imputation procedure is applied m times on the incomplete data, which produces m different imputed data sets. In (ii), statistical analysis is performed for each of the imputed data sets such that m statistical results are generated. Finally (iii), the m statistical results are pooled to one final result.

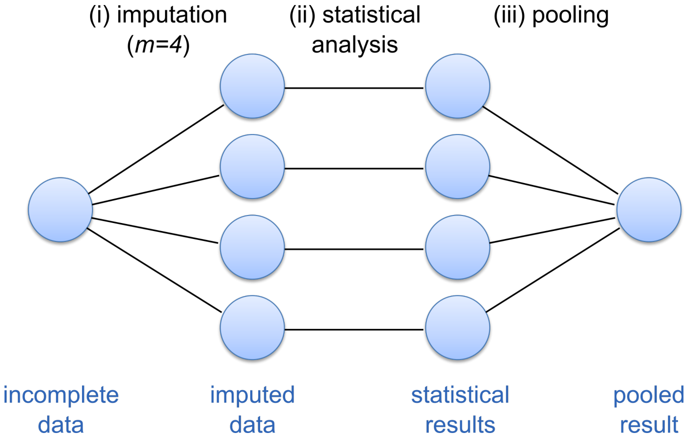


Figure 2.5: Multiple imputation concept (figure adapted from [36]). The given incomplete data is imputed $m = 4$ times, followed by statistical analysis of each of the four imputed data sets. This results in four different statistical results, which need to be combined into one single result.

For pooling statistical estimates, a routine established by Rubin [102, 103], available within the R package *mice*, version 2.25 was used. The number of imputations was set to $m = 20$ for all multiple imputation procedures. In a modified version of multiple imputation, we imputed the incomplete data set m times and pooled the m imputed values by calculating their average across the data sets. Finally, statistical analysis was performed only once on the averaged imputed values.

2.3 Pathway analysis

Metabolites are commonly assigned to sets of compounds, termed “pathways”, according to their biochemical or structural properties. In our work, we mainly used pathway annotations provided by Metabolon, Inc.: each metabolite with known chemical structure is assigned to exactly one *super-pathway* representing metabolite classes such as “Lipid” or general metabolic processes such as “Energy”, and one *sub-pathway* representing biochemical subclasses or processes within a super-pathway such as “Lysolipid” or “TCA cycle”, respectively.

Pathway enrichment

In our study [104], we investigated whether significantly many metabolites of a certain pathway tend to occur exclusively in a certain body fluid. To this end, we performed Fisher’s exact test based on the following contingency table:

	In pathway	Not in pathway
Unique in fluid	a	b
Not unique in fluid	c	d

where a is the number of metabolites in the investigated pathway, which are unique in a fluid (only measured in this fluid); and b is the number of metabolites uniquely occurring in the respective fluid, but which do not belong to the given pathway. Accordingly, c is the number of metabolites in the pathway, which have been measured in at least two body fluids. Finally, d is the number of metabolites not in the pathway and not uniquely occurring in the investigated fluid. The hypergeometric probability of getting the observed numbers on the null hypothesis was given by

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{p}{a+c}}. \tag{2.6}$$

Similarly, in another analysis, we aimed to identify whether metabolites uniquely occurring in a body fluid are related to a given phenotype.

Pathway representation

A more general approach for pathway analysis was based on pathway representatives. A representative is a variable that aggregates all concentrations in the set of metabolites into a single value. For each pathway a representative was defined either by the *eigenmetabolite* [105] or the *average* approach. In the former, a principal component analysis (PCA) for a given pathway was performed after scaling all variables to a mean of 0 and a variance of 1. The first principal component (*eigenmetabolite*), which is the direction explaining most of the data variation, was used as the representative variable. In the *average* approach all variables were also scaled such that the mean is 0 and the variance is 1. The pathway representative was then defined as the average of all variable values in the given pathway (aggregated z-score).

The pathway representatives represent a new data matrix, for which any univariate or multivariate statistical approach of interest could be performed. For instance, in the case of phenotype association analysis, a regression model was used to estimate the association of each pathway representative with the given phenotype.

2.4 Network inference

In *omics* research, networks are widely used to describe relationships between molecular components. In general, a network is a collection of nodes typically visualized as circles, which are connected by edges visualized as lines between two nodes. Although this may also be correct for a "graph" and these two terms are used indistinguishably in literature, we understand a graph as the structural information of a $p \times p$ adjacency matrix A, where p is the number of variables in the given data set. The graph consists of nodes presenting the variables in the adjacency matrix, and edges corresponding to the entries in A:

$$a_{ij} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

A network is then a graph with additional node and edge attributes (e.g., edge weights), which are not contained in the adjacency matrix.

Biological networks can either be knowledge-driven (connections between the nodes are based on prior knowledge), or data-driven (the edges are statistically inferred from data). Since prior knowledge on interactions between molecular entities is often incomplete, in our studies, we used Gaussian graphical models (GGMs) for network inference.

Gaussian graphical models

Graphical models are a class of probability distributions, and a Gaussian graphical model is a graphical model in which the joint distribution of the random variables is Gaussian. GGMs are based on partial correlations to estimate the conditional dependencies in multivariate Gaussian distributions. That is, the partial correlation of two variables is the association of these variables controlled for all other available variables. Thus, a GGM is a symmetric partial correlation matrix representing an undirected network, which contains only direct correlations. Here, the variables of the correlation matrix are represented by nodes, and two nodes are linked by an edge, if the partial correlation between the two respective variables is above a given significance threshold. The partial correlation $\widetilde{r}_{xy}$ between two variables x and y can be calculated from the inverse of the covariance matrix $\left(\omega_{xy}\right) = \mathbf{\Sigma}^{-1}$:

$$\widetilde{r}_{xy} = \frac{-\omega_{xy}}{\sqrt{\omega_{xx}\omega_{yy}}} \tag{2.8}$$

For high-dimensional data with low sample sizes ($n \ll p$), regularized partial correlations proposed by Schäfer and Strimmer [106] can be estimated. They make use of a shrinkage covariance estimator based on the Ledoit-Wolf lemma [107] for the calculation of the optimal shrinkage intensity λ^* on the empirical correlations $R = (r_{xy})$ instead of using the sample covariance matrix.

$$\lambda^* = \frac{\sum_{x \neq y} var(r_{xy})}{\sum_{x \neq y} r_{xy}^2} \tag{2.9}$$

This shrinkage intensity is used to calculate the shrinkage covariance matrix, which is positive defined even for small sample sizes. Partial correlations are then calculated by applying formula 2.8 on the inverse of the shrinkage covariance matrix. To assess significance of partial correlations, a mixture model was fitted to the partial correlations resulting in two-sided p-values for the null hypothesis of no correlation.

The estimation of partial correlations and the corresponding p-values were performed using the R package *GeneNet* version 1.2.13. The partial correlation matrix was filtered based on the corresponding p-values after multiple testing correction. We applied Bonferroni correction controlling for the family-wise error rate at $\alpha = 0.05$. Thus, partial correlations with a p-value greater than $\alpha/\binom{p}{2}$, where $\binom{p}{2} = p(p-1)/2$ is the number of tests performed, were set to zero. In addition, we also excluded all partial correlations for which the Pearson correlation was not significant after Bonferroni correction to avoid statistical artifacts.

Pathway-based modularity

Biochemical validity of a GGM was assessed by analyzing whether edges connect metabolites from the same or from between different pathways. The assumption was that metabolite correlations mainly occur within a pathway since these metabolites share common biological and biochemical properties and therefore, participate in the same metabolic processes. Let $(P_1, ..., P_q)$ be the non-overlapping sets of nodes, with P_v containing the nodes assigned to a pathway v. Here, we used pre-defined pathway annotations provided by Metabolon Inc. (see section 2.3). Then the pathway-based modularity can be mathematically described by

$$Q = \sum_{v=1}^{q} \left[\frac{A(P_v, P_v)}{A(P, P)} - \left(\frac{A(P_v, P)}{A(P, P)} \right)^2 \right], \tag{2.10}$$

where $A(P_v, P_w)$ is the number of edges between the two node sets P_v and P_w. A high Q value corresponds to high pathway-based modularity in the GGM, that is, edges tend to occur between metabolites of the same pathway rather than between metabolites from different pathways.

Phenotype embedding and visualization

The GGMs were visualized as networks with nodes corresponding to the metabolites and edges representing partial correlations after filtering. Edge width represented the absolute partial correlation strength. To contextualize a phenotype, e.g. clinical parameters or disease status, with metabolic correlations, the nodes in the inferred network were colored according to the association of the respective metabolite with the given phenotype, which was obtained from a differential analysis (Figure 2.6). The colored networks were interactively visualized with Cytoscape, version 2.8.3 or yEd, version 3.12.2.

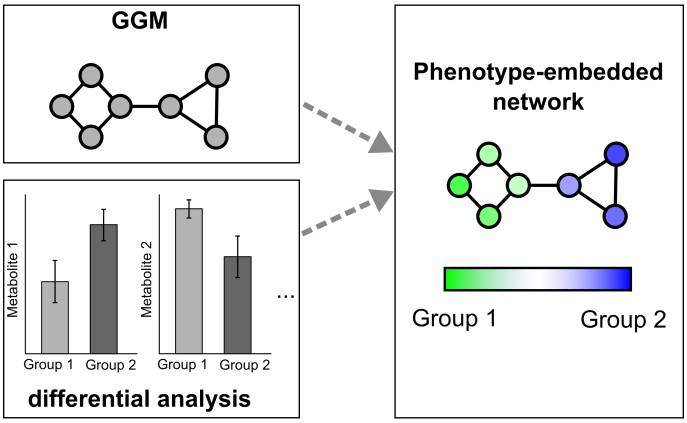


Figure 2.6: Phenotype embedding into networks. A network reflecting statistically inferred metabolic processes is inferred by estimating partial correlations between the metabolites, followed by significance filtering. In a differential analysis, associations of the single metabolites with a given phenotype are determined. The nodes in the network are then colored according to these phenotype associations.

2.5 Phenotype-driven module identification

In our work [108], we developed a network-based method - implemented in the R package *MoDentify* - for the phenotype-driven identification of functional modules on different layers of resolution; from fine-grained metabolites to global pathways. At metabolite level, the network is inferred based on Gaussian graphical models (see section 2.4). At pathway level, we first performed module representation (see section 2.3), followed by GGM estimation. Based on the given network, *MoDentify* searches for functional modules that are highly associated with the phenotype of interest by score maximization.

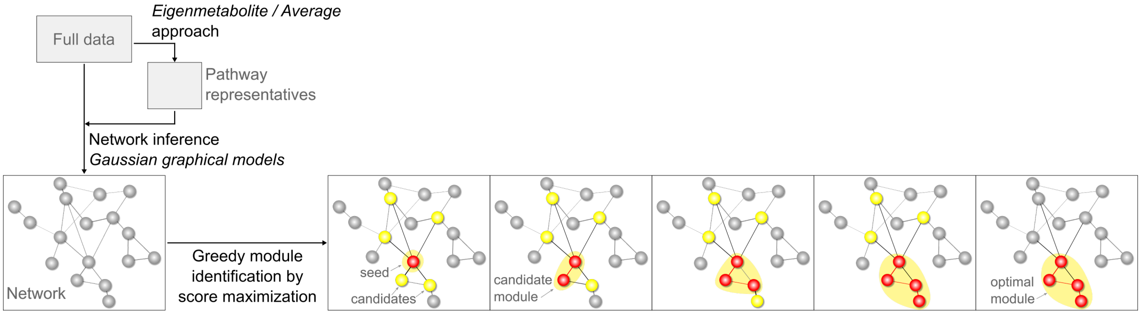

Full data
Eigenmetabolite / Average approach
Pathway representatives
Network inference
Gaussian graphical models
Network
Greedy module identification by score maximization
seed
candidates
candidate module
optimal module

Figure 2.7: Network-based module identification (figure from *MoDentify* package vignette [108]). The first step is the generation of a network using Gaussian graphical models (GGMs). Optionally, a pathway network can also be inferred by first defining the pathway representatives with the *eigenmetabolite* or *average* approach and subsequently, estimating a GGM based on the pathway representatives. The second step is the application of a greedy search procedure on the inferred network to identify an optimal module by score maximization. From an initial seed node, the algorithm extends the candidate module along the network edges, until no further score improvement is possible.

Module search

Given a phenotype vector P containing the condition for each individual (either continuous such as age or binary such as case *vs.* control), an $n \times p$ data matrix, a network represented by an adjacency matrix containing the p metabolites as nodes and vectors with node and edge attributes, and a scoring function, a greedy approach is applied to search for functional modules by score maximization. The algorithm performs the following procedure iteratively, starting with a seed node as the initial candidate module (Figure 2.7):

Let M be the candidate module. The neighborhood of M, defined as the set of nodes connected to any node in M in the given network is determined. Successively, each neighbor is then added to M and the score of the extended module is calculated. The neighbor i that meets the following requirements is finally added to M:

1. Score improvement: the score of the extended module $M \cup i$ is higher than the score of the original module M.
2. Maximal score improvement: there is no other neighbor, which leads to higher score improvement than i.
3. The score of the extended module $M \cup i$ is higher than the scores of all of its single components.

The new candidate module is then $M \cup i$ and the described procedure is repeated. The algorithm terminates if no score improvement is possible. The significance of the final module is assessed at $\alpha = 0.05$ after applying Bonferroni multiple testing correction for p tests. After the algorithm is applied to all seed nodes as initial candidate modules, all overlapping final modules are merged in a consolidation step and the respective modules scores are re-calculated.

Module scoring

Given a candidate module M, the score was obtained from the multivariable linear regression model

$$R_M = \beta_{M,0} + \beta_{M,1} \times P + \sum_{i=1}^{|C|} (\beta_{M,i+1} \times c_i) + \epsilon_M, \tag{2.11}$$

where R_M is the module representative, $\beta_{M,0}$ is the intercept, $\beta_{M,i}$ is the regression coefficient for the respective independent variable, P is the given phenotype, $C = \{c_1, ..., c_{|C|}\}$ is a set of covariates, and ϵ_M is a normally distributed error term. R_M was calculated using the *average* method as described in section 2.3. If M consisted of multiple pathways, then

R_M was calculated based on the union set of all metabolites from the respective pathways (without using the pathway representatives). The module score was defined as the negative logarithm of the p-value corresponding to $\beta_{M,1}$.

Module visualization

Besides R data structures and flat files, *MoDentify* enables the interactive visualization of the modules within the given network in the open source software Cytoscape [109]. One of the major advantages of our package is the direct call of Cytoscape from within R *via* the *RCytoscape* package [110] without the usual cumbersome exporting of data files from R and re-importing them into Cytoscape. The visualized network contains all necessary node and edge attributes, including different node colors and node sizes, which depict the membership of a node in a certain module and its association with the given phenotype from a classical, single-molecule association analysis, respectively. Moreover, significance of the phenotype association is indicated by diamond-shaped nodes.

Chapter 3

Summary of contributed articles

A summary of all first-author publications that resulted from my doctoral studies is provided below, sorted by the research questions introduced in section 1.4. Shared first authorships are indicated by * symbols in bibliographic nominations.

I. *Which missingness patterns can be observed in real metabolomics data and how can these missing values be handled appropriately?*

   Kieu Trinh Do*, Simone Wahl*, Johannes Raffler, Sophie Molnos, Michael Laimighofer, Jerzy Adamski, Karsten Suhre, Konstantin Strauch, Annette Peters, Christian Gieger, Fabian J Theis, Harald Grallert, Gabi Kastenmüller, Jan Krumsiek (2017), **Characterization of missingness in untargeted MS-based metabolomics data and evaluation of missing data handling strategies.** *Under review* in *Metabolomics*

   In mass-spectrometry based untargeted metabolomics data typically 20-30% of the data is missing. Missing values are either excluded (complete case analysis, CCA) or imputed (replaced by a reasonable value). The handling of these missing values as a data quality control step is of crucial importance, since all downstream statistical analyses will be affected. We statistically evaluated the best missing values handling strategy in a controlled setting based on artificial data, and also assessed biological validity of the data for each strategy. To this end, this study covers four parts: the investigation of missing values mechanisms in real MS-based metabolomics data, the evaluation of imputation methods based on artificial data that was generated based on the gained insights, the evaluation of imputation methods based on biochemical pathways, and the evaluation of imputation methods based on associations of metabolites with quantitative trait loci.

   We analyzed possible underlying mechanisms of missing values in the KORA F4 study (see section 2.1). We found that 62% of the metabolites contain missing values that most probably occurred due to a limit of detection (LOD). However, there was no single fixed LOD, but rather a blurred detection limit for all metabolites. We also observed that the amount of missing values differ substantially across rundays, most likely reflecting shifts in instrument sensitivity. Finally, we found that the majority of metabolites display an inverse relationship of runday mean and runday-specific amount of missing values (see sections 2.2.1 and 2.2.1), indicating a runday-dependent LOD-based mechanism, which might give rise to the observed blurred overall LOD.

We evaluated 31 imputation approaches (see section 2.2.3) in a framework consisting of three evaluation schemes. In the first scheme, we aimed to evaluate imputation methods in a controlled setting by simulating incomplete data that reflected the real-data properties based on the obtained insights from the previous analysis on real data (see section 2.2.2). The imputation performance was assessed based on the ability of the approaches to achieve unbiased statistical estimates and valid hypothesis test results after conducting correlation and regression analyses. We observed that KNN-based imputation on the observations with variable pre-selection (KNN-obs-sel) and Multiple Imputation by Chained Equations (MICE) performed best in this evaluation scheme (see section 2.2.3).

In the second evaluation scheme, we conducted biologically-driven analyses using real data from the KORA F4 cohort. We investigated how accurately real biochemical pathways can be reconstructed in data-driven metabolic networks inferred from the imputed data. To this end, a Gaussian graphical model is estimated after imputation (see section 2.4), and pathway-based modularity, indicative of the partition of the data-driven network into pre-defined pathways, was assessed from the fraction of metabolite-metabolite correlations occurring within pathways compared to across pathways (see section 2.4). High modularity suggested biological validity of the GGM and served as a quality criterion for the imputation method in this analysis. Imputation with MICE resulted in a GGM with the highest modularity, closely followed by imputation with KNN-obs-sel.

In the final biological evaluation scheme, we explored the ability of imputation methods to preserve effects of genetic variants on metabolite levels while increasing statistical power. We investigated the association of 18 SNP-metabolite pairs, for which a functional relationship was evident from previous studies. An imputation strategy was assumed to perform well if there was statistical power gain while the effect size of the SNP-metabolite association was preserved after imputation. Again, we found KNN-obs-sel to perform well throughout all pairs.

Overall, for the first time, we presented a detailed description of missing values mechanisms in MS-based metabolomics data. We performed comprehensive evaluation of imputation strategies based on both statistical and biological readouts. KNN-based imputation on observations with variable pre-selection consistently performed best in all three evaluation schemes and is recommended for missing values handling in future MS-based data.

My contribution: This publication was a joint work with the co-first author Simone Wahl. The general idea to analyze missing values in metabolomics data originates from Jan Krumsiek. Together with him, Gabi Kastenmüller, and Simone Wahl, I planned the study design. I performed data preprocessing and ensured data quality control. I had the idea to analyze the limit of detection property of missingness, and to look for influences of runday effects. To this end, I designed the descriptive analysis and performed all statistical analysis and created and interpreted all result figures. I had the idea to compare imputation approaches to find the best strategy to handle missing values. To this end, I performed imputation on the real data and evaluated the methods based on biochemical pathways. I created and interpreted all result figures for this part. The idea to analyze metabolite-SNP pairs originated from Gabi Kastenmueller, and Johannes Raffler performed the statistical analysis. Interpretation of the results was done by Gabi Kastenmüller, Jan Krumsiek, Johannes Raffler and me. The statistical evaluation of imputation methods based on simulated data originated from Simone Wahl and she initially performed all computational analyses for this part. I adapted the code and created and interpreted the result figures after she left Helmholtz Zentrum Munich. Finally, I wrote the complete draft of the manuscript, which I then finalized based on comments from Jan Krumsiek, Gabi Kastenmüller, and Simone Wahl.

II. *How can metabolic processes within and between different body fluids be investigated, and how can we use these processes to analyze (patho-) physiological outcomes?*

Kieu Trinh Do, Gabi Kastenmüller, Dennis O Mook-Kanamori, Noha A Yousri, Fabian J Theis, Karsten Suhre, Jan Krumsiek (2015), **Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva.** *J Proteome Res, 14: 1183-1194*

For an integrated picture of metabolic interactions and physiological processes in the human body, a simultaneous analysis of multiple body fluids or tissues is essential, but yet missing. In this work, we addressed this challenge by inferring networks that reflect a comprehensive view of multi-fluid metabolic interactions. The analysis was based on untargeted metabolomics measurements for plasma, urine, and saliva samples from the QMDiab cohort (see section 2.1).

With this comprehensive view we aimed to answer three questions: (i) How similar are the fluids in terms of metabolic composition and correlations? (ii) How do fluids interact metabolically? And (iii), how can phenotype information be integrated into these metabolic processes and which insights can be gained?

To address question (i), we first compared the metabolic composition of the fluids by analyzing the occurrence of metabolite classes (see section 2.3) in the fluids. We found that certain groups of metabolites, such as lipids, were fluid-specific (e.g., they significantly occurred in only one body fluid) while others, such as amino acids, were shared across fluids (e.g., they occurred significantly often in at least two fluids), indicating strong exchange of these metabolites between the body fluids. To compare the within-fluid correlation structure, a Gaussian graphical model (GGM) was estimated for each body fluid (see section 2.4). The fluid-specific GGMs were superimposed for a direct comparison of the fluids. While plasma had the highest number of metabolite correlations, saliva showed the lowest number of correlations. Blood and urine are physiologically connected through excretion processes in the kidneys; and as expected, these two fluids were more similar to each other than to saliva. To systematically check the biological validity of the statistically derived metabolite correlations, we performed Fisher's exact test to analyze whether correlating metabolites were assigned to the same pathways (see section 2.3). In all fluids, a substantial high number of edges occurred within a pathway, indicating that the inferred networks were biologically reasonable. Finally, we provided an in-depth view of the networks by discussing arbitrarily extracted example subnetworks, including purines, amino acids, and fatty acids as well as free acylcarnitines.

To address question (ii), a GGM for all metabolites from all fluids was estimated to obtain the cross-fluid metabolite correlations. We found a considerable higher fraction of edges between plasma and urine than between plasma and saliva, or urine and saliva. This was expected, since blood and urine are physiologically connected through excretion and re-absorption processes in the kidneys. In contrast, the correlations found between urine and saliva might reflect stable concentrations throughout the body, e.g., small changes between food intake and excretion. We repeated the biochemical validity check for the cross-fluid GGM and observed a remarkably high fraction of edges within pathways. Moreover, correlations occurring between fluids were mainly between the same metabolites, pointing to excretion, transport, and diffusion processes. Finally, we provided an in-depth description of the cross-fluid GGM by discussing subnetworks containing cotinine, saccharin, as well as cortisol and cortisone.

To address question (iii), we generated a consolidation of three types of information. Statistical correlation between metabolites were combined with direct comparison of the fluids as already obtained for question (2). In addition, we integrated the association of metabolites to the four phenotypes, type 2 diabetes (T2D), age, gender, and BMI into the cross-fluid GGM. To this end, nodes were depicted as pie charts that reflected the strength of phenotype associations obtained from a multivariable linear regression model (see section 2.4). With this approach we were able to visually identify metabolic modules that were related to one of the phenotypes, e.g., a group of correlating monosaccharides associated with T2D. Moreover, we could directly compare the phenotypes and their effects on metabolism. For instance, steroid hormones were associated with age mainly in urine, and with sex mainly in blood. We also found that metabolites associated with T2D and age were significantly shared between the fluids, suggesting widespread effects of these phenotypes on metabolism in the human body. Finally, we discussed phenotype related subnetworks spanning steroid hormones, monosaccharides, the drug metformin, and amino acids and derivatives.

Overall, we introduced a network-based framework for the analysis of multi-fluid metabolomics data and their associations with clinical outcomes. Our approach is generic and therefore, widely applicable, which was demonstrated by its application on type 2 diabetes [111], insulin-like growth factor I [IGF-I) [112], and restless legs syndrome and Parkinson's Disease [68] in follow-up studies.

My contribution: The original idea to analyze multi-fluid metabolomics data based on network inference was from Karsten Suhre and Jan Krumsiek. I had the idea to create overlaid within-fluid networks and a combined cross-fluid network. I performed data preprocessing and ensured data quality control. For the publication, I planned the design of the study and performed all computational analyses. All visualization ideas originated from me and I developed and applied all necessary steps. I created and interpreted all result figures and wrote the first complete draft of the publication, which I then finalized based on comments from Karsten Suhre, Jan Krumsiek, and Gabi Kastenmüller.

III. *How can interpretable functional modules be identified in a systematic and phenotype-driven manner?*

Kieu Trinh Do, Maik Pietzner, David Rasp, Nele Friedrich, Matthias Nauck, Thomas Kocher, Karsten Suhre, Dennis O Mook-Kanamori, Gabi Kastenmüller, Jan Krumsiek (2017), **Phenotype-driven identification of modules in a hierarchical map of multi-fluid metabolic correlations.** *NPJ Syst Biol Appl. 2017;3:28*

Phenotype associations typically span sets of correlating metabolites, termed functional modules. Module identification algorithms are established for *omics* data, but none of the methods considered the fact that there are different scales of phenotype associations. For phenotypes with only few metabolite associations (”sparse” effects), the identification and interpretation of modules is usually straightforward. However, there are also phenotypes associating with more than half of the metabolome (”dense” effects), which complicates biological interpretation due to the sheer quantity of associated compounds. The aim of this work was to develop an approach for the systematic identification of metabolic modules associated with a given phenotype. In particular, these modules should be well interpretable for all types of phenotype effects. To this end, our idea was to adapt the metabolic resolution of the module identification to the scales of phenotype effects.

To obtain a multi-level view on metabolism, we generated a hierarchical map reflecting multi-fluid human metabolism at different resolution levels using plasma, urine, and saliva untargeted metabolomics data from the SHIP-TREND cohort (see section 2.1). We inferred networks depicting metabolite-metabolite and pathway-pathway correlations within and between the three body fluids. The metabolite networks were inferred by estimating Gaussian graphical models (see section 2.4). For pathway-pathway interactions we generated two networks – one for “sub-pathways” and one for “super-pathways” (see section 2.3). The sub-pathway network was inferred by estimating a GGM on the sub-pathway *eigenmetabolites* (see section 2.3). The super-pathway network was generated by collapsing the sub-pathway network. These three networks formed a hierarchical map, which reflected human metabolism at three different layers of resolution.

We obtained a general overview of the hierarchical map by comparing plasma, urine, and saliva in terms of metabolite composition and correlation structure. The results suggested diverse metabolic processes in the fluids, most likely due to substantially different physiological roles of each fluid. In addition, we confirmed observations from our previous work that plasma and urine were tightly connected, while urine and saliva shared only few edges [104]. Cross-fluid correlations were mainly between biochemically closely related molecules, most probably due to transport and exchange processes between fluids.

Next, we developed a network-based greedy approach for the identification of functional modules (see section 2.5). A module is defined as a group of metabolic entities (metabolites or pathways) that is stronger associated with a given phenotype then all of its single components. We applied the module identification approach to insulin-like growth factor (IGF-I) representing phenotypes with sparse effects, and gender representing phenotypes with dense effects, at all levels of the hierarchical map. Thereby, we illustrated that for the sparse effects, modules at the fine-grained metabolite level were well interpretable, while at the global pathway levels no modules were identified. In contrast, for the dense phenotype associations, we identified 73 modules at metabolite level, which were unraveled to 13 sub-pathway modules, alleviating biological interpretation. The results for gender were successfully replicated using the QMDiab cohort (see section 2.1). Despite substantial differences in study design, metabolomics measurement, and power, half of the modules found in SHIP-TREND were also identified in QMDiab.

With this work we introduced a systematic approach for the identification of phenotype-driven modules. In particular, we showed that phenotype effects are in fact on different scales (sparse *vs.* dense), and module interpretation can be substantially enhanced if the metabolic resolution is adapted to these scales. Moreover, we demonstrated that the identified modules provide deeper insights into mechanistic aspects of phenotype associations, and a holistic view can only be obtained when multiple body fluids are simultaneously analyzed. Finally, our approach showed considerable increase in statistical power compared to classical association analysis.

My contribution: I had the idea to identify modules based on multi-fluid metabolic networks. Together with Jan Krumsiek, we decided to apply the module identification on multiple layers of resolution. I developed, implemented, and applied all computational approaches. I performed data preprocessing and ensured data quality control. I created and interpreted all result figures and wrote the first complete draft of the manuscript, which I then finalized based on comments from Jan Krumsiek.

Do KT, Rasp D, Kastenmüller G, Suhre K, Krumsiek J (2017), **MoDentify: a tool for phenotype-driven identification of modules based on hierarchical networks** *Under review* in *Oxford Bioinformatics*

The aim of this work was to provide the developed approach for phenotype-driven module identification reported in our previous publication [104] to the scientific community.

To this end, we implemented an R package named *MoDentify* for the phenotype-driven module identification at different layers of resolution 2.5. Thus, the algorithm can detect functional modules for a given phenotype at the fine-grained metabolite level as well as at coarser pathway levels. *MoDentify* consists of four main steps: (i) network inference, (ii) module identification, (iii) module scoring, and (iv) module visualization.

We implemented two approaches for (i): data-driven networks can either be estimated as Gaussian graphical models using partial correlations (see section 2.4), or as Pearson correlation networks. At the fine-grained level, the nodes in the resulting network correspond to metabolites, while at the global level, a node represents a whole pathway. Edges depict significant (partial) correlations between two nodes. Pathway networks can be inferred using pathway representative variables (see section 2.3). In addition, it is also possible to use an external network as a basis for the module identification procedure.

Based on the network, a greedy approach is performed to identify optimal modules associated with the phenotype of interest (ii). Given an initial candidate module (seed node), an iterative procedure is performed: the neighborhood of the candidate module is scanned, and the candidate module is extended by the node leading to the highest score improvement. The score of a candidate module is obtained from a multivariable linear regression model (iii) with the module representative (see section 2.3) as response and the given phenotype and optional covariates as predictors (see section 2.5). The algorithm terminates if no further score improvement is possible. Finally, identified modules can be visualized interactively within the underlying network (iv). To this end, the open source software Cytoscape can be directly called from within R (see section 2.5).

With *MoDentify*, we provided a freely available and easy-to-use R software, which is widely applicable for all types of data due to its generic character.

My contribution: I developed the algorithm and implemented it in MATLAB. I supervised a Bachelor student to translate the algorithm to R and to create the backbone of the R package *MoDentify*. I debugged, extended and finalized *MoDentify*. I wrote the package vignette. I prepared the example data set and made it publicly available. I applied the method to the example data and created all result figures. I wrote the first complete draft of the Application Note, which I then finalized based on comments from Jan Krumsiek.

Chapter 4

Discussion and future perspectives

4.1 Discussion

Metabolomics, the study of metabolic profiles at a global level, is frequently used to identify patterns associated with various (patho-) physiological outcomes. Technical advancements in this field have led to substantially increased complexity in measured data since much more compounds in more than one type of sample can be measured. Although the biological insights obtainable from this complex data facilitates fundamental understanding of metabolism and its link to clinical parameters and disease outcomes, the data complexity poses a considerable challenge to both appropriate data quality control, and the statistical approaches for data analysis and result interpretation.

We contributed to improved data quality by investigating the best approach to handle missing values. We provided a comprehensive overview of missing values patterns in real MS-based data, and offered imputation guidance under consideration of both statistical and biological evaluation schemes; two aspects that have not yet been addressed [97].

We contributed to more powerful metabolomics data analysis by developing a network-based approach to extract biological insights from complex multi-fluid metabolomics data and their links to clinical endpoints [33]. To the best of our knowledge, this was the first study involving the systematic, unsupervised and large-scale analysis of multiple body fluids based on human *in vivo* samples. Our approach is generic, and therefore, can be applied to any other metabolomics data set as demonstrated in three follow-up studies [68, 111, 112].

As an extension of our previous work, we moved from *visual* investigation to *systematic* assessment of phenotype associations in the context of complex (multi-fluid) metabolism. We developed a network-based greedy procedure to systematically and automatically identify functional modules – groups of correlating metabolic entities that are driven by a given phenotype [104]. With this work we also contributed to easy result interpretation of complex metabolomics data by adapting our approach to different layers of metabolic resolution such that results are easily interpretable for any scale of phenotype association;

from sparse associations such as for asthma and insulin-like growth factor I [64, 112], to dense associations such as for gender and BMI[90, 91, 95]. To the best of our knowledge, we were the first to address the challenge of searching for phenotype-driven modules in metabolic networks at different layers of resolution. In particular, we allow to consider cross-fluid communication by analyzing multiple body fluids simultaneously. Our approach has been shown to be statistically much more powerful than common association methods (e.g., t-test, regression analysis, correlation analysis). Importantly, with the increased statistical power we identified links between phenotype and metabolism, which have been hidden before. In a follow-up publication, we made our generic approach available to the scientific community by implementing an open-source and easy-to-use R package called *MoDentify*.

Despite our relevant contributions to improved metabolomics data quality control, data analysis, and data interpretation, our studies could be extended in several directions: (1) We found runday-effects on the occurrence of missing values in MS-based data. To exploit this information in the imputation process, we developed algorithms to explicitly consider global or runday-specific LOD-based mechanisms based on a truncated distribution. Unexpectedly, these methods did not achieve better imputation performances compared to other imputation approaches, even in simulated data with only LOD-based missingness. A possible reason may lie in the low number of observations available within rundays. Since the developed methods rely on maximum likelihood estimation to reconstruct the truncated distribution, low sample sizes limit the statistical power as have been shown in previous studies [42, 98]. Another possible reason is the omission of the data covariance structure. Taking the correlation between the variables into account may improve imputation performance, however, the specification of multivariate truncated distributions can be fairly challenging and might be addressed in future studies.

(2) We analyzed occurrence and handling of missing values in KORA F4. It would be also interesting to investigate whether missingness in data from a another cohort, from other sample types, or from other measurement platforms show similar patterns, and whether the same imputation performances can be expected. In addition, we performed our analyses independent of the other preprocessing steps (see section 1.2.2), but it would be interesting to investigate whether the handling of missing values is also influenced by normalization, transformation, and outlier handling. This aspect will be discussed in more detail in section 4.2.

(3) To generate an atlas of multi-fluid metabolic correlations in the human body [33], we used QMDiab, which at that time was the only available study with human multi-fluid metabolomics data. This study is a type 2 diabetes case-control cohort with non-fasting participants. Although non-fasting samples may diminish the statistical power to find intrinsic metabolic correlations, the correlations that persist in the non-fasting state can be assumed to be robust and stable throughout different metabolic states. With this in mind, investigating multiple body fluids under different dietary conditions may provide additional biological insights that can not be extracted with a single measurement, as piloted by the HUMET study from Krug *et al.* comprising 15 healthy men undergoing

various nutritional challenges [12].

(4) We condensed the complex metabolic correlations at metabolite-level to a more global and more interpretable pathway level. To this end, we inferred a network based on pathway representatives. Following the *eigengene* approach, we defined a pathway representative as the first principal component of its metabolic intensities (*eigenmetabolite*). For the majority of pathways, the degree of variance captured by the eigenmetabolite is high (>50%), but for very heterogeneous pathways such as "Phenylalanine and Tyrosine Metabolism" or "Food Component / Plant" the eigenmetabolite may not be representative. Instead of relying on pathway representatives, more sophisticated approaches such as canonical correlation analysis [113] or O2-PLS [114] could be adapted to model the relationship between entire pathways.

(5) To identify modules, we follow a greedy procedure by score maximization. A well-known problem of this local optimization technique is its inability to identify global minima (or global maxima in our case) if a local movement would increase the cost (or decrease the score in our case). Thus, with this approach, we prohibited "bridging" nodes – nodes that decrease the phenotype association if added to the candidate module, but which would connect to other nodes that would substantially increase the phenotype association. One possible solution to this limitation would be the use of global optimization techniques such as simulated annealing [115].

(6) In all projects, we analyzed metabolic pathways relying on pathway annotations provided by Metabolon Inc. These annotations were analogous to KEGG pathways, but were non-overlapping (each metabolite is assigned to exactly one pathway), which might not fully reflect real metabolic pathways. Additional analysis based on other available pathway annotations such as from HMDB [80], MetaCyc [81], or Recon [116] may strengthen the validity of our approaches; however, mapping metabolites to database entries could result in substantial loss of information since a large number of measured metabolites will not be covered in these databases.

(7) Multi-fluid data can be extremely useful for the identification and analysis of easier accessible, and therefore less invasive alternatives for known biomarkers (*surrogate marker*). For instance, in our study we found that salivary 1,5-anhydroglucitol is strongly correlated with the same metabolite measured in blood. And indeed, this salivary metabolite turned out to be a promising candidate as a non-invasive marker for short-term glycemic control [9]. This interesting aspect will be discussed in more detail in section 4.3.

4.2 Metabolomics preprocessing

Although preprocessing of metabolomics data – normalization, transformation, outlier and missing values handling, and scaling – substantially affects all downstream statistical analyses and result interpretation, these aspects are often omitted and there is no standard workflow available. Several studies [28–30, 41, 117–119], including our work on metabolomics missing values imputation [97], have investigated single preprocessing steps; however, these single steps influence each other, thus, not only the inherent effects of each step, but also their combination and the order of application have effects on the final data quality. One example is the order of application of probabilistic quotient normalization and missing values imputation: Normalization *before* imputation may require the exclusion of variables with missing values in order not to distort the estimation of the dilution factor. However, this approach is accompanied by massive loss of information for untargeted metabolomics data. In contrast, normalization *after* imputation would make use of all available data, but depending on the performance of the imputation method, the estimated dilution factor could be strongly biased. This example demonstrates that the entire preprocessing procedure as a whole and the influence of the single steps on each other need to be investigated to obtain a comprehensive standardized preprocessing pipeline for metabolomics studies.

A successful launch of such a study would require the evaluation of data quality based on artificial data where the ground truth is known and preprocessing can be performed in a completely controlled setting. However, the simulation of appropriate artificial data for such a study poses several challenges. First, simulated data should reflect a real data situation as best as possible. To this end, data characteristics driving each preprocessing step, such as technical or biological variation for normalization, and missing values patterns for imputation, must be identified. This is a rather challenging and sometimes even an unfeasible task. Second, these data characteristics can be fairly different between different types of data (e.g., MS-based data *versus* NMR data), or even between data sets of the same type (e.g., MS-based data measured in the 1990s and MS-based data measured only recently). That is, there is no guarantee that the simulated data reflects other data than the data at hand. Finally, the more data properties are considered the more complex is the simulation process.

4.3 Surrogate marker identification in multi-fluid data

In clinical medicine and research, *biomarkers* are medical signs with diagnostic or predictive power for a clinical endpoint. In some cases, the measurement of such a biomarker is technically challenging or fairly expensive. An alternative with the same descriptive and predictive power for such a biomarker, but which is easier and cheaper to obtain – here we will call it a *surrogate marker* – would be extremely convenient. An eligible surrogate for a known biomarker would be the same compound, a very similar compound, or a higly correlated compound that is measured in an easier accessible body fluid, for which reason multi-fluid data are exceptionally useful for the identification of surrogate markers. In our studies, we analyzed blood, urine, and saliva samples of the same individuals. Even more than blood, urine and saliva are samples that can be collected in the least invasive way, for which reason biomarkers in these fluids have considerable clinical advantages. For instance, a marker in saliva would enable a systematic screening for a given disease by dentists and oral hygienists with marginal efforts and expenses [9].

Often, biomarkers for a certain clinical endpoint are known from previous studies, but the clinical variable itself is not available in the data at hand, such that an association analysis of a candidate surrogate marker with the clinical variable is not feasible. Therefore, an interesting question is whether it is possible to propose a surrogate marker only based on its association with the known biomarker. Or formulated differently: if a biomarker is strongly associated with a given phenotype, and a candidate marker is strongly associated with the biomarker, under which condition is it possible to infer that the candidate marker is also associated with the phenotype? To answer this question, we started a pilot study in form of a master's thesis analyzing the occurrences of correlation triples in multi-fluid metabolomics data [120] with the conclusion that although Pearson correlations are not always transitive, the best candidate markers tend to be metabolites with very strong correlations to the given biomarker. However, much further work is needed to identify useful and clinically applicable surrogate markers. For instance, it could be possible that the surrogate marker of a metabolic biomarker is not a single other metabolite, but a set of different compounds.

4.4 Conclusion

Due to technical advancements and lower costs, a vast amount of high-dimensional and complex multi-fluid metabolomics data exists for which statistical approaches for appropriate data handling, analysis, and interpretation are in need. In this thesis, we addressed these needs by extensively evaluating existing approaches and developing novel strategies for preprocessing and statistical analysis of complex data. To the best of our knowledge, we were the first to systematically investigate multi-fluid metabolomics data from human *in vivo* samples. In particular, our approaches have been successfully applied in many follow-up studies and will continue to be valuable for future studies in biomedical research, aiding to move towards precision medicine.

Bibliography

[1] Mete Civelek and Aldons J. Lusis. Systems genetics approaches to understand complex traits. Nature Reviews Genetics, 15(1):34–48, January 2014. ISSN 1471-0056. doi: 10.1038/nrg3575. URL http://www.nature.com/nrg/journal/v15/n1/full/nrg3575.html.

[2] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews Molecular Cell Biology, 13(4):263–269, April 2012. ISSN 1471-0072. doi: 10.1038/nrm3314. URL http://www.nature.com/nrm/journal/v13/n4/full/nrm3314.html.

[3] Caroline H. Johnson, Andrew D. Patterson, Jeffrey R. Idle, and Frank J. Gonzalez. Xenobiotic Metabolomics: Major Impact on the Metabolome. Annual Review of Pharmacology and Toxicology, 52(1):37–56, 2012. doi: 10.1146/annurev-pharmtox-010611-134748. URL https://doi.org/10.1146/annurev-pharmtox-010611-134748.

[4] Jan Krumsiek, Jörg Bartel, and Fabian J Theis. Computational approaches for systems metabolomics. Current Opinion in Biotechnology, 39:198–206, June 2016. ISSN 0958-1669. doi: 10.1016/j.copbio.2016.04.009. URL http://www.sciencedirect.com/science/article/pii/S0958166916301173.

[5] Bo Peng, Hui Li, and Xuan-Xian Peng. Functional metabolomics: from biomarker discovery to metabolome reprogramming. Protein & Cell, 6(9):628–637, September 2015. ISSN 1674-800X. doi: 10.1007/s13238-015-0185-x. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4537470/.

[6] G. A. Nagana Gowda, Shucha Zhang, Haiwei Gu, Vincent Asiago, Narasimhamurthy Shanaiah, and Daniel Raftery. Metabolomics-Based Methods for Early Disease Diagnostics: A Review. Expert review of molecular diagnostics, 8(5):617, September 2008. ISSN 10.1586/14737159.8.5.617. doi: 10.1586/14737159.8.5.617. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3890417/.

[7] Marta Esteban and Argelia Castaño. Non-invasive matrices in human biomonitoring: A review. Environment International, 35(2):438–449, February 2009. ISSN 0160-4120. doi: 10.1016/j.envint.2008.09.003. URL http://www.sciencedirect.com/science/article/pii/S0160412008001992.

[8] David T. W. Wong. Salivaomics. Journal of the American Dental Association (1939), 143(10 Suppl):19S–24S, October 2012. ISSN 1943-4723.

[9] Dennis O. Mook-Kanamori, Mohammed M. El-Din Selim, Ahmed H. Takiddin, Hala Al-Homsi, Khoulood A. S. Al-Mahmoud, Amina Al-Obaidli, Mahmoud A. Zirie, Jillian Rowe, Noha A. Yousri, Edward D. Karoly, Thomas Kocher, Wafaa Sekkal Gherbi, Omar M. Chidiac, Marjonneke J. Mook-Kanamori, Sara Abdul Kader, Wadha A. Al Muftah, Cindy McKeon, and Karsten Suhre. 1,5-Anhydroglucitol in Saliva Is a Noninvasive Marker of Short-Term Glycemic Control. The Journal of Clinical Endocrinology & Metabolism, 99(3):E479–E483, January 2014. ISSN 0021-972X. doi: 10.1210/jc.2013-3596. URL http://press.endocrine.org/doi/abs/10.1210/jc.2013-3596.

[10] Souhaila Bouatra, Farid Aziat, Rupasri Mandal, An Chi Guo, Michael R. Wilson, Craig Knox, Trent C. Bjorndahl, Ramanarayan Krishnamurthy, Fozia Saleem, Philip Liu, Zerihun T. Dame, Jenna Poelzer, Jessica Huynh, Faizath S. Yallou, Nick Psychogios, Edison Dong, Ralf Bogumil, Cornelia Roehring, and David S. Wishart. The Human Urine Metabolome. PLoS ONE, 8(9), September 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0073076. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3762851/.

[11] Aihua Zhang, Hui Sun, Xiuhong Wu, and Xijun Wang. Urine metabolomics. Clinica Chimica Acta, 414(Supplement C):65–69, December 2012. ISSN 0009-8981. doi: 10.1016/j.cca.2012.08.016. URL http://www.sciencedirect.com/science/article/pii/S0009898112004135.

[12] Susanne Krug, Gabi Kastenmüller, Ferdinand Stückler, Manuela J. Rist, Thomas Skurk, Manuela Sailer, Johannes Raffler, Werner Römisch-Margl, Jerzy Adamski, Cornelia Prehn, Thomas Frank, Karl-Heinz Engel, Thomas Hofmann, Burkhard Luy, Ralf Zimmermann, Franco Moritz, Philippe Schmitt-Kopplin, Jan Krumsiek, Werner Kremer, Fritz Huber, Uwe Oeh, Fabian J. Theis, Wilfried Szymczak, Hans Hauner, Karsten Suhre, and Hannelore Daniel. The dynamic range of the human metabolome revealed by challenges. The FASEB Journal, 26(6):2607–2619, June 2012. ISSN 0892-6638, 1530-6860. doi: 10.1096/fj.11-198093. URL http://www.fasebj.org/content/26/6/2607.

[13] Jeremy K. Nicholson, John Connelly, John C. Lindon, and Elaine Holmes. Metabonomics: a platform for studying drug toxicity and gene function. Nature Reviews. Drug Discovery, 1(2):153–161, February 2002. ISSN 1474-1776. doi: 10.1038/nrd728.

[14] Nathan Blow. Metabolomics: Biochemistry's new look. Nature, 455(7213):697–700, 2008. ISSN 0028-0836. doi: 10.1038/455697a. URL http://www.nature.com/nature/journal/v455/n7213/full/455697a.html.

[15] Rima Kaddurah-Daouk, Bruce S. Kristal, and Richard M. Weinshilboum. Metabolomics: A Global Biochemical Approach to Drug Response and Disease. Annual Review of Pharmacology and Toxicology, 48(1):653–683, 2008. doi:

10.1146/annurev.pharmtox.48.113006.094715. URL http://dx.doi.org/10.1146/annurev.pharmtox.48.113006.094715.

[16] Marissa Fessenden. Metabolomics: Small molecules, single cells. Nature, 540(7631): 153–155, December 2016. ISSN 0028-0836. doi: 10.1038/540153a. URL http://www.nature.com/nature/journal/v540/n7631/full/540153a.html.

[17] T. W. M. Fan. 2.35 - Metabolomics-Edited Transcriptomics Analysis (Meta). In Charlene A. McQueen, editor, Comprehensive Toxicology (Second Edition), pages 685–706. Elsevier, Oxford, 2010. ISBN 978-0-08-046884-6. doi: 10.1016/B978-0-08-046884-6.00239-6. URL https://www.sciencedirect.com/science/article/pii/B9780080468846002396.

[18] Ho-Youn Kim, Hae-Rim Kim, and Sang-Heon Lee. Advances in Systems Biology Approaches for Autoimmune Diseases. Immune network, 14(2):73–80, April 2014. ISSN 1598-2629. doi: 10.4110/in.2014.14.2.73.

[19] Karsten Suhre. Metabolic profiling in diabetes. Journal of Endocrinology, 221(3): R75–R85, June 2014. ISSN 0022-0795, 1479-6805. doi: 10.1530/JOE-14-0024. URL http://joe.endocrinology-journals.org/content/221/3/R75.

[20] Christopher B. Newgard. Metabolomics and Metabolic Diseases: Where Do We Stand? Cell Metabolism, 25(1):43–56, January 2017. ISSN 1550-4131. doi: 10.1016/j.cmet.2016.09.018. URL http://www.cell.com/cell-metabolism/abstract/S1550-4131(16)30503-4.

[21] Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Frontiers in Bioengineering and Biotechnology, 3, March 2015. ISSN 2296-4185. doi: 10.3389/fbioe.2015.00023. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350445/.

[22] Jan Krumsiek, Karsten Suhre, Anne M Evans, Matthew W Mitchell, Robert P Mohney, Michael V Milburn, Brigitte Wägele, Werner Römisch-Margl, Thomas Illig, Jerzy Adamski, Christian Gieger, Fabian J Theis, and Gabi Kastenmüller. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS genetics, 8(10):e1003005, 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003005.

[23] Olav M. Kvalheim, Frode. Brakstad, and Yizeng. Liang. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. Analytical Chemistry, 66(1):43–51, January 1994. ISSN 0003-2700. doi: 10.1021/ac00073a010. URL http://dx.doi.org/10.1021/ac00073a010.

[24] Emily G. Armitage, Helen L. Kotze, and Kaye J. Williams. Correlation-based network analysis of cancer metabolism: A new systems biology approach in metabolomics. Springer, May 2014. ISBN 978-1-4939-0615-4. Google-Books-ID: p2cgBAAAQBAJ.

[25] G. E. P. Box and D. R. Cox. An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2):211–252, 1964. ISSN 0035-9246. URL http://www.jstor.org/stable/2984418.

[26] P. Filzmoser. A multivariate outlier detection method. In State University, pages 18–22, 2004.

[27] Henning Redestig and Ivan G. Costa. Detection and interpretation of metabolite–transcript coresponses using combined profiling data. Bioinformatics, 27(13): i357–i365, July 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr231. URL http://bioinformatics.oxfordjournals.org/content/27/13/i357.

[28] Emily Grace Armitage, Joanna Godzien, Vanesa Alonso-Herranz, Ángeles López-Gonzálvez, and Coral Barbas. Missing value imputation strategies for metabolomics data. Electrophoresis, 36(24):3050–3060, December 2015. ISSN 1522-2683. doi: 10.1002/elps.201500352.

[29] Olga Hrydziuszko and Mark R. Viant. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics, 8(1):161–174, October 2011. ISSN 1573-3882, 1573-3890. doi: 10.1007/s11306-011-0366-4. URL http://link.springer.com/article/10.1007/s11306-011-0366-4.

[30] Piotr S. Gromski, Yun Xu, Helen L. Kotze, Elon Correa, David I. Ellis, Emily Grace Armitage, Michael L. Turner, and Royston Goodacre. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. Metabolites, 4(2): 433–452, June 2014. doi: 10.3390/metabo4020433. URL http://www.mdpi.com/2218-1989/4/2/433.

[31] Jianguo Xia, Nick Psychogios, Nelson Young, and David S. Wishart. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research, 37(Web Server issue):W652–W660, July 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp356. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703878/.

[32] Haiying Chen, Sara A. Quandt, Joseph G. Grzywacz, and Thomas A. Arcury. A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection. Environmental Health Perspectives, 119(3):351–356, March 2011. ISSN 0091-6765. doi: 10.1289/ehp.1002124. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059998/.

[33] Kieu Trinh Do, Gabi Kastenmüller, Dennis O. Mook-Kanamori, Noha A. Yousri, Fabian J. Theis, Karsten Suhre, and Jan Krumsiek. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. Journal of Proteome Research, 14(2):1183–1194, February 2015. ISSN 1535-3907. doi: 10.1021/pr501130a.

[34] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. Bioinformatics (Oxford, England), 17(6):520–525, June 2001. ISSN 1367-4803.

[35] Ralf Steuer, Katja Morgenthal, Wolfram Weckwerth, and Joachim Selbig. A gentle guide to the analysis of metabolomic data. Methods in Molecular Biology (Clifton, N.J.), 358:105–126, 2007. ISSN 1064-3745. doi: 10.1007/978-1-59745-244-1_7.

[36] Stef van Buuren and Karin Groothuis-Oudshoorn. Flexible Multivariate Imputation by MICE | BibSonomy, 1999. URL `https://www.bibsonomy.org/bibtex/137c5545c8d791a83353823e983338e4e`.

[37] Gerhard Tutz and Shahla Ramzan. Improved methods for the imputation of missing data by nearest neighbor methods. Computational Statistics & Data Analysis, 90:84–99, 2015. ISSN 0167-9473. doi: 10.1016/j.csda.2015.04.009. URL `http://www.sciencedirect.com/science/article/pii/S0167947315001061`.

[38] Florian Buettner, Victoria Moignard, Berthold Göttgens, and Fabian J. Theis. Probabilistic PCA of censored data: accounting for uncertainties in the visualization of high-throughput single-cell qPCR data. Bioinformatics, page btu134, March 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu134. URL `http://bioinformatics.oxfordjournals.org/content/early/2014/03/29/bioinformatics.btu134`.

[39] Laleh Haghverdi, Florian Buettner, and Fabian J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics, 31(18):2989–2998, September 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv325. URL `https://academic.oup.com/bioinformatics/article/31/18/2989/241305`.

[40] Sandra L. Taylor, L. Renee Ruhaak, Karen Kelly, Robert H. Weiss, and Kyoungmi Kim. Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. Briefings in Bioinformatics, February 2016. ISSN 1477-4054. doi: 10.1093/bib/bbw010.

[41] Riccardo Di Guida, Jasper Engel, J. William Allwood, Ralf J. M. Weber, Martin R. Jones, Ulf Sommer, Mark R. Viant, and Warwick B. Dunn. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics, 12, 2016. ISSN 1573-3882. doi: 10.1007/s11306-016-1030-9. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831991/`.

[42] Jasmit S. Shah, Shesh N. Rai, Andrew P. DeFilippis, Bradford G. Hill, Aruni Bhatnagar, and Guy N. Brock. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. BMC Bioinformatics, 18, February 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1547-6. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319174/`.

[43] Helena Gibbons, Aoife O'Gorman, and Lorraine Brennan. Metabolomics as a tool in nutritional research. Current Opinion in Lipidology, 26(1):30–34, February 2015. ISSN 1473-6535. doi: 10.1097/MOL.0000000000000140.

[44] Dohoon Kim, Brian P. Fiske, Kivanc Birsoy, Elizaveta Freinkman, Kenjiro Kami, Richard Possemato, Yakov Chudnovsky, Michael E. Pacold, Walter W. Chen, Jason R. Cantor, Laura M. Shelton, Dan Y. Gui, Manjae Kwon, Shakti H. Ramkissoon, Keith L. Ligon, Seong Woo Kang, Matija Snuderl, Matthew G. Vander Heiden, and David M. Sabatini. SHMT2 drives glioma cell survival in the tumor microenvironment but imposes a dependence on glycine clearance. Nature, 520(7547):363–367, April 2015. ISSN 0028-0836. doi: 10.1038/nature14363. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533874/`.

[45] Jeny Ghartey, Jamie A. Bastek, Amy G. Brown, Laura Anglim, and Michal A. Elovitz. Women with preterm birth have a distinct cervicovaginal metabolome. American Journal of Obstetrics and Gynecology, 212(6):776.e1–776.e12, June 2015. ISSN 0002-9378. doi: 10.1016/j.ajog.2015.03.052. URL `http://www.sciencedirect.com/science/article/pii/S0002937815003336`.

[46] Annalaura Mastrangelo, Emily G. Armitage, Antonia García, and Coral Barbas. Metabolomics as a tool for drug discovery and personalised medicine. A review. Current Topics in Medicinal Chemistry, 14(23):2627–2636, 2014. ISSN 1873-4294.

[47] Daniel Krug and Rolf Müller. Secondary metabolomics: the impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products. Natural Product Reports, 31(6):768–783, June 2014. ISSN 1460-4752. doi: 10.1039/c3np70127a.

[48] Christopher B. Newgard, Jie An, James R. Bain, Michael J. Muehlbauer, Robert D. Stevens, Lillian F. Lien, Andrea M. Haqq, Svati H. Shah, Michelle Arlotto, Cris A. Slentz, James Rochon, Dianne Gallup, Olga Ilkayeva, Brett R. Wenner, William S. Yancy, Howard Eisenson, Gerald Musante, Richard S. Surwit, David S. Millington, Mark D. Butler, and Laura P. Svetkey. 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, April 2009. ISSN 1550-4131. doi: 10.1016/j.cmet.2009.02.002. URL `http://www.sciencedirect.com/science/article/pii/S1550413109000400`.

[49] Cristina Menni, Gabriella Kastenmüller, Ann Kristin Petersen, Jordana T. Bell, Maria Psatha, Pei-Chien Tsai, Christian Gieger, Holger Schulz, Idil Erte, Sally John, M. Julia Brosnan, Scott G. Wilson, Loukia Tsaprouni, Ee Mun Lim, Bronwyn Stuckey, Panos Deloukas, Robert Mohney, Karsten Suhre, Tim D. Spector, and Ana M. Valdes. Metabolomic markers reveal novel pathways of ageing and early development in human populations. International Journal of Epidemiology, 42(4): 1111–1119, August 2013. ISSN 1464-3685. doi: 10.1093/ije/dyt094.

[50] Peter Würtz, Ville-Petteri Mäkinen, Pasi Soininen, Antti J. Kangas, Taru Tukiainen, Johannes Kettunen, Markku J. Savolainen, Tuija Tammelin, Jorma S. Viikari, Tapani Rönnemaa, Mika Kähönen, Terho Lehtimäki, Samuli Ripatti, Olli T. Raitakari, Marjo-Riitta Järvelin, and Mika Ala-Korpela. Metabolic Signatures of Insulin Resistance in 7,098 Young Adults. Diabetes, 61(6):1372–1380, June 2012. ISSN

0012-1797. doi: 10.2337/db11-1355. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3357275/.

[51] Nicholette D. Palmer, Robert D. Stevens, Peter A. Antinozzi, Andrea Anderson, Richard N. Bergman, Lynne E. Wagenknecht, Christopher B. Newgard, and Donald W. Bowden. Metabolomic Profile Associated With Insulin Resistance and Conversion to Diabetes in the Insulin Resistance Atherosclerosis Study. The Journal of Clinical Endocrinology & Metabolism, 100(3):E463–E468, March 2015. ISSN 0021-972X. doi: 10.1210/jc.2014-2357. URL https://academic.oup.com/jcem/article/100/3/E463/2839996/Metabolomic-Profile-Associated-With-Insulin.

[52] Thomas J. Wang, Martin G. Larson, Ramachandran S. Vasan, Susan Cheng, Eugene P. Rhee, Elizabeth McCabe, Gregory D. Lewis, Caroline S. Fox, Paul F. Jacques, Céline Fernandez, Christopher J. O'Donnell, Stephen A. Carr, Vamsi K. Mootha, Jose C. Florez, Amanda Souza, Olle Melander, Clary B. Clish, and Robert E. Gerszten. Metabolite Profiles and the Risk of Developing Diabetes. Nature medicine, 17 (4):448–453, April 2011. ISSN 1078-8956. doi: 10.1038/nm.2307. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126616/.

[53] Walter E Gall, Kirk Beebe, Kay A Lawton, Klaus-Peter Adam, Matthew W Mitchell, Pamela J Nakhle, John A Ryals, Michael V Milburn, Monica Nannipieri, Stefania Camastra, Andrea Natali, Ele Ferrannini, and RISC Study Group. alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. PloS one, 5(5):e10883, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010883.

[54] Rui Wang-Sattler, Zhonghao Yu, Christian Herder, Ana C. Messias, Anna Floegel, Ying He, Katharina Heim, Monica Campillos, Christina Holzapfel, Barbara Thorand, Harald Grallert, Tao Xu, Erik Bader, Cornelia Huth, Kirstin Mittelstrass, Angela Döring, Christa Meisinger, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, Maren Carstensen, Lu Xie, Hisami Yamanaka-Okumura, Guihong Xing, Uta Ceglarek, Joachim Thiery, Guido Giani, Heiko Lickert, Xu Lin, Yixue Li, Heiner Boeing, Hans-Georg Joost, Martin Hrabé de Angelis, Wolfgang Rathmann, Karsten Suhre, Holger Prokisch, Annette Peters, Thomas Meitinger, Michael Roden, H.-Erich Wichmann, Tobias Pischon, Jerzy Adamski, and Thomas Illig. Novel biomarkers for pre-diabetes identified by metabolomics. Molecular Systems Biology, 8(1):615, January 2012. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2012.43. URL http://msb.embopress.org/content/8/1/615.

[55] Anna E. Thalacker-Mercer, Katherine H. Ingram, Fangjian Guo, Olga Ilkayeva, Christopher B. Newgard, and W. Timothy Garvey. BMI, RQ, Diabetes, and Sex Affect the Relationships Between Amino Acids and Clamp Measures of Insulin Action in Humans. Diabetes, 63(2):791–800, February 2014. ISSN 0012-1797, 1939-327X. doi: 10.2337/db13-0396. URL http://diabetes.diabetesjournals.org/content/63/2/791.

[56] Thomas J. Wang, Debby Ngo, Nikolaos Psychogios, Andre Dejam, Martin G. Larson, Ramachandran S. Vasan, Anahita Ghorbani, John O'Sullivan, Susan Cheng, Eugene P.

Rhee, Sumita Sinha, Elizabeth McCabe, Caroline S. Fox, Christopher J. O'Donnell, Jennifer E. Ho, Jose C. Florez, Martin Magnusson, Kerry A. Pierce, Amanda L. Souza, Yi Yu, Christian Carter, Peter E. Light, Olle Melander, Clary B. Clish, and Robert E. Gerszten. 2-Aminoadipic acid is a biomarker for diabetes risk. The Journal of Clinical Investigation, 123(10):4309–4317, October 2013. ISSN 0021-9738. doi: 10.1172/JCI64801. URL `http://www.jci.org/articles/view/64801`.

[57] Jeff Cobb, Andrea Eckhart, Alison Motsinger-Reif, Bernadette Carr, Leif Groop, and Ele Ferrannini. alpha-Hydroxybutyric Acid Is a Selective Metabolite Biomarker of Impaired Glucose Tolerance. Diabetes Care, 39(6):988–995, June 2016. ISSN 0149-5992, 1935-5548. doi: 10.2337/dc15-2752. URL `http://care.diabetesjournals.org/content/39/6/988`.

[58] Svati H. Shah, James R. Bain, Michael J. Muehlbauer, Robert D. Stevens, David R. Crosslin, Carol Haynes, Jennifer Dungan, L. Kristin Newby, Elizabeth R. Hauser, Geoffrey S. Ginsburg, Christopher B. Newgard, and William E. Kraus. Association of a Peripheral Blood Metabolic Profile With Coronary Artery Disease and Risk of Subsequent Cardiovascular Events. Circulation: Cardiovascular Genetics, 3(2):207–214, April 2010. ISSN 1942-325X, 1942-3268. doi: 10.1161/CIRCGENETICS.109.852814. URL `http://circgenetics.ahajournals.org/content/3/2/207`.

[59] Svati H. Shah, William E. Kraus, and Christopher B. Newgard. Metabolomic Profiling for the Identification of Novel Biomarkers and Mechanisms Related to Common Cardiovascular Diseases Form and Function. Circulation, 126(9):1110–1120, August 2012. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.111.060368. URL `http://circ.ahajournals.org/content/126/9/1110`.

[60] Zeneng Wang, Elizabeth Klipfell, Brian J. Bennett, Robert Koeth, Bruce S. Levison, Brandon DuGar, Ariel E. Feldstein, Earl B. Britt, Xiaoming Fu, Yoon-Mi Chung, Yuping Wu, Phil Schauer, Jonathan D. Smith, Hooman Allayee, W. H. Wilson Tang, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature, 472(7341):57–63, April 2011. ISSN 0028-0836. doi: 10.1038/nature09922. URL `http://www.nature.com/nature/journal/v472/n7341/full/nature09922.html`.

[61] W. H. Wilson Tang, Zeneng Wang, Leslie Cho, Danielle M. Brennan, and Stanley L. Hazen. Diminished Global Arginine Bioavailability and Increased Arginine Catabolism as Metabolic Profile of Increased Cardiovascular Risk. Journal of the American College of Cardiology, 53(22):2061–2067, June 2009. ISSN 0735-1097. doi: 10.1016/j.jacc.2009.02.036. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755213/`.

[62] Peter Würtz, Aki S. Havulinna, Pasi Soininen, Tuulia Tynkkynen, David Prieto-Merino, Therese Tillin, Anahita Ghorbani, Anna Artati, Qin Wang, Mika Tiainen, Antti J. Kangas, Johannes Kettunen, Jari Kaikkonen, Vera Mikkilä, Antti Jula, Mika Kähönen, Terho Lehtimäki, Debbie A. Lawlor, Tom R. Gaunt, Alun D. Hughes, Naveed Sattar, Thomas Illig, Jerzy Adamski, Thomas J. Wang, Markus Perola, Samuli

Ripatti, Ramachandran S. Vasan, Olli T. Raitakari, Robert E. Gerszten, Juan-Pablo Casas, Nish Chaturvedi, Mika Ala-Korpela, and Veikko Salomaa. Metabolite Profiling and Cardiovascular Event Risk: A Prospective Study of Three Population-Based Cohorts. Circulation, page CIRCULATIONAHA.114.013116, January 2015. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.114.013116. URL http://circ.ahajournals.org/content/early/2015/01/08/CIRCULATIONAHA.114.013116.

[63] David B. Liesenfeld, Nina Habermann, Robert W. Owen, Augustin Scalbert, and Cornelia M. Ulrich. Review of mass spectrometry-based metabolomics in cancer research. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 22(12):2182–2201, December 2013. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-13-0584. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912559/.

[64] J. S. Ried, H. Baurecht, F. Stückler, J. Krumsiek, C. Gieger, J. Heinrich, M. Kabesch, C. Prehn, A. Peters, E. Rodriguez, H. Schulz, K. Strauch, K. Suhre, R. Wang-Sattler, H.-E. Wichmann, F. J. Theis, T. Illig, J. Adamski, and S. Weidinger. Integrative genetic and metabolite profiling analysis suggests altered phosphatidylcholine metabolism in asthma. Allergy, 68(5):629–636, 2013. ISSN 1398-9995. doi: 10.1111/all.12110. URL http://onlinelibrary.wiley.com/doi/10.1111/all.12110/abstract.

[65] Bruce A. Luxon. Metabolomics in asthma. Advances in Experimental Medicine and Biology, 795:207–220, 2014. ISSN 0065-2598. doi: 10.1007/978-1-4614-8603-9_13.

[66] Kedir N. Turi, Lindsey Romick-Rosendale, Kelli K. Ryckman, and Tina V. Hartert. A review of metabolomics approaches and their application in identifying causal pathways of childhood asthma. Journal of Allergy and Clinical Immunology, May 2017. ISSN 0091-6749. doi: 10.1016/j.jaci.2017.04.021. URL http://www.sciencedirect.com/science/article/pii/S0091674917307443.

[67] Teklab Gebregiworgis, Helle H. Nielsen, Chandirasegaran Massilamany, Arunakumar Gangaplara, Jay Reddy, Zsolt Illes, and Robert Powers. A Urinary Metabolic Signature for Multiple Sclerosis and Neuromyelitis Optica. Journal of Proteome Research, 15(2):659–666, February 2016. ISSN 1535-3893. doi: 10.1021/acs.jproteome.5b01111. URL http://dx.doi.org/10.1021/acs.jproteome.5b01111.

[68] Eva C. Schulte, Elisabeth Altmaier, Hannah S. Berger, Kieu Trinh Do, Gabi Kastenmüller, Simone Wahl, Jerzy Adamski, Annette Peters, Jan Krumsiek, Karsten Suhre, Bernhard Haslinger, Andres Ceballos-Baumann, Christian Gieger, and Juliane Winkelmann. Alterations in Lipid and Inositol Metabolisms in Two Dopaminergic Disorders. PloS One, 11(1):e0147129, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0147129.

[69] Christopher B. Newgard. Interplay between Lipids and Branched-Chain Amino Acids in Development of Insulin Resistance. Cell Metabolism, 15(5):606–614, May 2012.

ISSN 1550-4131. doi: 10.1016/j.cmet.2012.01.024. URL http://www.sciencedirect.com/science/article/pii/S1550413112001039.

[70] Vanessa K. Ridaura, Jeremiah J. Faith, Federico E. Rey, Jiye Cheng, Alexis E. Duncan, Andrew L. Kau, Nicholas W. Griffin, Vincent Lombard, Bernard Henrissat, James R. Bain, Michael J. Muehlbauer, Olga Ilkayeva, Clay F. Semenkovich, Katsuhiko Funai, David K. Hayashi, Barbara J. Lyle, Margaret C. Martini, Luke K. Ursell, Jose C. Clemente, William Van Treuren, William A. Walters, Rob Knight, Christopher B. Newgard, Andrew C. Heath, and Jeffrey I. Gordon. Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice. Science, 341(6150):1241214, September 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1241214. URL http://science.sciencemag.org/content/341/6150/1241214.

[71] Hong Zheng, Christian C. Yde, Karina Arnberg, M&#xf8, Christian Lgaard, Kim F. Michaelsen, Larnkj&#xE6, Anni R, Hanne C. Bertram, Hong Zheng, Christian C. Yde, Karina Arnberg, M&#xf8, Christian Lgaard, Kim F. Michaelsen, Larnkj&#xE6, Anni R, and Hanne C. Bertram. NMR-Based Metabolomic Profiling of Overweight Adolescents: An Elucidation of the Effects of Inter-/Intraindividual Differences, Gender, and Pubertal Development, NMR-Based Metabolomic Profiling of Overweight Adolescents: An Elucidation of the Effects of Inter-/Intraindividual Differences, Gender, and Pubertal Development. BioMed Research International, BioMed Research International, 2014, 2014:e537157, March 2014. ISSN 2314-6133, 2314-6133. doi: 10.1155/2014/537157,10.1155/2014/537157. URL http://www.hindawi.com/journals/bmri/2014/537157/abs/,http://www.hindawi.com/journals/bmri/2014/537157/abs/.

[72] Danuta Dudzik, Marcin Zorawski, Mariusz Skotnicki, Wieslaw Zarzycki, Gabryela Kozlowska, Katarzyna Bibik-Malinowska, María Vallejo, Antonia García, Coral Barbas, and M. Pilar Ramos. Metabolic fingerprint of Gestational Diabetes Mellitus. Journal of Proteomics, 103:57–71, 2014. ISSN 1874-3919. doi: 10.1016/j.jprot.2014.03.025. URL http://www.sciencedirect.com/science/article/pii/S187439191400147X.

[73] Claudia Balderas, Francisco Javier Rupérez, Elena Ibañez, Javier Señorans, Julio Guerrero-Fernández, Isabel González Casado, Ricardo Gracia-Bouthelier, Antonia García, and Coral Barbas. Plasma and urine metabolic fingerprinting of type 1 diabetic children. Electrophoresis, 34(19):2882–2890, October 2013. ISSN 1522-2683. doi: 10.1002/elps.201300062.

[74] Saif Ullah Munshi, Bharat Bhushan Rewari, Neel Sarovar Bhavesh, and Shahid Jameel. Nuclear Magnetic Resonance Based Profiling of Biofluids Reveals Metabolic Dysregulation in HIV-Infected Persons and Those on Anti-Retroviral Therapy. PLoS ONE, 8(5), May 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0064298. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3655987/.

[75] Ji Won Kim, Sung Ha Ryu, Siwon Kim, Hae Won Lee, Mi-sun Lim, Sook Jin Seong, Suhkmann Kim, Young-Ran Yoon, and Kyu-Bong Kim. Pattern Recognition

Analysis for Hepatotoxicity Induced by Acetaminophen Using Plasma and Urinary 1h NMR-Based Metabolomics in Humans. Analytical Chemistry, 85(23):11326–11334, 2013. ISSN 0003-2700. doi: 10.1021/ac402390q. URL http://dx.doi.org/10.1021/ac402390q.

[76] Marianne C. Walsh, Lorraine Brennan, J. Paul G. Malthouse, Helen M. Roche, and Michael J. Gibney. Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. The American Journal of Clinical Nutrition, 84(3):531–539, September 2006. ISSN 0002-9165.

[77] Anne Sofie Korsholm, Thomas Nordstrøm Kjær, Marie Juul Ornstrup, and Steen Bønløkke Pedersen. Comprehensive Metabolomic Analysis in Blood, Urine, Fat, and Muscle in Men with Metabolic Syndrome: A Randomized, Placebo-Controlled Clinical Trial on the Effects of Resveratrol after Four Months' Treatment. International Journal of Molecular Sciences, 18(3), March 2017. ISSN 1422-0067. doi: 10.3390/ijms18030554. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5372570/.

[78] Aram Adourian, Ezra Jennings, Raji Balasubramanian, Wade M. Hines, Doris Damian, Thomas N. Plasterer, Clary B. Clish, Paul Stroobant, Robert McBurney, Elwin R. Verheij, Ivana Bobeldijk, Jan van der Greef, Johan Lindberg, Kerstin Kenne, Ulf Andersson, Heike Hellmold, Kerstin Nilsson, Hugh Salter, and Ina Schuppe-Koistinen. Correlation network analysis for data integration and biomarker selection. Molecular BioSystems, 4(3):249–259, February 2008. ISSN 1742-2051. doi: 10.1039/B708489G. URL http://pubs.rsc.org/en/content/articlelanding/2008/mb/b708489g.

[79] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, 40(Database issue):D109–114, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr988.

[80] David S. Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie-Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David. D. Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alim M. Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: the Human Metabolome Database. Nucleic Acids Research, 35(Database issue):D521–D526, January 2007. ISSN 0305-1048. doi: 10.1093/nar/gkl923. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1899095/.

[81] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong,

Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research, 42(D1):D459–D471, January 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1103. URL http://nar.oxfordjournals.org/content/42/D1/D459.

[82] Natalie C. Duarte, Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences, 104(6):1777–1782, February 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0610772104. URL http://www.pnas.org/content/104/6/1777.

[83] Nicolas Alcaraz, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, Josch Pauling, and Jan Baumbach. Efficient key pathway mining: combining networks and OMICS data. Integrative Biology: Quantitative Biosciences from Nano to Macro, 4(7):756–764, July 2012. ISSN 1757-9708. doi: 10.1039/c2ib00133k.

[84] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. Nature reviews. Genetics, 14(10):719–732, October 2013. ISSN 1471-0056. doi: 10.1038/nrg3552. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3940161/.

[85] Krzysztof Polanski, Johanna Rhodes, Claire Hill, Peijun Zhang, Dafyd J. Jenkins, Steven J. Kiddle, Aleksey Jironkin, Jim Beynon, Vicky Buchanan-Wollaston, Sascha Ott, and Katherine J. Denby. Wigwams: identifying gene modules co-regulated across multiple biological conditions. Bioinformatics, 30(7):962–970, April 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt728. URL http://bioinformatics.oxfordjournals.org/content/30/7/962.

[86] Yoo-Ah Kim, Dong-Yeon Cho, Phuong Dao, and Teresa M. Przytycka. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics (Oxford, England), 31(12): i284–292, June 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv247.

[87] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. Molecular Systems Biology, 3(1), January 2007. doi: 10.1038/msb4100180. URL http://msb.embopress.org/content/3/1/140. Mapping the pathways that give rise to metastasis is one of the key challenges of breast cancer research. Recently, several large-scale studies have shed light on this problem through analysis of gene expression profiles to identify markers correlated with metastasis. Here, we apply a protein-network-based approach that identifies markers not as individual genes but as subnetworks extracted from protein interaction databases. The resulting subnetworks provide novel hypotheses for pathways involved in tumor progression. Although genes with known breast cancer mutations are typically not detected through analysis of differential expression, they play a central role in the protein network by interconnecting many differentially

expressed genes. We find that the subnetwork markers are more reproducible than individual marker genes selected without network information, and that they achieve higher accuracy in the classification of metastatic versus non-metastatic tumors.

[88] Ali May, Bernd W. Brandt, Mohammed El-Kebir, Gunnar W. Klau, Egija Zaura, Wim Crielaard, Jaap Heringa, and Sanne Abeln. metaModules identifies key functional subnetworks in microbiome-related disease. Bioinformatics, page btv526, September 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv526. URL `http://bioinformatics.oxfordjournals.org/content/early/2015/09/24/bioinformatics.btv526`.

[89] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC systems biology, 5:21, 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-21.

[90] Kirstin Mittelstrass, Janina S. Ried, Zhonghao Yu, Jan Krumsiek, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, Alexey Polonikov, Annette Peters, Fabian J. Theis, Thomas Meitinger, Florian Kronenberg, Stephan Weidinger, Heinz Erich Wichmann, Karsten Suhre, Rui Wang-Sattler, Jerzy Adamski, and Thomas Illig. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. PLoS Genetics, 7(8), August 2011. ISSN 1553-7390. doi: 10.1371/journal.pgen.1002215. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154959/`.

[91] A Floegel, A Wientzek, U Bachlechner, S Jacobs, D Drogan, C Prehn, J Adamski, J Krumsiek, M B Schulze, T Pischon, and H Boeing. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. International Journal of Obesity (2005), 38(11):1388–1396, November 2014. ISSN 0307-0565. doi: 10.1038/ijo.2014.39. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4229626/`.

[92] Elisabeth Altmaier, Rebecca T. Emeny, Jan Krumsiek, Maria E. Lacruz, Karoline Lukaschek, Sibylle Häfner, Gabi Kastenmüller, Werner Römisch-Margl, Cornelia Prehn, Robert P. Mohney, Anne M. Evans, Michael V. Milburn, Thomas Illig, Jerzy Adamski, Fabian Theis, Karsten Suhre, and Karl-Heinz Ladwig. Metabolomic profiles in individuals with negative affectivity and social inhibition: A population-based study of Type D personality. Psychoneuroendocrinology, 38(8):1299–1309, August 2013. ISSN 0306-4530. doi: 10.1016/j.psyneuen.2012.11.014. URL `http://www.psyneuen-journal.com/article/S0306453012003836/abstract`.

[93] Chaoyang Hu, Jianxin Shi, Sheng Quan, Bo Cui, Sabrina Kleessen, Zoran Nikoloski, Takayuki Tohge, Danny Alexander, Lining Guo, Hong Lin, Jing Wang, Xiao Cui, Jun Rao, Qian Luo, Xiangxiang Zhao, Alisdair R. Fernie, and Dabing Zhang. Metabolic variation between japonica and indica rice cultivars as revealed by non-targeted metabolomics. Scientific Reports, 4:5067, May 2014. ISSN 2045-2322. doi: 10.1038/srep05067. URL `http://www.nature.com/srep/2014/140527/srep05067/full/srep05067.html`.

[94] Laura E. McNamara, Terje Sjöström, R. M. Dominic Meek, Richard O. C. Oreffo, Bo Su, Matthew J. Dalby, and Karl E. V. Burgess. Metabolomics: a valuable tool for stem cell monitoring in regenerative medicine. Journal of The Royal Society Interface, 9(73):1713–1724, August 2012. ISSN 1742-5689, 1742-5662. doi: 10.1098/rsif.2012.0169. URL http://rsif.royalsocietypublishing.org/content/9/73/1713.

[95] Jan Krumsiek, Kirstin Mittelstrass, Kieu Trinh Do, Ferdinand Stückler, Janina Ried, Jerzy Adamski, Annette Peters, Thomas Illig, Florian Kronenberg, Nele Friedrich, Matthias Nauck, Maik Pietzner, Dennis O. Mook-Kanamori, Karsten Suhre, Christian Gieger, Harald Grallert, Fabian J. Theis, and Gabi Kastenmüller. Gender-specific pathway differences in the human serum metabolome. Metabolomics, 11(6):1815–1833, 2015. ISSN 1573-3882. doi: 10.1007/s11306-015-0829-0. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4605991/.

[96] R. Holle, M. Happich, H. Löwel, H. E. Wichmann, and MONICA/KORA Study Group. KORA–a research platform for population based health research. Gesundheitswesen (Bundesverband Der Ärzte Des Öffentlichen Gesundheitsdienstes (Germany)), 67 Suppl 1:S19–25, August 2005. ISSN 0941-3790. doi: 10.1055/s-2005-858235.

[97] Kieu Trinh Do, Simone Wahl, Johannes Raffler, Sophie Molnos, Michael Laimighofer, Jerzy Adamsky, Karsten Suhre, Konstantin Strauch, Annette Peters, Christian Gieger, Claudia Langenberg, Isobel Stewart, Fabian J. Theis, Harald Grallert, Gabi Kastenmueller, and Jan Krumsiek. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. bioRxiv, page 260281, March 2018. doi: 10.1101/260281. URL https://www.biorxiv.org/content/early/2018/03/02/260281.

[98] Dennis R. Helsel. Less than obvious - statistical treatment of data below the detection limit. Environmental Science & Technology, 24(12):1766–1774, 1990. ISSN 0013-936X. doi: 10.1021/es00082a001. URL http://dx.doi.org/10.1021/es00082a001.

[99] Dennis R. Helsel. More than obvious: better methods for interpreting nondetect data. Environmental Science & Technology, 39(20):419A–423A, October 2005. ISSN 0013-936X.

[100] David B. Richardson and Antonio Ciampi. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. American Journal of Epidemiology, 157(4):355–363, February 2003. ISSN 0002-9262.

[101] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple Imputation by Chained Equations: What is it and how does it work? International journal of methods in psychiatric research, 20(1):40–49, March 2011. ISSN 1049-8931. doi: 10.1002/mpr.329. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/.

[102] Donald B. Rubin. Introduction. In Multiple Imputation for Nonresponse in Surveys, pages 1–26. John Wiley & Sons, Inc., 1987. ISBN 978-0-470-31669-6. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470316696.ch1/summary.

[103] Andrea Marshall, Douglas G Altman, Roger L Holder, and Patrick Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology, 9:57, July 2009. ISSN 1471-2288. doi: 10.1186/1471-2288-9-57. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727536/`.

[104] Kieu Trinh Do, Maik Pietzner, David Jnp Rasp, Nele Friedrich, Matthias Nauck, Thomas Kocher, Karsten Suhre, Dennis O. Mook-Kanamori, Gabi Kastenmüller, and Jan Krumsiek. Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations. NPJ systems biology and applications, 3:28, 2017. ISSN 2056-7189. doi: 10.1038/s41540-017-0029-9.

[105] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology, 1(1):54, November 2007. ISSN 1752-0509. doi: 10.1186/1752-0509-1-54. URL `http://www.biomedcentral.com/1752-0509/1/54/abstract`.

[106] Juliane Schäfer and Korbinian Strimmer. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology, 4(1), 2005. ISSN 1544-6115. doi: 10.2202/1544-6115.1175. URL `https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1175/sagmb.2005.4.1.1175.xml`.

[107] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2):365–411, February 2004. ISSN 0047-259X. doi: 10.1016/S0047-259X(03)00096-4. URL `http://www.sciencedirect.com/science/article/pii/S0047259X03000964`.

[108] Kieu Trinh Do, David J. N.-P. Rasp, Gabi Kastenmueller, Karsten Suhre, and Jan Krumsiek. MoDentify: a tool for phenotype-driven module identification in multilevel metabolomics networks. bioRxiv, page 275057, March 2018. doi: 10.1101/275057. URL `https://www.biorxiv.org/content/early/2018/03/04/275057`.

[109] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11):2498–2504, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.

[110] Paul T. Shannon, Mark Grimes, Burak Kutlu, Jan J. Bot, and David J. Galas. RCytoscape: tools for exploratory network analysis. BMC Bioinformatics, 14:217, July 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-217. URL `https://doi.org/10.1186/1471-2105-14-217`.

[111] Noha A. Yousri, Dennis O. Mook-Kanamori, Mohammed M. El-Din Selim, Ahmed H. Takiddin, Hala Al-Homsi, Khoulood A. S. Al-Mahmoud, Edward D. Karoly, Jan Krumsiek, Kieu Thinh Do, Ulrich Neumaier, Marjonneke J. Mook-Kanamori, Jillian Rowe, Omar M. Chidiac, Cindy McKeon, Wadha A. Al Muftah, Sara Abdul

Kader, Gabi Kastenmüller, and Karsten Suhre. A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. Diabetologia, 58(8):1855–1867, 2015. ISSN 0012-186X. doi: 10.1007/s00125-015-3636-2. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4499109/.

[112] Henrike Knacke, Maik Pietzner, Kieu Trinh Do, Werner Römisch-Margl, Gabi Kastenmüller, Uwe Völker, Henry Völzke, Jan Krumsiek, Anna Artati, Henri Wallaschofski, Matthias Nauck, Karsten Suhre, Jerzy Adamski, and Nele Friedrich. Metabolic fingerprints of circulating IGF-I and the IGF-I/IGFBP-3 ratio: a multifluid metabolomics study. The Journal of Clinical Endocrinology and Metabolism, page jc20162588, October 2016. ISSN 1945-7197. doi: 10.1210/jc.2016-2588.

[113] Harold Hotelling. RELATIONS BETWEEN TWO SETS OF VARIATES. Biometrika, 28(3-4):321–377, December 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL https://academic.oup.com/biomet/article/28/3-4/321/220073/RELATIONS-BETWEEN-TWO-SETS-OF-VARIATES.

[114] Johan Trygg. O2-PLS for qualitative and quantitative analysis in multivariate calibration. Journal of Chemometrics, 16(6):283–293, June 2002. ISSN 1099-128X. doi: 10.1002/cem.724. URL http://onlinelibrary.wiley.com/doi/10.1002/cem.724/abstract.

[115] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. Science (New York, N.Y.), 220(4598):671–680, May 1983. ISSN 0036-8075. doi: 10.1126/science.220.4598.671.

[116] Neil Swainston, Kieran Smallbone, Hooman Hefzi, Paul D. Dobson, Judy Brewer, Michael Hanscho, Daniel C. Zielinski, Kok Siong Ang, Natalie J. Gardiner, Jahir M. Gutierrez, Sarantos Kyriakopoulos, Meiyappan Lakshmanan, Shangzhong Li, Joanne K. Liu, Veronica S. Martínez, Camila A. Orellana, Lake-Ee Quek, Alex Thomas, Juergen Zanghellini, Nicole Borth, Dong-Yup Lee, Lars K. Nielsen, Douglas B. Kell, Nathan E. Lewis, and Pedro Mendes. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics, 12, 2016. ISSN 1573-3882. doi: 10.1007/s11306-016-1051-4. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896983/.

[117] Stefanie M. Kohl, Matthias S. Klein, Jochen Hochrein, Peter J. Oefner, Rainer Spang, and Wolfram Gronwald. State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics, 8(Suppl 1):146–160, June 2012. ISSN 1573-3882. doi: 10.1007/s11306-011-0350-z. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337420/.

[118] Ibrahim Karaman. Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis. Advances in Experimental Medicine and Biology, 965:145–161, 2017. ISSN 0065-2598. doi: 10.1007/978-3-319-47656-8_6.

[119] Sandra Castillo, Peddinti Gopalacharyulu, Laxman Yetukuri, and Matej Orešič. Algorithms and tools for the preprocessing of LC–MS metabolomics data. Chemometrics

and Intelligent Laboratory Systems, 108(1):23–32, August 2011. ISSN 0169-7439. doi: 10.1016/j.chemolab.2011.03.010. URL `http://www.sciencedirect.com/science/article/pii/S0169743911000608`.

[120] Linda Marchioro. Towards surrogate biomarkers. a study of correlation triples in multifluid metabolomics data. Master's Thesis. Technische Universitat Muenchen, 2015.

Appendix A

Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva.

Do KT, Kastenmüller G, Mook-Kanamori DO, Yousri NA, Theis FJ, Suhre K, J Krumsiek (2015), **Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva.** *J Proteome Res, 14: 1183-1194*

Journal of proteome research

Article

pubs.acs.org/jpr

Network-Based Approach for Analyzing Intra- and Interfluid Metabolite Associations in Human Blood, Urine, and Saliva

Kieu Trinh Do,[†] Gabi Kastenmüller,[‡,§] Dennis O. Mook-Kanamori,[∥,⊥,#] Noha A. Yousri,[∥,#] Fabian J. Theis,[†,¶] Karsten Suhre,[‡,∥] and Jan Krumsiek*,[†]

[†]Institute of Computational Biology and [‡]Institute of Bioinformatics and Systems Biology Helmholtz-Zentrum München, D-85764 Neuherberg, Germany

[§]German Center for Diabetes Research (DZD), D-85764 Neuherberg, Germany

[∥]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar

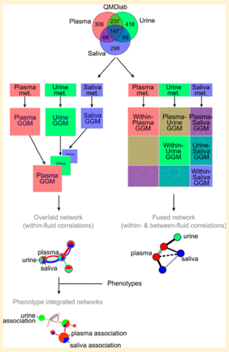
[⊥]Department of Clinical Epidemiology and [#]Department of Endocrinology, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands

[#]Department of Computer and Systems Engineering, Alexandria University, Alexandria 21525, Egypt

[¶]Department of Mathematics, Technische Universität München, 85748 Garching, Germany

Supporting Information

ABSTRACT: Most studies investigating human metabolomics measurements are limited to a single biofluid, most often blood or urine. An organism's biochemical pool, however, comprises complex transboundary relationships, which can only be understood by investigating metabolic interactions and physiological processes spanning multiple parts of the human body. Therefore, we here propose a data-driven network-based approach to generate an integrated picture of metabolomics associations over multiple fluids. We performed an analysis of 2251 metabolites measured in plasma, urine, and saliva, from 374 participants of the Qatar Metabolomics Study on Diabetes (QMDiab). Gaussian graphical models (GGMs) were used to estimate metabolite-metabolite interactions on different subsets of the data set. First, we compared similarities and differences of the metabolome and the association networks between the three fluids. Second, we investigated the cross-talk between the fluids by analyzing correlations occurring between them. Third, we propose a framework for the analysis of medically relevant phenotypes by integrating type 2 diabetes, sex, age, and body mass index into our networks. In conclusion, we present a generic, data-driven network-based approach for structuring and visualizing metabolite correlations within and between multiple body fluids, enabling unbiased interpretation of metabolomics multifluid data.



KEYWORDS: *multiple body fluids, multifluid, metabolomics, network inference, partial correlation, Gaussian graphical models, type 2 diabetes*

INTRODUCTION

Metabolomics is the holistic study of metabolic processes at a global level and has been used to analyze crucial biochemical mechanisms in various organisms, tissues, and conditions.[1] A metabolite profile represents a quantitative description of all relevant small molecules in a tissue sample or a body fluid. The metabolome is considered to represent an end point of complex genomic, transcriptomic, proteomic, physiological, environmental, and dietary processes, and thus provides a readout that is closely linked to the phenotype of interest.[2,3] Metabolomics is thus frequently used to identify patterns associated with pathophysiological states.[4−6] Moreover, because of its dynamic nature, metabolomics has gained importance as a key technology for systems biology.[1,7]

Most clinical studies on human metabolomics data have focused only on one body fluid type, typically blood or urine. However, an organism's biochemical pool comprises complex relationships that pass beyond the border of a single tissue or body fluid. A comprehensive approach, combining metabolomics analysis of two or more types of biosamples, should allow to generate an integrated picture of metabolic interactions and physiological processes that encompass multiple parts of the human body. Several studies already investigated metabolomics data from multiple fluids.[8−17] For example, Munshi et al.[16] performed multivariate analysis (PCA and PLS-DA) and pathway analysis on plasma, urine and saliva metabolomics data from HIV/AIDS patients. Multivariate analysis was also performed by Walsh et al.,[17] who collected plasma, urine, and saliva metabolomics samples to investigate the effect of acute dietary standardization on human metabolomics profiles. However, none of those studies performed a systematic investigation of the interactions across the fluids. Between-fluid relationships were studied by Adourian et al.,[18] who

Received: October 31, 2014

calculated a correlation network spanning plasma and liver tissue of rats to study circulating blood biomarkers related to liver tissue changes. Others analyzed metabolic processes across multiple tissues without metabolomics measurements. For example, Nyman et al.[19] used whole-body models to explore insulin signaling and glucose homeostasis. Shlomi et al.[20] predicted and compared human tissue-specific metabolism using gene and protein expression data that were integrated into Recon 1, a comprehensive reconstruction of the human metabolic network.[21] Bordbar et al.[22] developed a genome-scale multitissue model based on Recon 1, online databases and literature-curated data to study the metabolic differences in obese and obese T2D individuals. However, a simultaneous, systematic integration and network-based analysis of multifluid metabolomics data on human in vivo samples has not been performed yet.

The Qatar Metabolomics Study on Diabetes (QMDiab) was a case-control study with 374 participants, and the first to include plasma, urine, and saliva untargeted metabolomics measurement. The main objective of QMDiab was to shift the focus that has previously been on plasma and urine toward saliva for the identification of new biomarkers for type 2 diabetes. In particular, since saliva has been proposed as a promising resource for noninvasive diagnostics.[4,23] Although QMDiab was originally conducted for the investigation of T2D, in this paper, we focused more generally on multifluid data analyses.

We previously demonstrated that Gaussian Graphical Models (GGMs) applied to metabolomics data is capable of reconstructing biochemical pathways without use of prior knowledge.[24,25] The key idea behind GGMs is to use partial correlations rather than Pearson correlations to estimate conditional dependencies in multivariate Gaussian distributions, which enables to distinguish between direct and indirect interactions.[25] Network inference with GGMs is purely data-driven and does not require mapping of metabolites to known pathways available from public databases, such as KEGG[26] and METLIN.[27] In general, only a small fraction of metabolites can be mapped to corresponding database entries due to inconsistencies in biochemical annotations or lack of chemical identity of the measurement signal (usually referred to as *unknowns*[24]). These inconsistencies and *unknowns* can massively hamper the analysis of metabolic processes.[28−30] In contrast to knowledge-based methods, data-driven methods such as GGMs provide an unbiased alternative by avoiding metabolite mappings. In this study, we determined whether GGMs can also reveal biologically and medically relevant relations when applied to a metabolomics data set that covers three fluids in parallel: blood, urine and saliva. With a study such as the QMDiab available, here we introduce the application of GGMs to multifluid metabolomics data.

First, we systematically compared the metabolomes of the body fluids to investigate their differences and similarities. Subsequently, we raised three major questions for the analysis of metabolic correlations in multiple fluids. First, how similar are the networks derived from the body fluids? Second, how do metabolites and metabolic networks connect and interact between the body fluids? Third, since the primary goal of metabolomics research is to relate metabolic patterns to medically relevant phenotypes, we asked how the integration of such phenotype information into a multifluid network can be achieved and which insights can be obtained. We addressed these questions by evaluating two complementary approaches to construct networks consisting of within- and between-fluid correlations. We then integrate phenotype data of diabetes status, age, sex, and BMI into our visualizations to provide a comprehensive phenotype association analysis for all metabolites considered. The analysis workflow is shown in Figure 1.

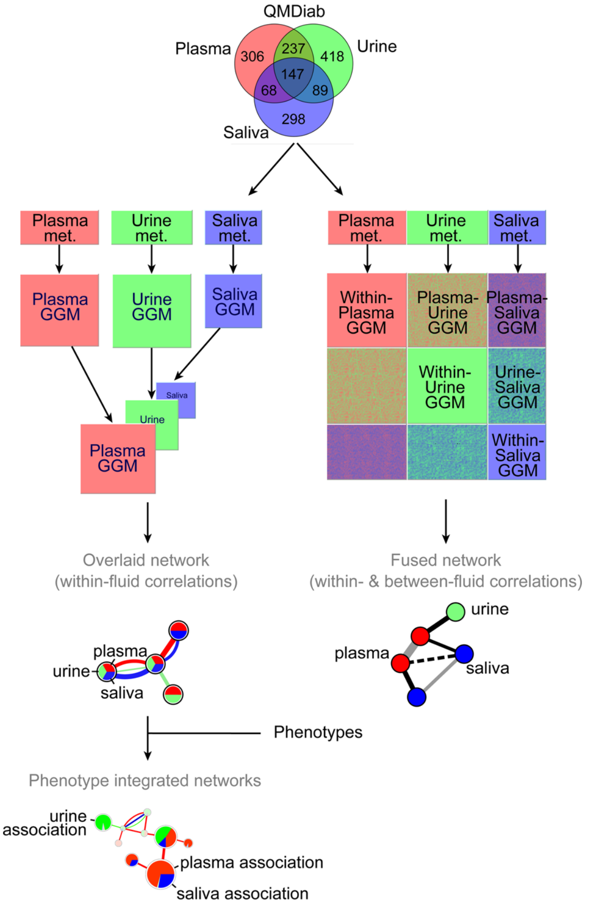


Figure 1. Study workflow. The QMDiab study included 758 metabolites in plasma, 891 metabolites in urine, and 602 metabolites in saliva, which resulted in a total of 2251 measured metabolites. For each body fluid, a single Gaussian graphical model (GGM) was generated. These GGMs were overlapped and visualized as a single network for comprehensive comparisons of multiple body fluids (overlaid network). In parallel, between-fluid metabolite correlations were investigated to explore the interactions of the three body fluids (fused network). Subsequently, metabolite associations with multiple phenotypes were integrated into the overlaid network (phenotype-integrated networks).

To the best of our knowledge, we are the first to simultaneously and systematically integrate multifluid metabolomics data by network inference on a human case-control study.

■ MATERIALS AND METHODS

Study Design

The data set was based on the QMDiab study, which was conducted in 2012 at the Dermatology Department of Hamad Medical Corporation and the Weill Cornell Medical College in Doha, Qatar. QMDiab included 374 males and females of Arab and Asian ethnicities aged 17−81 years. The study was approved by the Institutional Review Boards of HMC and Weill Cornell Medical College-Qatar (Research Protocol number 11131/11). Written informed consent was obtained from all participants. The study included 184 people as controls (defined by an absence of major systemic disorders) and 190 cases with a primary form of T2D. Cases with incomplete records were excluded, leaving 188 cases and 184 controls. Each participant provided a sample for nonfasting plasma, urine, and saliva, yielding a total of 1116 samples. Study enrollment was conducted between February and June 2012 during hospital visits as outpatients. General information, including age, sex, ethnicity, BMI, and any history of T2D, was obtained using questionnaires.[4]

Metabolomics Measurements

Samples were sent to Metabolon Inc. for untargeted metabolite identification and quantification. The analyses included LC/MS in both positive and negative modes and GC/MS, followed by metabolite identification and quantification. Details of the experimental procedures are provided in our previous paper.[4]

Overall, the Metabolon data set included 2251 measured metabolites, 1022 of which could be detected in only one, 394 in two, and 147 in all three body fluids, which resulted in 1563 unique metabolites. Overall, urine was the largest data set with 891 metabolites, followed by plasma with 758 metabolites and saliva with 602 metabolites. Approximately 45% of the measured compounds were chemically unidentified (unknowns). A detailed list of metabolites, including their associated pathways, is provided as Supporting Information Table S1.

Data Preprocessing

Metabolite concentrations were normalized by dividing by the median of each run day. Moreover, concentrations measured in urine and saliva were divided by the respective fluid's osmolality (osmoles of solute per kilogram: osmol/kg). Measurements from all three body fluids were log-transformed and standardized using *z*-scores. Metabolites with >80% missing values were excluded from the data set to prevent artifacts during partial correlation analysis and to preserve statistical power. Remaining missing values were imputed to the lowest concentration determined for a given metabolite, assuming that missing measurements are generally below the particular detection threshold. Metabolite measurements were defined as outliers if their values differed from the mean levels by more than four standard deviations over all samples. After filtering, the data set comprised 1951 metabolites (637 plasma, 825 urine, and 489 saliva compounds) in 372 samples (184 controls and 188 diagnosed diabetics).

Fluid-Specific Pathway Occurrence

We tested whether a metabolite class (superpathways) tends to occur uniquely in either one of the fluids. After excluding unknowns, there were eight classes in total: amino acids, carbohydrates, cofactors and vitamins, energy, lipid, nucleotide, peptide, and xenobiotics. We performed Fisher's exact tests on the two conditions: fluid-specificity (unique occurrence in one of the fluids) and pathway annotation. A positive association indicates that significantly many metabolites of a certain pathway occur exclusively in a fluid. In contrast, a negative association denotes that metabolites of a certain pathway tend to be shared by at least two body fluids. We corrected for multiple testing using a stringent Bonferroni level of significance of 0.05/(3*9) ≈ 0.001850.

Network Inference

Network inference was performed by estimating GGMs from the metabolite concentration data. GGMs are based on partial correlations, which represent the associations between two variables corrected for all remaining variables. We previously showed that these models could reconstruct metabolic pathways from metabolomics data sets.[24,25,57] Since the data set contained less samples than variables, a regularized GGM approach was used. We used "GeneNet", a shrinkage estimator-based approach for partial correlation calculation, which is freely available as an R package.[58] Links between metabolites were defined if both their Pearson correlations and their partial correlations were statistically significant at $\alpha = 0.05$ after Bonferroni corrections (correcting for $\binom{p}{2}$ tests, where p is the number of metabolites). For the overlaid, fused, and phenotype-associated correlation analyses, networks were inferred using different subsets of the data.

Phenotype Integration

To estimate the associations between metabolites and the four phenotypes, multivariate linear regression models were constructed as

$$\text{metabolite}_i \approx \beta_{i0} + \beta_{i1}\cdot\text{diabetes} + \beta_{i2}\cdot\text{age} + \beta_{i3}\cdot\text{sex} + \beta_{i4}\cdot\text{BMI} + \varepsilon_i$$

where i is the index of the metabolite, β_{i0} is the intercept, β_{i1}, ..., β_{i4} are the regression coefficients for each explanatory variable, and ε_i is a normally distributed error term. Since T2D and sex are binary variables, the linear regression corresponds to a *t* test. For each phenotype, the corresponding log p-values obtained from this regression analysis were mapped onto the network as node size and node pie charts. To explore the fluid-specificity of phenotype related metabolites, for each phenotype we performed Fisher's exact test on the two conditions fluid-specificity and phenotype association. We corrected for multiple testing with a Bonferroni significance cutoff of 0.05/(3*4) ≈ 0.0667.

■ RESULTS

Plasma, Urine, and Saliva Metabolome

Our analyses were based on QMDiab, which was conducted in 2012 in collaboration with the Dermatology Department of Hamad Medical Corporation in Doha, Qatar.[4] QMDiab included 374 males and females of various ethnicities and a wide age range between 17 and 81 years. The study included 184 individuals as controls (no major systemic disorders) and 190 cases with a primary form of diagnosed T2D.[4] Untargeted metabolomics measurements resulted in 758 plasma, 891 urine, and 602 saliva metabolites (Figure 1). After preprocessing, quality control steps, including normalization, treatment for missing values, and outlier detection, a total of 188 cases and 184 controls were included for further analysis. A total of 637 plasma, 825 urine, and 489 saliva metabolites passed our quality control procedures. For each metabolite, a superpathway and a subpathway annotation is available (Supporting Information Table S1). The first corresponds to the general metabolic class (e.g., amino acid, lipid, carbohydrate, etc.), while the latter represents the more specific metabolic pathway of the metabolite (oxidative phosphorylation, glycolysis, etc.). For GGM generation, all metabolite partial correlations calculated

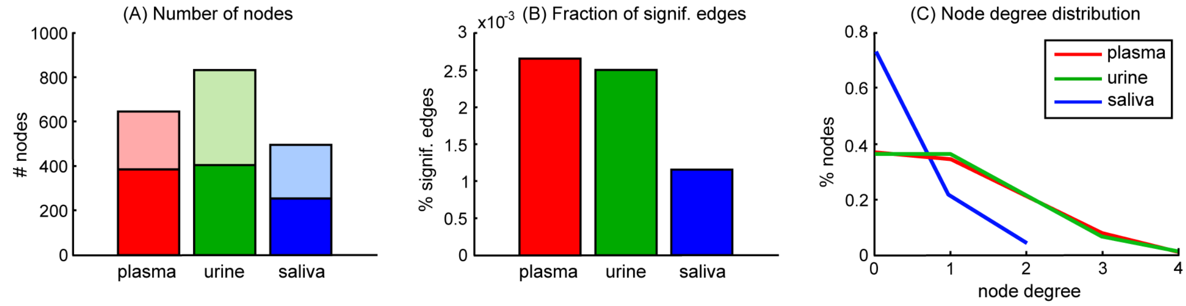


Figure 2. Graph statistics for fluid-specific GGMs. (A) Metabolites detected in only one body fluid were excluded from this analysis, which resulted in nearly equal numbers of metabolites for the urine and plasma data sets. Light-colored bars correspond to the original data set sizes and dark-colored bars are the data set sizes after exclusion. (B) Plasma had the highest proportion of significant metabolite correlations at $\alpha = 0.05$ after Bonferroni corrections, followed closely by urine. Saliva metabolite concentrations showed fewer correlations. (C) The highest node degrees were found for urine and plasma with degrees up to 4, whereas the maximum saliva node degree was 2. The saliva network was generally less connected for the selected significance cutoff, while distributions for plasma and urine were similar.

Table 1. Correlating Metabolites Sharing Super- and Sub-Pathway in Overlaid Network[a]

	Same super-pathway				Same sub-pathway		
	P-P	U-U	S-S		P-P	U-U	S-S
Overlaid network	79.57%	78.16%	80.65%	P	62.37%	57.47%	48.39%

[a] P = plasma, U = urine, S = saliva.

in this study were corrected for confounding effects of age, body mass index, sex, and T2D.

In our first analysis, we characterized the differences between the measured metabolomes of the three fluids. Urine had the highest number of quantified metabolites, half of which were also found in plasma or saliva. To assess whether metabolites from certain metabolic pathways (amino acid, carbohydrate, cofactors and vitamins, energy, lipid, nucleotide, peptide, and xenobiotics) tend to uniquely occur in either one of the fluids, we performed Fisher's exact tests for each superpathway/fluid combination (for details see Supporting Information Table S2). A positive association indicates that significantly many metabolites of a pathway occur exclusively in a fluid. In contrast, a negative association denotes that metabolites of the pathway are significantly underrepresented in the set of metabolites unique to the fluid. We observed that plasma lipids tend to be uniquely occurring in plasma and were not measured in urine or saliva. In contrast, lipids measured in urine or saliva are significantly shared. This might be partly due to the water insolubility of lipids, which need to be actively transported from blood to other fluids or tissues. For all body fluids, amino acids were significantly negatively associated with their exclusive occurrence in the corresponding fluid. The same applies for urinary and salivary carbohydrates as well as for plasma and urine xenobiotics. These negative associations indicate that the respective metabolite classes are being extensively shared between all three fluids. For plasma carbohydrates and saliva xenobiotics, no significant association could be observed.

Overlaid Network: Within-Fluid Correlations

To compare the network structure and pairwise partial correlation differences, we generated a GGM for each of the body fluids and overlaid them for direct visual comparison. We excluded all compounds detected in one fluid only to compare correlation differences and overlaps in metabolic interactions between the different body fluids. This resulted in nearly equal numbers of compounds in the urine and plasma subsets (Figure 2A). For each reduced data set, a metabolic network was inferred by generating GGMs.[25] Edges were drawn between metabolite pairs if both partial correlations and standard Pearson correlations were statistically significant at $\alpha = 0.05$ after Bonferroni correction. Interactive versions (Cytoscape sessions) of the networks are available as Supporting Information S3.

We observed substantial differences between the network connectivity of the three body fluids. The urine network comprised a slightly higher number of nodes than the plasma network (Figure 2A). Interestingly, urine showed a smaller proportion of significant edges (Figure 2B) compared to plasma. In contrast, saliva had the lowest number of nodes and the smallest proportion of significant correlations. In general, plasma and urine had higher node degrees than saliva, which also showed the highest proportion of metabolites not connected to any other node (Figure 2C). The node degrees in the urine and plasma networks were similarly distributed.

Comparing the partial correlation coefficients between the same metabolite pairs across different body fluids showed that, in general, there were fewer differences between plasma and urine than between plasma and saliva or urine and saliva (Supporting Information Figure S4). While the mean of the partial correlation differences of all comparisons approached zero, the variance of plasma vs urine was substantially lower than those for urine vs saliva and plasma vs saliva.

To systematically assess the biochemical validity of the three networks, we analyzed whether edges connect metabolites from the same super- and subpathways (Table 1). For this comparison, edges connected to an unknown metabolite were excluded. 78−81% of the correlating metabolites were annotated with the same superpathway, 48−63% also showed the same subpathway (Table 1). A detailed result list is available in Supporting Information Table S5. Although the plausibility of the remaining edges would need further manual investigation

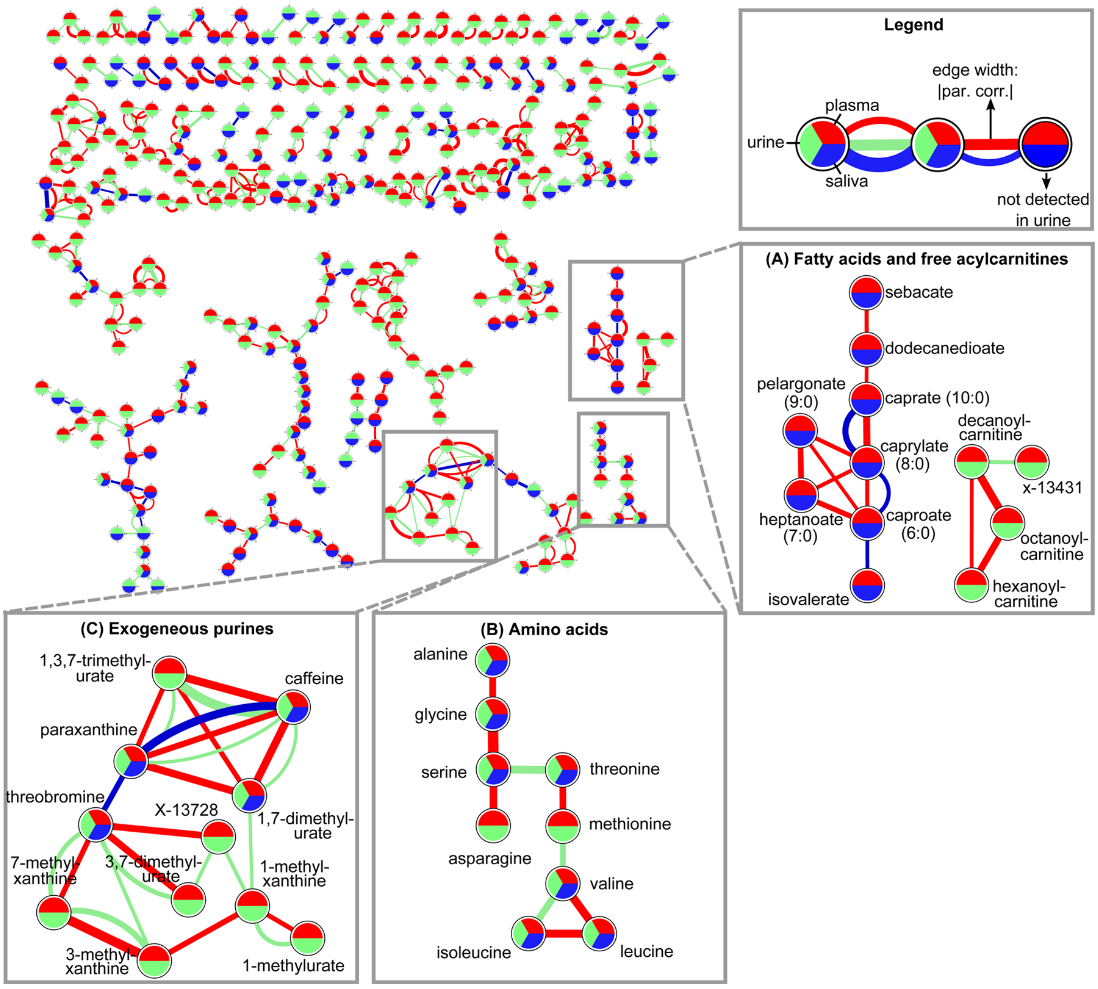


Figure 3. Overlaid network. This network represents an overlay of all three fluid-specific GGMs, thus only including within-fluid partial correlations. It consists of 464 metabolites measured in at least two body fluids, with a total of 435 edges. Metabolites without significant edges were excluded from the visualization. Pie chart area colors and edge colors correspond to the body fluid in which the metabolite or association was measured. Red = plasma, green = urine, blue = saliva. Panels A, B, and C show three exemplary subnetworks, which are discussed in the main text.

and can be addressed in future studies, the results show that for the most part the networks inferred for plasma, urine and saliva are biochemically valid.

To enable a visual comparison of the metabolic networks derived from the three body fluids, the three GGMs were superimposed to a single network comprising 464 nodes and 435 edges. This network (in the following termed overlaid network) represents the pairwise overlap of all within-fluid partial correlations (Figure 3). After systematically checking the single networks for biochemical validity above, we arbitrarily extracted three exemplary subnetworks to provide an in-depth view of the overlaid network.

The first subnetwork (Figure 3A) includes acylcarnitines and a set of free fatty acids (FFAs). Both fatty acids and their corresponding acylcarnitines were measured in plasma, whereas saliva only contained FFAs and urine only contained acylcarnitines. This was expected because only a relatively small amount of FFAs are secreted into urine due to their water insolubility. In contrast, their respective acylcarnitines were detected in urine in this study and also in previous studies.[31] This could have been due to the buffering function of carnitine, which plays an important role in controlling the acyl group pool in the body by excretion.[32,33] Moreover, acylcarnitines are water-soluble and thus more likely to be excreted in urine. Interestingly, in this subnetwork, we primarily observed significant partial correlations between metabolites in plasma and only a few in saliva and urine.

The second subnetwork includes a set of nine amino acids (Figure 3B). Methionine and asparagine were not detected in saliva, whereas the remaining amino acids, alanine, glycine, serine, threonine and the branched-chain amino acids valine, leucine, and isoleucine, were detected in all three body fluids. We only found significant correlations in either plasma or urine, but there was no overlap between these two fluids. Although detected in saliva, these amino acids showed no statistically significant associations in that fluid.

The third subnetwork shown in Figure 3 includes compounds from xenobiotics and caffeine metabolism. All metabolites were detectable in plasma and urine, with caffeine, paraxanthine, theobromine, and 1,7-dimethylurate also being detected in saliva. Associations were primarily observed in plasma and urine, which mirrors the breakdown of caffeine into its derivative compounds, further metabolism to *N*-methylxanthines, and finally, the excretion of these xenobiotics in

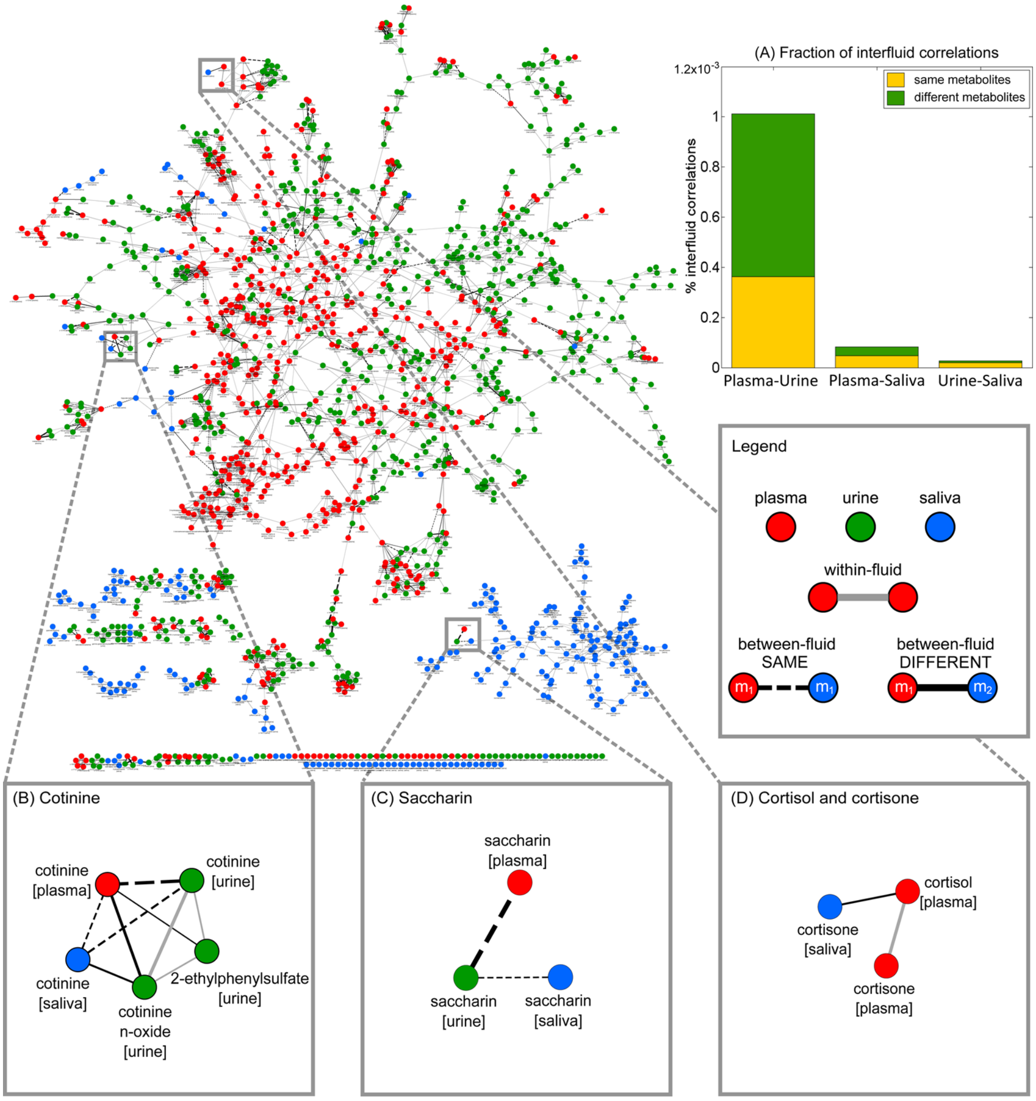


Figure 4. Fused network. The fused network comprises plasma (red), urine (green), and saliva (blue) nodes, which are connected by three types of edges. Gray edges indicate correlations between metabolites measured in the same body fluid and black edges represent between-fluid partial correlations. Solid black edges represent correlations between different metabolites and dashed edges indicate correlations between the same metabolites. Metabolites without significant edges were excluded from the visualization. (A) Proportions of between-fluid correlations for each fluid pair, were divided into correlations between either different or the same metabolites. (B, C, and D) Exemplary subnetworks, which are discussed in the main text.

urine. The plasma associations between these xenobiotics were already described for the German KORA cohort in a previous study.[24] Remarkably, there was also a strong partial correlation between saliva theobromine and paraxanthine, as well as between saliva paraxanthine and caffeine, reflecting the direct metabolization of caffeine to paraxanthine and partly to theobromine. The correlation of salivary caffeine and paraxanthine was previously reported and their ratio was suggested as a potential marker for evaluating CYP1A2 metrics and liver function.[34]

The overlaid network contains systematic signatures of metabolic pathways and interesting subnetworks for deeper analysis. In the next step, we focus on the interactions between the different body fluids.

Fused Network: Between-Fluid Correlations

To study how the body fluids are connected and interact, we generated a network consisting of both within- and between-fluid correlations. In contrast to the overlaid network, identical metabolites measured in different body fluids were here considered as individual nodes. The final network (fused network) comprised 1951 nodes and 2947 significant edges (Figure 4).

The fused network showed a substantial higher fraction of metabolic correlations between plasma and urine than between plasma and saliva or urine and saliva. The total number of significant between-fluid edges for plasma and urine was 536, which represented 93.5% of all between-fluid correlations in this network (Figure 4A).

We again checked the edges of the fused network for biochemical validity, defined by shared pathway annotations of

Table 2. Correlating Metabolites Sharing Super- and Sub-Pathway in the Fused Network[a]

	Same super-pathway				Same sub-pathway			
		P	U	S		P	U	S
Fused network	P	80.08%	93.38%	92.86%	P	58.33%	87.42%	92.86%
	U		71.92%	100%	U		46.13%	100%
	S			77.14%	S			40%

[a] P = plasma, U = urine, S = saliva.

the correlating metabolites. There is a slightly lower fraction of within-fluid correlating pairs in the fused than in the overlaid network (Table 2) sharing the same super- and subpathways (71−81% and 40−59%, respectively). In contrast, the fraction of shared annotations for between-fluid pairs is remarkably high. 92−100% of correlating metabolites measured in different fluids were assigned to the same superpathway, and 87−100% also share the same subpathway. These results indicate that almost the whole network in general consists of biochemically valid edges. Note that the overlaid network was estimated based on a subset of the data set (only nonunique metabolites), while the fused network consists of all metabolites. In contrast to Pearson correlations, the correlation structure can be slightly different if partial correlations are calculated on different underlying data sets. Therefore, the overlaid network and the within-fluid part of the fused network are not identical, albeit highly similar.

Between-fluid associations can occur between measurements for the "same" metabolites (e.g., plasma glucose and urine glucose) or between "different" metabolites (e.g., plasma glucose and urine mannose). 191 same- and 341 different-pairs showed significant partial correlations (Supporting Information Table S6). Partial correlations in the top list of between-fluid correlations occurred primarily between the same metabolites. Moreover, many of the correlated "different" metabolites were chemically related, for example, by glucuronidation (salicylurate vs salicyluric acid) or by oxidation (cotinine vs cotinine *N*-oxide). Examining the between-fluid top list revealed that for plasma−urine correlations, the highest partial correlation between metabolites of known identity was found for the same-metabolites homostachydrine, which is widely used by people as a dietary supplement.[35] The second highest plasma−urine correlation was found for the same-metabolites 1,2-propanediol (propylene glycol), a synthetic organic chemical that can be found in cosmetics, pharmaceuticals (e.g., lorazepam), and other products.[36] The high between-fluid partial correlation between both homostachydrine and 1,2-propanediol may reflect the unchanged excretion of these exogenous compounds in urine. For plasma−saliva edges, the highest absolute partial correlation was found for tryptophan betaine, while cotinine correlates strongest between urine and saliva. The corresponding plasma−urine and urine−saliva partial correlations were also significant, although these correlation values were lower.

The high fraction of between-fluid metabolite pairs sharing the same super- and subpathway annotations and in particular, the more frequent occurrence of between-fluid correlations between the same metabolites observed in the fused network is biologically plausible, because fluids exchange parts of their biochemical pools by excretion or transport processes in the kidneys and salivary glands.

Similar to the overlaid network, we provide a zoom-in view of three arbitrarily selected subnetworks consisting of biochemically related molecules.

The first subnetwork included the tobacco metabolite cotinine and 2-ethylphenyl sulfate, a metabolite from benzoate metabolism (Figure 4B). Cotinine is the main derivative of nicotine, which is an exogenous alkaloid occurring naturally in many plants. Nicotine enters the organism for instance during active or passive smoking or nicotine containing chewing gum. In the body it is mainly converted to cotinine.[37] In our study, plasma and urine cotinine are strongly correlated with each other, while their correlation to the salivary compound is weaker but still significant. These between-fluid edges probably reflect the unchanged excretion of cotinine after nicotine metabolization.

The second subnetwork contained saccharin (see Figure 4C), an artificial sweetener from the family of aromatic heterocyclic compounds, which is not metabolized and excreted by the kidneys.[38] As mentioned above, exogenous compounds are expected to be excreted unchanged, accounting for the reasonable correlations between plasma, urine, and saliva. Urinary and salivary saccharin are significantly correlated, while interestingly we did not observe a significant correlation between the plasma and saliva compounds.

The third subnetwork included the steroid hormone cortisol and its inactive form cortisone (Figure 4D). Cortisol plays an important role in the human body stress response by its effects on intermediary metabolism.[39] In this study, cortisol was only detected in plasma, whereas cortisone was detected in all three body fluids. Hydrophobic, insoluble steroid hormones like cortisol are converted in the liver to their water-soluble inert forms for urinary excretion. This could explain why cortisol was not detected in urine. In our study, plasma cortisol levels were significantly correlated with saliva cortisone levels, which is in accordance with previous findings that salivary cortisone rather than cortisol is a surrogate for free serum cortisol.[39,40]

Phenotype Integration into the Multifluid Network

As a final step in our study, we provide a showcase of how to analyze associations of multifluid metabolites and phenotypes. We determined the statistical associations of each metabolite to the four phenotypes BMI, age, sex and type 2 diabetes (T2D) by linear regression (linear regression results can be found in Supporting Information Table S7).

Similar to our analysis of the basic multifluid metabolomes, we first addressed the question of whether the metabolites significantly associating with a phenotype tend to occur exclusively in one body fluid. Plasma and urine metabolites associating with type 2 diabetes show a negative association with respect to their exclusive occurrence. That is, significantly many T2D associating metabolites occur in at least two body fluids. The same trend can be observed for age-associated

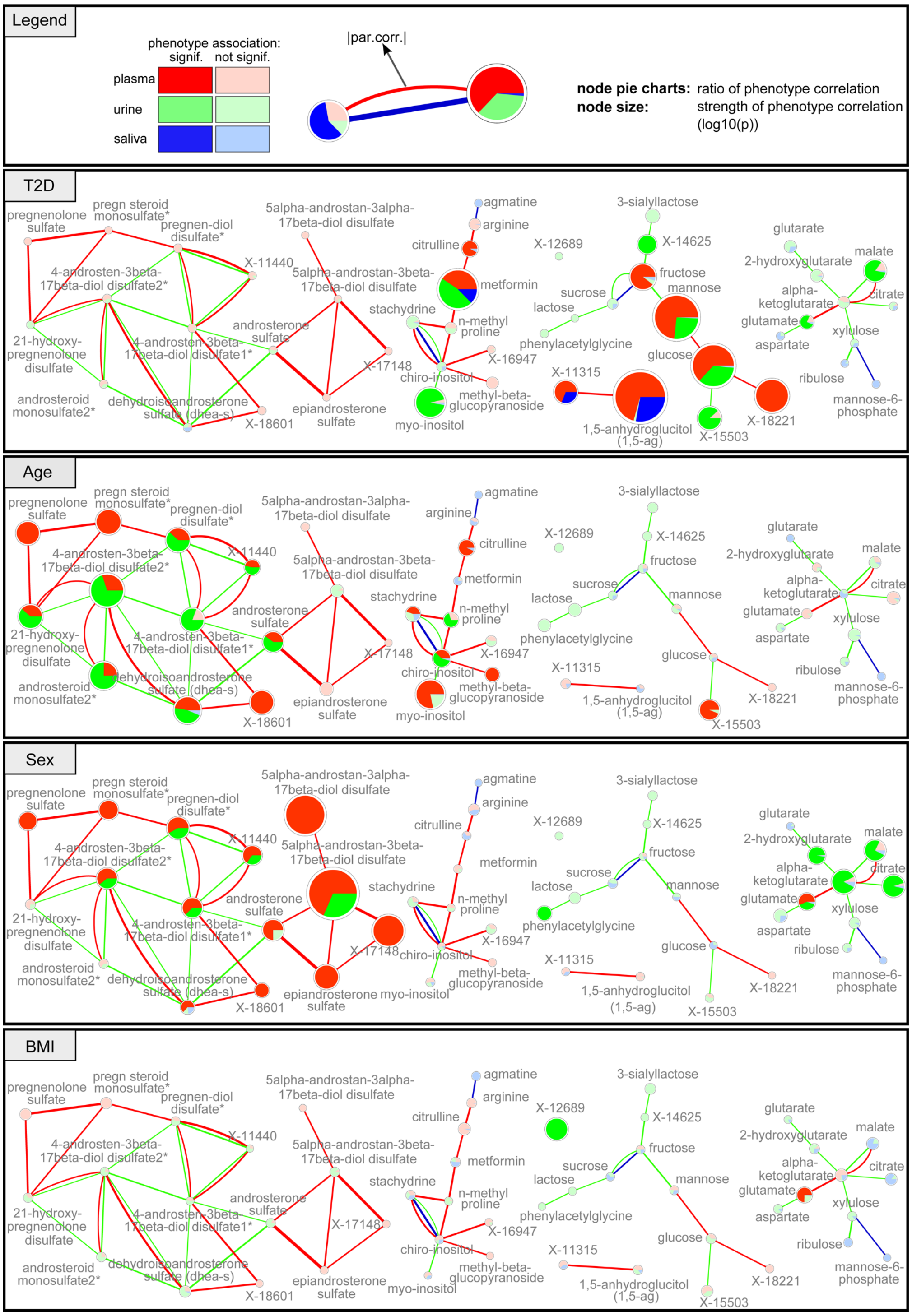


Figure 5. Overlaid networks with phenotype integration. We integrated phenotypic information for diagnosed T2D, age, sex, and BMI into the overlaid network, which included all metabolites for this analysis. Pie chart areas represent the ratios of $-\log_{10}$ p-values for the phenotype associations for each fluid. The node size corresponds to the lowest p-value (i.e., strongest association in all three fluids). The same subnetworks comprising a total of 48 nodes and 64 edges are shown for all phenotypes.

metabolites measured in plasma (see Supporting Information Tables S2). For sex and BMI, no significant associations were observed.

As a next step, we integrated these phenotype associations into a multifluid network by regenerating the overlaid GGM from Figure 3, this time including all metabolites into the analysis. The new network was a consolidation of three types of information: (i) statistical associations between measured metabolites; (ii) the metabolite assignment to different body fluids for direct visual comparisons of plasma, urine, and saliva;

and (iii) the statistical associations of all metabolites to the four phenotypes.

In the following, we present four subnetworks that included steroids, amino acids, carbohydrates, and keto acids for all four phenotypes (Figure 5). These subnetworks were manually selected such that the heterogeneity of the phenotypic association is best visible. Full networks can be found in Supporting Information S3.

The first subnetwork (Figure 5) included steroid hormones around dehydroisoandrosterone sulfate (DHEA-S) and showed significant associations with age and sex, but not with T2D and BMI. Several components of this subnetwork tended to have a stronger association with age in urine, whereas associations with sex were higher in plasma. DHEA-S is an endogenous steroid hormone that is produced from cholesterol primarily in the adrenal glands, gonads, and brain. It is a precursor for testosterone, androstenedione, estradiol, and estrone and is a marker of aging.[41] The subnetwork partly captures the steroid hormone biosynthesis pathway. All significant partial correlations between plasma compounds were also significant for the same urine metabolites. In contrast, some edges were urine-specific and did not occur in plasma, although this metabolite was detected in plasma. DHEA-S was also measured in saliva but showed no significant associations with any phenotype.

The second subnetwork included monosaccharides, which were strongly associated with T2D only. As previously observed in one of our studies performed with the QMDiab data set,[4] a significant T2D association was found for plasma and salivary 1,5-anhydroglucitol (1,5-AG), while this metabolite showed no significant association in urine. Plasma 1,5-AG, a clinical biomarker for short-term glycemic control,[42] was highly partially correlated with X-11315, an unknown metabolite that was negatively associated with T2D in both plasma and saliva. The saliva within-fluid partial correlation between these two metabolites was not significant. Other monosaccharides, such as glucose, mannose, fructose, and their related metabolites, were also highly associated with diabetes. In contrast to 1,5-AG, these molecules showed very high associations in plasma and were also significantly associated in urine. Although these monosaccharides were also detected in saliva, their salivary concentrations were not statistically associated with T2D.

The third subnetwork included the T2D drug metformin, amino acids and their derivatives, and several carbohydrates. Metformin was detected in all body fluids and was highly associated with T2D in plasma and urine and also significantly in saliva but to a lesser extent. This drug was mainly measured in T2D cases, although it was also observed in a substantial number of controls (Supporting Information Table S8). We assume that this is due to the use of metformin as a prophylaxis against diabetes or for ovulation induction.[43] Metformin showed a strong plasma link to citrulline, which has also been statistically associated with diabetes and was described as a promising factor for therapies for obesity and diabetes due to its blood glucose-lowering properties.[44,45] A second T2D-related metabolite in this subnetwork was myo-inositol in urine, which is linked to its epimer chiro-inositol, which had no significant associations to any of the analyzed phenotypes. Interestingly, plasma myo-inositol concentrations were statistically correlated with age. Elevated urine myo-inositol levels in diabetes cases were previously reported by several groups.[46−48] However, to the best of our knowledge, its plasma concentrations have not been investigated with respect to age.

Amino acids, monosaccharides, and intermediates of the TCA cycle formed a fourth subnetwork. Significant associations with phenotypes were primarily found in urine. Malate and glutamate were only associated with T2D, while other intermediates of the TCA cycle (α-ketoglutarate, 2-hydroxglutarate, and citrate) were also associated with sex. These intermediates were detected in all body fluids, but the partial correlations were only significant for the urine metabolites. The carboxylic acid glutamate in plasma was associated with both sex and BMI values. The lowest p-value for BMI in the data set was detected for the unknown urinary metabolite X-12689.

Taken together, this section suggested an approach to analyze statistical associations of metabolites from all body fluids to multiple phenotypes. On the basis of four subnetworks we showed known and several novel fluid-specific phenotype-associated metabolite sets.

■ DISCUSSION

We developed a network-based approach to structure and visualize metabolite correlations in a multitissue metabolomics setting. We applied this approach to a three body fluid data set from the QMDiab cohort. To the best of our knowledge, this is the first study to address multifluid metabolomics data from a human case-control study using network inference.

First, we systematically investigated metabolites from which pathways tend to occur exclusively in a certain fluid. We found that many lipids are plasma-specific, which might be due to their water insolubility. In contrast, the significant occurrence of amino acids, carbohydrates and xenobiotics in at least two body fluids indicates a strong exchange of these metabolites between the fluids.

With the overlaid network (Figure 3), an overlay of the within-fluid plasma-, urine-, and saliva-GGM, we compared the body fluids in terms of within-fluid partial correlations between metabolites. Since one node represents the same metabolite from multiple body fluids, this type of visualization facilitates comparisons of metabolite level correlations that are either fluid-specific or found across all fluids. Systematically checking whether correlating metabolites in the network shared the same super- and subpathway annotations revealed that the major part of the overlaid network was biochemical valid. A small fraction of metabolite pairs not having the same superpathway remained, which need further investigations to determine whether they are statistical artifacts or due to actual biochemical or physiological processes. The overlaid network enables the analysis of physiological processes and pathways. We illustrated this approach with detailed descriptions of biologically interesting subnetworks that included FFAs, acylcarnitines, amino acids, and exogenous purines.

The fused network (Figure 4), a merged network comprising both within-fluid and between-fluid correlations, allows for the investigation of interfluid metabolic correlations. Again we checked systematically for biochemical validity and observed a substantially high fraction of between-fluid metabolite pairs with both the same super- and subpathway annotations. As a next step, we distinguished between correlations for the same and different metabolites measured in different body fluids and found that the former appeared substantially more frequently than the latter. Both the more frequent occurrence of partial correlations between same-metabolites and the described subnetworks comprising steroid hormones, saccharin and exogenous compounds showed that our GGM-based approach can infer biologically meaningful between-fluid correlations.

We found a higher overlap of within-fluid correlations between plasma and urine than between plasma and saliva and urine and saliva. In addition, we observed more plasma—urine than plasma—saliva and urine—saliva between-fluid correlations. These two observations reflect the close relationship between plasma and urine due to excretion processes in the kidneys. Urine can be considered as a subset of plasma, since in contrast to saliva, urine is not directly influenced by external factors. All substances excreted through urine are transported through blood into the kidneys to be filtered into the bladder to form urine. Metabolite correlations that were found for plasma, but not for urine, could be explained, e.g., by rapid absorption processes or homeostasis maintenance in blood. Blood is a tightly controlled fluid due to homeostasis, whereas urine is a waste product for which a biochemical balance cannot be expected. Moreover, complex reabsorption processes may substantially alter the correlation patterns between metabolites. Saliva is produced by the salivary glands, which are perfused by blood capillaries to facilitate the entry of molecules from the systemic circulation into saliva.[49,50] Many compounds can be actively transported from blood to the salivary glands where they are secreted into saliva. However, serum constituents that are not part of saliva, such as drugs or hormones, can also enter this fluid by passive diffusion or by ultrafiltration.[49] Moreover, saliva can be considered to be substantially less well controlled than blood or urine, because it can be directly affected by external factors such as dietary intake or oral hygiene. For instance, the described presence of FFAs in saliva was most likely due to dietary fat intake. In contrast, plasma FFA correlations are probably most likely due to biochemical process in organ tissues. Urine and saliva are physiologically connected through blood. Thus, the links between saliva and urine metabolites possibly reflect stable concentrations throughout the body, for example, showing small changes between food intake and excretion. Such correlations would be expected to occur for molecules that are not metabolized in the body, as observed for instance, for cotinine, hippurate, catechol sulfate, and their derivatives.

In a third approach, we integrated phenotypic associations into our multifluid metabolomics analysis. At a general level, we observed that metabolites related to T2D and age are significantly shared between fluids. Investigating these associations in detail, we found single metabolites that were associated with only one of the analyzed phenotypes in certain body fluids, such as T2D-associated monosaccharides (e.g., glucose, mannose) and metformin, or sex-related intermediates of the TCA cycle (e.g., citrate, malate). Moreover, we identified a series of metabolites with fluid-dependent associations with two different phenotypes. For example, urine myo-inositol was associated with T2D, whereas its plasma concentrations were associated with age. Similarly, the DHEA-S cluster was found to have stronger associations with age for urine, while sex associations were higher for plasma. We confirmed previous findings with regard to markers, such as 1,5-AG and DHEA-S, for T2D and age, respectively. Furthermore, we identified previously unreported associations (e.g., plasma myo-inositol association with age), which can be investigated in future studies.

There are also metabolites that have been previously reported to associate with one of the phenotypes, but did not reach significance in our study. For instance, plasma branched-chain amino acids were observed to be associated with T2D in several studies.[51] In our data, only a significant association with urine BCAAs could be found, which is corroborated by other studies.[52] The lack of association in plasma may be explainable by the nonfasting state of the study participants, which could blur the signal. As already observed in the HuMet study[53] nonfasting states, as in our study, can substantially affect the behavior of plasma BCAA concentrations in healthy men.

Our study could be extended in several directions. (1) It may be interesting to assess the differences between body fluid networks statistically by performing tests for differences of dependent, nonoverlapping correlations as suggested by Raghunathan et al.[54] (2) The multifluid analysis should be replicated in population studies with larger sample sizes. For example, the Study of Health in Pomerania (SHIP) included more than 1000 samples with several follow-up studies that includes deep phenotyping and collection of blood, urine, and saliva samples.[55] Moreover, in contrast to case-control studies, population studies include a wide range of metabolic states, which may provide a more unbiased exploration of body fluid correlations. (3) As already discussed, our analyses were performed on plasma, urine and saliva samples from nonfasting study participants, possibly diminishing statistical power to find intrinsic metabolic correlations. Nevertheless, correlations that persist despite increased variability in the nonfasting state can be considered as robust and stable throughout different metabolic states. (4) Investigating multiple body fluids under controlled dietary challenging conditions, for example, after intake of a standardized meal following a prolonged fasting period, may provide further associations that cannot be detected from a single measurement time point.[56] A pilot study that included 15 people was published by Krug et al. (HuMet[53]), which comprised timeline data for blood, urine, and breath samples for 15 young, healthy men who were undergoing various nutritional challenges. (5) For a small fraction of the within- and between-fluid correlations found in our data, we could not determine direct biochemical explanations. Further studies that investigate the actual metabolically active organs involved will be needed for this analysis. This may include studies with animal models to investigate tissue metabolomics for the brain, liver, muscles, and other organs. (6) Knowledge of interactions between fluids can facilitate the search for alternatives for known biomarkers in potentially easier accessible body fluids (e.g., saliva markers for known plasma markers). For instance, in QMDiab we identified salivary 1,5-anhydroglucitol as a noninvasive surrogate marker for short-term glycemic control, based on its strong correlation to the same metabolite in plasma.[4] In general, the identification of biomarkers would be addressed by conventional regression approaches rather than GGMs, which are an exploratory tool for the investigation of fundamental metabolic mechanisms and pathways.

In conclusion, we have presented a data-driven network-based framework to analyze metabolic associations using multifluid data. Since our approach is generic, it can be applied to any other metabolomics data set that includes multifluid or multitissue data, as well as phenotypic information for the samples.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Metabolites with pathway annotations, fluid-specific occurrence of pathways, fluid-specific occurrence of phenotype associated metabolites, pathway enrichment of phenotype associated metabolites, all networks as Cytoscape sessions, comparison of absolute partial correlation differences, biochemical validity, same versus different between-fluid correlations, phenotype associations for all metabolites, and metformin measurement. The Matlab code for the analyses is available upon request. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49 89 3187-3641. Fax: +49 89 3187-3369. E-mail: jan.krumsiek@helmholtz-muenchen.de.

Notes

The statements made herein are solely the responsibility of the authors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by grants from the German Federal Ministry of Education and Research (BMBF), by BMBF Grant No. 01ZX1313C (project e:Athero-MED) and Grant No. 03IS2061B (project Gani_Med). Moreover, the research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-Health-F5-2012) under grant agreement no. 305280 (MIMOmics) and from the European Research Council (starting grant "LatentCauses"). J.K. is supported by a grant from the Helmholtz Postdoctoral Programme, "Initiative and Networking Fund". K.S. and D.O.M.-K. are supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. At Weill Cornell Medical College in Qatar, we express our appreciation to Zeinab El-Din and Walaa El Maraghy for their assistance in data collection, to Abeer Gohar and Sherryl Payra for their support with administrative questions regarding the project, to Thomas Doyle for organizing the sample shipment and overseeing lab safety, and to Jillian Rowe, Wafaa Sekkal Gherbi, Omar M. Chidiac, Marjonneke J. Mook-Kanamori, Sara Abdul Kader, Wadha A. Al Muftah, and Cindy McKeon, who all contributed to the clinical study. Furthermore, we thank all the medical and support staff at the dermatology department of Hamad Medical Corporation, in particular MDs Mohammed M. El-Din Selim, Ahmed H. Takiddin, Hala Al-Homsi, Khoulood A. S. Al-Mahmoud, and Amina Al-Obaidli, from the clinics of whom the patients were enrolled. Most of all, we thank all the study participants for their contribution to this diabetes research study.

■ REFERENCES

(1) Kaddurah-Daouk, R.; Kristal, B. S.; Weinshilboum, R. M. Metabolomics: A Global Biochemical Approach to Drug Response and Disease. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 653−683.

(2) Patti, G. J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: The Apogee of the Omics Trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.

(3) Blow, N. Metabolomics: Biochemistry's New Look. *Nature* **2008**, *455*, 697−700.

(4) Mook-Kanamori, D. O. 1,5-Anhydroglucitol in Saliva Is a Non-Invasive Marker of Short-Term Glycemic Control. *J. Clin. Endocrinol. Metab.* **2014**, DOI: 10.1210/jc.2013-3596.

(5) Kim, H.-Y.; Kim, H.-R.; Lee, S.-H. Advances in Systems Biology Approaches for Autoimmune Diseases. *Immune Netw.* **2014**, *14*, 73−80.

(6) Suhre, K. Metabolic Profiling in Diabetes. *J. Endocrinol.* **2014**, *221*, R75−85.

(7) Weckwerth, W. Metabolomics in Systems Biology. *Annu. Rev. Plant Biol.* **2003**, *54*, 669−689.

(8) Vitkin, E. Peer Group Normalization and Urine to Blood Context in Steroid Metabolomics: The Case of CAH and Obesity. *Steroids* **2014**, DOI: 10.1016/j.steroids.2014.07.003.

(9) Zhang, H.; et al. Metabolomic Analysis of Biochemical Changes in the Plasma and Urine of Collagen-Induced Arthritis in Rats after Treatment with Huang−Lian−Jie−Du−Tang. *J. Ethnopharmacol.* **2014**, DOI: 10.1016/j.jep.2014.03.007.

(10) Zheng, H.; et al. NMR-Based Metabolomic Profiling of Overweight Adolescents: An Elucidation of the Effects of Inter-/Intraindividual Differences, Gender, and Pubertal Development. *BioMed. Res. Int.* **2014**, *2014*, 537157.

(11) Dudzik, D.; et al. Metabolic fingerprint of Gestational Diabetes Mellitus. *J. Proteomics* **2014**, *103C*, 57−71.

(12) Yao, W.; et al. Integrated Plasma and Urine Metabolomics Coupled with HPLC/QTOF-MS and Chemometric Analysis on Potential Biomarkers in Liver Injury and Hepatoprotective Effects of Er−Zhi−Wan. *Anal. Bioanal. Chem.* **2014**, DOI: 10.1007/s00216-014-8169-x.

(13) Kim, J. W.; et al. Pattern recognition analysis for hepatotoxicity induced by acetaminophen using plasma and urinary ^{1}H NMR-based metabolomics in humans. *Anal. Chem.* **2013**, *85*, 11326−11334.

(14) Walsh, M. C.; et al. Impact of geographical region on urinary metabolomic and plasma fatty acid profiles in subjects with the metabolic syndrome across Europe: The LIPGENE study. *Br. J. Nutr.* **2014**, *111*, 424−431.

(15) Balderas, C.; et al. Plasma and urine metabolic fingerprinting of type 1 diabetic children. *Electrophoresis* **2013**, *34*, 2882−2890.

(16) Munshi, S. U.; Rewari, B. B.; Bhavesh, N. S.; Jameel, S. Nuclear Magnetic Resonance Based Profiling of Biofluids Reveals Metabolic Dysregulation in HIV-Infected Persons and Those on Anti-Retroviral Therapy. *PLoS One* **2013**, DOI: 10.1371/journal.pone.0064298.

(17) Walsh, M. C.; Brennan, L.; Malthouse, J. P. G.; Roche, H. M.; Gibney, M. J. Effect of Acute Dietary Standardization on the Urinary, Plasma, and Salivary Metabolomic Profiles of Healthy Humans. *Am. J. Clin. Nutr.* **2006**, *84*, 531−539.

(18) Adourian, A.; et al. Correlation Network Analysis for Data Integration and Biomarker Selection. *Mol. Biosyst.* **2008**, *4*, 249−259.

(19) Nyman, E.; et al. A Hierarchical Whole-body Modeling Approach Elucidates the Link between in Vitro Insulin Signaling and in Vivo Glucose Homeostasis. *J. Biol. Chem.* **2011**, *286*, 26028−26041.

(20) Shlomi, T.; Cabili, M. N.; Herrgård, M. J.; Palsson, B. Ø.; Ruppin, E. Network-Based Prediction of Human Tissue-Specific Metabolism. *Nat. Biotechnol.* **2008**, *26*, 1003−1010.

(21) Duarte, N. C.; et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 1777−1782.

(22) Bordbar, A.; et al. A Multi-Tissue Type Genome-Scale Metabolic Network for Analysis of Whole-Body Systems Physiology. *BMC Syst. Biol.* **2011**, *5*, 180.

(23) Wong, D. T. W. Salivaomics. *J. Am. Dent. Assoc. 1939* **2012**, *143*, 19S−24S.

(24) Krumsiek, J.; et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genet.* **2012**, *8*, No. e1003005.

(25) Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F. J. Gaussian Graphical Modeling Reconstructs Pathway Reactions from High-Throughput Metabolomics Data. *BMC Syst. Biol.* **2011**, *5*, 21.

(26) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets. *Nucleic Acids Res.* **2012**, *40*, D109−114.

(27) Smith, C. A.; et al. METLIN: A Metabolite Mass Spectral Database. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(28) Duren, W. MetDisease—Connecting Metabolites to Diseases via Literature. *Bioinformatics* **2014**, DOI: 10.1093/bioinformatics/btu179.

(29) Mendes, P. Metabolomics and the Challenges Ahead. *Brief. Bioinform.* **2006**, *7*, 127.

(30) Gagneur, J.; Jackson, D. B.; Casari, G. Hierarchical Analysis of Dependency in Metabolic Networks. *Bioinforma. Oxf. Engl.* **2003**, *19*, 1027−1034.

(31) Rinaldo, P.; Cowan, T. M.; Matern, D. Acylcarnitine Profile Analysis. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2008**, *10*, 151−156.

(32) Arrigoni-Martelli, E.; Caso, V. Carnitine Protects Mitochondria and Removes Toxic Acyls from Xenobiotics. *Drugs Exp. Clin. Res.* **2001**, *27*, 27−49.

(33) Melegh, B.; Kerner, J.; Bieber, L. L. Pivampicillin-Promoted Excretion of Pivaloylcarnitine in Humans. *Biochem. Pharmacol.* **1987**, *36*, 3405−3409.

(34) Perera, V.; Gross, A. S.; Xu, H.; McLachlan, A. J. Pharmacokinetics of Caffeine in Plasma and Saliva, and the Influence of Caffeine Abstinence on CYP1A2 Metrics. *J. Pharm. Pharmacol.* **2011**, *63*, 1161−1168.

(35) Lacefield, G.; Henning, J.; Rasnake, M.; Collins, M. *Alfalfa—The Queen of Forage Crops, AGR-76*; Cooperative Extension Service, University of Kentucky; Lexington, KY, 2014.

(36) West, R.; Banton, M.; Hu, J.; Klapacz, J. The Distribution, Fate, and Effects of Propylene Glycol Substances in the Environment. *Rev. Environ. Contam. Toxicol.* **2014**, *232*, 107−138.

(37) Benowitz, N. L.; Jacob, P.; Fong, I.; Gupta, S. Nicotine Metabolic Profile in Man: Comparison of Cigarette Smoking and Transdermal Nicotine. *J. Pharmacol. Exp. Ther.* **1994**, *268*, 296−303.

(38) Swithers, S. E. Artificial Sweeteners Produce the Counterintuitive Effect of Inducing Metabolic Derangements. *Trends Endocrinol. Metab.* **2013**, *24*, 431−441.

(39) Perogamvros, I.; Keevil, B. G.; Ray, D. W.; Trainer, P. J. Salivary Cortisone Is a Potential Biomarker for Serum Free Cortisol. *J. Clin. Endocrinol. Metab.* **2010**, *95*, 4951−4958.

(40) Lee, S.; et al. Simultaneous Quantitative Analysis of Salivary Cortisol and Cortisone in Korean Adults Using LC-MS/MS. *BMB Rep.* **2010**, *43*, 506−511.

(41) Cappola, A. R.; et al. DHEAS Levels and Mortality in Disabled Older Women: The Women's Health and Aging Study I. *J. Gerontol. A. Biol. Sci. Med. Sci.* **2006**, *61*, 957−962.

(42) True, M. W. Circulating Biomarkers of Glycemia in Diabetes Management and Implications for Personalized Medicine. *J. Diabetes Sci. Technol.* **2009**, *3*, 743−747.

(43) Elnashar, A. M. The Role of Metformin in Ovulation Induction: Current Status. *Middle East Fertil. Soc. J.* **2011**, *16*, 175−181.

(44) Wu, G.; et al. Arginine Metabolism and Nutrition in Growth, Health and Disease. *Amino Acids* **2009**, *37*, 153−168.

(45) Lucotti, P.; et al. Beneficial Effects of a Long-Term Oral L-Arginine Treatment Added to a Hypocaloric Diet and Exercise Training Program in Obese, Insulin-Resistant Type 2 Diabetic Patients. *Am. J. Physiol. Endocrinol. Metab.* **2006**, *291*, E906−912.

(46) Hong, J. H.; et al. Urinary Chiro- and Myo-Inositol Levels As a Biological Marker for Type 2 Diabetes Mellitus. *Dis. Markers* **2012**, *33*, 193−199.

(47) Clements, R. S., Jr.; Reynertson, R. Myoinositol Metabolism in Diabetes Mellitus. Effect of Insulin Treatment. *Diabetes* **1977**, *26*, 215−221.

(48) Kawa, J. M.; Przybylski, R.; Taylor, C. G. Urinary Chiro-Inositol and Myo-Inositol Excretion Is Elevated in the Diabetic db/db Mouse and Streptozotocin Diabetic Rat. *Exp. Biol. Med.* **2003**, *228*, 907−914.

(49) Kaufman, E.; Lamster, I. B. The Diagnostic Applications of Saliva—A Review. *Crit. Rev. Oral Biol. Med.* **2002**, *13*, 197−212.

(50) Farnaud, S. J. C.; Kosti, O.; Getting, S. J.; Renshaw, D. Saliva: Physiology and Diagnostic Potential in Health and Disease. *Sci. World J.* **2010**, *10*, 434−456.

(51) O'Connell, T. M. The Complex Role of Branched Chain Amino Acids in Diabetes and Cancer. *Metabolites* **2013**, *3*, 931−945.

(52) Salek, R. M.; et al. A metabolomic Comparison of Urinary Changes in Type 2 Diabetes in Mouse, Rat, and Human. *Physiol. Genomics* **2007**, *29*, 99−108.

(53) Krug, S.; et al. The Dynamic Range of the Human Metabolome Revealed by Challenges. *FASEB J.* **2012**, *26*, 2607−2619.

(54) Raghunathan, E. T.; Rosenthal, R.; Rubin, D. B. Comparing Correlated but Nonoverlapping Correlations. *Psychol. Methods* **1996**, *1*, 178−183.

(55) Völzke, H.; et al. Cohort Profile: The Study of Health in Pomerania. *Int. J. Epidemiol.* **2011**, *40*, 294−307.

(56) Mathew, S.; et al. Metabolomics of Ramadan Fasting: An Opportunity for the Controlled Study of Physiological Responses to Food Intake. *J. Transl. Med.* **2014**, *12*, 161.

(57) Shin, S.-Y.; et al. An Atlas of Genetic Influences on Human Blood Metabolites. *Nat. Genet.* **2014**, *46*, 543−550.

(58) Opgen-Rhein, R.; Strimmer, K. From Correlation to Causation Networks: A Simple Approximate Learning Algorithm and Its Application to High-Dimensional Plant Gene Expression Data. *BMC Syst. Biol.* **2007**, *1*, 37.

Appendix B

Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations.

Do KT, Pietzner M, Rasp D, Friedrich N, Nauck M, Kocher T, Suhre K, Mook-Kanamori DO, Kastenmüller G, Krumsiek J (2017), **Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations.** *NPJ Syst Biol Appl. 2017;3:28*

ARTICLE OPEN

Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations

Kieu Trinh Do[1], Maik Pietzner[2], David JNP Rasp[1], Nele Friedrich[2,3], Matthias Nauck[2,3], Thomas Kocher[4], Karsten Suhre[5,6], Dennis O. Mook-Kanamori[6,7,8], Gabi Kastenmüller[5,9] and Jan Krumsiek[1,9]

The identification of phenotype-driven network modules in complex, multifluid metabolomics data poses a considerable challenge for statistical analysis and result interpretation. This is the case for phenotypes with only few associations ('sparse' effects), but, in particular, for phenotypes with a large number of metabolite associations ('dense' effects). Herein, we postulate that examining the data at different layers of resolution, from metabolites to pathways, will facilitate the interpretation of modules for both the sparse and the dense cases. We propose an approach for the phenotype-driven identification of modules on multifluid networks based on untargeted metabolomics data of plasma, urine, and saliva samples from the German Study of Health in Pomerania (SHIP-TREND) study. We generated a hierarchical, multifluid map of metabolism covering both metabolite and pathway associations using Gaussian graphical models. First, this map facilitates a fundamental understanding of metabolism within and across fluids for our study, and can serve as a valuable and downloadable resource. Second, based on this map, we then present an algorithm to identify regulated modules that associate with factors such as gender and insulin-like growth factor I (IGF-I) as examples of traits with dense and sparse associations, respectively. We found IGF-I to associate at the rather fine-grained metabolite level, while gender shows well-interpretable associations at pathway level. Our results confirm that a holistic and interpretable view of metabolic changes associated with a phenotype can only be obtained if different layers of metabolic resolution from multiple body fluids are considered.

npj Systems Biology and Applications (2017)3:28 ; doi:10.1038/s41540-017-0029-9

INTRODUCTION

Metabolomics is the study of metabolic profiles at a global level. The metabolome is a readout of the biochemical transformations that involve small molecules in a body fluid or organ, and it reflects a snapshot of the state of a biological system.[1,2] Therefore, metabolomics has frequently been used to identify patterns associated with various pathophysiological states in humans, such as diabetes mellitus,[3,4] cardiovascular disease,[5,6] and Alzheimer's disease.[7–9]

Most published metabolomics studies focused on only one body fluid, usually blood or urine; however, phenotypes usually have links to metabolism in multiple fluids simultaneously. For example, we reported multifluid associations for type 2 diabetes in two recent studies.[10,11] With continuous technical advancements and decreasing costs, datasets with simultaneous metabolomics measurement should become available rapidly, as can be seen by the increasing research in this field.[12–16]

Phenotype associations in such large-scale, heterogeneous metabolomics datasets can be expected to be substantially complex, spanning functional modules, possibly across multiple fluids (Fig. 1). Functional modules are commonly defined as groups of correlating entities that are functionally coordinated, coregulated, or generally driven by a common biological process.[17] Systematic module identification algorithms are well established for omics data,[17–22] but have rarely been applied to high-throughput metabolomics data. A few metabolomics studies proceeded toward this objective by finding clusters in metabolite correlation networks, and by subsequently performing enrichment analyses with respect to a certain phenotype;[23–26] however, none of these studies performed a systematic phenotype-driven module search. Moreover, these analyses were performed in only one single fluid.

The identification and interpretation of modules for phenotypes that show rather few ('sparse') associations with metabolomics data are usually straightforward; however, phenotypes such as gender or BMI have been described to associate with more than a third to half of the blood metabolome.[26–28] A module search would lead to numerous results covering the majority of the metabolic network ('dense' associations), thereby impeding interpretation by their sheer quantity (Fig. 1). To solve this, we suggest performing association analysis and module identification at a coarser level, by grouping metabolites into their common pathways (defined as groups of metabolites with common biochemical and biological properties based on prior knowledge). The general idea is that while sparse phenotypic associations can only be detectable at the metabolite level, modules of dense

[1]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany; [2]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany; [3]DZHK (German Center for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany; [4]Department of Restorative Dentistry, Periodontology, Endodontology, Preventive and Pediatric Dentistry, Unit of Periodontology, University Medicine Greifswald, Greifswald, Germany; [5]Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, Neuherberg, Germany; [6]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar; [7]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands; [8]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands and [9]German Center for Diabetes Research (DZD), Neuherberg, Germany
Correspondence: Jan Krumsiek (jan.krumsiek@helmholtz-muenchen.de)

Received: 24 April 2017 Revised: 18 August 2017 Accepted: 25 August 2017
Published online: 21 September 2017

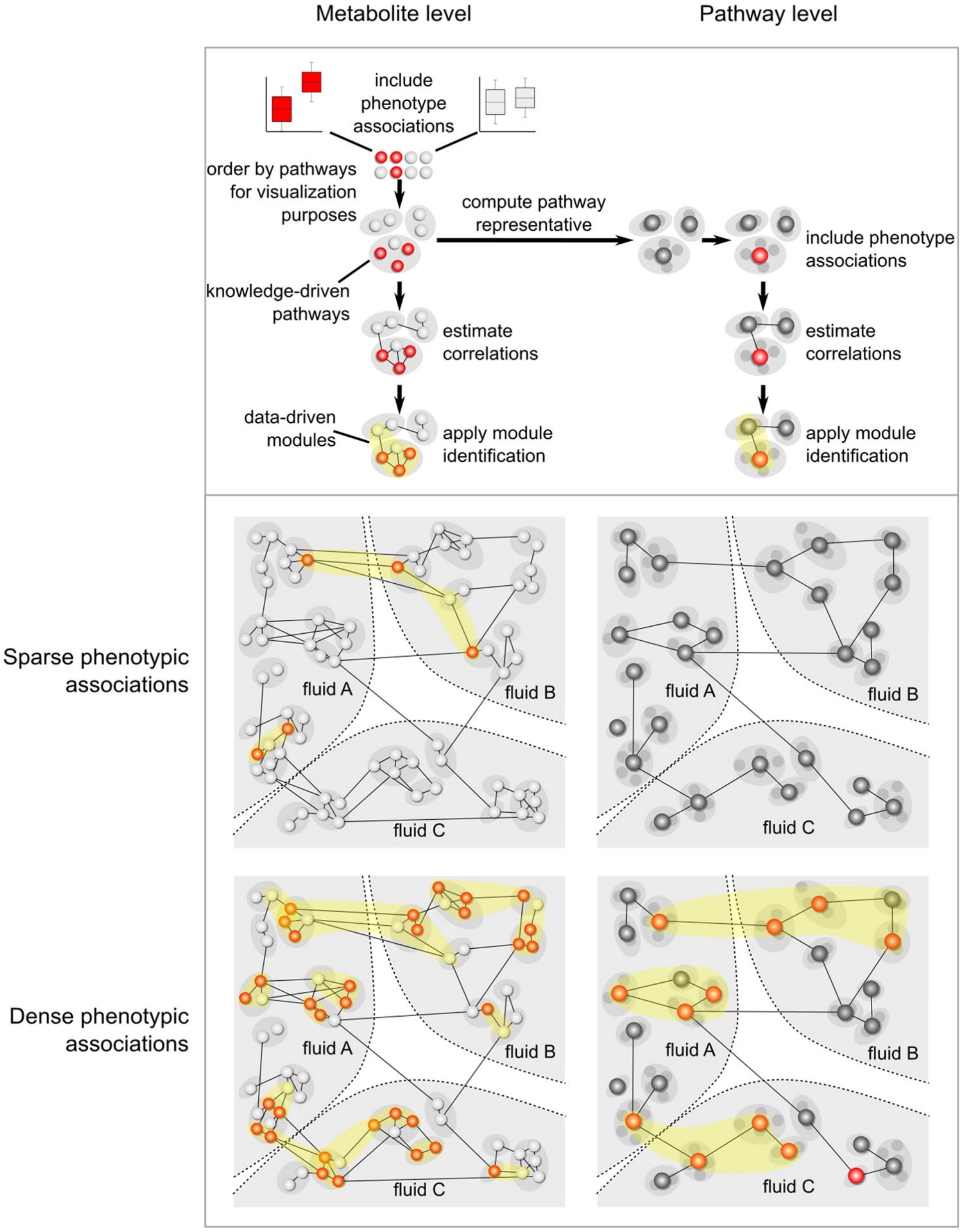

Metabolite level
Pathway level
include phenotype associations
order by pathways for visualization purposes
compute pathway representative
include phenotype associations
knowledge-driven pathways
estimate correlations
estimate correlations
data-driven modules
apply module identification
apply module identification
Sparse phenotypic associations
Dense phenotypic associations
fluid A
fluid B
fluid C

Fig. 1 Concepts of sparse and dense phenotype associations in metabolic networks. The figure depicts the concepts of sparse (top) and dense (bottom) phenotypic associations in metabolite (left) and pathway (right) networks. Metabolites, represented as nodes, can be grouped by knowledge-driven pathway information for visualization purposes. In addition, the nodes can be colored according to their phenotype associations (e.g. determined by a *t*-test). Network inference is performed to create a network, where an edge between two nodes represents their statistical correlation. Based on this network a module identification approach is applied to search for groups of correlating entities that are related to a phenotype of interest. For the pathway analysis, metabolites of the same pathway are aggregated to generate a pathway representative, which again can be colored according to phenotype associations. A pathway network is generated by connecting two pathway representations that are statistically correlated. Finally, a module identification approach is applied on this pathway network

phenotypic associations might be easier to interpret at the pathway levels.

In this study, we present a method for the systematic phenotype-driven identification of modules from multifluid metabolomics data, operating both at the single metabolite and at the pathway level. Specifically, we created a hierarchical map of multifluid metabolomics correlations as a template for the underlying metabolic network. Based on this network, we automatically extracted modules associating with two example phenotypes.

To create this hierarchical map, we generated data-driven multifluid networks from blood, urine, and saliva metabolomics data of the German Study of Health in Pomerania (SHIP-TREND) cohort.[29] Specifically, we estimated Gaussian graphical models (GGMs) based on partial correlations at the metabolite level and at two pathway levels: 'super-pathways' representing metabolite classes such as 'Lipid' or general metabolic processes such as 'Energy' and 'sub-pathways' representing biochemical subclasses or processes within a super-pathway such as 'Lysolipid' or 'TCA cycle,' respectively. The three networks (metabolite, sub-pathway, and super-pathway) together depict the hierarchical map.

Moreover, we developed a module search algorithm inspired by Chuang et al.[20] and applied it to serum measurements of insulin-like growth factor I (IGF-I) and gender. IGF-I is a growth hormone with high sequence homology to insulin. It participates in

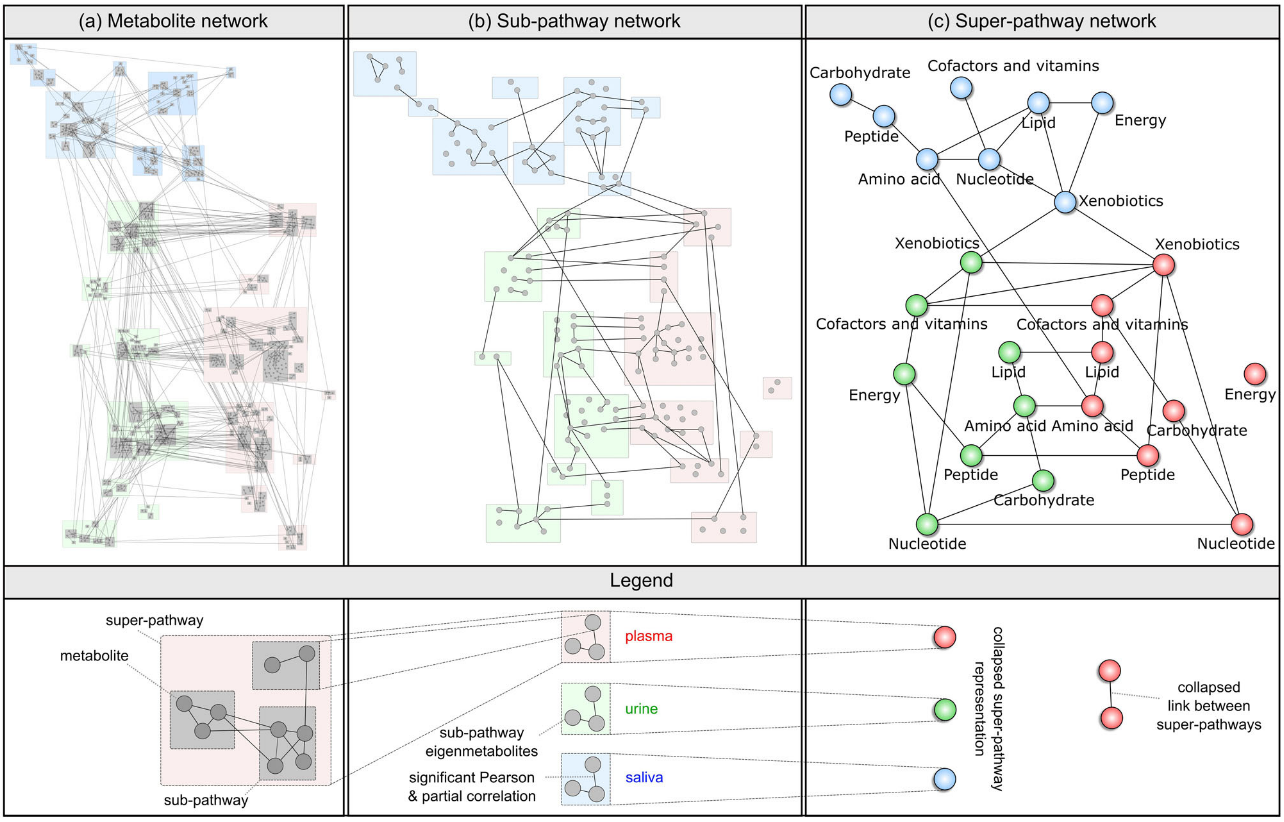


Fig. 2 Hierarchical map of multifluid metabolic processes at **a** metabolite, **b** sub-pathway, and **c** super-pathway levels. In the metabolite and sub-pathway network, edges were drawn if both partial correlation and Pearson correlation were significant at $\alpha = 0.05$ after Bonferroni correction for multiple testing. The super-pathway network **c** was generated by collapsing the sub-pathway GGM, i.e. drawing an edge between two super-pathways whenever at least one pair of their sub-pathways was connected. Note that all three networks share the same overall layout

numerous biochemical processes, in particular in the stimulation of cell growth and proliferation, and has been found to be associated with various disorders such as diabetes, cardiovascular diseases, and cancer.[30–34] Despite its key roles in various biochemical processes, mapping IGF-I–metabolite associations onto a metabolite network in a previous study on the same dataset from the SHIP cohort resulted in only a relatively small number of blood and urine metabolites.[34] Thus, IGF-I here serves as a trait with sparse associations. For gender associations, on the other hand, we found associations with a major part of the metabolic network,[26] thus representing a trait with dense associations.

RESULTS

Our analysis was based on data from the SHIP-TREND cohort. The dataset comprised 906 individuals, 512 females and 394 males, for which fasting plasma, urine, and saliva samples were available. Untargeted metabolomics measurements were performed by ultra-high liquid-phase chromatography coupled with tandem mass spectrometry (UPLC-MS/MS). Data preprocessing included run-day normalization, dilution factor normalization (for urine and saliva), log transformation, outlier handling, and handling of missing values. After preprocessing, 610 known metabolites and 387 metabolites, whose chemical structures had not been identified yet, were available for further analysis. For each metabolite, knowledge-based pathway annotations from the metabolomics platform (Metabolon Inc.) were used. Each known metabolite was annotated with one of 73 'sub-pathways', which represent metabolic pathways or biochemical subclasses of the compounds (e.g., 'Branched-chain amino acid', 'Lysolipid', 'Glycolysis'). In addition, each sub-pathway was assigned to one of eight broad 'super-pathways' ('Amino acid', 'Lipid', 'Carbohydrate', 'Nucleotide', 'Peptide', 'Energy', 'Cofactors and vitamins', and 'Xenobiotics'). These pathway annotations have been frequently used in previous studies that investigated data from the same platform (see e.g. refs. 35–37). Metabolites, their annotations, and a comparison of the measured metabolite pools between fluids can be found in Supplementary Information S1.

Pathway representation and generation of the hierarchical map

We generated the hierarchical metabolic map by inferring three networks, representing the metabolic processes at three decreasing levels of granularity (Fig. 2): The first comprised multifluid correlations between single metabolites based on a GGM, a correlation-based network inference approach. Note that unknown metabolites were used to estimate the metabolite network, but were excluded from this view. To generate a sub-pathway network, a GGM was calculated based on sub-pathway *eigenmetabolites*. The majority of these *eigenmetabolites* showed a high degree of explained variance for their respective metabolites (Supplementary Information S2), and thus were reasonable statistical representatives of the pathways. To generate the super-pathway network, the sub-pathway GGM was *collapsed* by connecting any two super-pathways that showed at least one connection in the sub-pathway network. This procedure was chosen instead of calculating GGMs on the corresponding super-pathway *eigenmetabolites* due to the substantially high heterogeneity of most of the super-pathways (e.g., the very broadly defined 'Lipid' super-pathway). This is reflected by low-explained variances for super-pathway *eigenmetabolites* (Supplementary

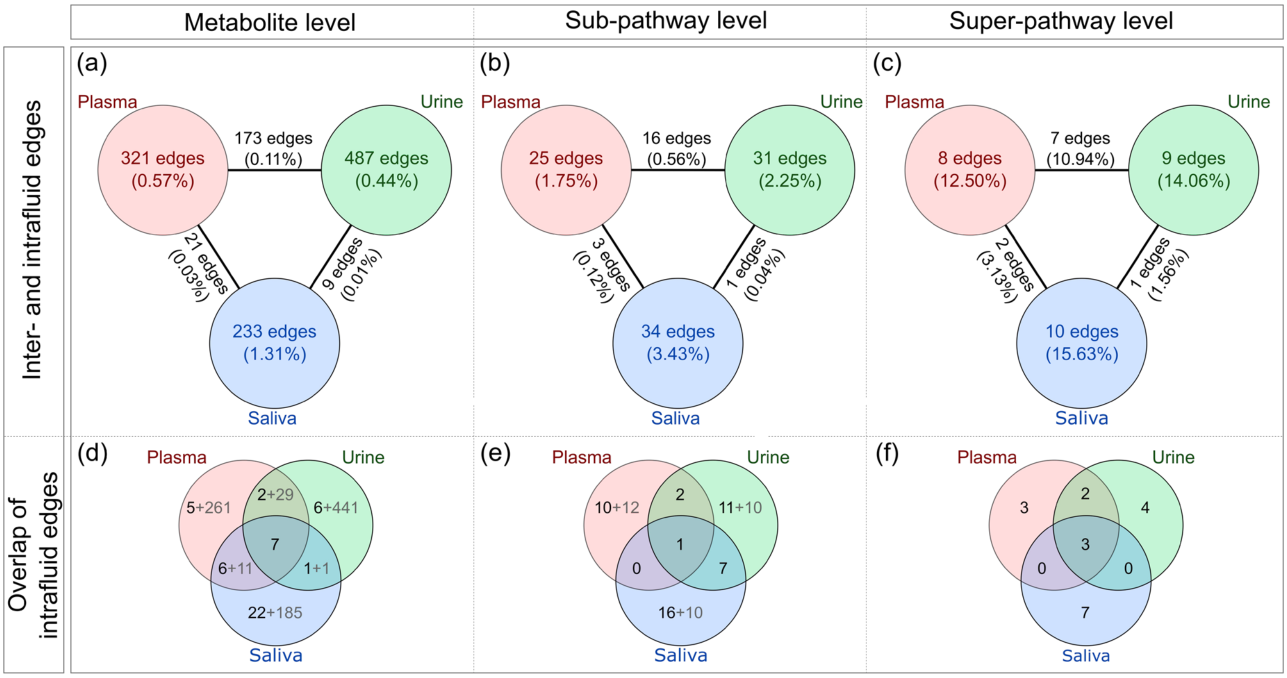


Fig. 3 Global structure of the hierarchical map. **a**–**c** Absolute number and percentage of significant intra-fluid and interfluid edges. The percentage is calculated as the number of edges divided by all possible edges. **d**–**f** Number of intrafluid edges occurring in only one fluid or shared across two or all three fluids. Black numbers correspond to links between metabolites, sub-pathways, or super-pathways that were measured in all the three body fluids, while gray numbers represent metabolites and pathways that occur in at most two body fluids

Information S2), which would not suffice as true pathway 'representatives'. Note that unknowns were excluded from the pathway analysis since these metabolites could not be assigned to a sub-pathway or super-pathway.

Interactive versions of the networks as yEd *.graphml* files, as well as corresponding correlation matrices, are available in Supplementary Information S3. Detailed lists of all correlation coefficients and the associated pathways can be found in Supplementary Information S4.

At the most fine-grained level, the hierarchical map contained 335 plasma, 473 urine, and 189 saliva metabolites, with a total of 1244 edges between them (1041 intrafluid and 203 interfluid edges, Fig. 2a, Fig. 3a). The sub-pathway GGM comprised 54 plasma, 53 urine, and 45 saliva *eigenmetabolites*, and, in total, 110 edges out of which 90 were within fluids and 20 were across fluids (Fig. 2b, Fig. 3b). The coarsest level represented by the *collapsed* super-pathway network consisted of 24 nodes, and 27 intrafluid and 10 interfluid edges (Fig. 2c, Fig. 3c). In general, we observed most correlations to be intrafluid in all the three networks. Since, in particular, salivary metabolomics measurements can be dependent on the oral hygiene of the study participants, we investigated whether the hierarchical map was influenced by the participants' teeth brushing behavior. Overall, we found only marginal differences in the correlation structures, which were mainly based on statistical variance rather than biologically driven by oral hygiene (Supplementary Information S5).

Similarities and differences of correlation structures across body fluids

To obtain a general overview of the hierarchical map, we explored it at the highest level of body fluids considering two aspects: (i) How similar are the intrafluid correlation structures when comparing the three different fluids? (ii) How can the crosstalk between fluids be characterized?

Similarity of body fluids. For all the three fluids, we determined the fluid-specific correlations, i.e., those exclusively occurring in only one body fluid, and the correlations shared between at least two fluids (Figs. 3d–f). At all levels, the number of fluid-specific edges far exceeds the number of shared edges.

At the metabolite level, 266, 447, and 207 intrafluid edges were exclusively found in plasma, urine, and saliva, respectively. A pairwise comparison of the fluids yielded 57 edges that occurred in at least two body fluids, with the majority (31) shared between plasma and urine (Fig. 3d). Plasma and saliva shared 17 correlations, whereas urine and saliva shared only 2 correlations. Overall, 50 and 77% of the fluid-specific metabolite edges occurred within the same sub-pathway and super-pathway, respectively, while for correlations that can be found in at least two fluids 80% were observed within sub-pathways and 90% in super-pathways (Supplementary Information S4). This indicates that, if correlations are shared across fluids, the two correlating metabolites more often act in similar biochemical processes compared to exclusive correlations. Comparing all the three fluids simultaneously, an overlap can only be reasonably analyzed for metabolites that were also measured in all the three of them. Inspecting only edges between such metabolites (black numbers in Fig. 3d) left 49 intrafluid edges, of which seven occurred in all the three fluids.

At the sub-pathway level, we found 69 fluid-specific and 10 shared edges (Fig. 3e). Only one edge occurred in all fluids ('Fatty Acid Metabolism (also BCAA Metabolism)' with 'Fatty Acid Metabolism (Acyl Carnitine)'). Two edges were observed in both plasma and urine, and interestingly, urine and saliva shared seven edges (Supplementary Information S4), all of which were within the same respective super-pathway. Overall, at the super-pathway level, more fluid-specific edges (14) were observed compared to the shared edges (5) (Fig. 3f).

Crosstalk between fluids. We investigated the crosstalk between the fluids by analyzing the interfluid correlations in the hierarchical map (Figs. 3a–c). In total, there were 203 crossfluid edges at the metabolite level (Fig. 3a). A vast majority of edges was observed between plasma and urine (173), while there were only 21 and 9 edges between plasma and saliva and urine and saliva, respectively. In total, 98 of these 203 edges (75 plasma-urine, 19 plasma-saliva, and 4 urine-saliva) were between the same metabolites measured in different fluids, for example, between plasma betaine and urine betaine (Supplementary Information S4). At the sub-pathway level, we found 20 crossfluid correlations, *collapsed* to 10 interfluid links between super-pathways (Fig. 3b). The majority of sub-pathway and super-pathway edges could again be observed between plasma and urine. Except one ('Tocopherol Metabolism' and 'Food Component/Plant'), all cross-fluid edges were between the same sub-pathways. Six out of eight plasma super-pathways were linked to the respective same super-pathway node in urine, reflecting the aforementioned strong connection between those two fluids (Fig. 3c). In contrast, plasma and saliva, as well as urine and saliva were connected by only a few links.

Summarizing the results from this section, we found the majority of intrafluid correlations to be fluid-specific at all levels, providing evidence for substantial discordance of correlation structures across the different fluids. All edges, in particular, the shared correlations, occurred mainly between entities of the same pathways. Our results also indicated that plasma and urine are both more similar and more strongly connected to each other than to saliva, while saliva has a higher similarity and more connections to plasma than to urine. In general, crossfluid correlations were mostly observed between the same pathways, pointing toward a substantial impact of transport and exchange processes on the metabolomes between the fluids.

Phenotype-driven module identification procedure

We developed a procedure to identify modules associated with a phenotype at different levels of the hierarchical map. The algorithm is graphically outlined in Supplementary Information S6. Briefly, given a network, a phenotype variable, a scoring function, and a seed (=starting) node; a greedy search algorithm identifies an optimal module by score maximization. The optimal module is determined by extending candidate modules along its network edges, until no further score improvement can be achieved. Each candidate module is scored by the negative logarithmized *p*-value of a regression-based association of a representative value of all metabolites in the module with the phenotype (see Methods). Notably, a single metabolite is scored by its univariate association with the phenotype. In a final consolidation step, overlapping optimal modules, for instance, those obtained from neighboring seed nodes, are identified and combined into a maximal module.

We followed a conservative multiple testing correction approach: To be significant, the *p*-value of a module had to be lower than the significance level of 0.05 divided by the number of network nodes (Bonferroni correction at node level). In addition, we required each module's score to be higher than the maximum score observed across all single components of the module.

The procedure was applied to two phenotypes at all the three levels (metabolite, sub-pathway, and super-pathway): IGF-I as a phenotype with sparse associations, and gender as a trait with dense associations.

Phenotype-driven module identification for sparse associations: IGF-I

Associations of IGF-I with blood and urine metabolites in the SHIP-TREND dataset were investigated in a previous study by Knacke et al.[34] Here, we additionally integrated metabolomics measurements from saliva. Notably, in the work by Knacke et al., IGF-I associations were analyzed for males and females separately. In our study, however, the results of a module search stratified by gender were mostly covered by modules from a joint gender analysis, which is why the latter analysis was chosen. Furthermore, Knacke et al. used a more relaxed multiple testing correction (FDR at 0.05), while in this study we applied the conservative Bonferroni correction, since we expected a substantially increased statistical power for the module-identification approach.

At the metabolite level, our algorithm identified six modules associated with IGF-I (Fig. 4). For the sub-pathway network, we obtained only one module comprising plasma and urine 'Steroid' pathway metabolites (Supplementary Information S7). Furthermore, no modules were found at the coarsest level for super-pathways, confirming that IGF-I associations are rather sparse in the metabolic network. Therefore, we restricted the following analysis to the modules identified at the fine-grained metabolite level.

The six identified modules demonstrate that the module-identification algorithm enhances classical association analysis in several ways: It detected modules that (i) cover multiple pathways (see modules A and F) and (ii) span multiple body fluids (see modules B–E). (iii) Moreover, the algorithm was able to dissect apparently related but distinct processes. For example, modules C and E were in close proximity in the network and both contained a steroid amongst unidentified metabolites; however, the identification of two distinct modules suggested that they reflect two different processes that are independently associated with IGF-I levels. (iv) The algorithm increased the statistical power in several cases. For modules A–C, none of the single metabolites inside the modules was significant, whereas the entire module showed a significant *p*-value. This can be attributed to the reduction of statistical noise when aggregating concentrations of multiple metabolites.

Beyond the advantages of our approach compared to classical association analysis, we found a series of biologically interesting results. Initially, we were able to confirm previously identified IGF-I associations. For instance, it has been reported that there is a complex interplay between sex hormones and IGF-I.[38,39] In our study, we identified a multifluid module containing a cluster of plasma and urine epiandrosterone and androsterone metabolites (module B). IGF-I has also been linked to the maintenance of physiological mitochondrial function via regulation of the expression of the mitochondrial pyrimidine nucleotide carrier PNC1.[34,40] The association of single blood metabolites from the pyrimidine pathway with IGF-I has already been reported by Knacke et al. In the present study, we additionally observed that the aggregation of several blood and urine metabolites from this pathway (module D) yielded a considerably lower *p*-value than the single components, further supporting the link between pyrimidines and IGF-I. Both modules B and D contain metabolites from plasma and urine, indicating that not only the concentration levels of the respective metabolites in these fluids but also their crossfluid transport processes might be associated with IGF-I.

We also detected IGF-I associations, that to the best of our knowledge, have not been reported previously. We found a saliva module (A) comprising three amino acids, 2-hydroxyglutarate, a lipid, and laurylsulfate, a xenobiotic, each of which alone was not significantly associated with the phenotype. Associations of these metabolites with IGF-I have not previously been reported, in particular not in human saliva. Module F contained the xenobiotic 2-ethylhexanoate (EHA) and the fatty acid caprylate (8:0), neither of which has been reported with IGF-I to date; however, in this case, the module score seems to be mainly driven by EHA, while the fatty acid only contributes marginally to the score.

Finally, we investigated the effects of oral hygiene on the modules identified for IGF-I by correcting for the teeth brushing behavior of the study participants in the module identification process. Exactly the same modules were found, indicating that oral hygiene has no effects on metabolic changes related to IGF-I (Supplementary Information S5).

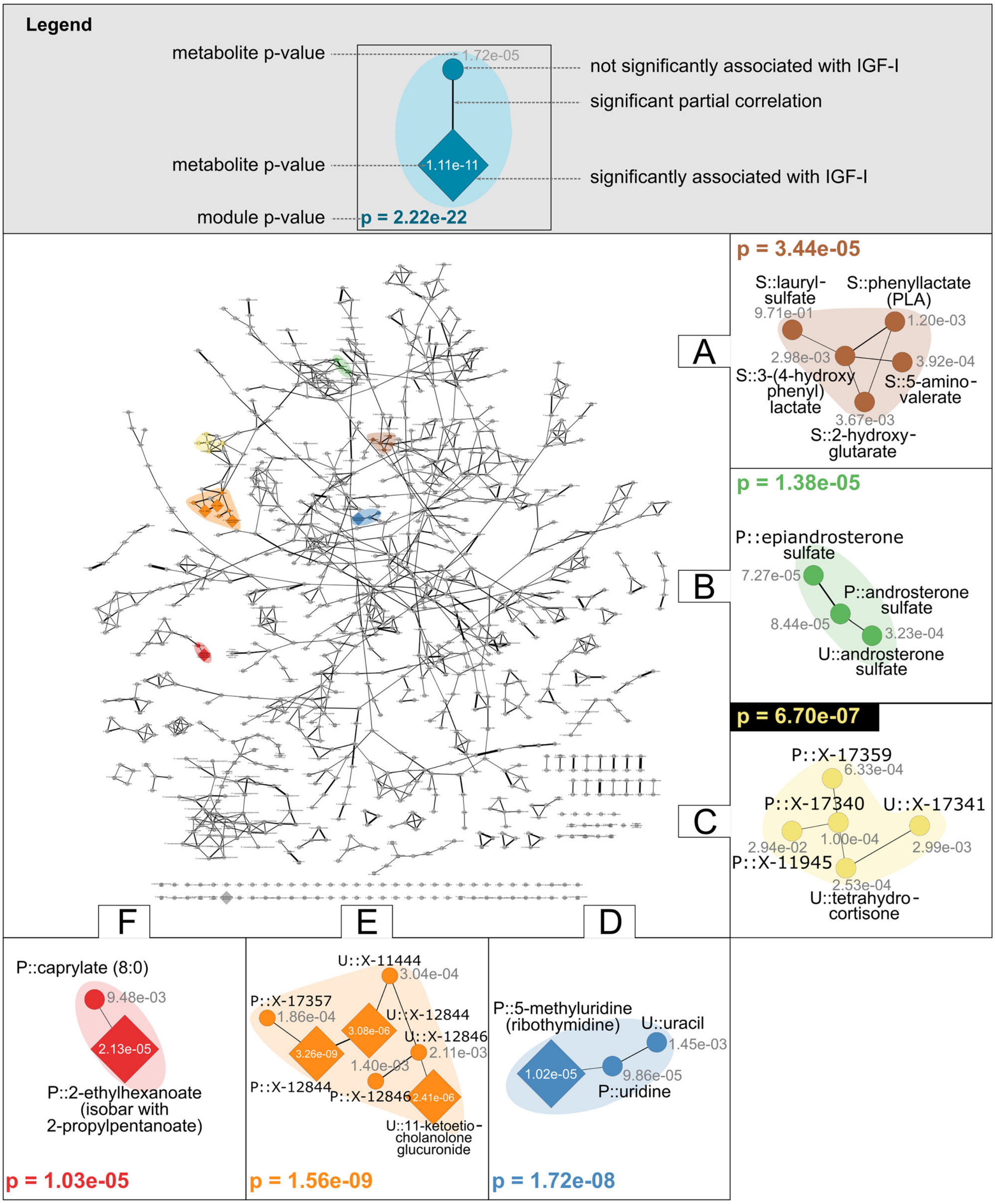


Fig. 4 IGF-I modules. This metabolite network is a relayouted version of the metabolite GGM in Fig. 2a. Edge widths reflect absolute partial correlation values. Each colored region corresponds to an identified IGF-I module. For readability, *p*-values are given in e-notation (e.g., 1.5e −5=1.5·10^{-5}). Node label prefixes P::, U::, and S:: indicate metabolites measured in plasma, urine, and saliva, respectively

Phenotype-driven module identification for dense associations: Gender

We next applied the module-identification algorithm with gender as phenotype, representing a trait with dense associations. As expected, for the metabolite network, we found a high number (73) of gender-associated modules (Supplementary Information S8). At the sub-pathway level, we identified 13 regulated modules (Fig. 5). Finally, at the super-pathway level, two modules indicating associations at a very global level were detected (Supplementary Information S8): the first module comprised plasma ‘Amino acid’

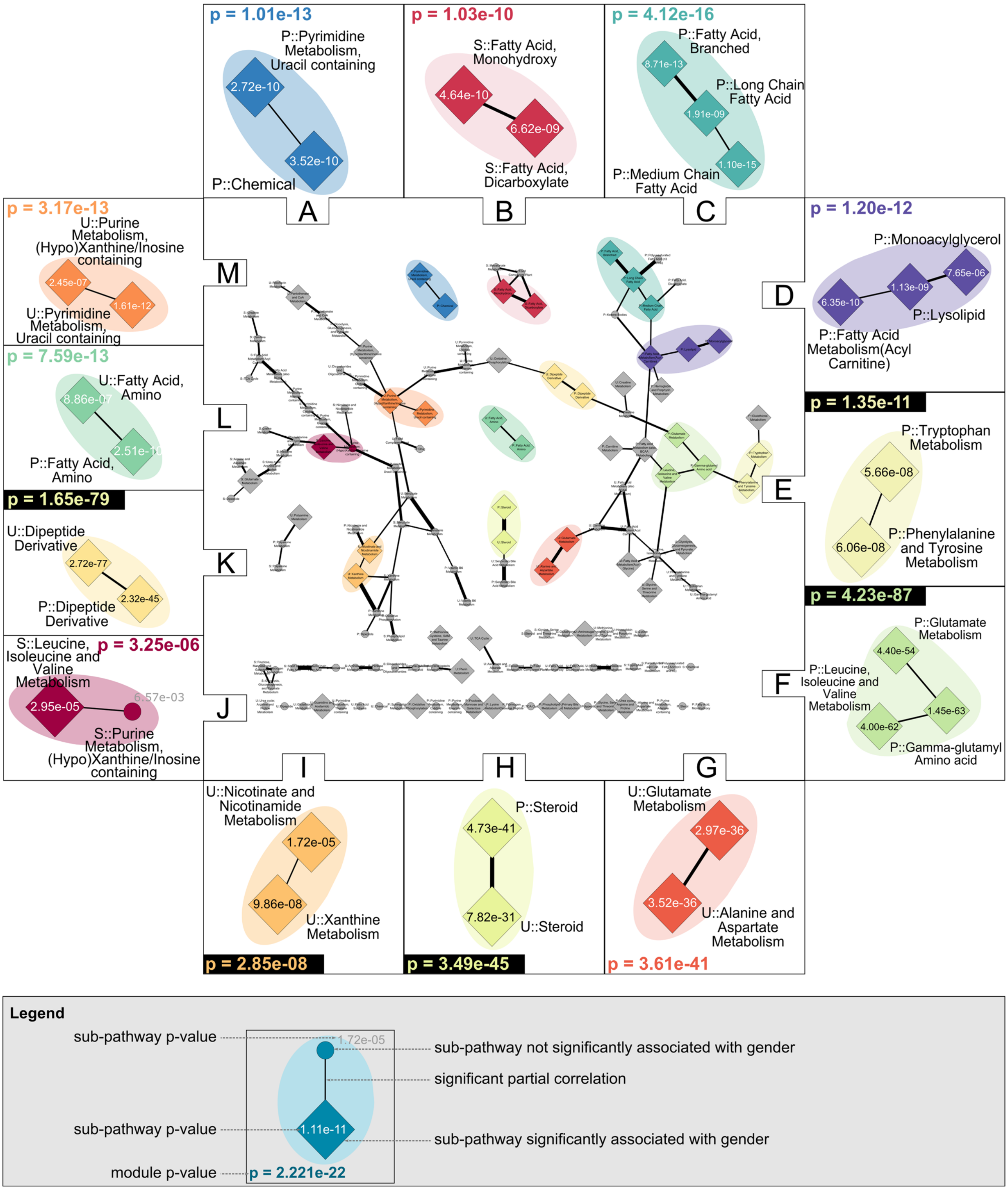


Fig. 5 Gender modules. This sub-pathway network is a relayouted version of the sub-pathway GGM in Fig. 2b. Edge widths reflect absolute partial correlation values. Each colored region of this network corresponds to one identified module. For readability, *p*-values are in e-notation (e.g., 1.5e−5=1.5·10^{-5}). Node label prefixes P::, U::, and S:: indicate metabolites measured in plasma, urine, and saliva, respectively

and 'Peptide', and the second module consisted of saliva 'Carbohydrate' and 'Amino acid'. Herein, we focus on modules detected at the sub-pathway level, which seemed to be an appropriate compromise between the metabolite and super-pathway levels.

From the 13 sub-pathway modules, 9 were within one fluid only (Fig. 5). We observed three multifluid modules comprising both plasma and urine sub-pathways (H, K, L). These results again demonstrate the strength of our approach to find phenotype-associated processes that span multiple pathways and even multiple body fluids. We were also able to reveal more subtle, non-obvious phenotype associations as shown in module J, where the combination of a gender-associated and a non-associated pathway led to a lower *p*-value than each pathway alone. In addition, the results show that the proposed method was able to link processes that appeared to be unrelated, since they were assigned two different pathways. This is shown in module I, which consisted of the plasma sub-pathways 'Nicotinate and Nicotinamide Metabolism' and 'Xanthine Metabolism'. Both pathways contain metabolites related to coffee metabolism. The former covers caffeine derivatives, while the latter consists of trigonelline, an alkaloid found in coffee. Interestingly, the module identification approach recognized the phenotype-driven relationship between these two pathways and grouped them into one gender module.

Similar to the IGF-I results, the method identified both previously reported and novel phenotype associations. A well-known metabolome–gender association, the steroid pathway, was thereby detected as a multifluid module, spanning the plasma and urine pathways (Module H). We also detected three intrafluid lipid modules (B–D), showing multiple processes within this pathway in which men and women differ. Modules C and D comprised pathways from blood, while module B consisted of salivary 'Fatty acid, monohydroxy' and 'Fatty acid, Dicarboxylate'. Several metabolites of these pathways in blood and saliva were found to associate with gender in previous studies,[10,26,41,42] but in addition, we were able to show that all pathways in saliva had a lower *p*-value when considered in combination. In module K, we found an association of histidyl peptides with gender, confirming previous findings in human muscle tissue that the female gender is associated with reduced levels of such peptides.[43,44] Moreover, we here illustrated that this sexual dimorphism can be observed across human blood and urine. Besides confirming known gender associations, we found a module (L) comprising plasma and urinary lipids of 'Fatty acid, Amino', which to the best of our knowledge, have not been reported before.

We investigated whether our identified modules could be replicated in the Qatar Metabolomics Study on Diabetes (QMDiab[3]). Since the set of measured metabolites differ between SHIP-TREND and QMDiab due to different profiling platforms, we only considered metabolites measured in both cohorts for an appropriate comparison. We generated a new hierarchical map based on the reduced SHIP-TREND dataset comprising 752 metabolites in total (490 known and 262 unknown), which were grouped into 134 different sub-pathways. The module search algorithm was run at the sub-pathway level of this newly generated hierarchical map for both SHIP-TREND and QMDiab. One-third of the gender-associated modules identified in the reduced SHIP-TREND were replicated in QMDiab (Supplementary Information S9). One factor accounting for this observation was the differing number of samples, i.e., the power of the cohorts. SHIP-TREND included 906 individuals with metabolomics measurement of all three body fluids, whereas QMDiab comprised a total of 372 participants. Finally, SHIP-TREND and QMDiab also differed in the study design. In SHIP-TREND, samples of fasting individuals were collected; whereas in QMDiab, the participants were nonfasting. The SHIP-TREND was conducted in West Pomerania, Germany, whereas the individuals in the QMDiab cohort were mainly of Arab and Asian ethnicity. Moreover, in contrast to SHIP-TREND, which was designed as a healthy cohort, QMDiab was a case-control study for type 2 diabetes. Despite substantial differences in study design and metabolomics measurement, QMDiab is, to the best of our knowledge, the only available cohort comprising metabolomics measurements in plasma, urine, and saliva, and therefore the only cohort available for replication of our results. Moreover, the replication of one-third of the results despite the substantial differences between the cohorts indicated that these results are very robust and generalizable.

We investigated the effects of oral hygiene on the gender-related modules. However, this analysis might be statistically unfeasible, because teeth brushing frequency was significantly associated with gender. Nevertheless, only one module was omitted when we corrected for the effects of oral hygiene (Supplementary Information S5).

DISCUSSION

In this paper, we presented an approach for the phenotype-driven identification of modules associated with phenotypes at multiple scales. To this end, a hierarchical, multifluid view of metabolism at three levels of granularity was generated and analyzed for metabolomics data from plasma, urine, and saliva of 906 participants in the SHIP-TREND cohort. A hierarchical module-identification procedure was then applied to this map for IGF-I measurements and gender representing phenotypes with 'sparse' and 'dense' associations, respectively.

The hierarchical map serves as a template of human metabolism for the module identification approach. But in addition, it allows to obtain a fundamental understanding of biochemical processes captured within and across body fluids. At all levels and as expected, the majority of correlations occurred within the same fluid. Moreover, most network edges were fluid-specific, that is, solely occurred in only one fluid, suggesting diverse metabolic processes in the fluids. This can probably be attributed to the substantially different physiological roles of each fluid, capturing metabolism at various levels. Analyzing the crosstalk between fluids, correlations were mainly observed between plasma and urine, followed by plasma and saliva, while only a few edges were found between urine and saliva. The strong link between plasma and urine was expected and reflects their close relationship through the excretion and reabsorption processes in the kidneys. Blood and saliva are also physiologically connected through the salivary glands. Finally, the weak urine–saliva crossfluid correlation might reflect an indirect connection of these fluids through blood. Overall, nearly half of the crossfluid correlations were observed between the same metabolites (e.g., plasma betaine and urine betaine). In the sub-pathway network, crossfluid correlations mostly connected the same sub-pathways (e.g., plasma 'Xanthine Metabolism' and saliva 'Xanthine Metabolism'). Such correlations between biochemically closely related molecules may arise due to transport and exchange processes across the fluids.

We then performed a module identification approach based on the hierarchical map. We found that IGF-I was associated with rather local parts of the metabolic network, while at the more global level (sub-pathways and super-pathways) fewer modules were detected. In contrast, for gender, we identified a large number of modules (73) in the fine-grained metabolite network. At the sub-pathway level, these numerous modules were fused into 13 sub-pathway modules, facilitating biological interpretation by providing a better overview over parts of metabolism affected by gender. Two modules were detected at the coarse super-pathway level, but did not promote biological interpretation in this case.

For both IGF-I and gender, we could confirm previously reported associations. In addition, our analyses extended these findings to multiple fluids. For example, we could extend the association between IGF-I and plasma pyrimidine metabolites to urine, which has already been reported by Knacke et al.[34] Moreover, to the best of our knowledge, in this study, IGF-I associations were analyzed in saliva for the first time. For gender, for instance, we showed that the association with histidyl peptides appears across human blood and urine. For both the phenotypes, the identification of multifluid modules suggested that not only the concentration levels of the respective metabolites in the corresponding fluids, but also the transport and exchange processes between the fluids were associated with the phenotypes. These types of findings could only be leveraged through a module search on multifluid data.

Our results demonstrated an increase in statistical power for the phenotype-driven module identification approach compared to classical analysis. We found significant modules comprising components that were not significantly associated with the phenotype when the single components were considered alone (e.g., IGF-I modules A–C). Moreover, by combining metabolite groups from different body fluids, statistical power is also substantially increased compared to the results from just a single fluid (e.g., gender modules H, K, L). This increase is most likely due to the reduction in statistical noise while aggregating measurements of multiple metabolites. Another major advantage of the module approach lies in overcoming borders of pathway definitions, which are inherently arbitrary for any commonly available metabolite-centric or process-centric pathway annotations. Our algorithm recognizes the phenotype-driven interplay of pathways and merges them, if appropriate, as shown in gender module I. This module reflected the well-known gender association with metabolites from caffeine metabolism,[45] which were originally assigned to two different pathways according to subpathway definitions in this study.

The present study could be extended in several directions. (i) We used pathway annotations provided by the metabolomics platform, which are analogous to KEGG pathways.[46] For future studies, other pathway definitions, such as in the HMDB[47] or MetaCyc[48] could be used; however, since a large number of measured metabolites will not be covered in those databases,[49] this approach would currently result in a substantial loss of information. (ii) Following the *eigengene* approach,[50] we defined a pathway representative as the first principal component of its metabolites. To capture a higher degree of variance explained, multivariate association methods, such as canonical correlation analysis[51] and O2-PLS[52] could be adapted to model the relationships between the two pathways. (iii) Our module identification approach is only suitable for finding modules where all components show the same direction of association with the phenotype (all positively or all negatively associated), while opposing effects will cancel out. A possible solution to this restriction would be the use of the multivariate modeling approaches mentioned in the previous point. (iv) We used data from the Qatar Metabolomics Study on Diabetes (QMDiab, Supplementary Information S9) to replicate the gender results. It would be interesting to also replicate our IGF-I results in a suitable cohort. To the best of our knowledge, no dataset comprising human plasma, urine, and saliva metabolomics data, as well as measurements of IGF-I are currently available besides the SHIP study. Moreover, to the best of our knowledge, QMDiab is the only available cohort comprising metabolomics measurements of plasma, urine, and saliva from the same individuals. (v) We applied our module identification approach to a multifluid metabolomics dataset. Owing to the rapid progress in high-throughput technologies, other omics data types have become readily available. It would be particularly interesting to include SNPs or transcripts, for instance, into the network.

In conclusion, we introduced a hierarchical map, that besides serving as a template of human metabolism for the module identification algorithm, can also be a valuable, downloadable resource for future studies, since it allows for a fundamental understanding of the complex correlation structure within and across multiple body fluids. Based on this map, we proposed an approach for the phenotype-driven identification of modules spanning multiple pathways and multiple body fluids. These modules provide deeper insights into mechanistic aspects of phenotype associations. Importantly, our module approach is generic, and therefore widely applicable. An R implementation of the algorithm is freely available as supplementary material for this paper. It can be used directly for any other dataset, given the presence of a data matrix, annotations of the respective variables, and a phenotype.

MATERIALS AND METHODS

Study cohort, metabolomics, and IGF-I measurement

Metabolomics data were obtained from the Study of Health in Pomerania (SHIP-TREND), conducted between 2008 and 2011 in West Pomerania, Germany, with 4420 participants. The study was approved by the local ethics committee and conformed to the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants. Details about sample acquisition and experimental procedures can be found elsewhere.[29,34] Briefly, metabolomics measurements were performed for a subset of 1000 participants without self-reported diabetes. The dataset included 561 females and 439 males with an age distribution of 50.14 ± 13.17 (mean ± SD) and 50.08 ± 14.24, and a BMI distribution of 26.99 ± 5.12 and 27.85 ± 3.7, respectively. Fasting (≥8 h) plasma and urine samples were collected between 07:00 and 12:00 am. Blood was sampled from the cubital vein of subjects in a supine position. Samples were stored at −80 °C. Stimulated saliva was collected with a commercially available collection system (Salivette®). The subjects chewed a plain cotton roll for exactly 1 min to stimulate salivation. The rolls with the absorbed saliva were placed into the Salivette® and immediately centrifuged at 1000×*g* for 20 min at 4 °C to remove food remnants, insoluble material, and cell debris. The resulting supernatant was stored at −80 °C. Samples were analyzed on an untargeted metabolomics platform established by Metabolon Inc. (Durham, USA) with ultra-high liquid-phase chromatography coupled with tandem mass spectrometry (UPLC-MS/MS) in both positive and negative modes. The measurements were performed at the Genome Analysis Center, Helmholtz Zentrum, Munich, yielding a total of 1665 metabolites across all fluids, of which 1190 represented unique metabolites. Blood IGF-I concentrations were determined by automated two-site chemiluminescent immunoassays on the IDS-iSYS kit (Immunodiagnostic Systems, Boldon, UK).

Preprocessing and quality control

To correct for daily variations of the metabolomics platform, raw ion counts of each metabolite were rescaled by their respective median value on the run day. To ensure valid medians, metabolites with fewer than three measured values for more than the half of the run days were filtered out. This procedure resulted in 1317 total (475, 558, and 284 metabolites for plasma, urine, and saliva, respectively) and 991 unique metabolites from all three body fluids. Probabilistic quotient normalization (PQN) was then applied to urine samples to account for diurnal variation. PQN has previously been shown to be superior to common creatinine scaling.[53] PQN was, moreover, used to normalize saliva measurements for dilution variations. For the PQN procedure, first a 'pseudo-sample' (reference) was calculated as the mean of all metabolites with no missing entries for all participants (131 urine and 37 saliva metabolites). Subsequently, a dilution factor was estimated as the median quotient between the reference and each sample. Finally, all measurements were divided by the respective dilution value. Of note, urine creatinine and the estimated urinary dilution factor were substantially correlated ($r = 0.91$, $p < 0.001$) within the SHIP-TREND data (Supplementary Information S1).

All metabolite levels and serum IGF-I measurements were $\log_2$-transformed. Multivariate outlier detection (using only metabolites with no missing values across all samples) was performed separately for all fluids using an algorithm proposed by Filzmoser et al. (2008),[54] implemented in the *pcout* function within the R package *mvoutlier*. Briefly, this algorithm calculates an outlier score for each sample using principal component analysis and the Mahalanobis distance on a robustly scaled data matrix. Default parameters were used for the identification process, and the exclusion criterion was set to 4 SD. As a result, 13, 8, and 16 samples from plasma, urine, and saliva, respectively, were excluded from further analyses. After these preprocessing steps, the dataset comprised 906 individuals for which fasting plasma, urine, and saliva samples, as well as IGF-I measurements were available.

Since for the network inference procedure a fully observed data matrix is required, missing values were imputed by the following procedure: All metabolites with more than 20% missing values (320) were excluded from the dataset to avoid false positive results and to preserve statistical power. The first step of imputation was performed per run day on the log-transformed raw data (before normalization). Following the assumptions that missing values occur due to a detection threshold and that metabolites are log-normally distributed, each missing value was replaced by a random value drawn from the censored part of a normal distribution reconstructed by maximum-likelihood estimation.[55,56] To ensure robust parameter estimation for the truncated normal distributions, this procedure was only applied to metabolites for which the respective run day contained more than 10 nonmissing concentration values. Remaining missing values were imputed with the *mice* R package (version 2.22) with predictive mean matching as an elementary imputation model. Note that we also used the stringent threshold of 20% to exclude variables with missing values, because estimating (partial) correlations based on too many imputed values that in turn were generated by mice using the covariance structure of the data might introduce unwanted bias. The final imputed dataset consisted of 906 samples and 997 metabolites.

Metabolite pathway representation

Each metabolite with known chemical structure (610 metabolites) was annotated with one of the 73 sub-pathways (such as 'Lysolipid', 'TCA Cycle', 'Glycolysis', 'Branched-chain amino acid'), and one of eight more general super-pathways ('Amino acid', 'Lipid', 'Carbohydrate', 'Nucleotide', 'Peptide', 'Energy', 'Cofactors and vitamins', and 'Xenobiotics'). The remaining 387 metabolites have unknown chemical structure (*unknowns*), and thus, cannot be assigned to any pathway for which reason they were excluded from the pathway analyses. A detailed list of metabolites and their annotated pathways is provided in Supplementary Information S1.

For each sub-pathway, a principal component analysis was performed after scaling all variables to a mean of 0 and a variance of 1. The first principal component was used as a representative value for the entire set of metabolites in the pathway. These *eigenmetabolites*[23,50,57] were then subjected to the network inference procedure below.

Network inference

Two networks were inferred using GGMs, one for metabolite concentrations (all metabolites) and one for the sub-pathway *eigenmetabolites* (unknowns excluded) using the *GeneNet* R package, version 1.2.12. GGMs are based on partial correlations, which represent the linear associations between two variables corrected for all remaining variables in multivariate Gaussian distributions. We included age, gender, and BMI as standard covariates into the model. Edges between metabolites or sub-pathways were assigned if both their Pearson correlations and their partial correlations were statistically significant with $\alpha = 0.05$ after the Bonferroni correction for $\binom{p}{2}$ tests, where p is the number of metabolites or sub-pathways, respectively.

To obtain a global view of connections between the super-pathways, the sub-pathway GGM was *collapsed* into a super-pathway network. To this end, a link between two nodes was drawn if there was at least one connection between any two sub-pathways assigned to the two respective super-pathways in the underlying sub-pathway GGM.

Module identification algorithm

Module representatives. For a *candidate module M*, a representative value R_M is defined as the average of scaled intensities (average z-score) of all metabolites in M. If M consists of sub-pathways, then the representative is calculated as the mean z-score of all metabolites in the set union of all sub-pathways. Notably, for pathway network estimation, a pathway representative was defined as the sub-pathway *eigenmetabolite* based on the assumption that pathway components share common chemical and biological properties. In contrast, we chose to use mean z-scores as module representatives, since modules are considerably more heterogeneous.

Scoring function. The score of each candidate module M is obtained from the multivariable linear regression model

$$R_M \sim \beta_{M,0} + \beta_{M,1} \cdot P + \beta_{M,2} \cdot \text{gender} + \beta_{M,3} \cdot \text{age} + \beta_{M,4} \cdot \text{BMI} + \in_M \qquad (1)$$

where R_M is the aforementioned representative value, $\beta_{M,0}$ is the intercept, $\beta_{M,1},\ldots,\beta_{M,4}$ are the regression coefficients for each independent variable, P is the phenotype of interest, and ϵ_M is a normally distributed error term. The module score is then defined as the negative logarithmized p-value of the coefficient $\beta_{M,1}$, which represents the magnitude of phenotype association. Notably, the score of a single component equals its negative logarithmized p-value from a univariate analysis. Furthermore, the scoring function for gender does not contain gender as a covariate.

Module identification. Given the scoring function and an initial node (seed node), a greedy search procedure is performed to identify an *optimal module*. In every iteration, each neighboring node of the *candidate module* is added and the score of the extended module is calculated. The neighbor leading to the highest score improvement is then added to the module. Furthermore, a neighbor is only added if the score of the new module is higher than the scores of all single components. The algorithm terminates if no further improvements can be made. In a final step, overlapping *optimal modules* from different seed nodes are combined into a single module (*maximal module*), which is rescored by the scoring function.

For the identification of IGF-I-associated and gender-associated modules, the procedure was applied to all three networks, namely, the metabolite, the sub-pathway, and the super-pathway networks. To assess the significance of the modules, a conservative multiple testing correction procedure was used with a significance level of $\alpha = 0.05$ after the Bonferroni correction for the total number of nodes in the underlying network. The proposed algorithm is visually described in Supplementary Information S6 and available as R code in Supplementary Information S10. An example of how to execute the R scripts in S10 is explained in S11.

DATA AVAILABILITY

The data that support findings of this study are available from the Ernst-Moritz-Arndt-Universität Greifswald but restrictions apply to the availability of these data (the informed consent given by the study participants does not cover data posting in public databases), which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Ernst-Moritz-Arndt-Universität Greifswald or can be directly applied for via www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php?lang=ger.

Code availability

An implementation of the generic approach and an example script is freely available as supplementary material (Supplementary Information S10 and S11).

ACKNOWLEDGEMENTS

We would like to thank Bianca Schmick from Genome Analysis Center for expert technical assistance. This work was supported in part by grants from the German Federal Ministry of Education and Research (BMBF), by BMBF Grant No. 01ZX1313C (project e:Athero-MED), Grant No. 03IS2061A and 03IS2061B (project Gani_Med). Gani_Med was also supported by BMBF Grant No. 01GI0883 (project COSYCONET), and BMBF IntegraMent. Moreover, the research leading to these results has received funding from the European Union's Seventh Framework Program [FP7-Health-F5-2012] under grant agreement no. 305280 (MIMOmics). K.S. is supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. D.O.M.-K. is supported by the Dutch Science Organization (ZonMW-VENI Grant 916.14.023). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

This study was designed by K.T.D., G.K., and J.K. M.P., N.F., M.N., and T.K. performed sample preparation, data acquirement, and data quality control for SHIP-TREND. K.S. and D.O.M.-K. performed sample preparation and data acquirement for QMDiab. K.T. D. performed the computational and statistical analyses. D.J.N.P.R. implemented the approach in R. K.T.D. and J.K. wrote the primary manuscript. All authors approved the final manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Systems Biology and Applications* website (https://doi.org/10.1038/s41540-017-0029-9).

Competing interests: The authors declare no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Weckwerth, W. Metabolomics in systems biology. *Annu. Rev. Plant Biol.* **54**, 669–689 (2003).
2. Wang, Y., Liu, S., Hu, Y., Li, P. & Wan, J. -B. Current state of the art of mass spectrometry-based metabolomics studies – a review focusing on wide coverage, high throughput and easy identification. *RSC Adv.* **5**, 78728–78737 (2015).
3. Mook-Kanamori, D. O. et al. 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *J. Clin. Endocrinol. Metab.* **99**, E479–E483 (2014).
4. Urpi-Sarda, M. et al. Metabolomics for biomarkers of type 2 diabetes mellitus: advances and nutritional intervention trends. *Curr. Cardiovasc. Risk Rep.* **9**, 1–12 (2015).
5. Rhee, E. P. & Gerszten, R. E. Metabolomics and cardiovascular biomarker discovery. *Clin. Chem.* **58**, 139–147 (2012).
6. Jensen, M. K. et al. Novel metabolic biomarkers of cardiovascular disease. *Nat. Rev. Endocrinol.* **10**, 659–672 (2014).
7. Han, X. et al. Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. *PLOS ONE* **6**, e21643 (2011).
8. Sato, Y. et al. Identification of a new plasma biomarker of Alzheimer's disease using metabolomics technology. *J. Lipid Res.* **53**, 567–576 (2012).
9. González-Domínguez, R., García-Barrera, T. & Gómez-Ariza, J. L. Metabolomic study of lipids in serum for biomarker discovery in Alzheimer's disease using direct infusion mass spectrometry. *J. Pharm. Biomed. Anal.* **98**, 321–326 (2014).
10. Do, K. T. et al. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *J. Proteome Res.* **14**, 1183–1194 (2015).
11. Yousri, N. A. et al. A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. *Diabetologia* **58**, 1855–1867 (2015).
12. Kim, J. W. et al. Pattern recognition analysis for hepatotoxicity induced by acetaminophen using plasma and urinary 1H NMR-based metabolomics in humans. *Anal. Chem.* **85**, 11326–11334 (2013).
13. Munshi, S. U., Rewari, B. B., Bhavesh, N. S. & Jameel, S. Nuclear magnetic resonance based profiling of biofluids reveals metabolic dysregulation in HIV-infected persons and those on anti-retroviral therapy. *PLoS ONE* **8**, e64298 (2013).
14. Vitkin, E. et al. Peer group normalization and urine to blood context in steroid metabolomics: the case of CAH and obesity. *Steroids* **88**, 83–89 (2014).
15. Dudzik, D. et al. Metabolic fingerprint of gestational diabetes mellitus. *J. Proteom.* **103**, 57–71 (2014).
16. Walsh, M. C. et al. Impact of geographical region on urinary metabolomic and plasma fatty acid profiles in subjects with the metabolic syndrome across Europe: the LIPGENE study. *Br. J. Nutr.* **111**, 424–431 (2014).
17. Mitra, K., Carvunis, A. -R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14**, 719–732 (2013).
18. Polanski, K. et al. Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics* **30**, 962–970 (2014).
19. Kim, Y. -A., Cho, D. -Y., Dao, P. & Przytycka, T. M. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* **31**, i284–292 (2015).
20. Chuang, H. -Y., Lee, E., Liu, Y. -T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
21. May, A. et al. metaModules identifies key functional subnetworks in microbiome-related disease. *Bioinformatics* **32**, 1678–1685 (2016).
22. Martignetti, L., Calzone L., Bonnet E., Barillot E., Zinovyev A. (2016) ROMA: representation and quantification of module activity from target expression data. Front. Genet. 7:18 (2016).
23. DiLeo, M. V., Strahan, G. D., Bakker, Mden & Hoekenga, O. A. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLOS ONE* **6**, e26683 (2011).
24. Fukushima, A., Kusano, M., Redestig, H., Arita, M. & Saito, K. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Syst. Biol.* **5**, 1 (2011).
25. Ried, J. S. et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Hum. Mol. Genet.* **23**, 5847–5857 (2014).
26. Krumsiek, J. et al. Gender-specific pathway differences in the human serum metabolome. *Metabolomics* **11**, 1815–1833 (2015).
27. Mittelstrass, K. et al. Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* **7**, e1002215 (2011).
28. Floegel, A. et al. Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: findings from a population-based study. *Int. J. Obes. 2005* **38**, 1388–1396 (2014).
29. Völzke, H. et al. Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
30. Pollak, M. The insulin and insulin-like growth factor receptor family in neoplasia: an update. *Nat. Rev. Cancer* **12**, 159–169 (2012).
31. Ren, J. & Anversa, P. The insulin-like growth factor I system: physiological and pathophysiological implication in cardiovascular diseases associated with metabolic syndrome. *Biochem. Pharmacol.* **93**, 409–417 (2015).
32. Li, D. -H., He, Y. -C., Quinn, T. J. & Liu, J. Serum insulin-like growth factor-1 in patients with De Novo, drug Naïve parkinson's disease: a meta-analysis. *PLoS ONE* **10**, e0144755 (2015).
33. Aguirre, G. A., Ita, J. R., Garza, R. G. & Castilla-Cortazar, I. Insulin-like growth factor-1 deficiency and metabolic syndrome. *J. Transl. Med.* **14**, 3 (2016).
34. Knacke, H. et al. Metabolic fingerprints of circulating IGF-I and the IGF-I/IGFBP-3 ratio: a multi-fluid metabolomics study. *J. Clin. Endocrinol. Metab.* **101**, 4730–4742 (2016).
35. Krumsiek, J. et al. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* **8**, e1003005 (2012).
36. Nieman, D. C., Shanely, R. A., Gillitt, N. D., Pappan, K. L. & Lila, M. A. Serum metabolic signatures induced by a three-day intensified exercise period persist after 14 h of recovery in runners. *J. Proteome Res.* **12**, 4577–4584 (2013).
37. Poisson, L. M. et al. A metabolomic approach to identifying platinum resistance in ovarian cancer. *J. Ovarian Res* **8**, 13 (2015).
38. Kanbur-Oksüz, N., Derman, O. & Kinik, E. Correlation of sex steroids with IGF-1 and IGFBP-3 during different pubertal stages. *Turk. J. Pediatr.* **46**, 315–321 (2004).
39. Meinhardt, U. J. & Ho, K. K. Y. Modulation of growth hormone action by sex steroids. *Clin. Endocrinol.* **65**, 413–422 (2006).
40. Floyd, S. et al. The insulin-like growth factor-I–mTOR signaling pathway induces the mitochondrial pyrimidine nucleotide carrier to promote cell growth. *Mol. Biol. Cell* **18**, 3545–3555 (2007).
41. Santosa, S. & Jensen, M. D. The sexual dimorphism of lipid kinetics in humans. *Front. Endocrinol.* **6**, 103 (2015).
42. Saito, K. et al. Gender- and age-associated differences in serum metabolite profiles among Japanese populations. *Biol. Pharm. Bull.* **39**, 1179–1186 (2016).
43. Everaert, I. et al. Vegetarianism, female gender and increasing age, but not CNDP1 genotype, are associated with reduced muscle carnosine levels in humans. *Amino Acids* **40**, 1221–1229 (2010).
44. Jung, S. et al. Carnosine, anserine, creatine, and inosine 5′-monophosphate contents in breast and thigh meats from 5 lines of Korean native chicken. *Poult. Sci.* **92**, 3275–3282 (2013).
45. Temple, J. L. & Ziegler, A. M. Gender differences in subjective and physiological responses to caffeine and the role of steroid hormones. *J. Caffeine Res* **1**, 41–48 (2011).
46. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–114 (2012).
47. Wishart, D. S. et al. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
48. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
49. Bartel, J. et al. The human blood metabolome-transcriptome interface. *PLoS Genet.* **11**, e1005274 (2015).
50. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54 (2007).
51. Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
52. Trygg, J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J. Chemom.* **16**, 283–293 (2002).
53. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* **78**, 4281–4290 (2006).
54. Filzmoser, P., Maronna, R. & Werner, M. Outlier identification in high dimensions. *Comput. Stat. Data Anal.* **52**, 1694–1711 (2008).
55. Richardson, D. B. & Ciampi, A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am. J. Epidemiol.* **157**, 355–363 (2003).
56. Nie, L. et al. Linear regression with an independent variable subject to a detection limit. *Epidemiology* **21**, S17–S24 (2010).
57. Wahl, S. et al. Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Med.* **13**, 48 (2015).

Appendix C

MoDentify: a tool for phenotype-driven identification of modules in high-throughput data.

Do KT, Rasp D, Kastenmüller G, Suhre K, Krumsiek J (2017), **MoDentify: a tool for phenotype-driven identification of modules in high-throughput data.** *Under review* in *Oxford Bioinformatics*

This work is under review in *Oxford Bioinformatics*. The version appended here is the PDF on *bioRxiv*, a preprint repository.

MoDentify: a tool for phenotype-driven module identification in multilevel metabolomics networks

Kieu Trinh Do[1], David J.N.-P. Rasp[1], Gabi Kastenmüller[2,3], Karsten Suhre[4], and Jan Krumsiek[1,2,5*]

[1]Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany

[2]German Center for Diabetes Research (DZD), Neuherberg, Germany

[3]Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum, Neuherberg, Germany

[4]Department of Physiology and Biophysics, Weill Cornell Medical College - Qatar, Education City, Doha, Qatar

[5]Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

*To whom correspondence should be addressed.

Abstract

Summary: Metabolomics is an established tool to gain insights into (patho)physiological outcomes. Associations of metabolism with such outcomes are expected to span functional modules, which are defined as sets of correlating metabolites that are coordinately regulated. Moreover, these associations occur at different scales, from entire pathways to only a few metabolites, which is an aspect that has not been addressed by previous methods. Here we present *MoDentify*, a freely available R package to identify regulated modules in metabolomics networks at different layers of resolution. Importantly, *MoDentify* shows higher statistical power than classical association analysis. Moreover, the package offers direct visualization of results as interactive networks in Cytoscape. We present an application example using a complex, multifluid metabolomics dataset. Owing to its generic character, the method is widely applicable to any dataset with a phenotype variable, a data matrix, and optional pathway annotations.

Availability and Implementation: *MoDentify* is freely available from GitHub:
https://github.com/krumsiek/MoDentify
The package vignette contains a detailed tutorial of the analysis workflow.

Contact: jan.krumsiek@helmholtz-muenchen.de

1 Introduction

Associations with phenotypic parameters and clinical endpoints in large-scale, heterogeneous metabolomics datasets are complex. They typically span entire functional modules, which are defined as groups of correlating molecules that are functionally coordinated, coregulated, or generally driven by a common biological process (Mitra *et al*, 2013). The systematic identification of modules is often based on networks, where nodes correspond to the molecules under investigation, and edges represent the correlations or associations between two molecules. Modules are commonly identified as highly connected parts of the network that contain nodes that are coordinately associated with a given phenotype.

Systematic module identification algorithms are well established for various types of *omics* data (Polanski *et al*, 2014; Chuang *et al*, 2007;

May *et al*, 2016; Martignetti *et al*, 2016). However, they have scarcely been applied to metabolomics data. Moreover, none of these methods consider that phenotype associations can occur at different scales, ranging from global associations spanning entire pathways or even sets of pathways (e.g., "dense" associations between metabolomics and gender or BMI), to localized associations with only a few metabolites (e.g., "sparse" associations between metabolomics and insulin-like growth-factor I levels or asthma) (Do *et al*, 2017). For sparse associations, the identification and interpretation of modules is usually straightforward. However, modules for dense phenotype associations at the metabolite level are challenging to interpret due to their overwhelming number. To facilitate interpretation, the plethora of information at the fine-grained metabolite level can be condensed to a hierarchically superordinate level, such as a pathway network. Here, nodes correspond to entire pathways, edges represent pathway relationships, and modules reflect phenotype-associated processes covering sets of pathways.

We have recently introduced a module identification algorithm for multifluid metabolomics data (Do *et al*, 2017). The approach was applied to blood concentrations of insulin-like growth factor (IGF-I) and gender as examples of sparse and dense phenotype associations, respectively. We here present *MoDentify*, a free R package implementing the approach for general use. In particular, *MoDentify* offers (i) the estimation of data-driven networks based on Gaussian graphical models (GGMs), (ii) module identification at both fine-grained metabolite level and more global pathway levels, and (iii) visualization of the identified modules in an interactive network through Cytoscape (Shannon *et al*, 2003). *MoDentify* increases statistical power compared with classical association analysis due to the reduction of statistical noise and can easily be applied to any type of quantitative data because of its generic character.

2 Description

MoDentify identifies network-based modules that are highly affected by a phenotype of interest. The underlying network is either directly inferred from the data at the single metabolite or pathway level (see below) or can be provided from an external source.

Network inference: *MoDentify* estimates either classical Pearson correlation networks or GGMs using the *GeneNet* R package (Opgen-Rhein & Strimmer, 2007). GGMs are based on partial correlations, which represent associations between two variables corrected for all remaining variables in multivariate Gaussian distributions (Krumsiek *et al*, 2011). An important property of GGMs compared with Pearson correlation networks is their sparsity, because only direct correlations are included. At the fine-grained level, the GGM consists of nodes corresponding to metabolites and edges representing significant partial correlations between two nodes after multiple testing correction. At the pathway level, the GGM consists of nodes corresponding to entire pathways (sets of metabolites), whereas edges represent significant partial correlations between two pathways. To estimate a correlation network between pathways, representative values are computed for each pathway (see **Pathway representation**). Alternatively, a network from an external source can be provided. Importantly, all nodes in the network must be measured in the given dataset.

Pathway representation: As stated above, in addition to regular network inference, *MoDentify* can build a network of interacting pathways. To this end, a new variable is defined as a representative for each pathway, which aggregates the total abundance of metabolites from the pathway into a single value. *MoDentify* provides two approaches for pathway representation:

(1) *eigenmetabolite* approach: For each pathway, a principal component analysis (PCA) is performed after scaling all metabolites to a mean of 0 and a variance of 1. The first principle component – also termed *eigenmetabolite* – is used as a representative value for the entire set of variables in the pathway (Langfelder and Horvath 2007).

(2) *average* approach: All variables are first scaled to mean 0 and variance 1. Subsequently, the pathway representative is calculated as the average of all variable values in the pathway.

MoDentify computes the amount of explained variances explained by each eigenmetabolite per pathway to facilitate the choice between these two approaches. If explained variances are high, the *eigenmetabolite* approach should be used; otherwise, the *average* approach might be the more appropriate choice.

Module identification: To identify functional modules, *MoDentify* uses a score maximization approach. Given a network, a scoring function, and a starting node (seed node) as the initial candidate module, the algorithm identifies an optimal module by score maximization. For a given candidate module, the following procedure is performed: Each neighboring node of the module is subsequently added to the candidate module, and the score of the extended module is calculated (see **Module scoring**). The neighbor resulting in the highest score improvement is finally added to the candidate module, if the new module score is higher than the score of each of its single components. The procedure is repeated until no further score improvements can be made, yielding the optimal module for the given seed node. In an optional consolidation step, overlapping optimal modules from different seed nodes are combined into one module, which is reevaluated by the scoring function.

Module scoring: The score of a candidate module is obtained from the multivariable linear regression model

$$R = \beta_0 + \beta_1 \times P + \sum_{i=1}^{|C|} \beta_{i+1} \times c_i + \epsilon$$

where R is the module representative defined by the *eigenmetabolite* or *average* approach (see **Pathway representation**), β_0 is the intercept, β_i are the regression coefficients for the respective independent variables, P is the phenotype of interest, C is an optional set of covariates c_i, and ϵ is a normally distributed error term. The module score is then defined as the negative log-transformed *p*-value of β_1. The significance of the modules is assessed by correcting for the total number of nodes in the underlying network.

Module visualization: *MoDentify* offers visualization of the identified modules within an interactive network in the open source software Cytoscape (Shannon *et al*, 2003). The network contains different node colors and node sizes, which depict the membership of a node in a certain module and its association with the given phenotype from a classical, single-molecule association analysis, respectively. Moreover, significance of the phenotype association is indicated by diamond-shaped nodes. In addition to returning R data structures and producing flat-file results, one of the main advantages of our module visualization is the direct call of Cytoscape from within R *via* the *RCytoscape* package (Shannon *et al*, 2013) for external visualization, without cumbersome exporting of data files from R and re-importing them into Cytoscape.

3 Application example

We demonstrate the easy use of *MoDentify* on plasma, urine, and saliva metabolomics data from the Qatar Metabolomics Study on Diabetes (QMDiab) (Mook-Kanamori *et al*, 2014), aiming to identify functional modules associated with type 2 diabetes (T2D). The multifluid dataset comprises mass spectrometry-based metabolomics measurements for 190 diabetes patients and 184 healthy controls of Arab and Asian ethnicities aged 17–81 years. The dataset consists of 1524 metabolites. For each metabolite, two levels of pathway annotation are available. The preprocessed QMDiab data (normalized, log-transformed, missing values handled, and scaled) are integrated within the *MoDentify* package. The dataset is also available from the following figshare repository via the following link https://doi.org/10.6084/m9.figshare.5904022.

MoDentify was applied to the QMDiab dataset at both metabolite and pathway levels. The following code with default parameters produces a list of metabolite modules associated with T2D, as well as interactive visualization of the modules in the underlying network in Cytoscape (Figure 1A). Here, we only show code for the application of *MoDentify* at the metabolite level. Code for application at the pathway level (Figure 1B) can be found in the package vignette, available from the GitHub repository.

```
# Load MoDentify
library(MoDentify)

# Network inference
met.graph <- generate.network(data = qmdiab.data, annotations = qmdiab.annos)

# Module identification
modules.summary <- identify.modules(graph = met.graph, data = qmdiab.data,
                                    annotations = qmdiab.annos,
                                    phenotype = qmdiab.phenos$T2D)

# Module visualization
draw.modules(graph = met.graph, summary = modules.summary)
```

By default, `generate.network` estimates partial correlations between metabolites and assigns edges using a significance threshold of $\alpha = 0.05$ after Bonferroni multiple testing correction. `identify.modules` searches network modules for the given phenotype, where the default module representation approach is the *average* approach, $\alpha = 0.05$ is set for significance filtering, and Bonferroni multiple testing correction is applied. The output structure `modules.summary` contains a list of modules with their components and scores, which can be visualized within an interactive network in Cytoscape using `draw.modules`.

MoDentify identified 36 modules for T2D at the metabolite level (Figure 1A). Many of these modules consist of metabolites that are not significantly associated with T2D if considered alone. However, in interplay with other metabolites, they form a module that is more strongly associated with T2D than all of its single components. This increased statistical power in *MoDentify* can be attributed to the reduction of statistical noise when aggregating module components and allows the detection of links between metabolites and phenotype that would have been missed with classical association analysis. *MoDentify* found several modules containing metabolites from at least two fluids. For instance, one module (orange in Figure 1A) comprises the three vitamin B derivatives plasma pantothenate (vitamin B_5) and pyridoxate (vitamin B_6), and urine riboflavin (vitamin B_2). Although pyridoxate and riboflavin are not related to T2D when analyzed alone, they form a module in combination with pantothenate that is significantly associated with the phenotype. This module corroborates previous observations that vitamin B levels in blood and urine are associated with T2D (Nix *et al*, 2015; Unoki-Kubota *et al*, 2010; Valdés-Ramos *et al*, 2015). In addition, the results indicate that not only the concentration levels in blood and urine but also exchange processes between the two fluids are linked to T2D as well. At the pathway level (Figure 1B), six modules were detected. These modules show the interplay of multiple pathways in diabetes. For instance, one module comprises plasma metabolites from glutathione and histidine metabolism and urinary metabolites from histidine metabolism (yellow in Figure 1B). Although histidine and glutathione were shown to be related to diabetes in previous studies (Kimura *et al*, 2013; Sekhar *et al*, 2011), the identified module suggests that histidine and glutathione metabolism as well as the secretion of histidine derivatives might be part of the same process in T2D.

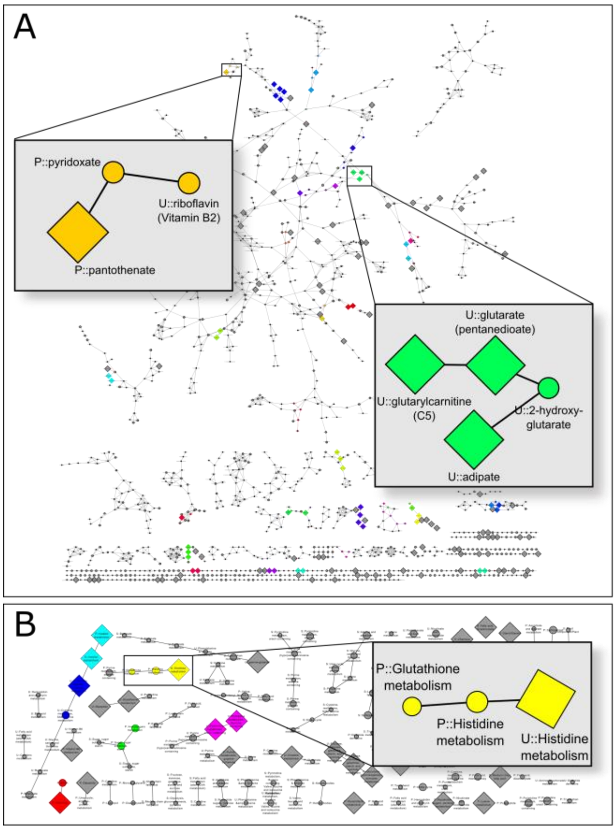


Figure 1 Visualization of identified modules for type 2 diabetes. The metabolomics networks with embedded modules at metabolite level (**A**) and pathway level (**B**) are screenshots of the interactive versions in Cytoscape produced by *MoDentify*. Zoom-ins have been added to highlight examples for *MoDentify*'s increased statistical power and its ability to extract biologically valuable insights. Round nodes correspond to metabolic entities not significantly associated with T2D when considered alone. Diamond nodes represent metabolic entities significantly related to T2D.

4 Conclusion

To the best of our knowledge, *MoDentify* implements the first approach for the systematic identification of phenotype-driven modules at different layers of resolution. To this end, the algorithm allows the estimation of data-driven networks based on Pearson or partial correlations. Optionally, a network from an external source can be provided. To facilitate result interpretation for different scales of phenotype associations, *MoDentify* enables the module search at both fine-grained metabolite level and more global pathway levels. Owing to the increased statistical power of the approach, novel links between clinical parameters and molecular levels can be detected. We presented an application example using a complex multifluid metabolomics dataset, but owing to its generic character, this approach can be applied for any quantitative dataset.

Acknowledgments

We wish to thank the study participants and research team of the QMDiab study for their contributions. The QMDiab study was approved by the Institutional Review Boards of HMC and WCM-Q under research protocol number 11131/11. All study participants provided written informed consent.

Funding

KD is supported by a grant (Grant no. 01ZX1313C, project e:Athero-MED) from the German Federal Ministry of Education and Research (BMBF). JK has received funding from the European Union's Seventh Framework Program (FP7-Health-F5-2012) under grant agreement no. 305280 (MIMOmics). GK is supported by funding from the National Institute of Aging (NIH/NIA) under grant no. 1RF1AG057452-01 (Metabolic Network Analysis of Biochemical Trajectories in Alzheimer's Disease) and from Qatar National Research Fund grant NPRP8-061-3-011 (Rational Metabolic Engineering as a New Tool for Targeted Intervention in Human Metabolism). All statements are made on behalf of the authors and do not necessarily represent the views of the research agencies. KS is supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation.

Conflict of Interest: none declared.

References

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140

Do,K.T. *et al.* (2017) Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations. *NPJ Syst. Biol. Appl.*, **3**, 28

Kanehisa,M. *et al*.(2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–114

Kimura,K. *et al*. (2013) Histidine Augments the Suppression of Hepatic Glucose Production by Central Insulin Action. Diabetes 62: 2266–2277

Krumsiek,J. *et al.* (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.*, **5**, 21

Martignett,L. *et al.* (2016) ROMA: Representation and quantification of module activity from target expression data. *Front. Genet.*, **7**, 18

May,A. *et al*. (2016) Metamodules identifies key functional subnetworks in microbiome-related disease. *Bioinforma. Oxf. Engl*., **32**, 1678–1685

Mitra K, *et al*. (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.*, **14**, 719–732

Mook-Kanamori DO, *et al*. (2014) 1,5-anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *J. Clin. Endocrinol. Metab*., **99**, E479–E483

Nix WA, *et al*. (2015) Vitamin B status in patients with type 2 diabetes mellitus with and without incipient nephropathy. *Diabetes Res. Clin. Pract.*, **107**, 157–165

Opgen-Rhein,R. and Strimmer,K .(2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.*, **1**, 37

Polanski K, *et al*. (2014) Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics* **30**, 962–970

Sekhar RV, *et al.* (2011) Glutathione synthesis is diminished in patients with uncontrolled diabetes and restored by dietary supplementation with cysteine and glycine. *Diabetes Care* **34**, 162–167

Shannon, P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13: 2498–2504

Shannon,P.T. *et al*. (2013) RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics* **14**, 217

Swainston,N.*et al.* (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12**, Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896983/ [Accessed August 7, 2017]

Unoki-Kubota,H. *et al*. (2010) Pyridoxamine, an inhibitor of advanced glycation end product (AGE) formation ameliorates insulin resistance in obese, type 2 diabetic mice. *Protein Pept. Lett.*, **17**. 1177–1181

Valdés-Ramos,R. *et al*. (2015) Vitamins and type 2 diabetes mellitus. *Endocr. Metab. Immune Disord. Drug Targets* **15**, 54–63

Appendix D

Characterization of missingness in untargeted MS-based metabolomics data and evaluation of missing data handling strategies.

Do KT*, Wahl S*, Raffler J, Molnos S, Laimighofer M, Adamski J, Suhre K, Strauch K, Peters A, Gieger C, Langenberg C, Stewart I, Theis FJ, Grallert H, Kastenmüller G, Krumsiek J (2017), **Characterization of missingness in untargeted MS-based metabolomics data and evaluation of missing data handling strategies.** *Under review* in *Metabolomics*

This work is under review in *Metabolomics.* The version appended here is the PDF on *bioRxiv*, a preprint repository.

Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies

Kieu Trinh Do[1¶], Simone Wahl[2,3,4¶], Johannes Raffler[5], Sophie Molnos[2,3,4], Michael Laimighofer[1], Jerzy Adamski[6,7], Karsten Suhre[9], Konstantin Strauch[10,11], Annette Peters[2,3], Christian Gieger[2,3], Claudia Langenberg[12], Isobel D. Stewart [12], Fabian J. Theis[1,13], Harald Grallert[2,3,4], Gabi Kastenmüller[4,5#], Jan Krumsiek[1,4,14#]

1 Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany, **2** Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **3** Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **4** German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany, **5** Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, Neuherberg, Germany, **6** Institute of Experimental Genetics, Genome Analysis Center Helmholtz Zentrum München, Neuherberg, Germany, **7** Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan, Germany, **8** German Center for Cardiovascular Disease Research (DZHK e.V.), partner-site Munich, Germany, **9** Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar, **10** Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany, **11** Chair of Genetic Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany, **12** MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, **13** Department of Mathematics, Technische Universität München, Garching, Germany **14** Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

[¶] These authors contributed equally to this work.

[#] **Corresponding authors**:

Dr. Gabi Kastenmüller, Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, Neuherberg, Germany, Phone: +49 89 3187-3578, Fax: +49 89 3187-3585, E-mail: g.kastenmueller@helmholtz-muenchen.de

Dr. Jan Krumsiek, Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany, Phone: +49 89 3187-3641, Fax: +49 89 3187-3369, E-mail: jan.krumsiek@helmholtz-muenchen.de

Abstract

BACKGROUND: Untargeted mass spectrometry (MS)-based metabolomics data often contain missing values that reduce statistical power and can introduce bias in epidemiological studies. However, a systematic assessment of the various sources of missing values and strategies to handle these data has received little attention. Missing data can occur systematically, e.g. from run day-dependent effects due to limits of detection (LOD); or it can be random as, for instance, a consequence of sample preparation.

METHODS: We investigated patterns of missing data in an MS-based metabolomics experiment of serum samples from the German KORA F4 cohort (n = 1750). We then evaluated 31 imputation methods in a simulation framework and biologically validated the results by applying all imputation approaches to real metabolomics data. We examined the ability of each method to reconstruct biochemical pathways from data-driven correlation networks, and the ability of the method to increase statistical power while preserving the strength of established genetically metabolic quantitative trait loci.

RESULTS: Run day-dependent LOD-based missing data accounts for most missing values in the metabolomics dataset. Although multiple imputation by chained equations (*MICE*) performed well in many scenarios, it is computationally and statistically challenging. K-nearest neighbors (*KNN*) imputation on observations with variable pre-selection showed robust performance across all evaluation schemes and is computationally more tractable.

CONCLUSION: Missing data in untargeted MS-based metabolomics data occur for various reasons. Based on our results, we recommend that *KNN*-based imputation is performed on observations with variable pre-selection since it showed robust results in all evaluation schemes.

Keywords: untargeted metabolomics, missing values imputation, limit of detection, batch effects, runday effects, *MICE*, K-nearest neighbor, mass spectrometry

Key messages

- Untargeted MS-based metabolomics data show missing values due to both batch-specific LOD-based and non-LOD-based effects.
- Statistical evaluation of multiple imputation methods was conducted on both simulated and real datasets.
- Biological evaluation on real data assessed the ability of imputation methods to preserve statistical inference of biochemical pathways and correctly estimate effects of genetic variants on metabolite levels.
- *KNN*-based imputation on observations with variable pre-selection and *K* = 10 showed robust performance for all data scenarios across all evaluation schemes.

Introduction

In epidemiological studies, metabolomics is an established tool that provides insights into disease mechanisms (1), as metabolite profiles generate a molecular readout that is closely linked to the (patho-)phenotype (2,3). Recent metabolomics studies have identified many metabolites as candidate biomarkers for various health conditions, such as diabetes (4–6) and cardiovascular diseases (7,8). Mass spectrometry (MS)-based metabolomics measurements can be performed either in a targeted or untargeted manner (9). In the former, only a limited number of already known and biochemically annotated metabolites are captured. In the latter, the measurements are not limited to predefined signals and offer discovery of novel compounds. While missing values in targeted MS-based data occur rarely, untargeted MS-based techniques typically produce 20-30% missing values, affecting more than 80% of the measured compounds (10–13).

There are various reasons why metabolite concentrations can be missing in an untargeted metabolomics dataset. First, it is possible that the molecules are truly absent from the sample, a situation that may occur e.g. for drug metabolites that only appear in a subset of people taking that medication. On the other hand, there are several technical reasons that could result in missing values, including: (i) instrument sensitivity thresholds, below which concentrations of a specific metabolite might not be detectable in a sample (i.e., below the limit of detection, LOD); (ii) matrix effects that impede the quantification of a metabolite in a sample through other co-eluting compounds and ion suppression; (iii) declining separation ability of the chromatographic column and increasing contamination of the MS instrument; and (iv) limitations in computational processing of spectra, such as poor selection and alignment of the spectral peaks across samples (14).

Commonly, observed patterns of missing data are categorized as either missing completely at random (MCAR), missing at random (MAR), or missing but not at random (MNAR) (15). In the MCAR category, the probability of missing values does not depend on observed or unobserved measurements. In contrast, the occurrence of MAR depends on other observed measurements (for instance, resulting from technical effects, such as overlapping peaks). MNAR describes the occurrence of missing values that depend on unobserved measurements (for instance, due to issues with the performance of the machine).

Although it is clear that the handling of missing values affects all downstream analyses, it is less clear how to appropriately handle their occurrence statistically. A simple *ad hoc* approach is known as complete case analysis (*CCA*), which only considers samples that do not contain any missing values in the metabolites analyzed in each statistical analysis step. However, missing data may occur in some systematic way (i.e., they are dependent on external factors). For example, if all cases in a case-control study have more missing data than the controls, removing observations that are missing will lead to bias in biological interpretation (16). Furthermore, *CCA* can cause severe loss of information and statistical power by excluding a majority of observations if multivariate methods, such as principal component analysis or partial correlation networks, are to be performed.

A widely used and flexible class of missing data strategies is imputation, which involves the replacement of missing values by reasonable substitute values. The most commonly used imputation approaches for metabolomics data assume that missing data occur because they are below the limit of detection (left-censoring, a variant of MNAR). Therefore, all missing entries of a metabolite are replaced by a low constant value, such as the actual LOD (if known), zero, or the smallest value found in the dataset for that metabolite (13). Another LOD-based substitution strategy assumes a parametric left-truncated normal distribution and performs likelihood-based parameter estimation on the observed values to reconstruct the truncated part of the distribution. Missing values are then replaced by numbers drawn from this estimated part (16,17). Additional imputation-based substitution approaches assume MCAR and replace missing values by the mean or median per metabolite (12). Advanced approaches use multivariate statistical methods for imputation, including multiple imputation by chained equations (MICE) (18) and K-nearest neighbors (KNN) imputation (19,20).

Several previous studies have investigated the occurrence and effects of different strategies for missing values in metabolomics data. Taylor *et al.* (21) reported that no single imputation method was universally superior, but constant substitution methods consistently showed poor performance. Gromski *et al.* (12) recommended imputation by Random Forests (RFs) for GC/MS metabolomics data after evaluating the outputs of supervised and unsupervised learning approaches. Di Guida *et al.* (15) investigated various combinations of different preprocessing steps to determine which were the most appropriate for univariate and multivariate analyses of UHPLC-MS metabolomics data. The authors recommended RF and *KNN*-based imputation for PCA and PLS-DA, respectively (15). Armitage *et al.* (10) studied missing values in CE/MS metabolomics data and reported *KNN* imputation to be more effective compared with simpler substitution-based imputation methods. Finally, in a study by Hrydziuszko and Viant (11), a *KNN*-based imputation approach also outperformed competing strategies in an investigation of direct infusion Fourier transform ion cyclotron resonance (DI-FTICR) MS-based metabolomics data.

Despite these advances in our understanding of the effects of imputation on metabolomics data analysis, several aspects have not been addressed by those previous studies. (i) A detailed statistical description of the patterns of missing values in MS-based metabolomics data has not yet been published. Most previous studies evaluated imputation strategies assuming only random or LOD-based missing values without assessing whether this applies to real metabolomics datasets. In particular, the influence of batch effects on the occurrence of missing values has not been investigated in any study. If a cohort comprises a large number of samples, the MS runs usually are spread across multiple days, which is known to influence metabolite measurements due to variation in instrument sensitivity. Here, the LOD itself is also expected to vary across run days, an assumption that has not been explicitly accounted for in any studies. (ii) In addition, a simulation framework that reflects realistic data situations is needed to provide an unbiased evaluation of strategies for handling missing values. Evaluation of previous studies has been biased in the sense that "complete" measured data (created by excluding all variables with missing values) with artificially introduced missing values were simulated, which most likely does not mirror realistic missing value patterns. (iii) Finally, biological validation and biochemical interpretation of the data have not been addressed in the majority of papers. Only Hrydziuszko *et al.* evaluated the ability of different imputation strategies to preserve metabolic differences between biological groups, which then were related to KEGG pathways (11).

In the present study, we analyzed patterns of missing data and evaluated the performance of various imputation strategies for untargeted MS-based metabolomics data from serum samples of the German Cooperative Health Research in the Region of Augsburg (KORA) F4 cohort. Data were measured on a typical, widely used untargeted MS-based metabolomics platform (Metabolon, Inc., USA) and should be representative of many untargeted population-scale metabolomics studies. The study consisted of three steps: (i) We described and analyzed patterns of missing values and their possible underlying mechanisms in a real untargeted metabolomics dataset. In particular, we investigated the occurrence of missing values within and across batches of measurements. (ii) The insights gained from these analyses were used to introduce realistic patterns of missing data into simulated data. We applied 31 imputation methods to the datasets and evaluated them with respect to their ability to achieve correct statistical estimates and hypothesis test results in various data scenarios. (iii) Finally, the imputation methods were applied to real metabolomics data (KORA F4), followed by two biologically-driven evaluation schemes. First, we assessed how accurately real biochemical pathways were reconstructed in data-driven correlation networks inferred from the imputed data. Second, we verified whether imputation led to a gain in statistical power, while preserving effects of genetic variants on metabolite levels. The study workflow is visualized in Figure 1.

Results

Characterization of missing data patterns in KORA F4 untargeted metabolomics data

We used an untargeted metabolomics dataset from the KORA F4 study, which was generated from fasting serum samples measured on three platforms: LC/MS in both positive (LC/MS+) and negative modes (LC/MS−), as well as a GC/MS platform. After log-transformation and outlier handling (see Methods), 1757 samples and 516 metabolites were available for analysis.

The dataset contained 19.41% missing values, with 416 (80.6%) metabolites and all observations showing at least one missing value. The majority (301) of these 416 metabolites had fewer than 10% missing values (Figure 2A). For only 9.9% (51) of the metabolites, more than 70% of the measurements were missing. The amount of missing values per observation ranged from 11.4% to 32.2%, with an average of 19.6% (Figure 2B).

LOD-based missing values

For metabolomics data, a common assumption is that missing values occur because of low concentrations that are below the limit of detection. To explore this assumption, we analyzed missing values of a metabolite using a second, strongly correlated metabolite, which we term the *auxiliary* metabolite. The auxiliary metabolite is defined as the metabolite with the highest correlation (r) to the given metabolite. Due to its strong correlation, we assume that insights into the pattern of missing values of a metabolite can be gained from the corresponding non-missing observations of its auxiliary metabolite. For example, assuming that metabolite A has missing values in certain observations for which its auxiliary metabolite B has measurements. If these measurements in B are low then a missing value in A most likely occurred because the actual concentrations were below the LOD. We required a minimum correlation of $r = 0.3$ for auxiliary metabolites, but other values gave qualitatively similar results (File S1).

Overall, an auxiliary metabolite was available for 56.6% of the metabolites. Of those, 62.0% showed a clear tendency for missing values to below the LOD (see Methods and File S1). An example for a clear LOD-tendency is shown for 7-methylxanthine in Figure 2C. This compound is a metabolite of caffeine metabolism that is correlated with 3-methylxanthine. The majority of observations with missing data in 7-methylxanthine showed low values for 3-methylxanthine, indicating that the 7-methylxanthine values were most probably below the LOD. An example for a metabolite pair that does not show an LOD-based missingness pattern is provided in Figure 2D for 1-arachidonoylglycerophosphocholine (1-AGPC) and its auxiliary metabolite 1-docosahexaenoylglycerophosphocholine (1-DGPC). Unlike the previous example, observations with missing data for 1-AGPC showed values varying over the whole range of 1-DGPC. Consequently, this suggests that LOD does not adequately explain the pattern of missing values for 1-AGPC. Scatterplots of investigated metabolites and their corresponding auxiliary metabolites, as well as boxplots of concentrations in the auxiliary metabolites for missing and non-missing observations in the investigated metabolites can be found in File S1.

Although the LOD-tendency was observed for many metabolites, there was no clear LOD threshold separating missing and observed measurements across all metabolites (Figure 2C), which would have been the case if LOD was the only underlying mechanism for missing data. Instead, the values of the auxiliary metabolites with missing values in the investigated metabolites were spread broadly over a range of lower values, indicating a blurred rather than a single fixed LOD for all metabolites.

Run day-dependent missing values

Batch (run day) effects also can drive systematic patterns of missing data due to daily variation in instrument sensitivity. To examine whether missing data depended on overall run day quality, we examined the amount of missing values per run day for each platform (LC/MS+, LC/MS–, or GC/MS). Subsequently, we investigated whether metabolites were affected differently by runday quality.

The KORA F4 samples were measured on 53 run days with 34 samples on average per day. If missing values were dependent on run day quality due to variation in instrument performance (e.g., caused by LC or GC column decline), we would expect there to be some days for which samples overall contained more ("bad" run day) or fewer ("good" run day) missing values compared with the average. Indeed, we observed such "bad" and "good" run days for all three platforms (Figure 3A). While the run day-specific amount of missing values tended to be correlated between LC/MS− and LC/MS+ (correlation of the run day-specific median of missing values between the two platforms was $r = 0.36$), there was no correlation between LC/MS+/− and GC/MS. This suggests that changes in instrument performance, rather than global effects (such as those that could originate from sample preparation) were responsible for differences in run day quality.

Although there was an overall effect of run day quality on the pattern of missing values, we observed considerable differences in the standard deviations (SD) of run day-specific missing values for metabolites with the same amount of missing data (Figure 3B). This suggests that metabolites were affected differently by run day quality. For example, the bile acid ursodeoxycholate (46% total missing data) showed relatively low variation in run day missing data (SD = 0.12) (Figure 3**Figure 3**C). However, for gamma-glutamylisoleucine (Figure 3D), a metabolite with a similar total amount of missing values (42%), the observed variation in missing data across run days was substantially larger (SD = 0.22).

Run day-dependent LOD mechanism

The observed run day-dependent pattern of missing data, together with the blurred LOD-based pattern, suggests that different run days may exhibit different LODs, which contributed to the blurred global LOD effect. To verify this, we calculated the correlation between run day mean and run day missingness for all metabolites. A histogram of the correlation coefficients is shown in Figure 4A. The majority of metabolites displayed a strong tendency for negative correlations. An example for run day-specific LODs is shown in Figure 4B–C: for 7-methylxanthine, the correlation of run day mean and the run day-specific amount of missing values is $r = -0.68$ (Figure 4B). Run days with low means tended to have a higher amount of missing values (Figure 4C). Density plots for all metabolites before and after run day normalization can be found in File S2.

Taken together, we observed that batch (run day) effects on the limit of detection can result in a blurred LOD-effect after run day normalization, which can explain patterns of missing values in most, but not all, metabolites.

Evaluation of imputation approaches in a simulation framework

As shown in the previous analyses, not all of the missing data in MS-based metabolomics studies can be attributed to run day-dependent LOD-based missing data. Thus, the optimal imputation approach should perform well across all possible patterns. We conducted a simulation study to compare statistical estimates between imputed and complete data. We simulated incomplete data according to the patterns of missing values observed in the real metabolomics data and imputed these data using various imputation approaches. We then evaluated these approaches for recovering correct statistical estimates after conducting correlation and regression analyses.

Simulation setup and evaluation criteria

We simulated six mechanisms for missing data derived from observations in the real data (see Methods, File S3, and Figure 5A–E): (i) *Fixed LOD,* as an extreme form of systematic missing values below a global LOD; (ii) *Probabilistic LOD*, where the probability of a missing value increases at lower values, which should resemble the blurred LOD-based patterns observed in the real data; (iii) *Run day-specific fixed LOD*, where LOD is assumed to vary across run days; (iv) *Run day-specific probabilistic LOD*, where a probabilistic form of LOD is assumed to occur across run days; (v) *Unsystematic (random) missingness*, for missing data with an unknown reason; and (vi) *Mixtures of LOD-based and unsystematic missingness*. Based on these 6 mechanisms, we created various parameter scenarios resembling realistic conditions. For each scenario, we conducted 250 simulations to assess whether the imputation methods could reconstruct statistical estimates of Pearson correlation, partial correlation, linear regression (results shown in File S3), and logistic regression. To this end, we calculated type 1 error as the proportion of simulations in which a significant estimate was obtained when the true correlation was equal to zero. In addition, we calculated power as the proportion of significant estimates when the true correlation was unequal to zero. We also estimated bias, which is shown in File S3. A detailed description of the simulation and evaluation framework is also provided in File S3.

Missing data handling strategies

We applied 31 imputation approaches (see Figure 5F; detailed descriptions in Methods and File S4) on the simulated data. Some were adapted to account for run day-specific missing values. The imputation approaches followed different concepts, which could have one of the following four properties or combinations thereof: (i) approaches that explicitly assume LOD-based missing values, (ii) approaches that consider run day-specific missing values, (iii) multivariate procedures using correlations among variables, and (iv) multiple imputation (MI) strategies. The MI approaches usually comprise imputation, analysis, and pooling steps. In the first step, the incomplete data are imputed *m* times to produce *m* complete datasets. Subsequently, statistical analysis is performed on each of the *m* complete datasets and then the *m* analyses are combined to one final result.

Simulation results

In the following, we evaluate the performance of the four imputation properties (i)–(iv) introduced above. Simulation results from other data scenarios, all variations of the imputation approaches used, and the combination of parameter settings are available in File S5.

Property (i): Methods that explicitly assume LOD-based missing values and perform imputation globally without taking run day information into account (*min*, Richardson & Ciampi (*RC*), imputation by truncated sampling (*ITS*)), showed inflated type 1 error rates and low power for both correlation and regression analysis. This was expected for three reasons. First, for a data scenario with run day-dependent probabilistic LOD-based missing values, these methods underestimate the LOD for most of the rundays and replace missing entries by too low values (Figure 6A). Second, for a data scenario with random missing values, they expectedly fail since the underlying assumption of an LOD is not met (Figure 6B). Finally, *min* and *RC* impute a metabolite by replacing all of its missing entries by a constant value, which substantially distorts the metabolite distribution (see File S5).

Property (ii): The LOD-based methods that take run days into account (*RC-R, ITS-R*) were expected to perform well in a simulated data scenario with run day effects (Figure 6A). Unexpectedly, we observed an inflated type 1 error rate and decreased power for all three statistical analyses (Pearson correlation, partial correlation, and logistic regression). *RC-R* and *ITS-R* assume that the observed values of a metabolite follow a truncated normal distribution, which is parametrized by maximum likelihood estimation (MLE), in order to replace missing values with randomly drawn values from the truncated part. The instability of MLE due to small sample sizes available within run days could explain the poor performance of these approaches. The same poor performance was observed for scenarios with a mixture of run day-dependent LOD-based and random missing values (Figure 6C). For the dataset with only random missing values, LOD- or run day-based approaches showed the expected strong reduction in power since here the underlying assumption of a truncated normal distribution is false (Figure 6B).

Property (iii): Multivariate approaches (imputation based on chained equations *(ICE)* and *KNN*-based imputation) take into consideration the correlation between variables or observations. *ICE* approaches had high power, but an increased type 1 error rate when missing value proportions increased (Figure 6). *KNN*-based imputation on observations with variable pre-selection and K = 10 (*KNN-obs-sel(10)*) was one of the best performing methods with high power and an overall marginal type 1 error rate, even for a high amount of missing values. The power for *KNN-obs* was also high, but it showed high type 1 error rate and therefore a poor ability to correctly identify truly absent associations. In contrast, *KNN-vars* had a low type 1 error rate, but decreased power, which became more pronounced at higher amounts of missing values.

Property (iv): Single imputation procedures often underestimate the variability of statistical estimates, resulting in inflated type 1 error rates. This should be avoided by approaches performing multiple imputations (MI). MI versions based on LOD- (*MITS*) and run day-effects (*MITS-R*) indeed had decreased type 1 error rates, although power was low (Figure 6). *MICE* with Bayesian linear regression (*MICE-norm*) or predictive mean matching (*MICE-pmm*) as imputation model showed negligible type 1 error rates and high power for all scenarios with up to 50% missing values. At higher amounts of missing data, the power decreased considerably, but the type 1 error remained marginal (File S5). A slight modification of the *MICE* algorithm applied widely in the metabolomics field (here termed *MICE-avg*) was performed on each imputed data, and comprised the pooling of the imputed data with subsequent statistical analyses rather than pooling the statistical estimates after analysis. This approach showed high power, but increased type 1 error rates, in particular for >30% missing values.

Taken together, when considering all patterns of missing data and all evaluation criteria, *KNN-obs-sel(10)* and *MICE-norm* were the most robust approaches. For higher amounts of missing data (≥50%), *MICE* showed a strong decrease in power with marginal type 1 error, whereas *KNN-obs-sel(10)* had only slightly increased type 1 error rates with high power.

Evaluation of imputation approaches on real MS-based metabolomics data

We conducted a biological evaluation of all approaches using the metabolomics data from the KORA F4 population study. An objective criterion for evaluation is challenging to construct, since the true values underlying the missing ones are unknown. We devised two indirect tests that assessed imputed values for biological validity. First, we assessed the ability of imputation methods to statistically reconstruct biochemical pathways in metabolomics data. Second, we evaluated the gain in statistical power while preserving the true effect size of genetic variants (SNPs) on metabolite levels.

Evaluation based on pathway modularity

GGMs are based on partial correlations and reflect conditional dependencies in multivariate Gaussian distributions (5,22). When applied to metabolomics data, they reconstruct a precise picture of the metabolic network, showing a modular topology with respect to known pathways. In other words, metabolites will tend to be correlated with other metabolites from the same biochemical pathway (5,22,23). We used this pathway-based modularity in a metabolic network as a quality criterion to indicate whether the imputation methods generally were capable of maintaining biochemically valid edges.

Each imputation strategy was applied to the KORA F4 metabolomics data, and a GGM was estimated for each obtained dataset. Subsequently, we used *a priori* pathway annotations from Metabolon Inc., where each metabolite was assigned to one pathway (e.g., branched-chain amino acids, lysolipids, xanthines) to calculate pathway-based modularity (Q), according to (22,24). This measure reflects the ratio of metabolite correlations within *versus* across pathways. A high Q value indicates a dense within-pathway correlation compared with cross-pathways. Variability was estimated by bootstrap resampling (see Methods).

Across all datasets, we obtained modularity values ranging from 0.384 to 0.434 (Figure 7A). Imputation methods that explicitly considered the LOD-based mechanism and their run day-specific versions (Figure 5, property (ii)) did not outperform alternative approaches. Multivariate, single imputation methods (property (iii)) yielded low Q values, except for *KNN-obs-sel*, which achieved the overall third best result (Q = 0.422 for K = 10) (Figure 5). The performance of *KNN*-based imputation methods strongly depended on the definition of neighbors (variables or observations) and on the number of these neighbors (K). The MI procedures (property (iv)) *MITS*, *MITS-R*, and *MICE-avg* performed poorly, whereas the networks generated on *MICE* imputed data showed the overall highest modularity (*Q* = 0.434 and *Q* = 0.424 for *MICE-norm* and *MICE-pmm*, respectively) (Figure 5). Overall, the three best performing approaches were *MICE-norm*, *MICE-pmm*, and *KNN-obs-sel(10)*.

Evaluation based on metabolite-SNP associations

Using KORA F4 data (n = 1750), we determined the ability of imputation methods to gain statistical power compared with complete case analysis (*CCA*, deleting samples with any missing values) while preserving the effect of genetic variants on metabolite levels in human blood. For the evaluation, we selected a set of metabolite-SNP associations from a previous genome wide association study (GWAS) in the KORA F4 and TwinsUK cohorts, for which a functional connection between the gene and the metabolite was biologically evident (Table S8) (25). For example, GOT2 (*rs4784054*), which was associated with concentrations of phenyllactate, encoded an enzyme that catalyzes the conversion of phenylalanine to phenylpyruvate, which is then converted to phenyllactate (25,26).

We investigated the gain in statistical power when using imputed datasets compared with the power obtained with *CCA* for 18 of such metabolite-SNP pairs, where the metabolite had between 10% and 70% missing values. Statistical power gain was calculated as the negative log10 of the ratio of the p-values estimated for the imputed data to the p-values estimated for *CCA* in corresponding linear regression models (detailed results in File S8 and Table S8). A high ratio indicates greater power for imputed data. As a second evaluation criterion, we calculated the log2 absolute ratio of the effect sizes obtained from the regression models for imputed data and those derived from *CCA* in KORA F4 (see Methods). A log2 ratio close to zero indicates that the imputation method was able to preserve effect sizes, whereas imputations yielding a highly negative or positive log2 ratios indicate underestimation or overestimation of the effect sizes, respectively.

Imputation with LOD-based methods (property (i)) yielded a gain in power for up to seven genetic associations of the 14 metabolites (Figure 7**Figure 7**). For two of these associations (tetradecanedioate and SLCO1B1; and hexadecanedioate and SLCO1B1), effect sizes were underestimated, and for the association between 1-methylurate and NAT2, the effect size was overestimated across all methods, except for *MITS-R*. Run day-specific imputation methods (property (ii)) performed well, with *ITS-R* yielding the highest number of associations (12) with greater statistical power, of which seven showed effect sizes similar to effect sizes derived from *CCA*. The best methods among multivariate approaches (property (iii) and (iv)) were *MICE-avg-norm*, *KNN-obs-sel(10), and KNN-obs-sel(20)*, all three of which generated a gain in statistical power for 12 associations. These methods also showed good performance in preserving genetic effects and did not show severe overestimation or underestimation of effect sizes. *MICE-norm/-pmm/-adjR* showed only moderate performance with a power gain for seven associations.

In an additional analysis, we used results from the EPIC-Norfolk cohort with n = 10 634 subjects (27), to assess the ability of imputation methods to preserve effects of genetic variants on metabolites. We hypothesized that the effect sizes would be estimated more accurately in this much larger dataset, and effect sizes obtained with KORA F4 imputed data should approximate effect sizes derived from EPIC-Norfolk. Overall, we observed that the majority of SNP-metabolite pairs showed either an overestimation or an underestimation of effect sizes across all imputation methods. This tendency might reflect differences between the cohorts KORA F4 and EPIC-Norfolk rather than differences between imputation strategies (see detailed results in File S7 and Table S8).

Overall, for nearly all metabolite-SNP pairs, this analysis showed that statistical power was increased by imputing missing values and the effect sizes could be preserved. *ITS-R*, *MICE-avg-pmm, KNN-obs-sel* with *K* = 10 and *K* = 20 were the imputation methods that generated the highest number of associations (12) and resulted in a gain in statistical power compared with *CCA*.

Discussion

In this study, we investigated patterns of missing data in a typical example of untargeted MS-based metabolomics data and their possible underlying mechanisms. Insights gained from these analyses were used to generate simulated data that reflected the real data situation for a comprehensive evaluation of 31 imputation methods. Finally, we applied the imputation strategies to real MS-based metabolomics data from the German KORA F4 study and evaluated them using biological validity measures.

For metabolomics data, an intuitive assumption is that missing data occur when metabolite concentrations fall below the machine's LOD. Indeed, we found evidence for systematic patterns of missing data due to LOD- and batch-effects for a large proportion of the analyzed metabolites. Missing data were found to be influenced by run day quality, although metabolites varied in their susceptibility to this effect. Finally, we found a negative correlation between run day mean and missing data per run day, further confirming LOD-based mechanism within run days. The existence of multiple run day-dependent LODs possibly accounted for the blurred rather than fixed global LOD observed in the data. It has been suspected that multiple detection limits arise from factors such as batch (run day) effects (27). However, to the best of our knowledge, this is the first time that these effects have been systematically explored so far.

We evaluated 31 imputation methods in an evaluation framework consisting of three schemes: (i) unbiased estimation of statistical estimates and hypothesis test results based on simulated data, (ii) statistical reconstruction of biochemical pathways in metabolic networks, and (iii) the ability to preserve effects of genetic variants on metabolite levels while allowing for a gain in statistical power.

MICE-norm was the best performing imputation method for evaluation scheme (i) and (ii), but it showed only moderate performances in the metabolite-SNP analysis. One major drawback of this method is that multiple imputations have to be performed, making these approaches statistically and computationally challenging. For *m* imputations, the desired statistical analyses must be performed on each of the *m* imputed datasets, and then the resulting *m* estimates must be combined to one statistical result. A widely applied alternative is to perform *m* multiple imputations and then combine the *m* complete datasets to one final dataset containing the average of the imputed values (*MICE-avg*). That is, *MICE-avg* does not require statistical estimates to be pooled, and therefore, it is much easier to apply. However, this simplicity is accompanied by an underestimation of metabolites' variances, resulting in poorer performance of statistical estimation (correlation and regression coefficients) and reconstruction of biochemical pathways.

A feasible, but better performing method was *KNN-obs-sel(10)*, which uses *KNN*-based imputation on observations with variable pre-selection and *K* = 10. This method ranked highly in all evaluation schemes. Other *KNN*-based imputation schemes, including *KNN*-based imputation on variables (*KNN-vars*) and on observations without variable pre-selection (*KNN-obs*), consistently showed poor performance across all evaluation schemes. Our results are in line with observations from previous studies, where *KNN*-based imputation performed well (10,11,15,28). However, we also observed that variations of *KNN* imputation lead to substantially different results, as in previous studies (20,28).

Although we observed LOD- and run day-based effects in real metabolomics data, methods that explicitly consider this information did not outperform competing approaches in the first two evaluation schemes. This is likely due to the fact that they perform imputation in a univariate manner without taking the correlation between the variables into account. Moreover, all of these LOD-based methods include maximum likelihood estimation in their imputation process, which was found to perform well only for larger sample sizes in previous studies (27,29). In our study, the number of observations within run days is limited, resulting in considerable instability of the MLE. LOD-based run day-dependent methods performed well with respect to gain in statistical power in the analysis of metabolites–SNP associations.

In summary, we have presented a detailed description of patterns of missing data in untargeted MS-based metabolomics data. In particular, we considered, for the first time, the effects of run days on systematic patterns of missing data. Our work showed that missing data occur in most cases due to LOD effects, which are moreover run day-dependent. Nevertheless, *MICE* and *KNN*-based imputation, methods that do not explicitly consider LOD-based effects, performed best when tested in both statistical and biological evaluation schemes. This is most likely because these methods take into account multivariate dependencies within the data. The two approaches are For future studies, we recommend *KNN*-based imputation on observations with *K* = 10, since it consistently performed well across all data scenarios and all evaluation schemes, and is computationally non-demanding for daily data analysis.

Material and Methods

Study cohort, metabolomics and genotype measurements

Data from 1768 fasting serum samples of the German Cooperative Health Research in the Region of Augsburg (KORA F4) population cohort (30) was used, comprising 910 females and 858 males. Age distribution was 60.53 ± 8.79 years for females and 61.20 ± 8.78 years for males. Body mass index (BMI) distribution was 27.88 ± 5.24 $\mathrm{kg/m^2}$ for females and 28.46 ± 4.29 $\mathrm{kg/m^2}$ for males.

Serum metabolomics measurements were performed on three platforms, LC/MS− (negative mode), LC/MS+ (positive mode), and GC/MS by Metabolon, Inc. (Durham, NC, USA). The 1768 serum samples were measured on 53 different run days, with 34 samples on average per run day. A total of 516 metabolites were quantified, of which 303 had an identified chemical structure. A more detailed description of sample acquisition, experimental procedures, and metabolite identification can be found in File S10.

Each known metabolite was annotated with one of 68 pathways by Metabolon, Inc. A full list of all measured metabolites, including pathway annotations, can be found in Table S9. For correlation analysis, data were normalized for run day-effects by dividing each metabolite by run day median. Since metabolite measurements were assumed to follow a log-normal distribution, the data were log-transformed for all statistical analyses. The run day-corrected and log-transformed data were used to determine outlier samples. Eleven individuals with a Mahalanobis distance (calculated across the complete dataset) greater than four SD from the mean were considered outliers and excluded from the dataset. For the biological evaluation schemes, age, sex, and BMI were used as standard covariates. Seven samples were excluded due to incomplete information in these phenotypes, resulting in 1750 individuals in total.

The KORA F4 cohort was genotyped using the Affymetrix Axiom platform. After quality control, genotype data (measured or imputed according to data from the 1000 genomes project, phase 1 version 3) were available for 1685 of the 1750 individuals.

Missing data in KORA F4

To explore the mechanism for the missing data of a given metabolite m, a second (auxiliary) metabolite m_{aux} was used. m_{aux} was defined as the metabolite with the strongest Pearson correlation to m (at least 0.3). An LOD-tendency was assumed if the average value of m_{aux} in samples with missing values in m was significantly lower than the average of m_{aux} in samples with measured values in m. Significance was assessed using Wilcoxon–Mann–Whitney tests with $\alpha = 0.05$ after Bonferroni correction for multiple testing.

For all correlation analyses, only metabolites with more than 10% and less than 70% overall missing values were considered.

In order to explore whether missing values varied among run days, the normalized proportions of missing values among the 53 run days were compared within each platform. For a metabolite m and a run day d, the normalized amount of run day-specific missing values was calculated as the number of missing values for m in d divided by the total number of samples measured in d, divided by the median value of missing data of m over all run days.

Simulation study

Insights gained from the analyses of missing values in real MS-based metabolomics data were used to create artificial data that best mirror reflected patterns of missing data. A brief overview of the simulation framework is provided below, and a detailed description can be found in File S3. For each set of parameters corresponding to a certain data situation, 250 random datasets were generated. For each dataset, two variables were simulated by drawing from a multivariate normal distribution, with sample sizes ranging from 100 to 1000, and with means equal to zero and covariance chosen such that variances were equal to one (representing scaled variables). The Pearson correlation between the two variables was ranged from 0 to 0.4. In addition, for the multivariate analyses and to evaluate imputation methods that apply to a multivariate strategy, auxiliary variables correlated with the two main variables were introduced. Their number and correlation strength were chosen to match the real data (for details, see File S3).

Simulated observations were randomly assigned to "run days" with the number of run days chosen such that each run day comprised 34 observations, according to the average number found for the real KORA F4 measurements.

A proportion of missing values (10%, 30%, 50%, and 70%) was introduced into the main variable pair according to different mechanisms derived from our observations in the KORA F4 Metabolon data (Figure 5, File S3).

We used the following parameter settings for the results in the main manuscript: moderate variability of missing data across run days (see File S3), uncorrelated run day-specific missing patterns of the metabolite pair, and varying association of the inverse relation between metabolite concentration and missing values, at $n = 250$ and in the presence of informative auxiliary metabolites. For Pearson and partial correlation analysis, both main variables had the same degree of missing data. For logistic regression analysis, the predictor variable had a mixture of 50% run day-dependent probabilistic LOD-based missing data and 50% non-systematic missing data. Results for more parameter settings can be found in File S5.

Imputation approaches

A variety of imputation methods (Figure 5**Figure 5**) were selected because they were reported in the context of metabolomics data or were developed and adopted to address characteristics in the current dataset.

Mean imputation (mean): All missing values of each incomplete variable are replaced by the average of the observed values of that metabolite. ***Minimum imputation (min):*** All missing values of each incomplete variable are replaced by the smallest observed value of that metabolite (5,13,16). ***Richardson & Ciampi (RC):*** Assuming that missing values occur due to LOD and the observed metabolite values follow a left-truncated normal distribution, maximum likelihood is used to estimate this distribution. A missing value x is then replaced by the expected value of x conditional on x being below the LOD, $E(x|x \leq LOD)$ (17). ***Imputation by truncated sampling (ITS):*** This is an extension of the *RC* method, where the missing values are replaced by randomly drawn values from the censored part of the estimated truncated normal distribution. ***Multiple imputation by truncated sampling (MITS):*** *ITS* is applied as described above, but multiple imputation is performed according to Rubin's rules (31) using the *R* package *mice*, version 2.25. These rules include: (i) the datasets are imputed m times, (ii) each of the m completed datasets is analyzed separately, and (iii) the m resulting estimates are combined using established procedures (31–33). The number of imputations was set to $m = 20$ for all methods. ***Runday-specific LOD-based methods (RC-R/ITS-R/MITS-R):*** The previously described methods *RC, ITS,* and *MITS* are applied within run days where at least 17 observations are available. In *RC-R*, the remaining missing values are set to the mean of all available expected values. For *ITS-R* and *MITS-R*, the remaining missing values are replaced using *ICE-norm* (see below). ***Imputation by chained equations (ICE-norm/-pmm/-adjR)*** was performed using the *R* package *mice*, version 2.25. It uses a repeated chain of equations through the incomplete variables, where in each imputation model, the respective incomplete variable is modeled as a function of the remaining variables (34–36). In *ICE-norm*, a Bayesian linear regression is used as the imputation model, whereas in *ICE-pmm* (predictive mean matching as imputation model), missing values are replaced by a random draw of measured values from other observations with the closest predicted values. In *ICE-adjR*, a model is specified with random intercept per run day, which aims to better utilize run day information. This model assumes that variable values (i.e., metabolite concentrations) have a run day-specific component, which varies randomly following a normal distribution. ***Multiple imputation by chained equations (MICE-norm/-pmm/-adjR)*** was performed using the *R* package *mice*, version 2.25: *MICE-norm, MICE-pmm,* and *MICE-adjR* consisted of $m = 20$ parallel imputation runs of *ICE-norm, ICE-pmm,* and *ICE-adjR,* respectively. Subsequently, the estimates are combined using Rubin's rules as described above for *MITS*. ***MICE average version (MICE-avg-norm/-pmm)***: *ICE-norm* or *ICE-pmm* is applied multiple ($m = 20$) times in parallel, followed by combining the m imputed datasets to one final dataset as the average of the imputed values. ***K-nearest neighbor imputation (KNN-var(K)/KNN-obs(K)/KNN-obs-sel(K)):*** In *KNN-var* and *KNN-obs*, missing values of each variable are replaced by the weighted average of pre-specified *K* nearest variables and observations, respectively. Distances to neighbors were defined as Euclidean distance and weights were chosen as e^{-d}, where d defines the distances between two variables or observations. In *KNN-obs-sel, KNN-obs* is performed by selecting the strongest correlated variables with $|\rho| \geq 0.2$, but it was constrained to a minimum of 5 and a maximum of 10 variables. The number of neighbors for *K* was set to 3, 5, 10, and 20.

More detailed descriptions of *RC*, *RC-R*, *ITS*, *MITS*, *ICE*, and *KNN*-based methods can be found in File S4. The two best performing methods, *KNN-obs-sel(K)* and *MICE* are available as R code in File S11.

Statistical evaluation of missing data handling strategies in the simulation study

Pearson correlation, partial correlation, linear regression, and logistic regression analysis were performed, and the ability of imputation methods to reconstruct true associations and unbiased hypothesis test results was evaluated. For logistic regression, a dichotomized variable was simulated by discretizing one of the simulated continuous variables: all values above the median were set to 1 and all values below the median were set to 0. This dichotomized variable was used as response and the remaining continuous variable as predictor. For MI strategies, the resulting (correlation or regression coefficient) estimates and their variances were combined using Rubin's rules. The obtained point estimates were then compared with the true underlying values by assessing the validity of hypothesis tests. To this end, type 1 error was calculated as the proportion of significant estimates (at α= 0.05) after imputation when there was no true effect. Power was calculated as the proportion of significant estimates (at α= 0.05) after imputation in the presence of a true effect. Detailed results can be found in File S5.

Evaluation based on pathway modularity

This analysis was based on pathway annotations from Metabolon Inc. (see Supporting Information S9). Each imputation strategy was applied to the KORA F4 metabolomics data, resulting in different imputed datasets. All unknown metabolites were excluded since these compounds were not assigned to a pathway. For each imputed dataset, a Gaussian graphical model (GGM) was estimated to infer a network using the *R* package *GeneNet*, version 1.2.12. In previous studies, we have demonstrated that these models correctly reconstruct biochemical pathways from the data (22,25,37). In the case of MIs, a GGM was estimated for each imputed dataset, followed by combining partial correlations using Rubin's rules after a Fisher Z-transformation. The network was constructed using partial correlations that are significantly different from zero after Bonferroni correction for $n * (n-1)/2$, where n is the number of metabolites.

The pathway-based network modularity measure Q (22,24) was calculated for each network as

$$Q = \sum_{i=1}^{|S|} \left[\frac{A(V_i,V_i)}{A(V,V)} - \left(\frac{A(V_i,V)}{A(V,V)} \right)^2 \right],$$

where $|S|$ is the total number of pathways, V is the set of all metabolites, and V_i describes the subset of metabolites annotated with pathway i. $A(V_i, V_j)$ is the number of edges between any two node sets V_i and V_j. The variance of Q was estimated non-parametrically using bootstrapping of the original dataset (R package *boot*, version 1.3-15) with 1000 runs.

Evaluation based on metabolite-SNP associations

Linear regression was performed using KORA F4 *CCA* and the results were compared with each other. For this analysis, we selected metabolite-SNP pairs for which (i) a genome-wide significant association could be identified in the meta-analysis of KORA F4 and TwinsUK cohorts in a previous GWAS (25) (summary statistics retrieved from http://www.gwas.eu); (ii) the proportion of each metabolite's missing values in KORA F4 was between 10% and 70%; (iii) the metabolite was measured in the EPIC-Norfolk cohort, which we used to further benchmark the preservation of effect sizes; and (iv) a functional connection between the genetic locus of the SNP and the metabolite (e.g., metabolite is a known substrate of the transporter) was evident according to manual curation of the GWAS results (Table S8). For each imputed dataset, 18 metabolite-SNP pairs were tested for genetic association using age- and sex-corrected linear regression models under the assumption of an additive genetic model $(\text{metabolite} \sim \beta_0 + \beta_1 \times \text{SNP} + \beta_2 \times \text{age} + \beta_3 \times \text{sex})$. To avoid spurious associations, metabolic data points greater than four SDs from the mean were removed prior to computing linear models. For MI approaches, the regression coefficients were pooled using Rubin's rules as provided by the *R* package *mice*, version 2.25. For each metabolite-SNP pair, the variance of the regression coefficients and p-values were estimated using bootstrapping.

To explore which imputation approaches increased statistical power, p-values obtained for the effect sizes based on imputed data were compared with p-values obtained from *CCA* by calculating their ratio as $r_p = \frac{-\log_{10}\left(\frac{p_{imp}}{p_{CCA}}\right)}{-\log_{10}(p_{CCA})}$, where p_{imp} was the p-value obtained for imputed data and p_{CCA} was the p-value derived from *CCA*. A ratio less than or equal to zero indicated either no power gain or a power loss, whereas a ratio greater than zero indicated a drop in p-value, which suggested that statistical power increased when imputation was performed.

In addition to statistical power gain, the imputation approaches should be able to preserve effect sizes compared to *CCA*. Standardized effect sizes obtained from the imputed data (β_{imp}) were compared with standardized effect sizes estimated for *CCA* (β_{CCA}) based on the KORA F4 data (n = 1750) and the EPIC-Norfolk data (n = 10 634), assuming estimates from the EPIC-Norfolk data to be close to true effects. We calculated the ratio $r_\beta = \log_2(|\frac{\beta_{imp}}{\beta_{CCA}}|)$, with a low ratio indicating a similar effect size between the imputed data and *CCA*. A highly negative or positive r_β indicates an underestimation or overestimation of the effect sizes in imputed data, respectively. A well performing imputation method is assumed to obtain high r_p and low absolute r_β.

Figures and Tables

Figure 1. Flow chart of the study design. Pre-processed KORA F4 metabolomics data were used to analyze patterns of missing values in the dataset. Possible underlying mechanisms were inferred and implemented in a simulation framework to generate data resembling the observed patterns. Based on these simulated data, imputation methods with different characteristics were applied and evaluated. Finally, the same imputation approaches were evaluated using KORA F4 metabolomics and genomics data.

Figure 2. Overall amounts of missing data and LOD effects. (A,B) The overall fraction of missing values across metabolites and observations, respectively. (C,D) Scatter plots and boxplots of selected metabolite pairs to illustrate missing data due to LOD and non-LOD effects, respectively. Blue - observed concentrations. Red - observed values of the auxiliary metabolite in observations with missing values of the investigated metabolite. Note that red data points are not part of the x-axis but were plotted in the same scatterplot for clarity. *corr* = correlation, *p* = p-value of correlation, $\boldsymbol{p_{Wst}}$ = p-value of Wilcoxon–Mann–Whitney test.

Figure 3. Run day-dependent effects on missing data. (A) Normalized amount of missing values per run day in each platform (LC/MS+, LC/MS−, GC/MS). For a given metabolite and run day, the normalized amount of missing data per run day was calculated as the number of missing values for the respective metabolite on the respective run day divided by the total number of observations for that run day, divided by the median amount of missing data of that metabolite over all run days. Thus, a normalized run day-missingness of 1 is the average run day-missingness for a given metabolite. Pearson correlation coefficients were calculated across all pairs of platforms. (B) Standard deviation of missing values across run days, depending on the total amount of missing data for each platform. Each dot in the plot shows the total proportion of missing values and the run day variation for one metabolite. (C)–(D) The distribution of the total amount of missing values is shown for a metabolite with moderate (ursodeoxycholate) and high (gamma-glutamylisoleucine) standard deviation.

Figure 4. Run day-dependent LOD. (A) Histogram of Pearson correlation coefficients of the percent of missing values and run day means. (B) Scatterplot of run day mean versus percent missing values, with 7-methylxanthine as an example of a negative correlation. (C) Run day distributions of 7-methylxanthine before run day normalization.

Figure 5. Mechanisms of missing data and imputation approaches used in the simulation study. (A)–(E) Mechanisms of missing values used in the simulation study, based on evidence from real metabolomics data. (F) Venn diagram of imputation methods showing different characteristics. Note that the figure contains complete case analysis (*CCA*), which is not an imputation method, and is noted in brackets. *CCA* and *mean* were placed outside the Venn diagram, as they do not comprise any of the four characteristics. LOD: limit of detection.

Figure 6. Simulation results for Pearson, partial correlation, and logistic regression analysis. Performance of imputation approaches in data scenarios where (A) both variables followed a run day-specific probabilistic LOD mechanism, (B) both variables showed non-systematic patterns of missing data, and (C) one variable with run day-specific probabilistic LOD-based missing data and the other variable showed non-systematic patterns of missing data. Type 1 error and power reflect the false positive and true positive rate of hypothesis testing, respectively. Note that power = 1 - type 2 error rate. Note further that due to readability issues, only KNN-based imputation methods with *K* = 3, 10, and 20 were included, whereas KNN imputation with *K* = 1 and 5 can be found in File S5.

Figure 7. Evaluation of imputation approaches on real data. (A) Pathway-based modularity for each imputation strategy. Modularity *Q* was calculated based on pathways. Vertical lines represent bootstrap-based confidence intervals (1000 times resampling). (B) The ability to gain statistical power and to preserve real metabolite-SNP associations after imputation. Circle color represents the ability of imputation methods to preserve effect sizes, with red and blue indicating possible overestimation and underestimation, respectively, and yellow corresponding to cases with good preservation of the association. Circle size depicts the gain in statistical power after imputation. The bigger the circle the higher the statistical power gain after imputation compared to *CCA*. Squares correspond to cases where no statistical power was gained. Note that due to readability issues, only KNN-based imputation methods with *K* = 3, 10, and 20 were included, whereas KNN imputation with *K* = 1 and 5 can be found in File S6 and Table S8.

Funding statement

This work was supported by grants from the German Federal Ministry of Education and Research (BMBF), by BMBF Grant no. 01ZX1313C (project e:Athero-MED) and Grant no. 03IS2061B (project Gani_Med). Moreover, the research leading to these results has received funding from the European Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement n° 305280 (MIMOmics) and from the European Research Council (starting grant “LatentCauses”). KS is supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. The KORA Augsburg studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF). Analyses in the EPIC-Norfolk study were supported by funding from the Medical Research Council (MC_PC_13048 and MC_UU_12015/1).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Fearnley LG, Inouye M. Metabolomics in epidemiology: from metabolite concentrations to integrative reaction networks. Int J Epidemiol. 2016 Apr 26;dyw046.

2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012 Apr;13(4):263–9.

3. Blow N. Metabolomics: Biochemistry's new look. Nature. 2008 Oktober;455(7213):697–700.

4. Mook-Kanamori DO, Selim MME-D, Takiddin AH, Al-Homsi H, Al-Mahmoud KAS, Al-Obaidli A, et al. 1,5-Anhydroglucitol in Saliva Is a Noninvasive Marker of Short-Term Glycemic Control. J Clin Endocrinol Metab. 2014 Jan 1;99(3):E479–83.

5. Do KT, Kastenmüller G, Mook-Kanamori DO, Yousri NA, Theis FJ, Suhre K, et al. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. J Proteome Res. 2015 Feb 6;14(2):1183–94.

6. Urpi-Sarda M, Almanza-Aguilera E, Tulipani S, Tinahones FJ, Salas-Salvadó J, Andres-Lacueva C. Metabolomics for Biomarkers of Type 2 Diabetes Mellitus: Advances and Nutritional Intervention Trends. Curr Cardiovasc Risk Rep. 2015 Feb 17;9(3):1–12.

7. Rasmiena AA, Ng TW, Meikle PJ. Metabolomics and ischaemic heart disease. Clin Sci. 2013 Mar 1;124(5):289–306.

8. Rhee EP, Gerszten RE. Metabolomics and Cardiovascular Biomarker Discovery. Clin Chem. 2012 Jan 1;58(1):139–47.

9. Wang JH, Byun J, Pennathur S. Analytical Approaches to Metabolomics and Applications to Systems Biology. Semin Nephrol. 2010 Sep 1;30(5):500–11.

10. Armitage EG, Godzien J, Alonso-Herranz V, López-Gonzálvez Á, Barbas C. Missing value imputation strategies for metabolomics data. Electrophoresis. 2015 Dec;36(24):3050–60.

11. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics. 2011 Oct 8;8(1):161–74.

12. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. Metabolites. 2014 Jun 16;4(2):433–52.

13. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W652–60.

14. Redestig H, Kobayashi M, Saito K, Kusano M. Exploring Matrix Effects and Quantification Performance in Metabolomics Experiments Using Artificial Biological Gradients. Anal Chem. 2011 Jul 15;83(14):5645–51.

15. Di Guida R, Engel J, Allwood JW, Weber RJM, Jones MR, Sommer U, et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics [Internet]. 2016 [cited 2017 Jan 13];12. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831991/

16. Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection. Environ Health Perspect. 2011 Mar;119(3):351–6.

17. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. Am J Epidemiol. 2003 Feb 15;157(4):355–63.

18. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007 Jun;16(3):219–42.

19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinforma Oxf Engl. 2001 Jun;17(6):520–5.

20. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015 Oktober;90:84–99.

21. Taylor SL, Ruhaak LR, Kelly K, Weiss RH, Kim K. Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. Brief Bioinform. 2016 Feb 19;

22. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol. 2011;5:21.

23. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013 Oct;14(10):719–32.

24. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004 Feb;69(2 Pt 2):026113.

25. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014 Jun;46(6):543–50.

26. Shrawder E, Martinez-Carrion M. Evidence of phenylalanine transaminase activity in the isoenzymes of aspartate transaminase. J Biol Chem. 1972 Apr 25;247(8):2486–92.

27. Helsel DR. More than obvious: better methods for interpreting nondetect data. Environ Sci Technol. 2005 Oct 15;39(20):419A–423A.

28. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. BMC Bioinformatics [Internet]. 2017 Feb 20 [cited 2017 Mar 16];18. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319174/

29. Helsel DR. Less than obvious - statistical treatment of data below the detection limit. Environ Sci Technol. 1990 Dezember;24(12):1766–74.

30. Holle R, Happich M, Löwel H, Wichmann HE, MONICA/KORA Study Group. KORA--a research platform for population based health research. Gesundheitswesen Bundesverb Ärzte Öffentl Gesundheitsdienstes Ger. 2005 Aug;67 Suppl 1:S19-25.

31. Rubin DB. Introduction. In: Multiple Imputation for Nonresponse in Surveys [Internet]. John Wiley & Sons, Inc.; 1987 [cited 2016 Feb 1]. p. 1–26. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470316696.ch1/summary

32. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009 Jul 28;9:57.

33. D'Angelo GM, Luo J, Xiong C. Missing Data Methods for Partial Correlations. J Biom Biostat [Internet]. 2012 Dec [cited 2016 Feb 28];3(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772686/

34. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681–94.

35. Van Hoewyk J, Lepkowski JM, Solenberger P, Raghunathan TE. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol. 2001 Aug 22;27(1):85–95.

36. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R | van Buuren | Journal of Statistical Software. J Stat Softw [Internet]. 2011 Dec 12 [cited 2016 Feb 28];45(3). Available from: https://www.jstatsoft.org/article/view/v045i03

37. Aichler M, Borgmann D, Krumsiek J, Buck A, MacDonald PE, Fox JEM, et al. N-acyl Taurines and Acylcarnitines Cause an Imbalance in Insulin Synthesis and Secretion Provoking β Cell Dysfunction in Type 2 Diabetes. Cell Metab. 2017 Jun 6;25(6):1334–1347.e4.

Supporting information captions

File S1. LOD tendency.

File S2. Runday-dependent densities in relation with missingness.

File S3. Simulation framework.

File S4. Imputation methods.

File S5. Simulation evaluation results.

File S6. Metabolite-SNP associations–beeswarm plots.

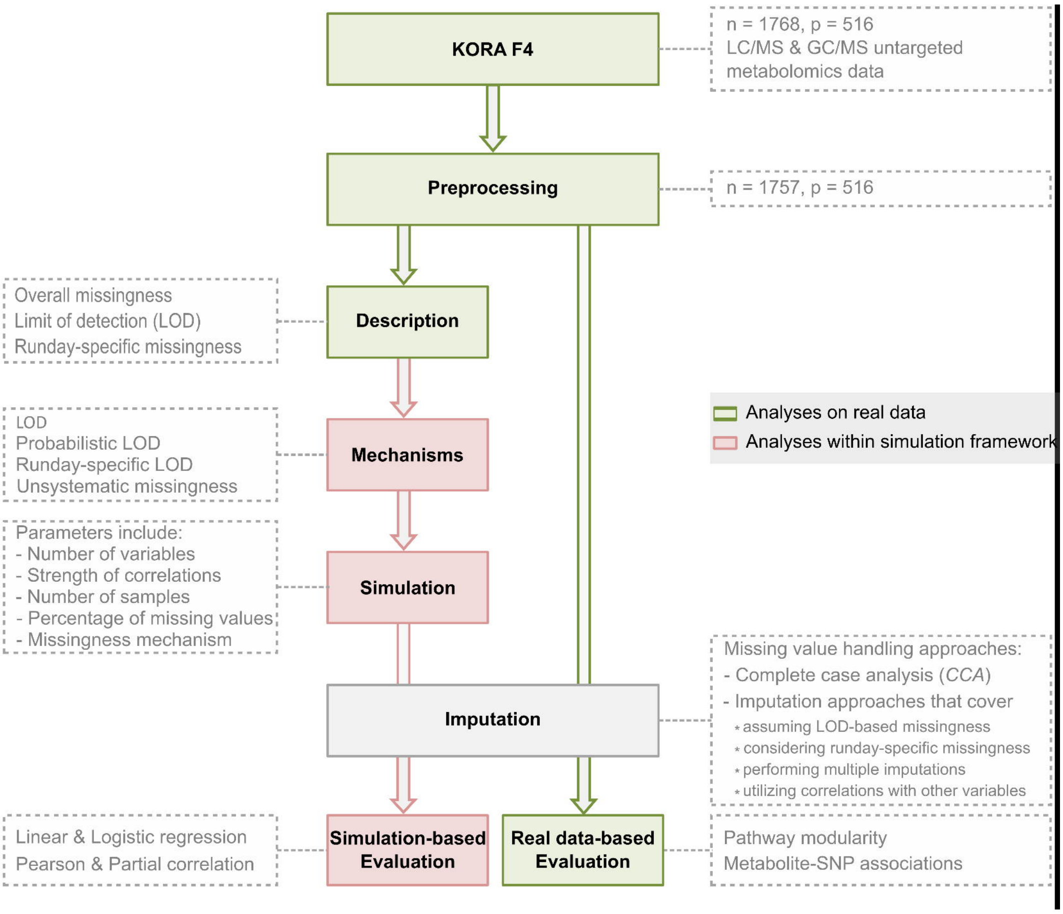
File S7. Metabolite-SNP associations compared with EPIC-Norfolk.

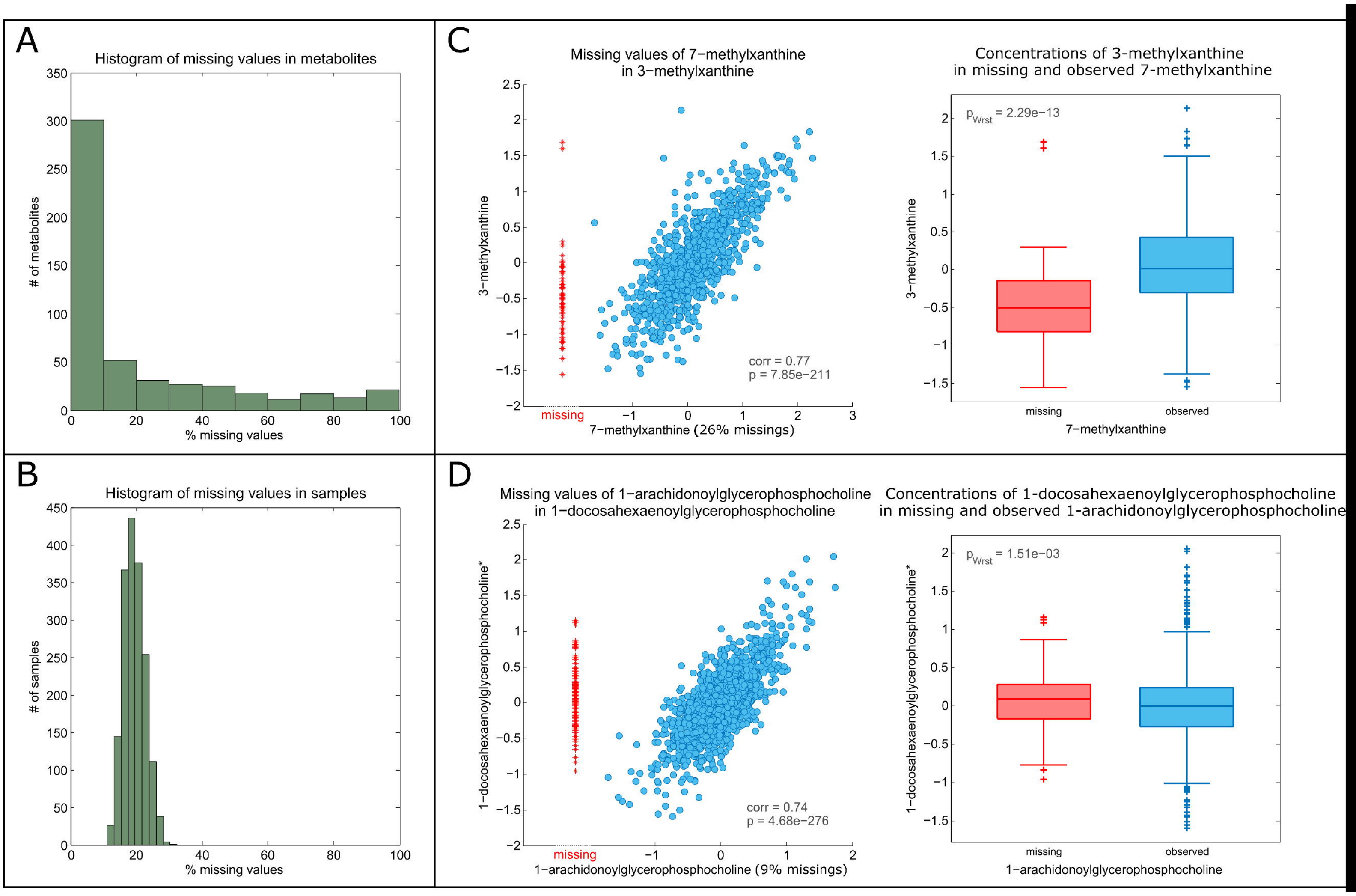
Table S8. Metabolite-SNP associations–linear regression results.

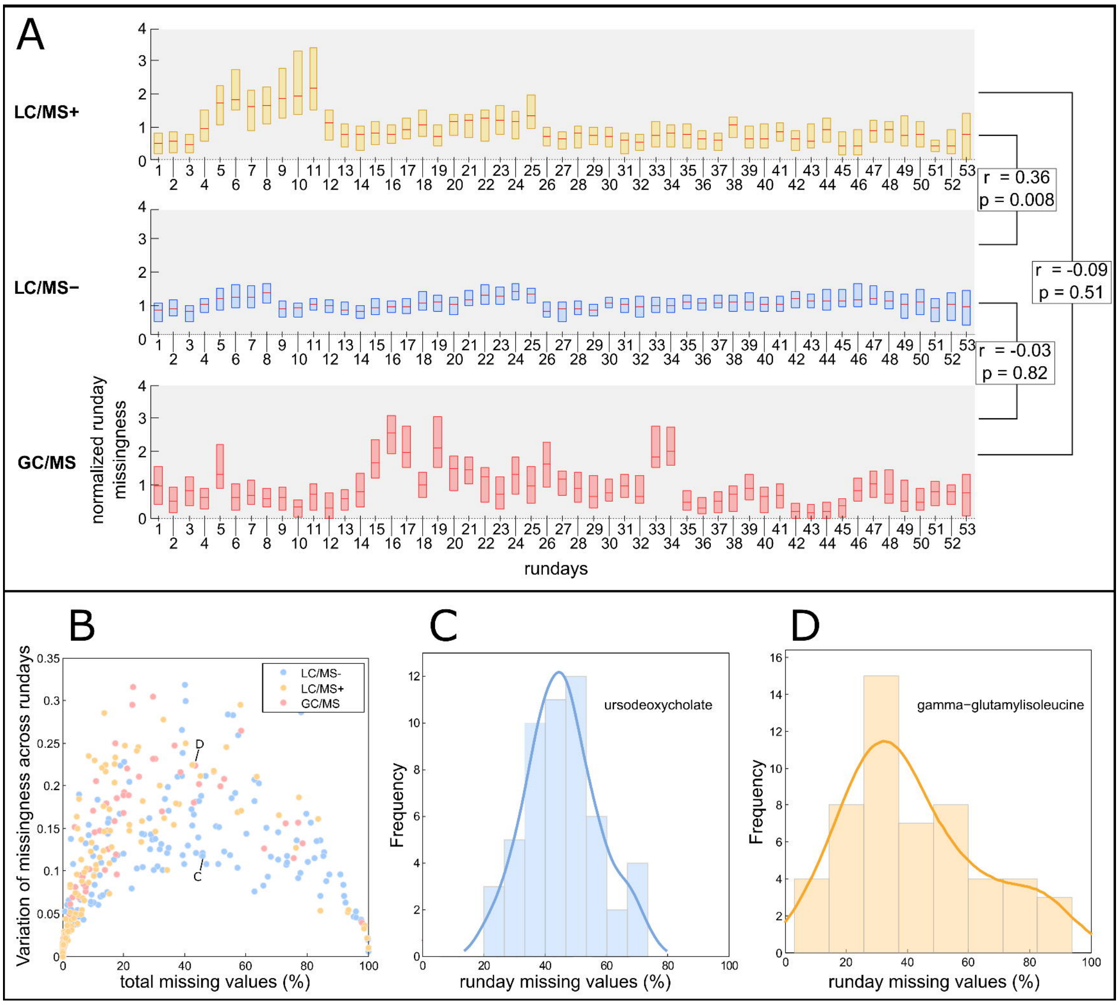
Table S9. KORA F4 annotations.

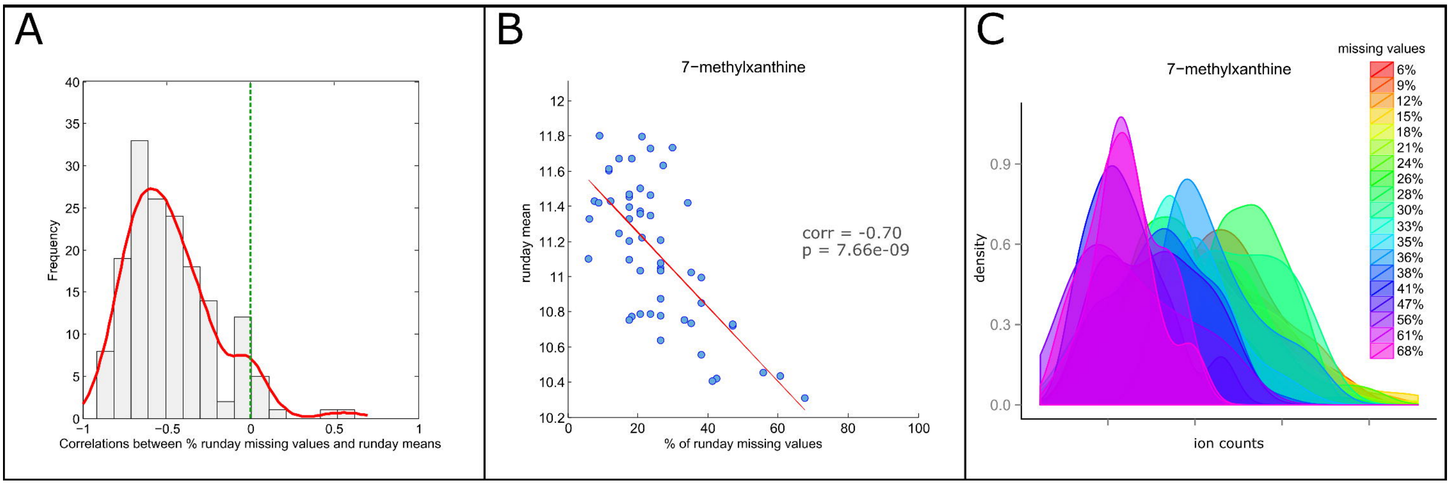
File S10. KORA F4 experimental setup.

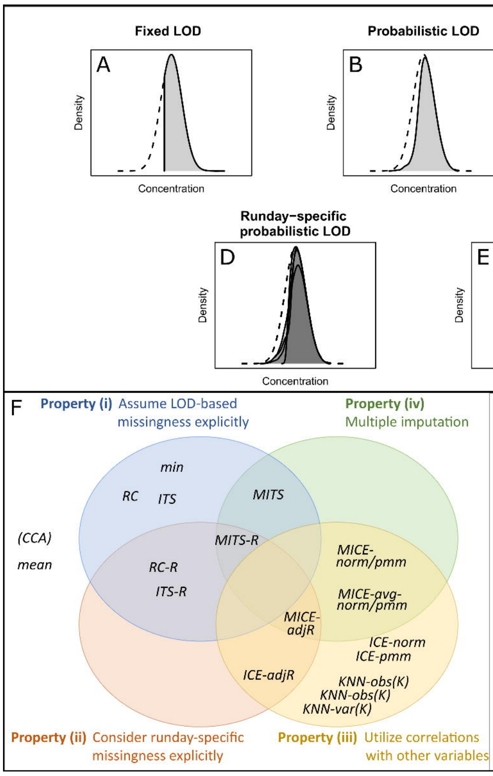
File S11. KNN-obs-sel and MICE imputation code.


KORA F4
n = 1768, p = 516
LC/MS & GC/MS untargeted metabolomics data
Preprocessing
n = 1757, p = 516
Description
Overall missingness
Limit of detection (LOD)
Runday-specific missingness
Mechanisms
LOD
Probabilistic LOD
Runday-specific LOD
Unsystematic missingness
Simulation
Parameters include:
- Number of variables
- Strength of correlations
- Number of samples
- Percentage of missing values
- Missingness mechanism
Analyses on real data
Analyses within simulation framework
Imputation
Missing value handling approaches:
- Complete case analysis (CCA)
- Imputation approaches that cover
⋆ assuming LOD-based missingness
⋆ considering runday-specific missingness
⋆ performing multiple imputations
⋆ utilizing correlations with other variables
Simulation-based Evaluation
Linear & Logistic regression
Pearson & Partial correlation
Real data-based Evaluation
Pathway modularity
Metabolite-SNP associations

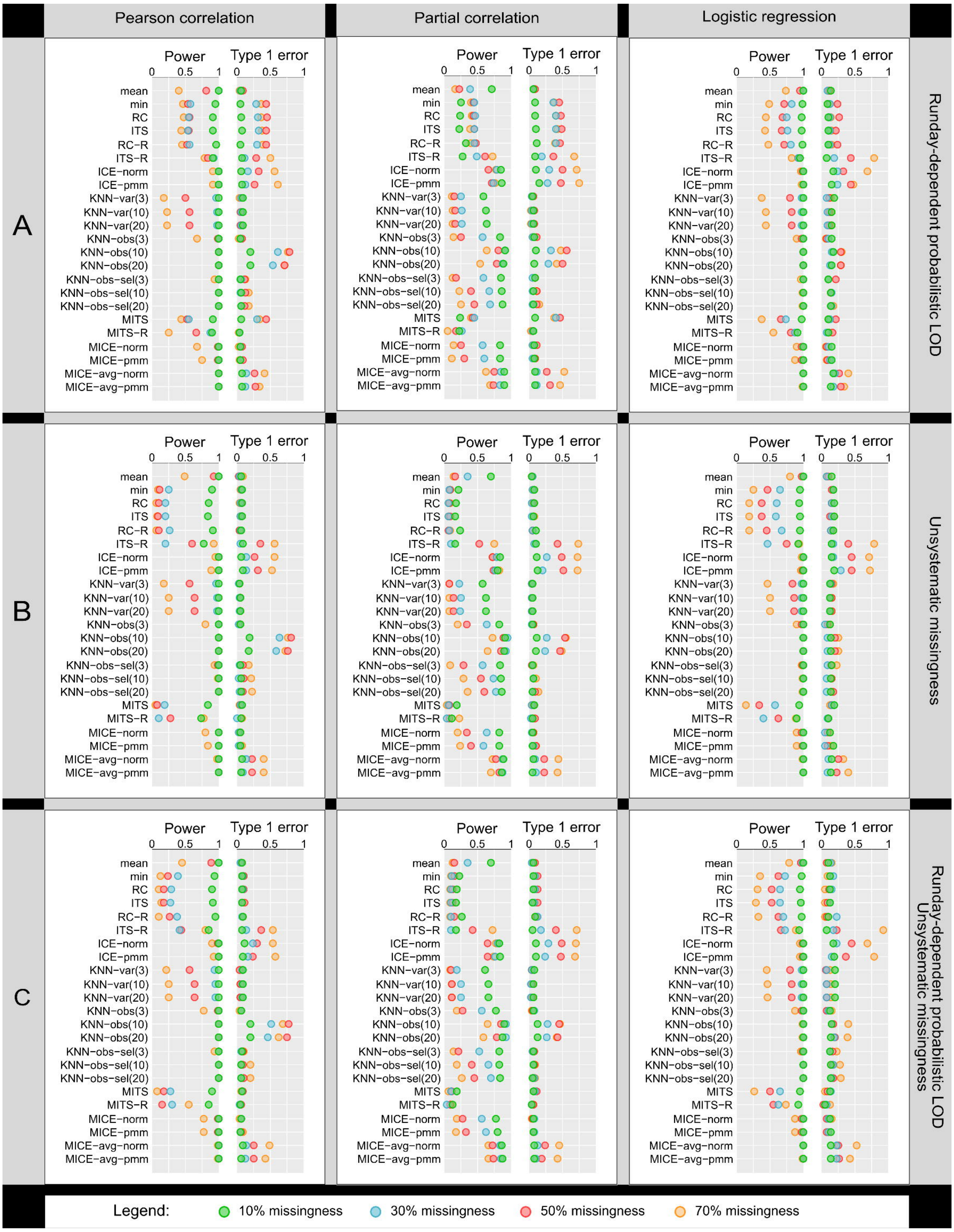

A
Histogram of missing values in metabolites
# of metabolites
0
50
100
150
200
250
300
350
0
20
40
60
80
100
% missing values
B
Histogram of missing values in samples
# of samples
0
50
100
150
200
250
300
350
400
450
0
20
40
60
80
100
% missing values
C
Missing values of 7−methylxanthine in 3−methylxanthine
3−methylxanthine
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
missing
−1
0
1
2
3
7−methylxanthine (26% missings)
corr = 0.77
p = 7.85e−211
Concentrations of 3-methylxanthine in missing and observed 7-methylxanthine
p_{Wrst} = 2.29e−13
3−methylxanthine
−1.5
−1
−0.5
0
0.5
1
1.5
2
missing
observed
7−methylxanthine
D
Missing values of 1−arachidonoylglycerophosphocholine in 1−docosahexaenoylglycerophosphocholine
1−docosahexaenoylglycerophosphocholine*
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
missing
−1
0
1
2
1−arachidonoylglycerophosphocholine (9% missings)
corr = 0.74
p = 4.68e−276
Concentrations of 1-docosahexaenoylglycerophosphocholine in missing and observed 1-arachidonoylglycerophosphocholine
p_{Wrst} = 1.51e−03
1−docosahexaenoylglycerophosphocholine*
−1.5
−1
−0.5
0
0.5
1
1.5
2
missing
observed
1−arachidonoylglycerophosphocholine

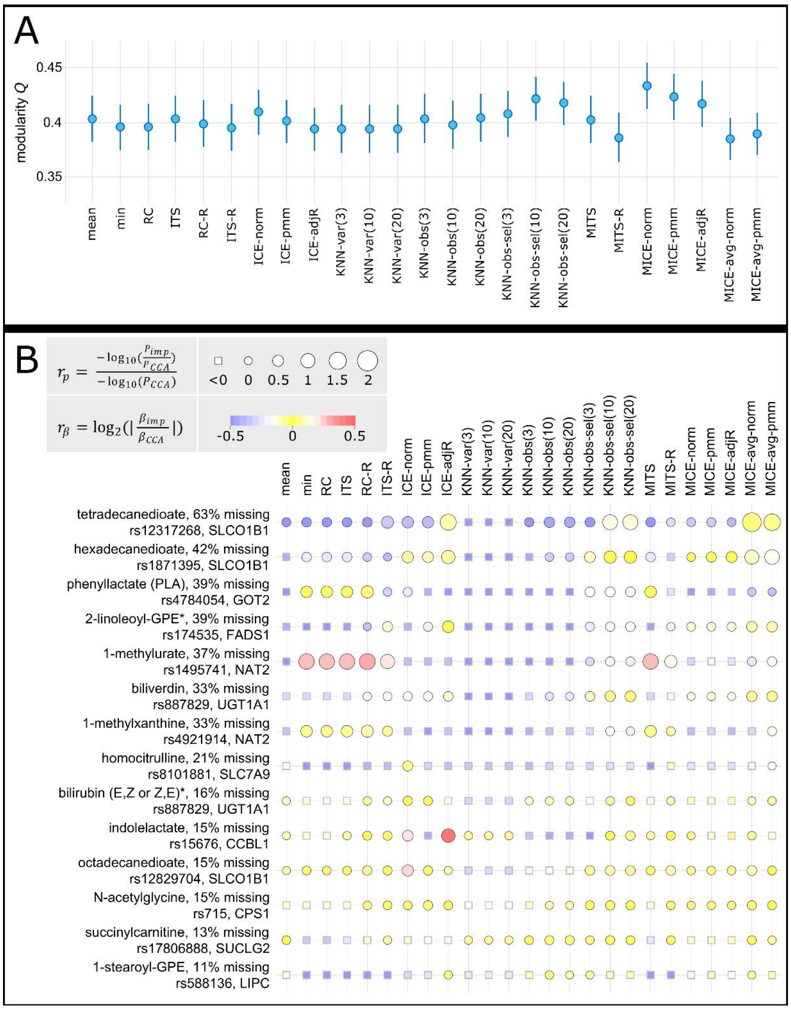

A
LC/MS+
LC/MS−
GC/MS
normalized runday missingness
rundays
r = 0.36
p = 0.008
r = -0.09
p = 0.51
r = -0.03
p = 0.82
B
Variation of missingness across rundays
total missing values (%)
LC/MS-
LC/MS+
GC/MS
D
C
C
ursodeoxycholate
Frequency
runday missing values (%)
D
gamma−glutamylisoleucine
Frequency
runday missing values (%)


A
40
35
30
25
20
15
10
5
0
Frequency
−1
−0.5
0
0.5
1
Correlations between % runday missing values and runday means
B
7−methylxanthine
12
11.8
11.6
11.4
11.2
11
10.8
10.6
10.4
10.2
runday mean
0
20
40
60
80
100
% of runday missing values
corr = -0.70
p = 7.66e-09
C
7−methylxanthine
missing values
6%
9%
12%
15%
18%
21%
24%
26%
28%
30%
33%
35%
36%
38%
41%
47%
56%
61%
68%
0.9
0.6
0.3
0.0
density
ion counts



CCA	Complete case analysis
min	Minimum imputation
RC	Richardson & Ciampi
ITS	Imputation by truncated sampling
MITS	Muliple ITS
RC-R	RC within rundays
ITS-R	ITS within rundays
MITS-R	MITS within rundays
mean	Mean imputation
ICE-norm	Imputation by chained equations using Bayesian regression imputation
ICE-pmm	ICE using predictive mean matching
ICE-adjR	ICE with random runday intercept
MICE-norm	Multiple ICE-norm, pooling statistics
MICE-pmm	Multiple ICE-pmm, pooling statistics
MICE-avg-norm	Multiple ICE-norm, pooling data
MICE-avg-pmm	Multiple ICE-pmm, pooling data
MICE-adjR	Multiple ICE-adjR
KNN-var(K)	*K*-nearest neighbor imputation per variable
KNN-obs(K)	KNN per observation
KNN-obs-sel(K)	KNN per observation using selected variables


Pearson correlation
Partial correlation
Logistic regression
A
B
C
Runday-dependent probabilistic LOD
Unsystematic missingness
Runday-dependent probabilistic LOD Unsystematic missingness
Power
Type 1 error
0
0.5
1
mean
min
RC
ITS
RC−R
ITS−R
ICE−norm
ICE−pmm
KNN−var(3)
KNN−var(10)
KNN−var(20)
KNN−obs(3)
KNN−obs(10)
KNN−obs(20)
KNN−obs−sel(3)
KNN−obs−sel(10)
KNN−obs−sel(20)
MITS
MITS−R
MICE−norm
MICE−pmm
MICE−avg−norm
MICE−avg−pmm
Legend:
10% missingness
30% missingness
50% missingness
70% missingness



A
modularity Q
0.45
0.4
0.35
mean
min
RC
ITS
RC-R
ITS-R
ICE-norm
ICE-pmm
ICE-adjR
KNN-var(3)
KNN-var(10)
KNN-var(20)
KNN-obs(3)
KNN-obs(10)
KNN-obs(20)
KNN-obs-sel(3)
KNN-obs-sel(10)
KNN-obs-sel(20)
MITS
MITS-R
MICE-norm
MICE-pmm
MICE-adjR
MICE-avg-norm
MICE-avg-pmm
B
$r_p = \frac{-\log_{10}(\frac{p_{imp}}{p_{CCA}})}{-\log_{10}(p_{CCA})}$
<0
0
0.5
1
1.5
2
$r_\beta = \log_2(|\frac{\beta_{imp}}{\beta_{CCA}}|)$
-0.5
0
0.5
mean
min
RC
ITS
RC-R
ITS-R
ICE-norm
ICE-pmm
ICE-adjR
KNN-var(3)
KNN-var(10)
KNN-var(20)
KNN-obs(3)
KNN-obs(10)
KNN-obs(20)
KNN-obs-sel(3)
KNN-obs-sel(10)
KNN-obs-sel(20)
MITS
MITS-R
MICE-norm
MICE-pmm
MICE-adjR
MICE-avg-norm
MICE-avg-pmm
tetradecanedioate, 63% missing rs12317268, SLCO1B1
hexadecanedioate, 42% missing rs1871395, SLCO1B1
phenyllactate (PLA), 39% missing rs4784054, GOT2
2-linoleoyl-GPE*, 39% missing rs174535, FADS1
1-methylurate, 37% missing rs1495741, NAT2
biliverdin, 33% missing rs887829, UGT1A1
1-methylxanthine, 33% missing rs4921914, NAT2
homocitrulline, 21% missing rs8101881, SLC7A9
bilirubin (E,Z or Z,E)*, 16% missing rs887829, UGT1A1
indolelactate, 15% missing rs15676, CCBL1
octadecanedioate, 15% missing rs12829704, SLCO1B1
N-acetylglycine, 15% missing rs715, CPS1
succinylcarnitine, 13% missing rs17806888, SUCLG2
1-stearoyl-GPE, 11% missing rs588136, LIPC