Applications of Data-Driven Network Analysis in Metabolomics

by

Gayatri Iyer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Alla Karnovsky, Chair
Professor Charles Burant
Professor Kayvan Najarian
Professor Maureen Sartor
Professor Kathleen Stringer

Gayatri Iyer

griyer@umich.edu

ORCID ID: 0000-0002-8100-0832

DEDICATION

Dedicated to my wonderful parents, Meena and Rajendran Iyer, my loving husband, Ashwin Hariharan, and my two littles, Cooper and Rishi...

ACKNOWLEDGEMENTS

I would like to begin by extending my utmost gratitude towards my advisor, Professor Alla Karnovsky, who has guided and supported me immensely throughout my graduate studies. I have been very fortunate to have been her student and have access to her vast knowledge and experience in bioinformatics. She has been a constant source of encouragement and ideas and it has been an absolute honor to be her doctoral student.

I would like to thank Dr. Karnovsky's group members Dr. Hani Habra, Christopher Patsalis, and Crystal Wen for their amazing help and useful discussions. A special thank you to William Duren and Janis Wigginton for their help with software-building.

I will be eternally grateful to my dissertation committee members, Professors Charles Burant, Kathleen Stringer, Maureen Sartor, and Kayvan Najarian for their constant support, understanding, and encouragement. They made my committee meetings a joy and a fantastic learning experience. I am thankful to Dr. Maureen Kachman and Dr. Charles Evans from the Metabolomics Core (MRC2) at the University of Michigan, Ann Arbor for generating and providing high quality data to work with.

I am particularly grateful to Dr. Jennifer LaBarre, Prof Stephen Goutman, Jonathan Boss, and Prof Michael Maile for making me a part of their research projects and giving me the freedom to analyze their data based on my own knowledge and expertise. They have been a pleasure to collaborate with and I look forward to more collaborative projects in the future!

I would like to extend my gratitude to the entire Department of Computational Medicine and Bioinformatics for accepting me into the department and providing a wonderful environment to grow as a graduate student and researcher. Particularly, I would like to thank Professor Margit Burmeister for providing valuable guidance and support as I navigated the PhD path and to be ever willing to support me when I applied to various graduate funding opportunities. Julia Eussen had been so amazing in patiently answering all my inane questions over the years! A special thank you to Marci Brandenburg for being so nice to work with on various projects and for the wonderful Tools and Tech seminar series!

I would like to thank the wonderful friends I have made during my PhD, whose presence made the hardships more bearable. Shriya Sethuraman, Catherine Barnier, Rucheng Diao, Santhoshi Krishnan – your company made all the department events that much more memorable. A very special thank you to my best friend Gurcharan Kaur from Biomedical Engineering. It was wonderful navigating our PhD journeys together – thank you GC for always being there and looking out for me!

My PhD journey would not have existed if not for the support and understanding from my lovely family. I am eternally grateful to my parents for always believing me, even when I doubted my own capabilities. I thank my husband for trusting and guiding me during my PhD. Last, but by no way the least, I want to thank my two babies Cooper and Rishi. Their innocent and pure love kept me going at my worst times. Their presence has made life so much more beautiful, and I hope I did them proud!

PREFACE

This dissertation describes the original work by Gayatri Iyer, under the guidance of my advisor Dr. Alla Karnovsky and various other research collaborators, detailed below.

Dr. Alla Karnovsky supervised my research, provided guidance in my various research projects, provided funding opportunities, and played a pivotal role in all the publications described in this dissertation. The MMIP lipidomics data (Chapter 2) and COVID-19 metabolomics data (Chapter 3) were generated at the Michigan Regional Comprehensive Metabolomics Research Core (MRC2) at the University of Michigan in Ann Arbor, directed by Prof Charles Burant and managed by Dr. Maureen Kachman.

Chapter 2. A version of this material has been published as **Iyer, Gayatri R.**, Janis Wigginton, William Duren, Jennifer L. LaBarre, Marci Brandenburg, Charles Burant, George Michailidis, and Alla Karnovsky. 2020. "Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations" *Metabolites* 10, no. 12: 479. Professor George Michailidis provided guidance and feedback on the *Filigree* software as well as for the code for feature aggregation and subsampling. Programmers William Duren and Janis Wigginton developed the *Filigree* software. Professor Charles Burant and Dr. Jen LaBarre provided the lipidomics data from the Michigan Mother Infant Pairs (MMIP) cohort and also provided valuable feedback on the biological interpretation of the analyses and the manuscript. Marci Brandenburg developed the user manual for Filigree and provided feedback on Filigree from aa user standpoint. I was responsible for exhaustively testing Filigree's implementation and features. I performed all the analyses described in the Chapter, created all Tables and Figures. I wrote the code for the subsampling portion of the DNEA algorithm that Filigree implements. The manuscript was written by Dr. Alla Karnovsky and me. The funding for this project was provided by Dr. Alla Karnovsky's and Professor George Michailidis's UO1 grant (NIH 1U01CA235487). Christopher Patsalis and I working on developing the DNEA R package. I wrote the feature aggregation and subsampling functions for the package.

Chapter 3. This material is currently unpublished and is being prepared as a manuscript for publication soon. Professors Michael Maile and Charles Burant spearheaded the study and provided the metabolomics data for the hospitalized COVID-19 patients. Dr. Heidi Iglay Reger provided us appropriately matched healthy control samples. Tanu Soni was responsible for data preprocessing and data cleaning. I performed all the analyses described in this chapter created all the Tables and Figures. Dr. Alla Karnovsky and I will write the manuscript that will follow this chapter.

Chapter 4. A version of this material has been published as Goutman, Stephen A., Jonathan Boss, **Gayatri Iyer**, Hani Habra, Masha G. Savelieff, Alla Karnovsky, Bhramar Mukherjee, and Eva L. Feldman. 2022. "Body mass index associates with amyotrophic lateral sclerosis survival and metabolomic profiles." *Muscle & Nerve*. Professors Stephen Goutman, Bhramar Mukherjee, and Eva Feldman designed and conceptualized the study, funded the project, and provided the metabolomics data for the ALS patients. Professor Stephen Goutman provided Tables 4.1 – 4.3. Jonathan Boss performed the survival analysis and formulated the BMI trajectory groups. He provided Figures 4.1 – 4.3 and Tables 4.4 and 4.5. Hani Habra performed the initial merging of the metabolomics datasets. Masha Savelieff helped in drafting the manuscript and organizing the Figures, Tables, and supplementary material. I performed all the analyses on the metabolomics data. I created Tables 4.6 – 4.10 and Figures 4.4 – 4.6. The metabolomics analysis sections of the manuscript were written by Dr. Alla Karnovsky and me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Metabolomics is a systems-wide study of small molecule metabolites. It provides a read-out of underlying cellular and biochemical events. Liquid Chromatography coupled with Mass Spectrometry (LC-MS) is one of the most common analytical platforms used to perform metabolomics studies. The analysis of LC-MS metabolomics data is a complex multi-step process. It involves data processing, normalization, followed by statistical analysis and functional interpretation. While several computational tools have been built to help perform these tasks, a major challenge remains linking alterations in metabolite levels to specific biological processes. In this dissertation, I develop and apply novel computational methods for the analysis and interpretation of metabolomics data, to help build testable hypotheses and derive novel biological insights.

Over the past decade, mapping and visualizing experimentally measured metabolites in the context of known biochemical pathways has become ubiquitous. However, pathway mapping is restricted to named metabolites from well-annotated biochemical pathways. Realizing the limitations of knowledge-based approaches, in Chapter Two, we developed a bioinformatics tool, *Filigree*, that provides a data-driven approach by inferring associations among metabolites directly from experimental measurements to construct metabolic networks. These associations can be quantified by 'partial correlations' that measure the conditional dependence between metabolites, thus eliminating spurious and non-informative interactions. In a high-dimensional setting (n << p), the partial correlation network is computed using the l1-regularized graphical lasso method. The Differential Network Enrichment Analysis (DNEA) algorithm that Filigree implements computes the network using a joint estimation method (JEM) which allows the use of all samples in both experimental groups by modifying the graphical lasso penalty term. The network is then clustered using consensus clustering to identify highly interconnected subnetworks; and finally, the enrichment of these subnetworks is determined using the

NetGSA algorithm. In addition, *Filigree* addresses common challenges that often arise in the analysis of “real world” metabolomics data like high dimensionality (n << p) and highly imbalanced experimental groups. To demonstrate *Filigree*’s applicability, I analyzed metabolomics datasets from type 1 and type 2 diabetes and lipidomics dataset from the Michigan Mother-Infant Pairs (MMIP) cohort and were able to identify previously known and some novel biochemical disruptions leading to an altered metabolic state.

In Chapter Three, I analyzed a COVID-19 metabolomics data to identify metabolic markers of disease severity. My analysis revealed that the plasma metabolome of COVID-19 patients and healthy controls is strongly influenced by clinical characteristics as well as anesthetic administration for intubation. There were distinct differences in the metabolic profiles of patients with mild and severe COVID-19. These differentiating metabolites included several acylcarnitines and acylglycerols and were better able to discriminate mild and severe COVID when compared to clinical risk factors.

In Chapter Four, I assessed the association of data-driven metabolic modules with the BMI trajectory of ALS (Amyotrophic Lateral Sclerosis) patients over 5- and 10-years preceding diagnosis. I showed that while individual metabolites do not show a significant association with BMI trajectory, metabolic modules obtained from partial correlation networks do, suggesting a nuanced relationship between BMI trajectory and the metabolome. Additionally, a subset of these metabolites was individually predictive of ALS survival as well, indicating a metabolic link between loss of BMI and ALS survival.

CHAPTER I

Introduction to Computational Metabolomics

1.1 Introduction to metabolomics

Metabolites are small (50 to 1500 Da) molecules that comprise the substrates, products, and intermediates of cellular metabolism. Metabolites play a crucial role in a variety of physiological processes including signaling[1,2], immune modulation[3], regulation of gene expression[4,5], and cofactor activity[6]. "Metabolomics" is the term given to the comprehensive and systematic study of the metabolome, with the aim to investigate the underlying physiological state of the system. The "metabolome" represents the repertoire of all metabolites in a biospecimen and thus provides a readout of the underlying cellular and biochemical events that reflect the genetic makeup, epigenetics, the microbiome, and environmental exposures, including diet. The metabolome can therefore be considered as a crucial link between the genotype and phenotype (**Figure 1.1,** obtained from Steur et al (2019)[7] and Carneiro et al (2019)[8]). Dynamic changes in metabolic processes occur on a timescale of a few seconds. These properties make the metabolome an attractive tool for the investigation of the system phenotype.

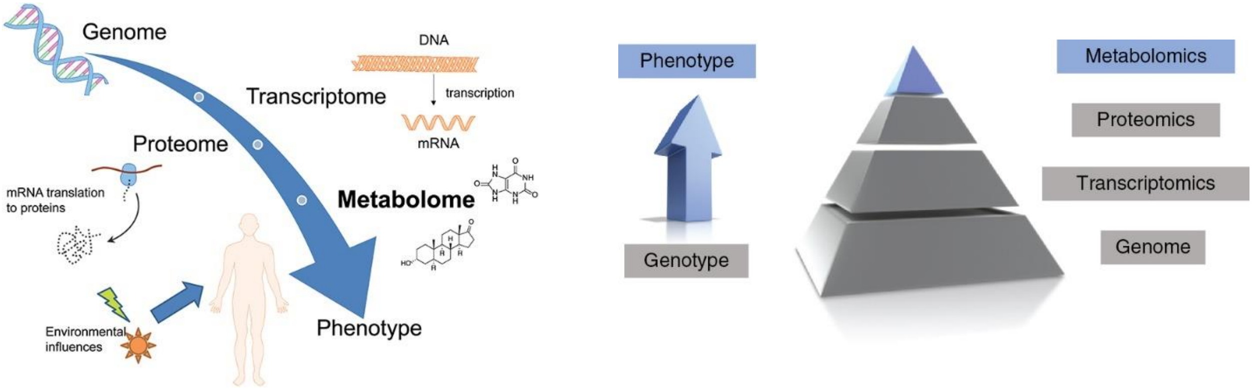


Figure 1.1: Overview of various omics technologies (left)[7] and the role of metabolomics in the "omics pyramid" (right)[8].

1.2 Metabolomics in biomarker discovery

Over the last decade, the field of metabolomics has become an integral part of basic, clinical, and translational research. Metabolomics has played a particularly crucial role in biomarker discovery in a variety of diseases, including several cancers[9–11], cardiovascular pathology[12,13], renal diseases[14,15], lung diseases[16], and diabetes[17–20]. One of the biggest advantages of utilizing metabolomics in a clinical setting for biomarker discovery is the ability to garner significant amount of information from non-invasive and relatively easy to obtain biological samples such as blood/plasma, urine, feces, saliva, and in some cases, even hair [21].

Assessing the levels of small molecule metabolites to ascertain the presence of disease is not a new concept. One of the earliest applications of metabolites as markers of disease was for diagnosing inborn errors of metabolism (IEM)[22]. Another classic example is the measurement of blood glucose levels to monitor diabetes. Similarly, serum creatinine is a marker for kidney function[23] while serum bilirubin, alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are markers of liver function[24].

Recent advances in sensitivity and accuracy of metabolomics assays had significant impact on biomarker discovery. These technologies have made it possible to identify multiple biomarkers for a disease, enabling better diagnosis. Metabolomics can be utilized to identify biomarkers of a disease after the disease has manifested itself i.e., metabolic differences are already apparent. In this case the identified biomarkers can be useful for diagnosis purposes. For example, sarcosine has been identified as a marker of prostate cancer progression[25] and trimethylamine N-oxide has been reported as a marker of cardiovascular disease[26]. On the other hand, metabolomics can also be used to identify predictive biomarkers of a disease before the onset of clinical symptoms. A panel of the three amino acids - isoleucine, tyrosine and phenylalanine, has been shown to be an effective marker of future development of type 2 diabetes[19], while the branched chain amino acids (BCAAs) isoleucine, leucine and valine have been shown to be markers of future development of pancreatic cancer[27]. In either scenario, the approach typically follows the “hypothesis-generating” model wherein the metabolic profiles of healthy and non-healthy individuals are compared, the most differentiating metabolites are identified leading to identification of dysregulated metabolic pathways.

The vast majority of biomarker studies rely on non-targeted metabolomics (described in subsequent sections) wherein the goal is to analyze as many metabolites as possible to arrive at a single or a panel of the most discriminating metabolites. Alternatively, targeted metabolomics (elaborated in subsequent sections) aims at identifying and quantifying a preselected category of metabolites within a sample. Selection criteria can be either based on a common chemical class (for example, amino acids or lipids) or based on specific biochemical pathways or a proposed hypothesis. Targeted and untargeted analyses are complementary and their integrative implementation in biomarker discovery reveals the true power of the metabolome in understanding complex biochemical pathways.

1.3 Metabolomics instrumentation and assays

Metabolites have highly diverse physical and chemical properties and are therefore classified into various biochemical classes (lipids, amino acids, peptides, sugars, fatty acids, organic acids, steroids, etc.). Owing to this chemical and structural diversity, the instrumentation and technologies applied to measure these metabolites are also varied and depend on the goal of the analyses.

Metabolomics experiments largely employ either one (or both, depending on the experiment design) of the two following analytical platforms:

1) Nuclear Magnetic Resonance (NMR) spectroscopy, and 2) Mass Spectrometry (MS), coupled with a sample separation technique such as Liquid Chromatography (LC-MS), Gas Chromatography (GC-MS), or Capillary Electrophoresis (CE-MS).

NMR spectroscopy exploits the unique energy signature emitted by a metabolite when subjected to electromagnetic radiation of a specified frequency in the presence of an external magnetic field to determine the molecular composition of a sample based on their chemical shift patterns. NMR spectroscopy is very advantageous in that it is a non-destructive technique and samples can be re-analyzed as needed. Additionally, NMR is a highly reproducible technique, requiring minimum effort for sample preparation and is routinely quantitative[28]. However, one of the biggest weaknesses of this technique is its low sensitivity; the low-throughput coverage of the metabolome also makes NMR less attractive in comparison to mass spectrometry-based metabolomics.

Mass Spectrometry allows the detection of very low abundant metabolites (picomolar range), making it an attractive alternative to NMR. A typical mass spectrometer consists of a sample-introduction system, ionization source, mass analyzer and ion detector. Molecules are separated and quantified based on their mass/charge (m/z) ratio.

For complex mixtures such as most biofluids, the mass spectrometry analysis is preceded by a chromatographic separation technique that reduces ion-suppression effects. The most common separation techniques used are liquid chromatography (LC) and gas chromatograph (GC), although capillary electrophoresis (CE) is also routinely employed. LC-MS is the most common technique applied in metabolomics. Liquid Chromatography consists of a non-polar stationary phase and a polar mobile phase. The analyte moves through the stationary phase (column) and gets adsorbed based on its physicochemical properties i.e., compounds with a higher affinity for the stationary phase will be retained in the column for longer and vice versa. Separation is achieved as compounds with differential affinity are eluted from the columns at different times. The eluting compounds can be characterized by their retention times. The most commonly employed LC modes in metabolomics are reversed-phase (RPLC) and hydrophilic-interaction liquid chromatography (HILIC). While HILIC is used for more polar compounds, RPLC is useful for separating less polar compounds. In gas chromatography, separation takes place in a gas phase. Thermally stable compounds are vaporized by bringing them to their boiling points. The temperature is gradually increased to vaporize different compounds at different times. The elution of compounds thus depends on their molecular weights as higher molecular weight compounds will have a higher boiling point. For non-volatile compounds, a derivatization step is typically employed to make them amenable to GC separation.

Metabolomics experiments typically involve either one of the following fundamentally complementary approaches depending on the goal of the study: targeted or untargeted analysis. Targeted metabolomics is a hypothesis-driven approach wherein the metabolites to be measured (typically < 200) are defined *a priori* based on chemical similarity or biochemical relationships. Because of this, metabolites can be quantified in absolute terms and with high precision. However, the drawback remains that only a small fraction of the metabolome is measured thus limiting novel findings. Recently, a targeted approach was employed to identify two panels of metabolic biomarkers of COVID-19. The first panel included the kynurenine/tryptophan ratio, lysoPC 26:0, and pyruvic acid discriminated controls from COVID-19 patients, while the second panel included C10:2, butyric acid, and pyruvic acid discriminated hospitalized and non-hospitalized COVID-19 patients [29].

Untargeted metabolomics aims to measure the “universe” of metabolites in the specimen. This is a hypothesis-generating approach that provides a holistic view of all the small molecules in the sample and has the potential to reveal novel and unanticipated perturbations. Untargeted data has a wealth of information that can be mined and has been used for biomarker discovery [30–32]. However, the data are very complex with the presence of a high proportion of unknown metabolites and ionization fragments and adducts (“metabolic features”). Compound identification typically requires tandem MS (MS/MS) analysis and can be cost and labor intensive. Further, simultaneous measurement of all metabolite classes is still challenging as several factors affect metabolite recovery, depending on the functional group of the metabolite [33]. Given the highly complex and redundant nature of untargeted metabolomics data, sophisticated computational tools are required for the analysis and interpretation. Some of the main considerations from a computational standpoint are big data processing, metabolite identification, statistical analyses, and biological/functional interpretation.

1.4 Description of metabolomics experiment workflow

A typical metabolomics experiment consists of the following steps: (i) experimental design and sample collection, (ii) sample preparation, (iii) data acquisition, (iv) data processing, (v) statistical analysis, and (vi) biological interpretation (**Figure 1.2**). These are detailed in the following sections.

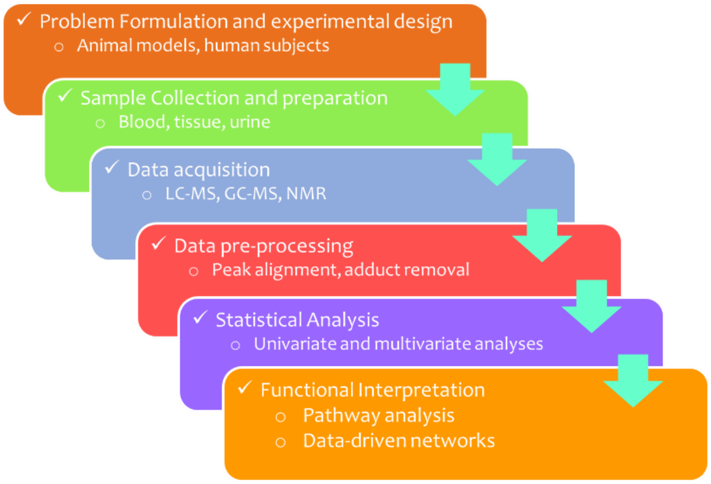

✓ Problem Formulation and experimental design
○ Animal models, human subjects
✓ Sample Collection and preparation
○ Blood, tissue, urine
✓ Data acquisition
○ LC-MS, GC-MS, NMR
✓ Data pre-processing
○ Peak alignment, adduct removal
✓ Statistical Analysis
○ Univariate and multivariate analyses
✓ Functional Interpretation
○ Pathway analysis
○ Data-driven networks

Figure 1.2: Typical Metabolomics experiment workflow

1.4.1 Experimental design

Experimental design is a crucial first step for a researcher about to embark upon a metabolomics study. A key consideration for experimental design includes the goal of the study, which can typically be comparing the metabolomes of the phenotypes of interest with that of controls, getting mechanistic insights into metabolic dysregulation, or characterizing the metabolome of a new specimen. This will then dictate what the source of the samples should be i.e., human or animal samples or models. The next decision to be made is whether the samples will be obtained from tissues, cells, biofluids, or cell cultures. Storage and stability of samples becomes crucial here as some variability in metabolite levels can be introduced based on whether freshly collected samples are used or freeze-thawed samples. Importantly, the number of samples and/or size of the experimental groups must be determined depending on the biological variability in the system. For example, samples harvested from controlled laboratory cell cultures will have far less variability when compared to human tissue samples obtained from a large epidemiological study. The latter will require higher number of samples for greater statistical power. Since the growth of cells in culture can be carefully controlled, a sample size of 3-5 per group can provide useful preliminary data, while epidemiological studies can require patient numbers in the thousands. Given the highly dynamic nature of the metabolome, the timing of sample collection is also a crucial consideration [34]. Typically, fasting samples are collected to minimize biases. Care needs to be taken to account for diurnal variability as well. Controlling for the effects of diet or the time of day of sample collection can also help to minimize variability that may otherwise confound true biological variations. Controlling for technical variations such as the containers utilized for storing samples, the anti-coagulant (in the case of plasma samples) used, and storage conditions can also help mitigate confounding effects. The next consideration is the choice of sample preparation and analytical platform. Finally, appropriate statistical and computational tools have to be selected based on the study objectives and the biological questions raised.

1.4.2 Sample preparation

The choice of sample preparation strategy will greatly contribute to the success of a metabolomics experiment as it influences not only the observed metabolite profile but also the quality of the data, which can in turn affect the biological interpretation [35]. An ideal sample preparation strategy will therefore involve maximizing the correlation between the observed and the true metabolome. Depending on the biospecimen being analyzed, sample preparation can involve several steps that include, but are not limited to, metabolite extraction, protein removal by precipitation, derivatization, evaporation, reconstitution, or dilution.

A well-designed experiment will also include quality control (QC) samples, the data from which are primarily utilized for normalization. QC samples can include pooled aliquots of all experimental samples and negative controls in the form of "blanks" that contain only the extraction solvent used. These QC samples are run intermittently along with experimental samples to avoid systemic biases in the measurements. Additionally, biological samples may be mixed with purified compounds (internal standards) to further reduce instrument driven technical artifacts.

1.4.3 Data acquisition

Two main platforms exist for measuring the levels of metabolites in the sample: mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. In order to effectively reduce the complexity of the biological sample and quantify sets of metabolites separately and sequentially, MS-based methods are usually preceded by a chromatographic separation step. The two commonly employed separation methods are High Performance Liquid Chromatography (HPLC) and Gas Chromatography (GC). HPLC-MS is preferred while analyzing more polar metabolites like amino acids, nucleotides, polyamines, etc while GC-MS is favored for analyzing non-polar compounds like lipids, eicosanoids, esters, etc. The latter requires some form of derivatization to make the sample amenable to analysis[36]. The focus of this dissertation will be on data obtained from HPLC-MS. The order of elution of analytes (represented by their "retention time") from a HPLC column is largely dependent on their physiochemical properties. Upon elution, the separated analytes are directly injected into the mass spectrometer, where they are instantly ionized to generate charged particles. Some commonly employed ionization methods include chemical ionization (CI), electrospray ionization (ESI), and Matrix-assisted laser desorption ionization (MALDI). ESI is one of the most widely used ionization technique in untargeted metabolomics studies, largely due to its applicability to a wide range of metabolites. In the next step, the charges particles migrate to the mass analyzer under high vacuum. Two main types of mass analyzers are commonly used: Time-Of-Flight (TOF) and Orbitrap. In TOF-MS, the mass of the charged ions is measured based on the time they take to traverse through a flight tube in an electric field. Ions with lower mass and higher charge tend to travel faster through the flight tube. On the other hand, in an Orbitrap, the mass of ions is calculated based on their oscillation frequencies when they are suspended in an electric field. Both QTOF and Orbitrap instruments are routinely used for untargeted metabolomics analysis. The mass spectral data information emerging from these instruments are then subjected to downstream processing.

1.4.4 Data preprocessing

LC-MS instruments generate large amounts of complex data on metabolic signals that require specialized tools and software for processing. Data preprocessing for MS typically includes noise reduction, baseline/retention time correction, normalization, peak alignment, peak detection and integration, peak quantification, and spectral deconvolution. Some of the commonly used open source software for data preprocessing include XCMS[37], MZmine3[38], MetaboAnalyst 5.0[39], MS-DIAL 4[40], and MAVEN[41]. Various commercial software also exist for preprocessing data from specific vendors; Agilent's MassHunter™, Thermo Fisher's Compound Discoverer™, and Bruker's MetaboScape®. Certain software addresses a specific step in the data preprocessing workflow, while others cover several or all the steps. For example, MZmize3 is designed to perform all of the preprocessing steps including noise filtering, peak detection, peak alignment, deisotoping, gap filling, normalization, and visualization, whereas XCMS, MAVEN and MetabolAnalyst do not perform deisotoping or allow the user to process the samples in batch mode. The choice of software largely depends on the specific application (eg.: targeted or untargeted data) and can have a significant impact on the results from downstream analysis. For example, the peak detection methods employed by XCMS and MZmine2 (centWave) have been shown to generate many false positive and false negative peaks, resulting in a higher initial feature count compared to other software[42]. Interpretation of the metabolomics data in a biological context must therefore be carefully considered based on the preprocessing software employed.

1.4.5 Statistical analysis

Statistical analyses of metabolomics data typically involve univariate and multivariate approaches.

The goal of univariate analyses is to identify individual metabolites that are most differentially abundant between the phenotypes of interest. For two-group data (both unpaired and paired), Student's *t*-tests and fold-change analysis are typically performed, while for multi-group data, one-way analysis of variance (ANOVA) and post hoc analysis are performed. The p-values obtained from these tests need to be corrected for multiple testing, given that a large number of metabolites are usually measured in a metabolomics experiment and statistical tests are performed for each metabolite. Commonly employed multiple testing correction methodologies include the Bonferroni correction and the Benjamin–Hochberg correction, also known as the false discovery rate (FDR) [43]. Additionally, one can also test the association of individual metabolites with phenotypes of interest in these statistical models. Here, the response variable is each metabolite's expression, and the predictors are the phenotypes (eg: age, BMI, gender). Such models can help in identifying potential confounders in the data and enable the researcher to correct for them by using the residuals from these models for downstream analyses.

Multivariate methodologies are particularly useful for exploratory data analysis as they assess the change in groups of metabolites simultaneously. Multivariate analysis can either be unsupervised or supervised. Unsupervised methods include principal component analysis (PCA) and cluster analysis (hierarchical, K-means clustering) and are helpful in deducing trends or patterns in the data and identify outliers. Supervised methods include partial least squares discriminant analysis (PLS-DA), Support Vector Machine

(SVM), Random Forest, k-nearest neighbor (KNN), and logistic regression (**Figure 1.3**, obtained from Bujak et al (2015)[44]). These are often useful in classification problems and aid in biomarker discovery[45].

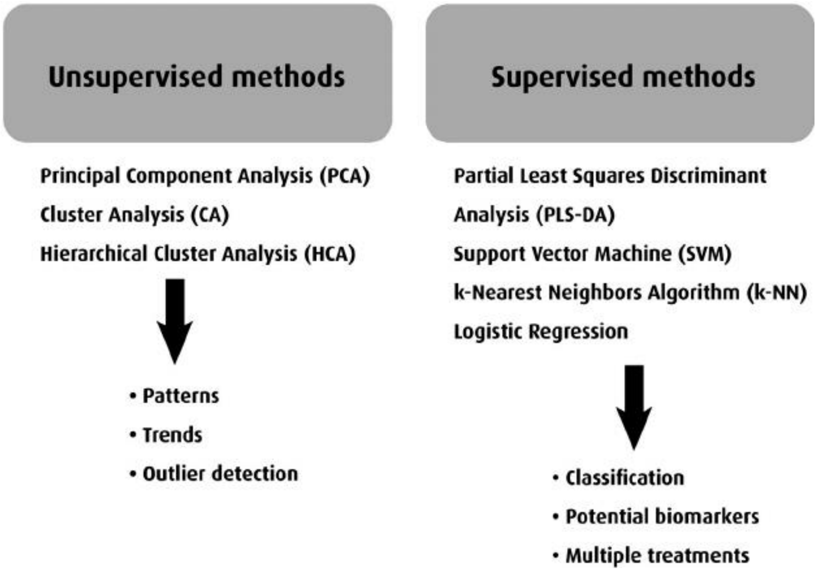

Unsupervised methods
Principal Component Analysis (PCA)
Cluster Analysis (CA)
Hierarchical Cluster Analysis (HCA)
• Patterns
• Trends
• Outlier detection
Supervised methods
Partial Least Squares Discriminant Analysis (PLS-DA)
Support Vector Machine (SVM)
k-Nearest Neighbors Algorithm (k-NN)
Logistic Regression
• Classification
• Potential biomarkers
• Multiple treatments

Figure 1.3: Supervised and Unsupervised multivariate analysis methods used in metabolomics[44].

1.4.5.1 Unsupervised methods

Principal Component Analysis (PCA)

PCA is a dimensionality-reduction method that transforms a large set of variables into a smaller set while retaining as much information as possible[46]. This is done by identifying a set of combinations of the variables that explain most of the variance in the data. PCA performs an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance comes to lie of the first coordinate (first principal component), the second greatest variance on the second coordinate, and so on. For a given dataset with n observations and p variables, the mathematical representation of PCA is as follows:

$$T = XP$$

where $X_{n\,x\,p}$ denotes the original dataset that has been standardized (mean 0 and unit variance), $T_{n\,x\,p}$ denotes the PC scores for all the subjects, and $P_{p\,x\,p}$ denotes the weights (i.e., loadings). For metabolomics data, observations refer to samples, and variables to metabolite abundance. In order to assess the contribution of a PC to the total sample variance, percentage of variance explained by the PC is calculated. This is done by dividing the eigenvalue of the corresponding PC by the sum of all the eigenvalues. The scree plot[47] shows the eigenvalues of the PCs and is useful in deciding how many PCs should be retained. The typical rule-of-thumb is to retain the PCs to the left of the “elbow” point in the curve of the scree plot after which the eigenvalues appear to level off.

PCA is arguably the most widely used exploratory analysis method to summarize complex metabolomics data. It has been extensively utilized for data reduction and to identify trends and patterns in the data that may correlate with other biological factors. PCA is also very useful in assessing data quality by identifying outlier samples and technical variation in the data. One of the biggest advantages of PCA is that it is not prone to over-fitting, however certain non-linear trends in the data are likely to be missed.

Hierarchical clustering

Hierarchical Clustering (HC)[48] partitions the dataset into a tree structure by building a hierarchy. Initially, all variables are treated as separate clusters and are merged using some similarity metric between pairs of variables. Euclidean distance is often used as a measure of dissimilarity for clustering. Other commonly used distance metrics include the Manhattan distance, Mahalanobis distance and maximum distance. The partitioning is represented as a dendogram and one can either decide how many clusters they desire and cut the tree accordingly or set a similarity cut-off and obtain the clusters. The main advantage of hierarchical clustering is that it does not require one to set the number of clusters *a priori*. This allows us to explore the structure of the data better and pick the appropriate number of clusters. However, HC is relatively sensitive to outliers in the data.

Hierarchical clustering, coupled with heatmaps, is immensely useful in visualizing and discovering the real structure of the metabolomics data. As with other unsupervised methods, the HC dendrograms provide a useful way to unearth trends in the data that may warrant further exploration.

K-means clustering

K-means clustering[49] is a centroid-based clustering method with the aim of partitioning n observations into k non-overlapping clusters. Unlike hierarchical clustering, the number of clusters k must be decided by the user. The method initiates k clusters in the space spanning the variables by randomly assigning k data points to each cluster (centroids). The algorithm then finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) choosing centroids based on the current assignment of data points to clusters. An extension of K-means is Fuzzy C-means clustering[50], where variables can be assigned to more than one cluster i.e., overlapping clusters.

K-means clustering is also a great tool for visualizing metabolomics datasets. Given that the number of clusters is predetermined, it is very helpful in gauging how the samples in the dataset cluster relative to the experimental group assignment. This can help in identifying possible biases in the data or some novel trends that can be further investigated.

1.4.5.2 Supervised methods

Partial least squares discriminant analysis (PLS-DA)

PLS-DA[51] can be thought of as the supervised version of PCA that provides a visual interpretation of complex datasets in a low-dimensional setting. The method aims to optimize the separation between groups of variables by maximizing the covariance between the data and the group membership by finding a linear subspace of the explanatory variables[52]. This new subspace permits the prediction of the grouping variable based on a reduced number of factors (PLS components, or latent variables). PLS-DA is fairly robust to highly collinear and noisy data[53]. It also provides variable importance scores (VIP scores) that can be used to rank variables based on their contribution to the classification[54]. However, PLS-DA is prone to over-fitting and the visual representation of the data must be interpreted with caution.

Support Vector Machine (SVM)

SVM is a popular machine learning algorithm used for both classification and regression[55]. In classification, the data is mapped onto a high-dimensional space to separate the two groups of samples into distinctive regions. The algorithm then identifies a n-dimensional hyperplane that distinctly classifies the data points by maximizing the distance between data points of both groups. Data points close to the separating hyperplane are termed ‘support vectors’ and they influence the position and orientation of the hyperplane. In order to avoid over-fitting of the model, oftentimes a small fraction of the training samples is allowed to be misclassified – this is referred to as soft margin SVM. The main advantage of SVM is that the kernel function can be chosen for both linear and non-linear separation. A 'kernel trick'[56] is applied for non-linear classification that transforms the input space into a high-dimensional feature space where the groups are linearly separable.

Random Forest (RF)

Random Forest[57] is another supervised classification algorithm. It consists of an ensemble of decision trees constructed by randomly sampling the input variables. It uses bootstrap sampling to create sets of randomly sampled variables and builds decision trees where each node's decision is based on a random set of features[58]. The final decision is based on majority voting after combining the decisions of all the trees. Random Forest has several advantages including its ability to work well with large datasets, provide feature importance score (mean decrease in accuracy), and handling missing values. Moreover, this method is fairly robust to overfitting and outliers as well.

K-nearest neighbor (KNN)

KNN[59] is another commonly used non-parametric, supervised classifier that assigns groups labels to a data point based on the group labels of the closest data points. Group labels are thus assigned based on majority vote. Euclidean distance is a commonly used distance metric for classification, but other metrics such as Manhattan distance, Minkowski distance, and Hamming distance can also be used. The value of K defines how many neighbors will be checked to determine the classification of a specific data point. Smaller values of K may have high variance, but low bias, while larger values of K may lead to high bias and lower variance. Cross-validation is typically applied to select an optimal value of K. KNN is a relatively easy algorithm to implement and has only a few hyperparameters to tune (value of K and distance metric). However, the method is prone to overfitting and does not work very well with high-dimensional data ($p \gg n$).

Logistic regression

A logistic regression model predicts the probabilities of a sample belonging to either of two groups for a set of metabolite peak intensities. If $P(A|X)$ and $P(B|X)$ are the probabilities of given sample belonging to group A and B respectively for an input data matrix X, then,

$$ln\frac{P(A|X)}{P(B|X)} = \beta_O + \beta_1 X_1 + \cdots \beta_p X_p$$

The response variable Y_i of the i-th sample is binary (0 or 1), corresponding to the two groups. Logistic regression also does not work well in a high-dimensional setting and typically requires a variable selection step preceding model fitting. A solution to this is a penalized logistic regression model[60], that has a built-in stepwise variable selection process. The tuning parameter is such a model it typically selected via cross-validation.

1.4.6 Biological interpretation

Statistical analysis of metabolomics data as described in the previous section result in a set of metabolites that are strongly associated with the disease or phenotype under study. The next step in the analysis pipeline is linking alterations in the levels of these metabolites to specific biological processes. This can be achieved by mapping and visualizing metabolites in the context of known biochemical pathways. Many bioinformatics tools exist that enable these analyses[61–64], several of which utilize the Functional Enrichment Testing (FET) approach, originally developed for gene expression data[65–67]. This helps to reduce data involving hundreds of altered genes or metabolites to smaller and more interpretable sets of altered biological ‘concepts’, helping generate testable hypotheses.

Functional Enrichment Testing can be broadly classified into two main types: (i) Over-representation Analysis (ORA), and (ii) Functional Class Scoring (FCS). Both these approaches have been directly borrowed from gene pathway analysis and are widely applied for metabolomics data.

1.4.6.1 Over-representation Analysis (ORA)

The goal of ORA is to gain insight into the underlying biological mechanisms and functional implications of a given set of metabolites. ORA performs a statistical test to assess whether the metabolite set is “enriched” with a specific annotation (e.g.: biological pathway) against a background set[68]. Briefly, the steps involved are as follows: (1) obtain a list of metabolites based on a separate statistical analysis (e.g.: t-test), (2) for each pathway, count the number of input metabolites that are part of that pathway, (3) repeat step 2 for a background set of metabolites (e.g.: all molecules which can be detected in the experiment), (4) assess when a pathway is over- or under-represented in the input set of metabolites[69]. The probability of observing at least k metabolites of interest in a pathway by chance is given by:

$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N is the size of background set, n denotes the number of metabolites of interest, M is the number of metabolites in the background set mapping to the i-th pathway, and k is the number of metabolites of interest which map to the i-th pathway[70].

Several statistical tests can be used to perform the analysis including chi-square, Fisher's exact test, binomial probability and hypergeometric distribution[71]. Pathways sets can be obtained from databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[72], BioCyc[73], Reactome[74], The Small Molecule Pathway Database (SMPDB)[75].

Despite its widespread application, ORA has certain limitations. First, because ORA does not consider any information regarding the extent of regulation of the input metabolites (e.g.: statistical significance, fold-change), it treats all metabolites equally, which is not always accurate. Second, ORA only considers significant metabolites that meet a certain threshold (e.g.: p-value < 0.05) and oftentimes, metabolites that are marginally less significant are missed, resulting in information loss. Third, ORA assumes that each metabolite is independent of the other. This is not always accurate and interactions among metabolites within and across pathways can manifest as change in expression levels. Similarly, ORA also assumes that pathways are independent of each other, which is inaccurate. Metabolic pathways are highly interconnected, and most metabolites are part of multiple pathways. The fact that these assumptions are not met leads to ORA often missing crucial topological differences of biological relevance.

1.4.6.2 Functional Class Scoring (FCS)

FCS, also known as Set Enrichment Analysis (SEA), hypothesizes that not just large changes in individual metabolite levels, but weaker coordinated changes in functionally related metabolites i.e., pathways, can also play an important role in biological mechanisms. First, a feature-level statistic (e.g.: t-test, ANOVA, correlation with phenotype) is computed for each metabolite. Unlike in ORA, all metabolites are taken into consideration without filtering them by a 'cut-off' Second, the feature-level statistic for all metabolites in a pathway is aggregated to compute a single pathway-level statistic. Examples of pathway-level statistics include the Kolmogorov-Smirnov statistic[76], the Wilcoxon rank sum[77], the maxmean statistic[78], or the sum, mean, or median of gene-level statistic[79]. Finally, the statistical significance of the pathway-level statistic is computed. FCS overcomes several limitations of ORA; however, it does still treat each pathway independently, which is rather inaccurate. Metabolites can be part of multiple pathways and there are clearly overlaps between pathways.

The concept of FCS has been routinely applied to gene expression data in the form of Gene Set Enrichment Analysis (GSEA)[76]. In its application to metabolomics data, it takes into consideration a quantitative measure associated with each metabolite (e.g., concentration, peak intensities). Metabolites are sorted by this quantitative measure and the consistency of each annotation/pathway is assessed in the top and bottom of the ranked link, compared to a background distribution.

Several computational tools exist that perform either ORA or FCS or both. These include, but are not limited to, Metabolite Set Enrichment Analysis (MSEA)[80], Metabolites Biological Role (MBRole)[81], Metabolite Pathway Enrichment Analysis (MPEA)[82], and Integrated Molecular Pathway Level Analysis (IMPaLA)[83].

1.4.6.3 Pathway Mapping and Visualization

One of the most intuitive approaches to interpreting metabolomics data is mapping the identified metabolites in the context of metabolic pathways or networks, usually obtained from reference metabolic pathway databases such as KEGG[72] or MetaCyc[84]. Such visualizations allow us to quickly explore the data in a biological context provided input metabolites are known. A myriad of software tools is available for pathway visualization that include various options for data input as well as statistical tests used for the comparison. These include MetaboAnalyst[39], 3Omics[61], Paintomics[63], PAPi[85], MetScape[64], MPEA[82], IMPaLa[83], PathVisio[86], MetaMapp[87], and MetExplore[88].

1.5 Network analysis of metabolomics data

Untargeted metabolomics offers an exciting avenue for detecting thousands of metabolic features simultaneously and identifying and characterizing novel metabolites. However, the large and complex nature of these datasets makes the biological interpretation challenging. Visualizing metabolites as connected entities is one approach to address these challenges. The connections between metabolites can be represented by informative relationships such as correlations, biochemical relatedness, or structural/chemical similarity. These connections can be formalized as networks where the nodes represent metabolic features, and the edges represent the context-dependent relationships between them.

Network analysis in metabolomics can be broadly classified into two main categories: knowledge-based and data-driven (**Figure 1.4**, obtained from Amara et al (2022)[89]).

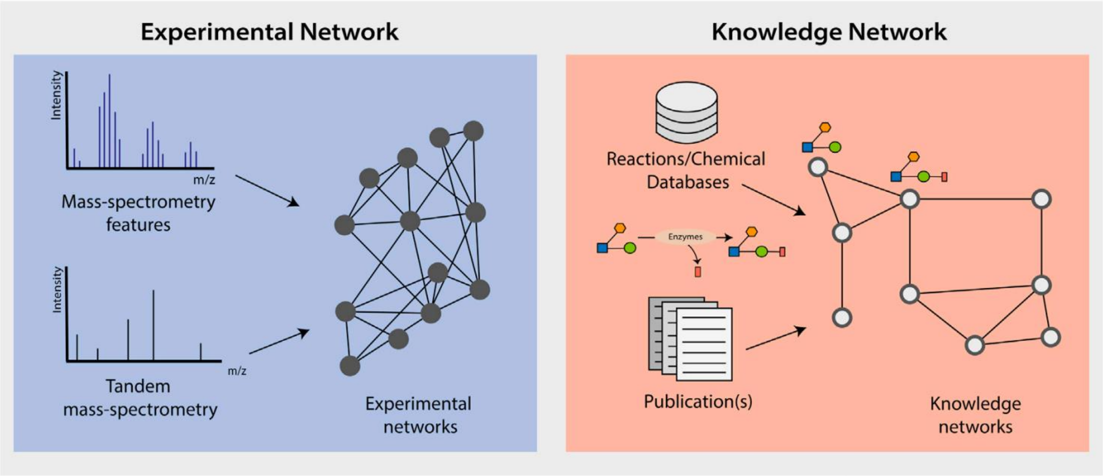


Figure 1.4: Two major types of networks generated and interpreted from metabolomics data.

Knowledge-based networks are generated based on prior knowledge of biological or biochemical relationships among the metabolites. Data-driven networks are generated directly from the experimental measurements based on relationships between metabolites in the data.

1.5.1 Knowledge-based networks

These networks provide a biological context to the interpretation and analysis of metabolomics data. A popular example of knowledge-based networks is Genome-Scale

Metabolic Networks (GSMNs). They utilize the existing knowledge of the metabolism of a particular organism captured in genome annotations and reaction databases. These networks are further refined by manual curation and exhaustive literature review. GSMNs encode information that map metabolites to reactions and the reactions to the corresponding genes and enzymes. For example, the human metabolic network Human 1 contains 13,417 reactions mapping to 4,164 metabolites[90]. GSMNs are very helpful in deriving directed and undirected graph representations of the system. Compound graphs (metabolites that are part of the same biochemical transformation are connected by an edge) and reaction graphs (a pair of reactions are connect if the product of one is the substrate of the other) are commonly generated graphs from GSMNs[91]. Graph representations of GSMNs make them amenable to graph-based analysis methods. For example, path searches have been used to infer metabolic pathways connecting metabolites of interest and for clustering and visualization of metabolomics data[92,93]. Additionally, centrality analysis has been applied to GSMNs as well to identify hub/driver metabolites in the network[94–96]. One of the biggest limitations to GSMNs is that they are heavily biased towards available genome annotations and knowledge of enzymatic reactions, making them relatively incomplete. While there have been several approaches developed to fill in these gaps in GSMNs[97,98], they are unlikely to capture all the metabolites identified in a metabolomics experiment.

Another example of knowledge-based networks are Chemical Ontology networks. These networks describe the structure of relationships among chemical compounds and thus provide a structured and formalized representation of chemical concepts. Here, connections between compounds do not represent a metabolic or biochemical relationship. Rather, they represent the relationship between compounds and broader chemical classes (eg: "fatty acids", "organic acids"). Chemical Ontology networks are therefore directed acyclic graphs due their hierarchical construction. Experimentally measured compounds typically constitute terminal nodes while the remaining nodes represent chemical classes, getting broader in scope as you go higher up the tree. These networks are also constructed via manual or semi-automated curation by domain experts. Such ontology networks are most useful in quantifying the relatedness between pairs of compounds based on their belonging to a shared chemical class. The ChEBI ontology[99] (60,329 compounds as of December 2022) and Gene Ontology (GO)[100] are commonly used ontology networks.

1.5.2 Data-driven networks

Data-driven networks are derived directly from the experimental measurements in an untargeted metabolomics data. Broadly, there can be different types of data-driven networks depending on the type of data used i.e., MS^1, MS^2, or MS^n and each of these deals with a different aspect of metabolic relationships. For example, in mass difference networks, the nodes represent metabolic features (m/z values), and edges represent mass differences that match a pre-defined biotransformation. These biotransformations can come from metabolic reaction databases such as KEGG[72], MetExplore[88], etc. and aim to identify potential biochemical reactions explaining the difference between m/z values. Similarly, adducts and feature networks capture mass differences between pairs of features that arise due to physicochemical transformations in the mass spectrometer i.e., non-biological mass differences. Spectral similarity networks depict the relationship between the MS^2 spectra of features based on certain spectral similarity measure such as cosine or modified cosine similarities[101]. Some of the most prominent tools for constructing spectra-based molecular networks include the Global Natural Product Social Molecular Networking (GNPS; http://gnps.ucsd.edu) community[102], the t-distributed stochastic neighbor embedding (t-SNE) algorithm-based software, MetGem[103], and the Python package Spec2Vec[104].

And finally, correlation networks represent the orchestrated or co-dependent changes in the abundance of metabolites i.e., metabolites that are associated within metabolic pathways tend to be correlated. All pairwise correlations between metabolites are encoded in a symmetric adjacency matrix. The adjacency matrix can be weighted (denoting the actual correlation coefficient values) or unweighted/binary, where two metabolites are linked only if their correlation value is higher than a particular threshold. The most computed correlation is Pearson's correlation coefficient. However, Pearson's correlation captures both direct and indirect associations between metabolites and the resulting networks are very dense, making them hard to analyze and interpret. Gaussian graphical models (GGMs) circumvent indirect associations by using partial correlations instead that capture only conditional dependencies, thereby resulting in more biologically meaningful networks[105]. However, computing a full partial correlation network requires that the sample size be at least as large as the number of features, which is rarely the case in untargeted metabolomics data. This limitation can be overcome by performing regularized estimation of the partial correlation networks[106]. Graphical lasso (Glasso)[107] and nodewise regression[108] are popularly used for performing regularized estimation. The Debiased Sparse Partial Correlation algorithm (DSPC)[109] builds partial correlation networks under the assumption that the number of true connections among the metabolites is much smaller than the available sample size, i.e., the true network is *sparse*. This allows the construction of partial correlation networks among a large number of metabolic features using fewer samples. The resulting network is thus a weighted one where nodes represent metabolites and edges represent partial correlation coefficients or the associated P-values[109].

Another popularly used tool for constructing correlation networks is the WGCNA (Weighted Gene Co-expression Network Analysis) R package[110,111]. WGCNA builds correlation networks based on the assumption that the underlying network has a scale-free topology. This is achieved by weighing the correlation coefficients by an exponent such that the degree distribution of the network follows a power-law. The final network is represented by the Topological Overlap Matrix (TOM) that represents the similarity between a pair of nodes based on number of shared neighbors by incorporating both direct and indirect relationships. The network is then clustered via hierarchical clustering to obtain modules. These modules can then be utilized to identify “hub”/driver genes within each module (“module eigengene”), perform association analysis with other clinical traits of interest, or perform pathway enrichment analysis of the genes in a specific module. While WGCNA was originally developed for transcriptomics data, it is being increasingly applied to study correlation networks from metabolomics data[112–116].

Data-driven correlation networks are therefore extremely useful in identifying novel perturbations in the system without the reliance on a priori knowledge of the relationships between the metabolites or pathway information, thus circumventing a multitude of challenges posed by knowledge-driven data analysis methodologies in metabolomics.

1.6 Dissertation outline

The metabolome provides a readout of the cellular and biochemical events that reflect the genetic and epigenetic makeup of an organism, as well as the microbiome and environmental exposures. One of the most common experimental designs in metabolomics aims to assess the differences in the metabolic repertoire under different conditions or disease states. Conventional data analysis approaches for these types of experiments involve some form of univariate analysis, followed by mapping of differentially abundant metabolites to the known biochemical pathways. However, univariate analysis methods ignore the interactions and associations between metabolites that may be characteristic of the phenotype(s) of interest, while mapping differentiating metabolites onto known pathways does not provide the ability to identify novel rewiring of pathways leading to metabolic dysregulation. Moreover, certain classes of compounds (eg: lipids) tend to be poorly represented in pathway databases, which limits the scope of application of this approach.

The goal of this dissertation is to begin addressing some of these limitations of knowledge-based enrichment analysis by employing data-driven network analysis approaches for metabolomics data. In Chapter Two, we introduce a Java-based bioinformatics tool, *Filigree*, that implements and extends the Differential Network Enrichment Analysis (DNEA) algorithm. *Filigree* recovers robust partial correlation networks from the input data even with limited sample size and highly imbalanced experimental group design. We tested *Filigree* with three metabolomics datasets pertaining to metabolic disorders/conditions and identified metabolic modules relevant to the condition(s) and associated with external traits. Such hypothesis-generating analyses can therefore aid in gaining deeper biochemical understanding from the data. In Chapter Three, I compared the plasma metabolome of COVID-19 patients with either mild or severe disease to that of healthy controls. My analysis revealed a strong association between the metabolic profiles and clinical traits. I identified metabolites that differentiated healthy controls and COVID-19 patients. I also identified several metabolites that differentiated mild and severe COVID-19. These metabolites performed much better than clinical characteristics (“risk factors”) in discriminating mild and severed COVID-19 groups. In Chapter Four, I assessed the association of data-driven metabolic modules with change in BMI over 5- and 10-years preceding diagnosis in ALS (Amyotrophic Lateral Sclerosis) patients. I identified eight metabolic modules (152 metabolites in total) that showed a strong association with BMI trajectory. A subset of these metabolites was also individually associated with ALS survival, suggesting a possible metabolic link between change in BMI over time and survival in ALS patients.

CHAPTER II

Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations

This chapter has been published as: **Iyer, G. R.**, Wigginton, J., Duren, W., LaBarre, J. L., Brandenburg, M., Burant, C., Michailidis G. & Karnovsky, A. (2020). "Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations." *Metabolites*. This chapter also includes the description of the DNEA R package and details my contributions into this work.

2.1 Introduction

The metabolome provides a readout of the underlying cellular and biochemical events that reflect individual genetic makeup[117], epigenetics[118], the microbiome[119], and environmental exposures, including diet[120,121]. Metabolic profiling has been successfully applied to biomarker discovery and the assessment of disease risk and progression in cancer[25,122], cardiovascular[13,123] and renal diseases[14,15], and type 1 (T1D)[17,18] and type 2 diabetes (T2D)[19,20].

Metabolism is interconnected through several major metabolic hubs, e.g., glucose-6-phosphate, pyruvate, acetyl-CoA, and malonyl-CoA. Beyond these central nodes, metabolic pathways have secondary rate-limiting steps that are often controlled by metabolites affecting multiple pathways (such as AMP, citrate, NAD, etc.), as well as by post-translational modifications of proteins regulating the pathway. Evaluating changes in the connectivity of the metabolome could help to understand how these pathways are affected in physiological and disease states.

Experimental design in metabolomics commonly involves assessment of metabolite levels in two or more disease conditions or experimental groups. Metabolomics data acquired from such experiments are amenable to univariate analysis, followed by pathway mapping and enrichment analysis. Enrichment analysis, originally developed for gene expression data, reduces data involving hundreds of altered genes or metabolites to smaller and more interpretable sets of altered biological 'concepts', helping generate testable hypotheses. The most common types of enrichment analysis are variants of over-representation analysis (ORA) or set enrichment analysis (SEA)[76]. In both cases, statistical tests are performed to assess the enrichment or depletion of a set of metabolites in a specific pathway against a background or reference set[69].

Several bioinformatics tools implementing the above data analysis workflow for metabolomics have been developed[68,124]. While overall this approach has proven to be extremely useful, each of the individual methods involved has limitations. First, univariate analysis considers only individual metabolites and does not account for the interactions between them. Indeed, biological constraints on metabolism result in many metabolites being highly correlated in biological samples (for instance, branched chain amino acids). Second, the application of metabolite pathway mapping and enrichment analysis is hampered by the low coverage of experimentally determined metabolites in biological pathway databases[125]. This is particularly true for lipids and secondary metabolites. The low coverage can in part be explained by the differences between chemistry-centric metabolomics experiments and genome-centric pathway databases. This problem is further compounded by the relatively small number of known metabolites measured in most experiments which limits both the statistical significance and overall reliability of analyses.

We present a user-friendly tool, *Filigree*, that overcomes many of the limitations of existing methods. *Filigree* implements our recently published differential network enrichment analysis (DNEA) method[126]. DNEA provides an alternative to traditional pathway-centric approaches by leveraging the underlying structure of the data and inferring associations among metabolites directly from experimental measurements. These associations can be quantified by partial correlations that measure the conditional dependence between metabolites, thus allowing elimination of spurious, non-informative associations. In lieu of predefined pathways, DNEA generates stable subnetworks comprised of biochemically and structurally related metabolites. It accounts for both changes in network structure and the differential abundance of metabolites when assessing significance of subnetworks, thus providing a systems level view of the data. To demonstrate the utility of *Filigree*, we applied it to previously published studies assessing the metabolome in the context of metabolic disorders (T1D and T2D) and the maternal and infant lipidome during pregnancy. *Filigree* is freely available at http://metscape.ncibi.org/*Filigree*.html.

2.2 Methods

2.2.1 Filigree application

The input to the tool is a plain text file containing per-sample unadjusted intensity values and group information. The output consists of three. csv files: (1) an 'edgelist' containing metabolite pairs and partial correlation values between them; (2) a 'nodelist' containing information about the differential status of each metabolite, along with its statistical significance and subnetwork membership, and (3) a NetGSA results file containing information about subnetworks, including number of edges/nodes and statistical significance of each subnetwork. These files can be easily imported into network visualization software such as Cytoscape for further exploration[127]. Additionally, the user can browse the interactive HTML files automatically generated by *Filigree*.

2.2.2 Extensions of DNEA Methodology

The DNEA method works particularly well both theoretically[128] and empirically[126] when group sizes are fairly balanced, and the number of metabolites is a low multiple of the sample size. However, in many applications the two groups under consideration may be grossly imbalanced or the number of samples severely limited. To that end, we developed several extensions to the DNEA methodology (described below) that improve its versatility, including (i) feature aggregation and (ii) group subsampling to attain more balanced sample sizes across groups.

2.2.2.1 Feature aggregation

Since network density and stability are strongly dependent on the ratio of features (metabolites and/or lipids) to samples[128], a preprocessing step to aggregate highly similar or redundant features may be appropriate. This step helps reduce the dimensionality of the data to promote the retrieval of more interpretable PCNs and is therefore highly recommended for datasets where the number of features is a high multiple of the number of samples.

We implemented an optional data preprocessing step for aggregation of highly similar or redundant features in the dataset in order to recover more stable PCNs. Feature aggregation performs optimally when data are log-transformed, but not auto-scaled. Several types of aggregation are possible: (1) a purely data-driven approach that collapses features with highly similar (Pearson) correlation profiles into singular features, (2) a purely knowledge-driven method that collapses chemically similar metabolites/lipids, or (3) a hybrid feature aggregation that collapses only features identified as chemically similar that also share a highly similar Pearson correlation profile. For options (2) or (3), the user may provide their own knowledge-based feature grouping file or can utilize the grouping file based on chemical similarities found in KEGG[72], HMDB[129] or LipidBlast[130]. For options (1) or (3), the user has the choice to view the features-to-sample-size ratio at various feature-aggregation tolerance values based on the correlation structure of the data. The user can then decide the extent of feature aggregation they wish to perform or can proceed with the recommended values. The output of this stage is a new data matrix where metabolites/lipids belonging to the same feature group are represented as singular features by computing their median intensity across all samples. The format of the new data matrix will be identical to that of the original input matrix.

2.2.2.2 Group subsampling

Highly imbalanced sample group sizes can result in PCNs where the smaller group is much sparser than the larger group, thus hindering interpretability of results. To address this issue, we modified the algorithm by using subsampling to create more balanced sample groups, leading to more stable and interpretable PCNs. The modified procedure is comprised of the following steps:

1) Determine size of smaller group (n_{min});
2) At every iteration of the stability selection (default value set to 500 iterations), create new data matrices for the two groups as follows:
    a) For the larger group, randomly sample $\alpha \times n_{min}$ samples without replacement.
    b) For the smaller group, randomly sample $\beta \times n_{min}$ samples without replacement. Additionally, in order to maintain some degree of randomness in the smaller group, $(1 - \beta) \times n_{min}$ samples are randomly chosen from this and added back.
3) Fit the training model for the new subsamples of the data at every iteration;
4) Obtain edge selection probabilities and retain edges with a selection probability of $> \tau$;
5) Use the selection probabilities as weights when estimating the partial correlation networks. Based on extensive experimentation, we recommend α = 1.3, β = 0.9 and τ = 0.9, but the practitioner can also experiment with other values.

2.2.3 Datasets

2.2.3.1 Mouse Model of T1D

Previous studies[131,132] have generated and examined GC-MS metabolomics data from non-obese diabetic (NOD) mice, some of which progressed to overt T1D (chronic hyperglycemia) while others avoided progression (normoglycemia). Metabolomics data containing 163 named metabolites from 71 mice (30 diabetic and 41 non-diabetic) were downloaded from the Metabolomics Workbench (Study ST000057). Age- and sex-adjusted data[132] were log-transformed and autoscaled to have zero mean and unit variance.

2.2.3.2 Framingham Heart Study (FHS) Offspring Cohort

The FHS Offspring Cohort is a longitudinal, community-based cohort that includes 3799 participants, aged 40–65 years, at the fifth quadrennial examination cycle 1991–1995 (baseline for our purposes)[133]. We downloaded plasma metabolite profiles (LC-MS/MS) for 956 subjects at baseline from the dbGaP database (https://www.ncbi.nlm.nih.gov/gap/). Approximately 10 years after the metabolomics analyses (2001–2005), subjects were re-recruited to be assessed for development of T2D, determined based on the following criteria: (1) fasting glucose $\geq$ 7 mmol/L, (2) 2-h glucose $\geq$ 11 mmol/L, and (3) consumption of oral hypoglycemics or insulin[134]. 674 subjects remained healthy while 100 subjects developed T2D (182 subjects had missing data in at least one of the variables). Age- and sex-adjusted data were log-transformed and autoscaled to have zero mean and unit variance.

2.2.3.3 Michigan Mother-Infant Pairs (MMIP) Cohort

The Michigan Mother-Infant Pairs (MMIP) cohort[135] evaluated the plasma lipidome in 106 pregnant women during the first trimester (M1), at the time of delivery (M3), and within infant umbilical cord blood (CB). Comprehensive lipidomics profiling identified 670 lipid species from 17 different classes. *Filigree* was used to perform pairwise analyses between: (i) M1 vs. M3; (ii) M1 vs. CB, and (iii) M3 vs. CB, classifying differences in the connectivity of subnetworks between time points. We used the feature aggregation functionality of *Filigree* to collapse highly correlated and chemically similar lipids into singular features, making the feature space comparable to the sample size.

2.2.4 Group Lasso Regression

Filigree subnetworks generated from pairwise comparisons (M1 vs. CB, M3 vs. CB and M1 vs. M3) were tested for their association with infant birth weight (BW) at individual time points (M1, M3 and CB) in a group lasso regression[136] model using the R package gglasso[137]. Group lasso is an extension of the traditional lasso regression methodology[138] that incorporates prior information about the grouping of variables. In contrast to lasso regression, variable selection is performed on an entire set of variables (or predictors) instead of individual variables. Let y be a vector of length N and X be an $N \times p$ matrix of features. Let the p features (or predictors) be divided into L groups such that there are p_l

predictors in group l. The matrix X_l therefore represents predictors from the l^{th} group with a coefficient vector β_l. β estimates are obtained by solving the optimization problem,

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \|\beta_l\|_2 \right)$$

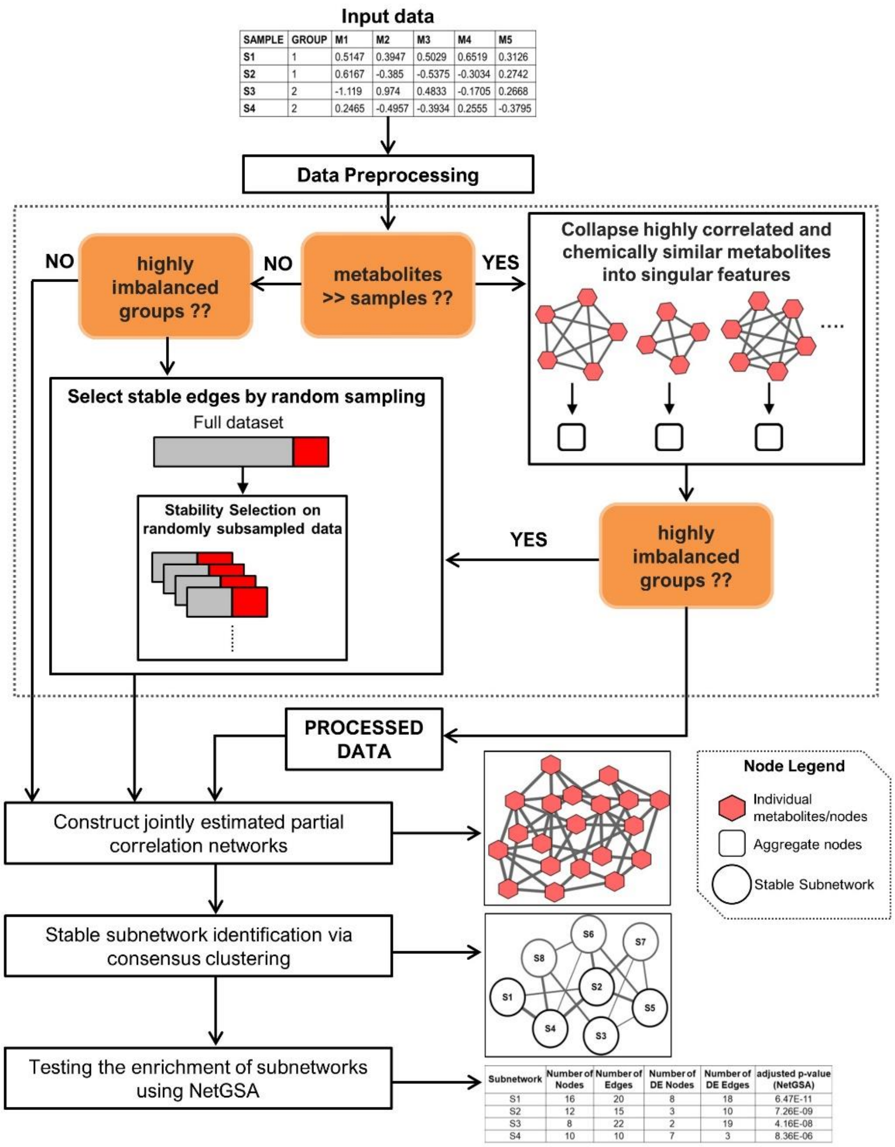
Here, $\|\cdot\|_2$ denotes the Euclidean norm and λ is the tuning parameter that controls the sparsity of the coefficients at the group level. It should be noted that this computation does not provide within-group sparsity, i.e., the coefficients of all the predictors in a group are either zero or non-zero. A range of 100 linearly increasing λ values is used (default), generated as a fraction of λ_{max}, the smallest λ value for which all the coefficients are zero. The strength of association between a group of predictors and the response variable is determined by the λ value corresponding to the entry of that group into the regression equation, with higher λ value corresponding to a stronger association. The group lasso model was run for 500 iterations (stability selection) for robustness. The statistical significance of subnetworks obtained from NetGSA was not taken into account while performing group lasso regression.

2.2.5 Data and Resource Availability

Mouse T1D metabolomics data analyzed during the current study are available in the Metabolomics Workbench repository (Study ST000057). Metabolomics data from the Framingham Heart Study Offspring Cohort analyzed during the current study are available on dbGaP (https://www.ncbi.nlm.nih.gov/gap/) with the study accession number phs000007.v29.p10 and dataset phenotypic identifiers 'pht002234.v5.p10:' (Metabolomics-HILIC), 'pht002894.v1.p10:' (Central Metabolomics-HILIC), 'pht002343.v4.p10:' (Metabolomics-Lipid Platform). Lipidomics data from the Michigan Mother-Infant Pairs Cohort (MMIP) analyzed during the current study are available in the Metabolomics Workbench repository (Project ID PR000386). *Filigree* is freely available at http://metscape.ncibi.org/*Filigree*.html. Scripts associated with the current analyses are available at https://github.com/griyer/Diabetes_manuscript_code.git.

2.3 Results and Discussion

The DNEA method[126] implemented in *Filigree* includes three main steps: (1) joint estimation of the partial correlation network (PCN) across two groups of samples, (2) unsupervised clustering of the resulting PCN using consensus clustering to obtain densely connected subnetworks, and 3) testing the subnetworks for enrichment using the NetGSA algorithm[139,140]. As mentioned in[126], the groups can correspond to treatment-control conditions, disease subtypes, etc. Further details of the DNEA algorithm are described in Supplementary Methods. **Figure 2.1** depicts our analysis pipeline and describes the *Filigree*/DNEA workflow.



Input data

SAMPLE	GROUP	M1	M2	M3	M4	M5
S1	1	0.5147	0.3947	0.5029	0.6519	0.3126
S2	1	0.6167	-0.385	-0.5375	-0.3034	0.2742
S3	2	-1.119	0.974	0.4833	-0.1705	0.2668
S4	2	0.2465	-0.4957	-0.3934	0.2555	-0.3795

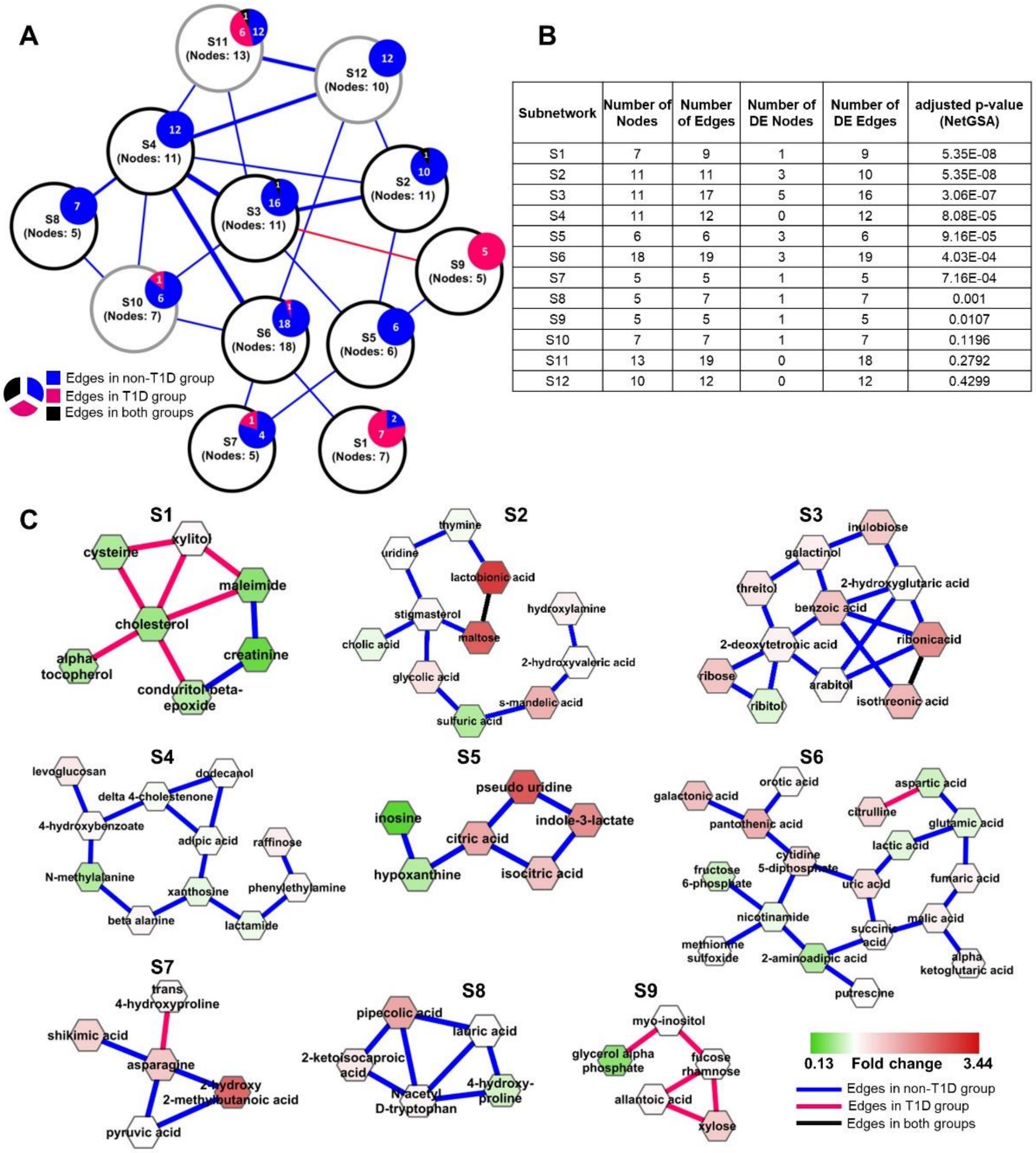
Subnetwork	Number of Nodes	Number of Edges	Number of DE Nodes	Number of DE Edges	adjusted p-value (NetGSA)
S1	16	20	8	18	6.47E-11
S2	12	15	3	10	7.26E-09
S3	8	22	2	19	4.16E-08
S4	10	10	7	3	8.36E-06

Data Preprocessing
NO
highly imbalanced groups ??
NO
metabolites >> samples ??
YES
Collapse highly correlated and chemically similar metabolites into singular features
....
Select stable edges by random sampling
Full dataset
Stability Selection on randomly subsampled data
YES
highly imbalanced groups ??
PROCESSED DATA
Node Legend
Individual metabolites/nodes
Aggregate nodes
Stable Subnetwork
Construct jointly estimated partial correlation networks
Stable subnetwork identification via consensus clustering
S6
S7
S8
S2
S1
S5
S4
S3
Testing the enrichment of subnetworks using NetGSA

Figure 2.1: Schematic representation the data analysis pipeline.

2.3.1 DNEA Analysis Reveals Dysregulation of Metabolite Networks in T1D vs. Non-Diabetic Mice

We utilized *Filigree* to perform DNEA analysis of the metabolomics data from NOD mice that either progressed or did not progress to overt T1D[131,132]. Plasma metabolites from T1D and non-diabetic NOD mice produced a PCN with stronger connectivity in the non-diabetic mice (**Figure 2.2A**). The subsequent analysis steps identified twelve stable subnetworks within the resulting PCN (**Figure 2.3**). Nine of the these were significantly differential between T1D and non-diabetic mice (FDR < 0.05) (**Figure 2.2B, 2.2C**).



Subnetwork	Number of Nodes	Number of Edges	Number of DE Nodes	Number of DE Edges	adjusted p-value (NetGSA)
S1	7	9	1	9	5.35E-08
S2	11	11	3	10	5.35E-08
S3	11	17	5	16	3.06E-07
S4	11	12	0	12	8.08E-05
S5	6	6	3	6	9.16E-05
S6	18	19	3	19	4.03E-04
S7	5	5	1	5	7.16E-04
S8	5	7	1	7	0.001
S9	5	5	1	5	0.0107
S10	7	7	1	7	0.1196
S11	13	19	0	18	0.2792
S12	10	12	0	12	0.4299

Figure 2.2: (A) Overview of T1D mouse model Filigree network showing associations between all the subnetworks. Each node represents a subnetwork with the overlaying pie charts showing the distribution of the intra-subnetwork edges. Inter-subnetwork edges are weighted by the total number of edges. Nodes with black outline are significantly differential by NetGSA **(B)** NetGSA output from Filigree showing subnetwork information and statistics. **(C)** Significantly differential subnetworks. Nodes are colored based on fold change (T1D over non-T1D).

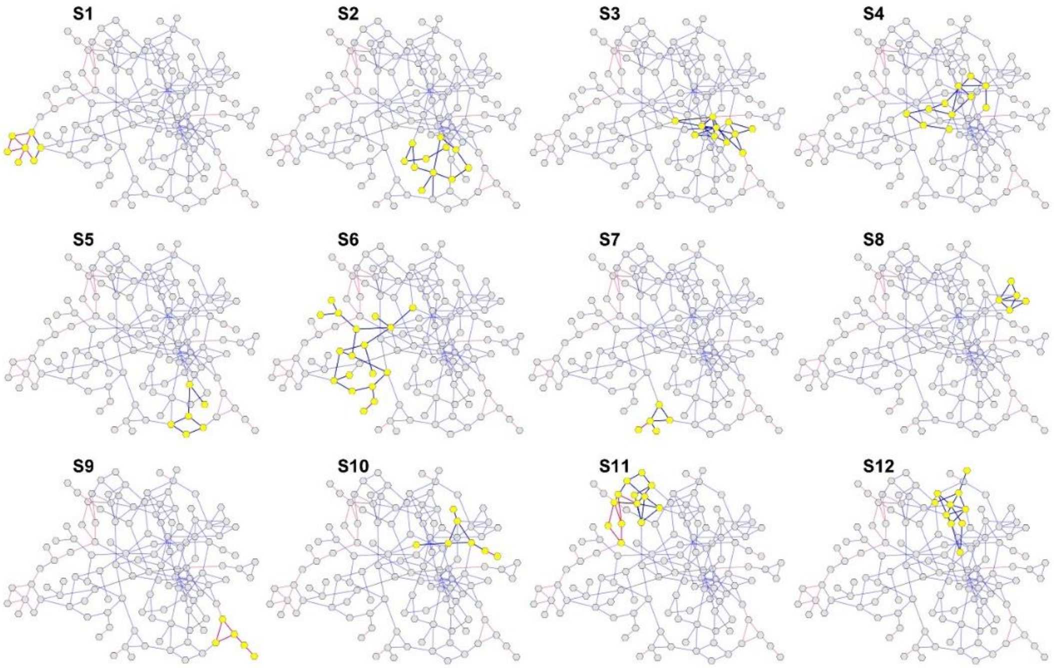


Figure 2.3: Filigree Partial Correlation Networks from T1D mouse model data highlighting all subnetworks.

Seven out of nine differential subnetworks contained edges present in non-diabetic mice that were disrupted in diabetic animals. Four out of these, S2, S3, S4, and S6, are highly interconnected. These subnetworks contain nucleobases, ribose and its reduction products, nucleic acids, amino acids, and also several sugars and sugar-related metabolites. (**Table 2.1**). We note that several edges connecting metabolites in these subnetworks represent oxidation/reduction reactions. For instance, galactinol, a sugar alcohol, is the reduction product of galactose and ribitol is a reduction product of ribose. This suggests that the connectivity between metabolites in these subnetworks is disrupted due to changes in redox potential that accompany the progression to T1D. Thus, a general decrease in the redox state of cells may contribute to the changes in the connectivity of metabolites seen in the plasma in T1D.

Table 2.1: Pathway information for the subnetworks in the T1D mouse model metabolomics data

Subnetwork	Pathway
Subnetwork 1	
cholesterol	Steroid biosynthesis; Primary bile acid biosynthesis
creatinine	Arginine and proline metabolism
cysteine	Cysteine and methionine metabolism; Glutathione metabolism; Aminoacyl-tRNA biosynthesis
xylitol	Pentose and glucuronate interconversions
Subnetwork 2	
2-hydroxyvaleric acid	Fatty acid degradation
cholic acid	Primary bile acid biosynthesis
sulfuric acid	Sulfur metabolism
thymine, uridine, sulfuric acid	Nucleotide metabolism
glycolic acid	Glyoxylate and dicarboxylate metabolism
hydroxylamine	Nitrogen metabolism
maltose	Carbohydrate metabolism
Stigmasterol	Steroid biosynthesis
Subnetwork 3	
2-hydroxyglutaric acid	Butanoate metabolism
arabitol, ribitol	Pentose and glucuronate interconversions
benzoic acid	Benzoate degradation
galactinol	Galactose metabolism/Carbohydrate metabolism
isothreonic acid	Ascorbate and aldarate metabolism
ribose	Pentose phosphate pathway
Subnetwork 4	
phenylethylamine	Phenylalanine metabolism
raffinose	Galactose metabolism/Carbohydrate metabolism
adipic acid	Caprolactam degradation
4-hydroxybenzoate	Benzoate degradation
delta-4-cholestenone	Steroid degradation
xanthosine, beta alanine	Nucleotide metabolism
levoglucosan	Carbohydrate metabolism (?)
Subnetwork 5	
citric acid, isocitric acid	Citrate cycle (TCA cycle); Glyoxylate and dicarboxylate metabolism
hypoxanthine, inosine, pseudouridine	Nucleotide metabolism
indole-3-lactate	Tryptophan metabolism
Subnetwork 6	
alpha ketoglutaric acid, fumaric acid, malic acid, succinic acid	Citrate cycle (TCA cycle)
cytidine-5-diphosphate, orotic acid, uric acid	Nucleotide metabolism
aspartic acid, citrulline, glutamic acid, putrescine	Arginine biosynthesis/Arginine Proline metabolism
fructose-6-phosphate, galactonic acid, lactic acid	Carbohydrate metabolism
2-aminoadipic acid	Lysine metabolism
nicotinamide, pantothenic acid	Vitamin metabolism
methionine sulfoxide	Cysteine and methionine metabolism
Subnetwork 7	
shikimic acid	Phenylalanine, tyrosine and tryptophan biosynthesis
trans-4-hydroxyproline	Arginine and proline metabolism

aspargine, pyruvic acid	Alanine, aspartate and glutamate metabolism
Subnetwork 8	
4-hydroxyproline	Arginine and proline metabolism
lauric acid	Fatty acid biosynthesis
N-acetyl D-tryptophan	Tryptophan metabolism
pipecolic acid	Lysine metabolism
2-ketoisocaproic acid	Valine, leucine and isoleucine metabolism
Subnetwork 9	
glycerol-alpha-phosphate	Glycerolipid metabolism
allantoic acid	Purine metabolism
fucose rhamnose	Fructose and mannose metabolism
xylose	Pentose and glucuronate interconversions
myo-inositol	Galactose metabolism; Ascorbate and aldarate metabolism; Inositol phosphate metabolism
Subnetwork 10	
serine; homoserine	Glycine, serine and threonine metabolism; Cysteine and methionine metabolism
capric acid	Fatty acid biosynthesis
4-hydroxybutyric acid	Butanoate metabolism
glutaric acid	Fatty acid degradation
Subnetwork 11	
alanine; cystine; glutamine; lysine; methionine; ornithine; oxoproline; proline; threonine; tryptophan; tyrosine	Amino acid(s) metabolism
fructose	Carbohydrate metabolism
Subnetwork 12	
palmitic acid; stearic acid; arachidonic acid; myristic acid; methylhexadecanoic acid	Fatty acid metabolism
glycerol	Glycerolipid metabolism
trehalose; tagatose	Carbohydrate metabolism

The association between the cellular redox state and the metabolome is further supported by S1 and S9, which contain predominantly diabetic edges (**Figure 2.2C**). In both subnetworks, the enrichment is driven primarily by the differential edges, while most metabolites (nodes) are not significantly differentially expressed and therefore would not be prioritized by univariate analysis (**Figure 2.2B**).

S1 consists of metabolites either directly or indirectly related to increased oxidative stress. Oxidative stress is a widely accepted complication accompanying the pathogenesis of diabetes by way of increased free radical (ROS) concentrations caused by hyperglycemia as well as decreased levels of major antioxidants such as glutathione[141], leading to significant damage to pancreatic islet beta cells responsible for insulin secretion[142]. Glutathione (gamma-glutamyl-cysteinyl-glycine) is a highly abundant tripeptide in the human body known to play a vital role in defense against oxidative stress as a free radical scavenger[143]. The bulk of the blood glutathione is found within erythrocytes (millimolar concentrations) while levels in the plasma tend to be in the micromolar range. Diminished levels of blood glutathione have been implicated both in T1D and in T2D[144–147]. While glutathione was not measured in this experiment, we speculate that reduced level of this metabolite can influence the levels of several S1 metabolites, including cysteine, cholesterol, creatinine, and xylitol. Cysteine, one of the three amino acid constituents of glutathione, is present in this subnetwork with lower levels in diabetic mice. It has been postulated that reduced levels of glutathione in type 1 diabetes is a consequence of increased utilization rather than decreased synthesis, thus resulting in reduced levels of cysteine[146]. A hub node of S1 is cholesterol. Counterintuitively, we see decreased levels of cholesterol in diabetic mice. This is likely due to the inhibitory effect of diminished glutathione on the enzyme HMG-CoA reductase, the rate-controlling enzyme in the cholesterol synthesis pathway (Malveonate pathway). Glutathione has been suggested to be one of the key activators of HMG-CoA reductase by maintaining the enzyme in its active, reduced sulfahydryl state[148–151]. Moreover, insulin has also been shown to be an activator of HMG-CoA reductase in a mechanism similar to glutathione[152]. Depleted glutathione also has an inhibitory effect on the enzyme creatine kinase (CK), responsible for the phosphorylation of creatine to phoshpocreatine, likely due to thiol oxidation of the sulfahydryl groups of the enzyme[153,154]. A reduction in CK activity leads to a decrease in phosphocreatine levels which further causes a decrease in creatinine levels, a product of phosphocreatine utilization. Consequently, we observe creatinine in subnetwork S1 at lower levels in diabetic mice. Additionally, xylitol, a five-carbon sugar alcohol and widely used sugar-substitute, has also been shown to serve as a glutathione-reducing compound *in vitro* and *in vivo*[155,156]. While we did not see a significant difference in the levels of xylitol between diabetic and non-diabetic mice, its potential association with glutathione is a possible reason for its presence in subnetwork S1. Finally, we see alpha-tocopherol (Vitamin E) in subnetwork S1. This is not unexpected as alpha tocopherol is a well-known potent antioxidant, similar to glutathione. It is therefore not surprising that we see lower levels of alpha-tocopherol in diabetic mice.

Several S1 metabolites are exogenous compounds often measured in plasma and urine. In general, these compounds are decreased in T1D mice and also have differential connectivity, suggesting that their metabolism is disrupted in T1D. Alternatively, exogenous compounds may not be easily absorbed in the intestine in T1D, potentially due to altered intestinal permeability. In T1D, there are marked changes in the intestinal morphology and expression of transporters[157] and increased intestinal permeability[158], altering the entry of exogenous substances with additional effects on cellular metabolism. These findings also support previously described disruptions in metabolism associated with T1D, including alterations in mitochondrial metabolism, increased oxidative stress, and changes in redox state[159]. Indeed, Fahrmann et al.[131] previously reported increased levels of sugar-related metabolites, branched chain amino acids, gluconic acid and nitric oxide-derived saccharic acid markers of oxidative stress in T1D mice. Our network-based approach confirms and extends the understanding of alteration in metabolism that occurs in T1D, including changes in the metabolism of nucleotides (S2–S5). Because these alterations are found in plasma, the tissue-specific origins of disruption in metabolism cannot be precisely localized.

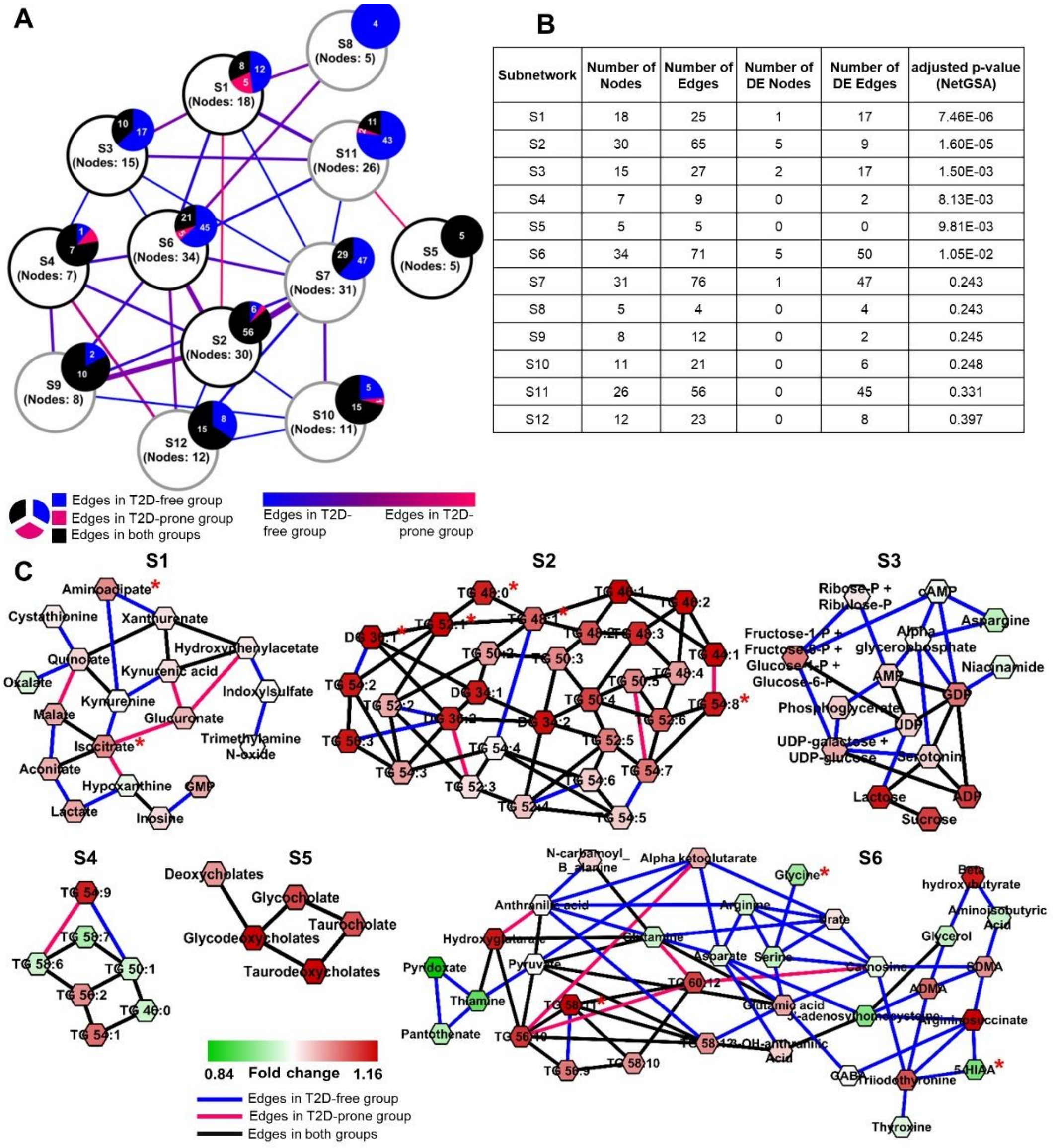
2.3.2 Connectivity of Metabolite Networks Differs between Non-Diabetics and Individuals Who Later Developed T2D from the Framingham Heart Study (FHS) Offspring Cohort

The FHS Offspring Cohort has been studied extensively and biomarkers for risk of cardiovascular disease and T2D have been identified[19,160]. We used DNEA to examine metabolomics data from 100 FHS subjects who developed T2D over the course of the subsequent twenty years (T2D-prone) and 674 subjects who remained non-diabetic (T2D-free). This highly imbalanced group distribution makes it difficult to recover robust and stable PCNs[128]. Statistical theory[161] suggests that subsampling approaches can reduce the bias towards the group with higher number of samples. We created a subsampling approach that allows a stable network topology to be obtained and reduces the number of edges in the non-diabetic group (described in *Methods*). The number of edges recovered with and without subsampling, within each group, is reported in **Table 2.2**.

Table 2.2: Number of edges discovered with and without subsampling the Framingham Heart Study Offspring Cohort T2D data

	Number of Edges		
	Non-diabetic	**Diabetic**	**Common**
Without subsampling	784	73	250
With subsampling	281	36	223

Our analysis identified substantial network differences between T2D-prone and T2D-free groups (**Figure 2.4A**). The algorithm identified twelve stable subnetworks (**Figure 2.5**) within the resulting PCN, with six subnetworks significantly differing between T2D-prone and T2D-free groups (FDR < 0.05) (**Figure 2.4B, 2.4C**). Similar to our findings in T1D, there were fewer edges in T2D-prone compared to T2D-free networks. This tendency is especially apparent in subnetworks S1, S3, and S6 (**Figure 2.4B**).



Subnetwork	Number of Nodes	Number of Edges	Number of DE Nodes	Number of DE Edges	adjusted p-value (NetGSA)
S1	18	25	1	17	7.46E-06
S2	30	65	5	9	1.60E-05
S3	15	27	2	17	1.50E-03
S4	7	9	0	2	8.13E-03
S5	5	5	0	0	9.81E-03
S6	34	71	5	50	1.05E-02
S7	31	76	1	47	0.243
S8	5	4	0	4	0.243
S9	8	12	0	2	0.245
S10	11	21	0	6	0.248
S11	26	56	0	45	0.331
S12	12	23	0	8	0.397

Figure 2.4: (A) Overview of the Framingham Heart Study Offspring Cohort T2D network showing associations between all the subnetworks. Each node represents a subnetwork with the overlaying pie charts showing the distribution of the intra-subnetwork edges. Inter-subnetwork edges are weighted by the total number of edges. Nodes with black outline are significantly differential by NetGSA. **(B)** NetGSA output showing subnetwork information and statistics. **(C)** Significantly differential subnetworks. Nodes are colored based on fold change (T2D-prone over T2D-free). Nodes marked with red asterisk (*) have been reported as T2D predictors by Merino and colleagues (2018).

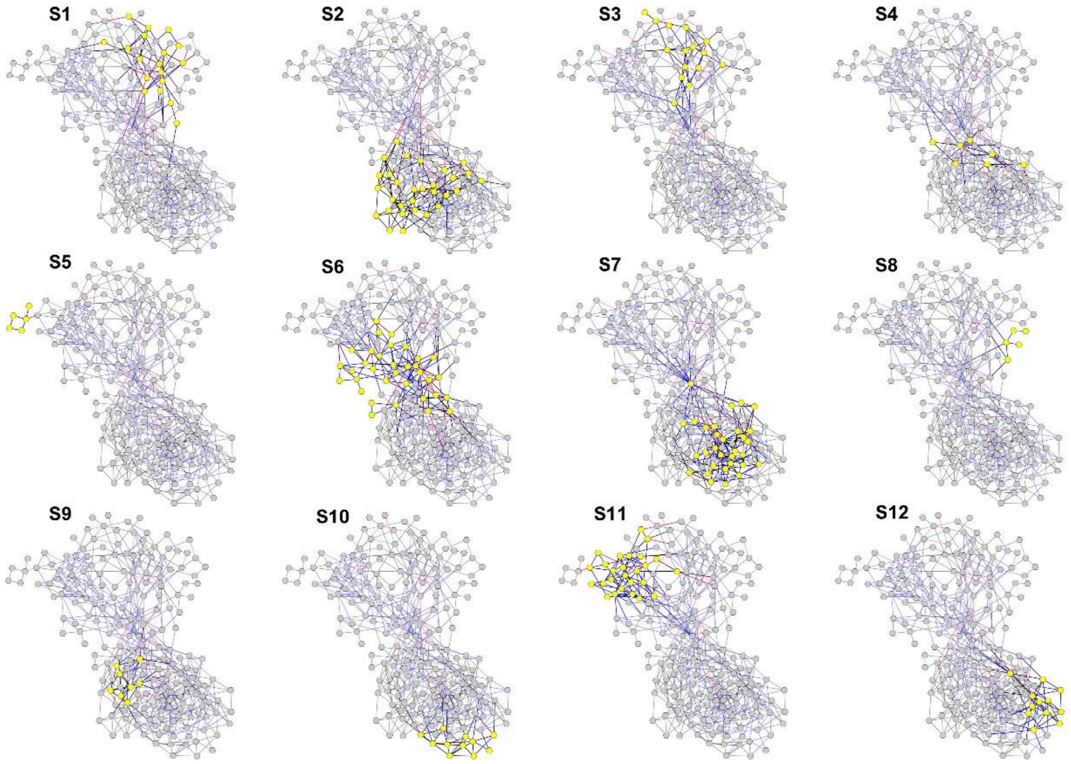


Figure 2.5: Filigree Partial Correlation Networks from Framingham Heart Study Offspring Cohort T2D data highlighting all subnetworks.

The most significant subnetwork (S1) includes intermediates of tryptophan, cysteine, lysine, tyrosine, and phenylalanine metabolism (**Table 2.2**; **Figure 2.6**). Dysregulation of tryptophan metabolism[162,163] and elevated level of 2-amnionadipic acid have been associated with the development of T2D[164]. Previous studies in the FHS Offspring Cohort found that branched chain and aromatic amino acids were positively associated with the risk of developing T2D[19]. The subnetwork containing branched chain amino acids (S11) is not significantly differential between groups (**Figure 2.7**), consistent with the findings of Merino and colleagues[134] who found that branched chain amino acids (BCAAs) were not predictive of T2D in this sample cohort, perhaps due to the relatively small differences in insulin resistance between the T2D-prone and T2D-free individuals in these data. Subnetwork S1 also includes several intermediates of purine metabolism (**Figure 2.6**). Increased levels of uric acid, the end-product of purine metabolism, is a common finding in obese T2D patients and has been implicated in the pathogenesis of metabolic syndrome disorders[165,166]. These latter studies suggest the role of hyperuricemia in increased mitochondrial oxidative stress. While uric acid was not measured in the FHS Offspring Cohort study, increases in GMP and hypoxanthine may reflect the upstream hyperuricemia in the T2D-prone subjects. Additionally, subnetwork S1 includes the TCA cycle metabolites malate, isocitrate and aconitate, which are all increased in T2D-prone subjects, suggesting alterations in mitochondrial metabolism.

Table 2.3: Pathway information for the subnetworks in the FHS Offspring Cohort metabolomics data

Subnetwork	Pathway
Subnetwork 1	
Isocitrate, Malate, Aconitate	Citrate cycle (TCA cycle); Glyoxylate and dicarboxylate metabolism
Lactate, Malate	Pyruvate metabolism
GMP, Hyproxanthine, Inosine	Purine metabolism
Glucuronate	Ascorbate and aldarate metabolism; Pentose and glucuronate interconversions
2-Hydroxyphenylacetate	Phenylalanine metabolism
Quinolinate	Nicotinate and nicotinamide metabolism
2-Aminoadipate	Lysine degradation
Cystathionine	Glycine, serine and threonine metabolism; Cysteine and methionine metabolism
Kynurenine	Tryptophan metabolism
Subnetwork 2	
DG 34:1, DG 34:2, DG 36:1, DG 36:2	Diglycerides
TG 44:1, TG 46:1, TG 46:2, TG 48:0, TG 48:1, TG 48:2, TG 48:3, TG 48:4, TG 50:2, TG 50:3, TG:4, TG 50:5, TG 52:1, TG 52:2, TG 52:3, TG 52:4, TG 52:5, TG 52:6, TG 54:2, TG 54:3, TG 54:4, TG 54:5, TG 54:6, TG 54:7, TG 54:8, TG 56:3	Triglycerides
Subnetwork 3	
UDP-galactose, UDP-glucose, Sucrose, Lactose, Glucose-1-phosphate, Glucose-6-phosphate, Fructose-1-phosphate, Fructose-6-phosphate	Sugar metabolism
cyclic adenosine monophosphate, adenosine monophosphate, adenosine diphophate, guanosine diphosphate, uridine diphosphate	Nucleotide metabolism
ribose phosphate, ribulose phosphate	Pentose phosphate pathway
Nicotinamide	Nicotinate and nicotinamide metabolism
alpha-glycerophosphate	Glycerolipid metabolism
Asparagine	Alanine, aspartate and glutamate metabolism
Serotonin	Tryptophan metabolism

Subnetwork 4	
TG 46:0, TG 50:1, TG 54:1, TG 54:9, TG 56:2, TG 58:6, TG 58:7	Triglycerides
Subnetwork 5	
glycocholate, taurocholate	Primary bile acids
deoxycholate, taurodeoxycholate, glycodeoxycholate	Secondary bile acids
Subnetwork 6	
TG 56:9, TG 56:10, TG 58:10, TG 58:11, TG 58:12, TG 60:12	Triglycerides
Glutamate, Arginine, Argininosuccinate, Aspartate, Glutamine, alpha-ketoglutarate, Aminoisobutyric acid, Pyruvate, Glycine, Serine, Carnosine, S-Adenosylhomocysteine, 5-Hydroxyindoleacetic acid, 3-Hydroxyanthranilic acid, anthranilic acid	Amino acid metabolism
Triiodothyronine, Thyroxine, Pyruvate	Tyrosine metabolism
Pyruvate, alpha-ketoglutarate	TCA cycle
Pantothenate, N-carbamoyl-beta-alanine, Pyridoxate	Vitamin metabolism
Subnetwork 7	
LPC 14:0, LPC 16:0, LPC 18:0, LPC 18:1, LPC 18:2, LPC 20:3, LPC 20:4, LPC 30:5, LPC 22:6	Lysophosphatidylcholines
LPE 16:0, LPE 18:0, LPE 18:1, LPE 18:2, LPE 20:4, LPE 22:6	Lysophosphatidylethanolamines
alpha-glycerophosphocholine, PC 32:1, PC 32:2, PC 34:1, PC 34:2, PC 34:3, PC 34:4, PC 36:1, PC 36:2, PC 36:3, PC 36:4, PC 38:4, PC 38:5, PC 40:6	Phosphatidylcholines
Subnetwork 8	
Hippurate	NA
Gentisate	Benzoate degradation
Indole propionate	Tryptophan metabolism
Fumarate + maleate + valerate	Tyrosine metabolism
Salicylurate	NA
Subnetwork 9	
TG 56:4, TG 56:5, TG 56:6, TG 56:7, TG 56:8, TG 58:8, TG 58:9	Triglycerides
PC 38:6	Phosphatidylcholines
Subnetwork 10	
CE 14:0, CE 16:0, CE 16:1, CE 18:0, CE 18:1, CE 18:2, CE 18:3, CE 20:3, CE 20:4, CE 20:5, CE 22:6	Cholesteryl Esters
Subnetwork 11	
Histidine, Phenylalanine, Methionine, Valine, Alanine, Lysine, Isoleucine, Leucine, Threonine, Tryptophan, Tyrosine, Proline	Amino acid metabolism; Aminoacyl-tRNA biosynthesis
Choline, Betaine, Threonine, NMMA	Glycine, serine and threonine metabolism
Hydroxyproline, Proline, Ornithine, Citrulline	Arginine and proline metabolism
Taurine	Taurine and hypotaurine metabolism
Xanthine, Xanthosine	Purine metabolism

Subnetwork 12	
SM 14:0, SM 16:0, SM 16:1, SM 18:0, SM 18:1, SM 22:0, SM 22:1, SM 24:0, SM 24:1	Sphingomyelins
PC 32:0, PC 38:2, PC 38:3	Phosphatidylcholines

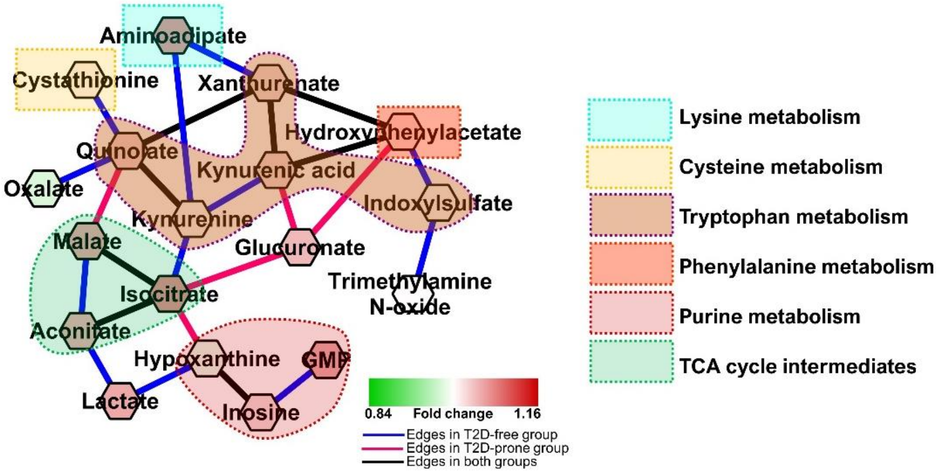


Figure 2.6: Subnetwork S1 in the Framingham Heart Study Offspring Cohort T2D data highlighting intermediates of various amino acids' metabolism and the TCA cycle

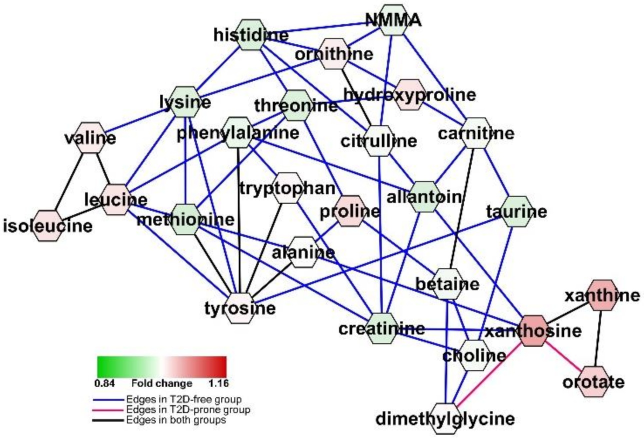


Figure 2.7: Branched chain amino acids-containing subnetwork (S11) in the Framingham Heart Study Offspring Cohort T2D data

Subnetwork S3 contains a higher proportion of edges in the non-diabetic group and is populated by sugars and sugar phosphates in the glycolysis and pentose shunt pathways, nucleotides, and sugar nucleotides. T2D-prone subjects have higher plasma levels of these sugars and sugar-derivatives than non-diabetic subjects. Taken together, the metabolite alterations seen in subnetworks S1 and S3 are indicative of widespread changes in the orderly flux of metabolites through mitochondria in diabetes-prone individuals. While not a new concept (reviewed in [167]), our results demonstrate the utility of the DNEA approach to provide insights into altered whole body metabolism using plasma metabolomics.

Subnetworks S2 and S4 were statistically significant in our analysis, even though the majority of edges are non-differential. These subnetworks are primarily made up of long-chain (C44-C58) polyunsaturated triglycerides (PUFA-TGs) with the additional inclusion of four diglyceride (DG) species (DG 34:1, DG 34:2, DG 36:1, DG 36:2), two saturated triglycerides (TG 46:0 and TG 48:0) and six monounsaturated triglycerides (TG 44:1, TG 46:1, TG 48:1, TG 50:1, TG 52:1, and TG 54:1). Most TG lipids, except TG 46:0, TG 50:1, TG 58:6, and TG 58:7, are present at higher levels in T2D-prone subjects. Overall, the enrichment of these two subnetworks is primarily driven by differential expression of the nodes. Increased plasma triglycerides have been reported as an independent predictor of T2D in several prospective cohort studies[168,169]. Additionally, triglycerides tend to be highly correlated with each other and typically form densely connected clusters in correlation networks[126]. The presence of a separate smaller triglyceride subnetwork (S4) may be due to the absence in the dataset of key triglyceride species that could link these subnetworks.

Subnetwork S5 exclusively contains bile acids with non-differential edges, suggesting that the differences between T2D-prone and T2D-free subjects in this case are driven by differential expression of the metabolites. Bile acids are the primary route of cholesterol catabolism and are synthesized by the oxidation of the latter by the action of the rate-limiting enzyme cholesterol 7 alpha-hydroxylase. Alterations in bile acid metabolism have been associated with T2D[170–174]. Additionally, obese T2D individuals have increased fasting and post-prandial total bile acid concentrations, due to increased enterohepatic circulation[175].

Subnetwork S6 contains several amino acids and their derivatives, TCA cycle intermediates, vitamin B metabolites and thyroid hormones (**Table 2.2**). In general, network connectivity was higher in the T2D-free group compared to the T2D-prone group. The levels of the individual amino acids and primary metabolites in this subnetwork are generally lower in T2D-prone group. Reductions in glycine and glutamine-to-glutamate ratio have been found in T2D subjects and in T2D-prone individuals[176]. The basis for the changes in arginine and aspartate levels, which are reduced in concert with other amino acids (save glutamate) in this network are less clear. We did not observe differential connectivity among the polyunsaturated fatty acid-containing triglycerides (PUFA-TGs). However, their levels were increased in the T2D-prone group, consistent with the overall increase in the TGs in the T2D-prone population.

Our analysis of the FHS Offspring Cohort metabolomics data supports many of the previous findings elucidating the role of changes in amino acid metabolism and increased oxidative stress in the prediction of T2D onset. Additionally, of the nineteen metabolites prioritized by Merino and colleagues (from the same dataset) that significantly improved T2D prediction in a model including traditional T2D risk factors[134], ten were part of our significantly differential subnetworks S1–S6 (**Table 2.3**). With these previously observed metabolite relationships as a foundation, our subnetworks can provide further biochemical context and help build on the understanding of metabolic changes that eventually lead to disease.

Table 2.4: Subnetwork assignments of the top 19 T2D predictors reported by Merino et al (2018). Predictors that are present in significant subnetworks (S1-S6) are highlighted in green.

Predictors	Subnetwork
2-Aminodipate	S1
Isocitrate	S1
DAG C36:1	S2
TAG C48:0	S2
TAG C48:1	S2
TAG C52:1	S2
TAG C54:8	S2
TAG C58:11	S6
5-Hydroxyindoleacetic acid	S6
Glycine	S6

PC C36:4	S7
LPC C18:2	S7
LPC C18:1	S7
CE C20:3	S10
L-Phenylalanine	S11
Taurine	S11
SM C24:0	S12
3-Methyladipic acid	NA
D-Glucose	NA

2.3.3 Subnetworks of Lipids Relate to Infant Birth Weight in the Michigan Mother-Infant Pairs (MMIP) Cohort

We used *Filigree* to analyze the MMIP dataset[135], comparing the lipidomes of women at different stages of pregnancy and their offspring (**Figure 2.8**). Capitalizing on the method's ability to identify functionally related metabolic modules, we sought to explore the association of subnetworks with infant birth weight (BW). Accordingly, we performed three pairwise comparisons (M1 vs. M3, M1 vs. CB, and M3 vs. CB). Since the dataset contained 670 lipids and 106 samples, we used the feature aggregation functionality of the tool (described in Methods) to reduce the dimensionality of the data. **Table 2.4** gives the reduced feature count for each of the comparisons and the percent of feature reduction. Overall, a 55–60% reduction was chosen, yielding feature counts comparable to the sample size. Most of the identified subnetworks were significantly enriched in each of the pairwise comparisons: 14/19 in M1, 19/20 in M3, and 9/12 in CB (**Table 2.4**). Consistent with our previous observations[126], lipids from the same or highly related classes were often found within the same subnetworks, such as diglycerides (DG) and triglycerides (TG), phosphatidylcholines (PC) and phosphatidylethanolamines (PE), and lysophosphatidylcholines (LPC) and lysophasphatidylethanolamines (LPE). Most subnetworks included differential edges at each time point, indicating changes in the connectivity of the lipidome during pregnancy.

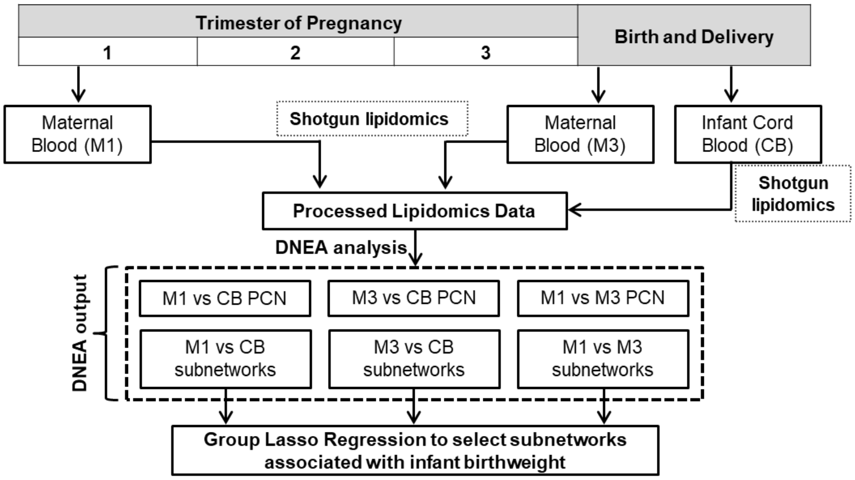


Figure 2.8: Michigan Mother-Infant Pairs (MMIP) study design. 106 pregnant women were monitored through the course of their pregnancy. Maternal plasma samples were collected at the first trimester (M1) and at time of delivery (M3), along with Cord Blood (CB). Data from subsequent lipidomics experiments was analyzed in a pairwise manner using Filigree and resulting subnetworks were tested for their association with infant birth weight in a group lasso regression model.

Table 2.5: Summary of the node-aggregation and identified subnetworks in each pairwise comparison of the MMIP lipidomics data

Comparison	Effective number of features	% reduction in feature space	Number of significant subnetworks (adj p-val < 0.05)	Total number of subnetworks identified
M1 – M3	298	55.45	14	19
M1 - CB	298	55.45	19	20
M3 - CB	286	57.25	9	12

Next, we assessed whether any of the identified subnetworks were associated with infant BW, which is of particular interest due to its relationship with future weight gain and risk for metabolic disease[177]. We used group lasso regression[136] (described in Methods) to model our *Filigree* subnetworks as predictors and Fenton BW[178] (BW normalized for gestation period and sex) as the outcome variable. In the M1 vs. CB comparison, two subnetworks containing LPC-LPE-PlasmenylPC (S18) and PC-TG (S12) components displayed strong association with BW (**Figure 2.9**). The LPC-LPE-PlasmenylPC subnetwork, composed of lipids with saturated, monounsaturated, and polyunsaturated fatty acid tails, showed a stronger association with BW in CB. Previous work has emphasized the relationship between CB LPCs and BW[135,179], but no previous studies have reported an association with PlasmenylPCs. Plasmalogen formation is primarily regulated by peroxisomes and it has been proposed that plasmologens are related to inflammation and oxidative stress[180], potentially explaining their association with BW. The PC-TG subnetwork displayed a stronger association with BW in M1. This network is composed of lipids that contain saturated fatty acid tails with 12–16 carbons. Our results expand on the previous analysis[135] that found minimal associations between the M1 lipidome and BW, emphasizing the advantage of our network-based approach. We hypothesize that lipids with saturated fatty acids play a role in establishing BW in the first trimester of pregnancy (8–14 weeks), highlighting the plasticity of the developing fetus in early gestation, responding potentially through epigenetic modifications[181]. Interestingly, the edges within the subnetwork diminish in CB, suggesting different connectivity between these saturated lipids at each time point, potentially due to changes in insulin sensitivity during pregnancy[182].

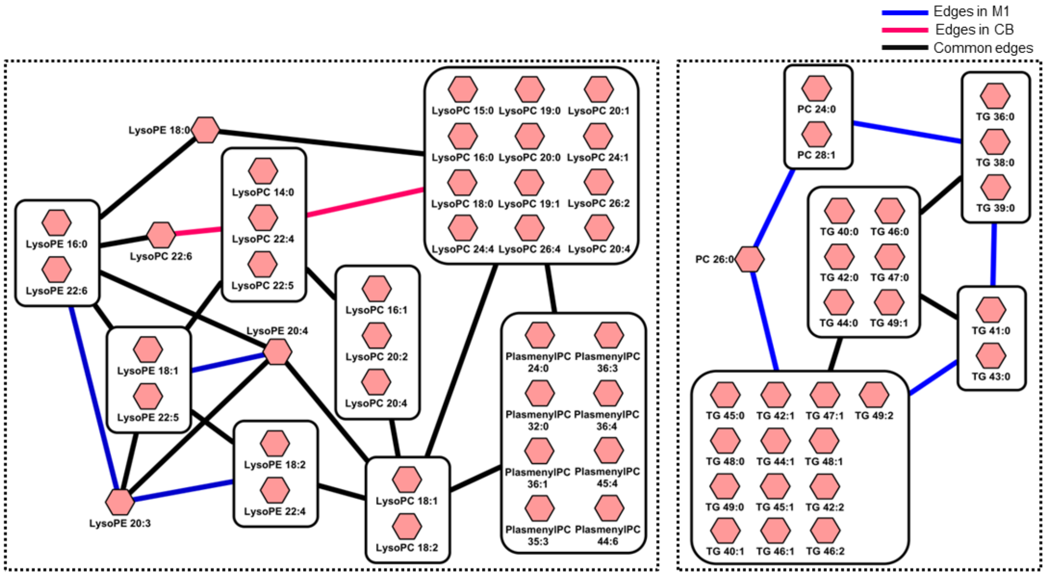


Figure 2.9: Top two M1 vs. CB subnetworks strongly associated with infant birth weight. LPC-LPE-PlasmenylPC subnetwork in infant Cord Blood and PC-TG subnetwork during the first trimester of the mother are strongly associated with infant birth weight. Large square nodes containing smaller nodes within them represent 'aggregated' nodes with their individual lipid species.

In the M3 vs. CB comparison, two subnetworks containing LPC-LPE (S6) and PC-PlasmenylPC-PlasmenylPE-DG-TG (S10) components displayed strong associations with BW (**Figure 2.10**). These subnetworks were associated with BW specifically in the CB, rather than maternal plasma (M3). The LPC-LPE subnetwork only includes one PlasmenylPC (PlasmenylPC 26:0), suggesting that plasmalogens are less strongly correlated with lysophospholipids in this comparison. Almost a complete overlap of lysophospholipids was observed between M1-CB S18 and M3-CB S6. The PC-PlasmenylPC-PlasmenylPE-DG-TG subnetwork contains lipids with long-chain and very long-chain polyunsaturated fatty acid tails. Previous work[135] has suggested the association between BW and CB polyunsaturated TGs and DGs. However, our approach additionally shows the interconnectivity between multiple lipid classes. Since polyunsaturated fatty acids are preferentially transferred from maternal to fetal circulation[183], our results may suggest a mechanism that modifies fetal growth and BW for optimal development. Previous studies using polyunsaturated fatty acid supplementation during pregnancy have yielded mixed results[184], warranting further analyses of the interconnectivity of these lipid classes and their relationship to BW.

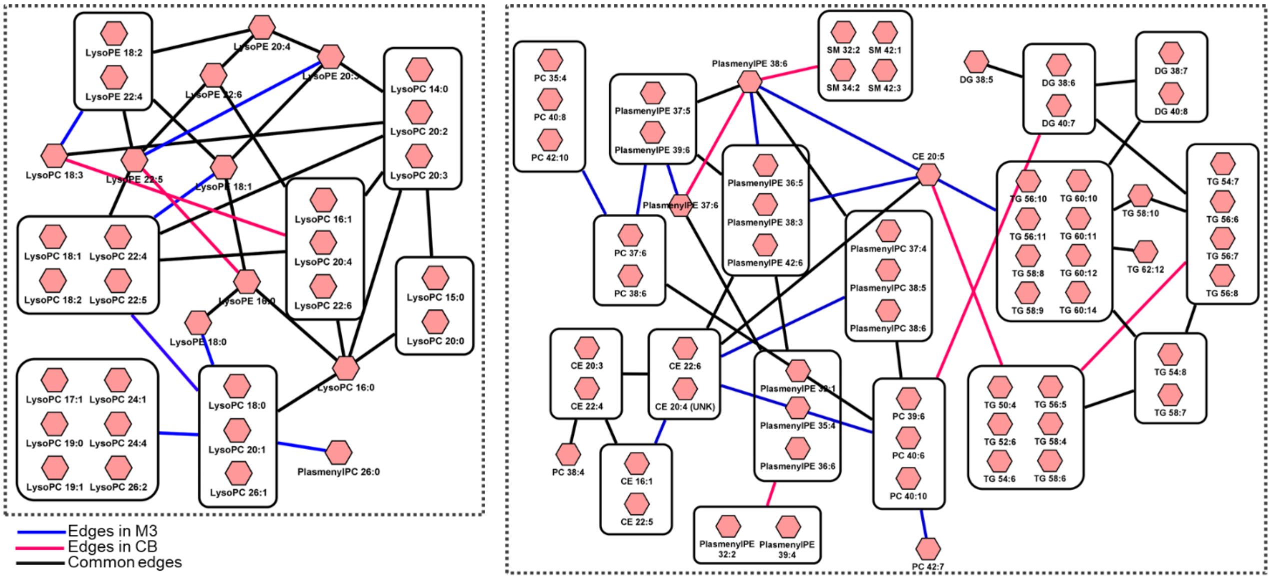


Figure 2.10: Top two M3 vs CB subnetworks strongly associated with infant birthweight. LPC-LPE and CE-PC-PlasmenylPC-PlasmenylPE-DG-TG subnetworks in infant Cord Blood are strongly associated with infant birthweight. Large square nodes containing smaller nodes within them represent 'aggregated' nodes with their individual lipid species.

Finally, in the M1 vs. M3 comparison, two subnetworks containing DG-TG (S7) and LPC-LPE (S14) components displayed strong associations with infant BW, led by maternal blood (M3) (**Figure 2.11**). The LPC-LPE subnetwork contains the same lysophospholipids as M1-CB S18 and M3-CB S6. These results suggest that maternal late gestation lysophospholipids are related to BW, potentially due to the active transport of lysophospholipids from maternal plasma to the CB by the major facilitator superfamily domain containing 2a (MFSD2a) protein[185]. Thus, enriched subnetworks obtained from the *Filigree* have meaningful biological significance and can be utilized to advance lipidomics data analysis by looking at their association with other phenotypes of interest.

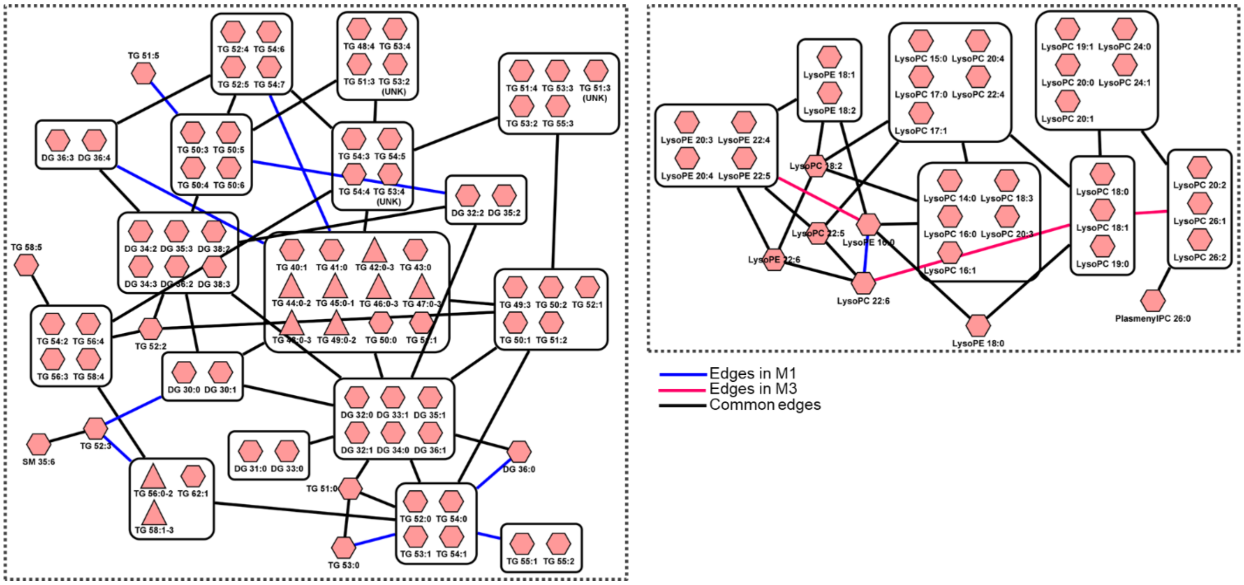


Figure 2.11: Top two M1 vs M3 subnetworks strongly associated with infant birthweight. DG-TG and LPC-LPE subnetworks during the third trimester of the mother are strongly associated with infant birthweight. Large square nodes containing smaller nodes within them represent 'aggregated' nodes with their individual lipid species. Triangular nodes represent a small group of triglycerides (2-3) with the same chain length and sequential unsaturation units.

In conclusion, we presented a novel bioinformatics approach for gaining new insights into high dimensional metabolomics data as implemented in our tool, *Filigree*. Our method helps overcome common challenges of pathway-based enrichment testing approaches, providing robust results even with limited sample sizes and highly imbalanced experimental group designs.

Currently, to the best of our knowledge, there is no other tool with comparable analysis pipeline. While partial correlation networks can be built with existing methodologies[109], *Filigree* provided a clear advantage in network estimation. In the T1D dataset, the number of metabolites far exceeded the number of samples, considerably restricting the number of statistically significant edges that could be recovered by other existing methods[109]. Our analysis also demonstrated that topology-based enrichment method implemented in *Filigree* is more powerful than traditional enrichment testing because it has the ability to provide information about changes in topology across the biological conditions.

In re-analyzing several existing datasets with *Filigree*, we observed a strong differential connectivity in metabolite networks in T1D and T2D and were also able to demonstrate various associations with infant BW in the lipidomes of pregnant women. *Filigree* is particularly useful as a hypothesis-generating tool. The results presented here suggest potential follow-up studies that could shed light on additional metabolic factors contributing to T1D and T2D and on potential lipidomic influences on BW during pregnancy.

2.4 DNEA: An R package for Data-Driven Network Enrichment Analysis of Metabolomics Data

While *Filigree* provides a user-friendly option for a casual user, the increasing size of metabolomics datasets involving larger number of compounds due to increased instrument resolution and larger number of samples, calls for more powerful computational solutions. To address this need, we developed an R package that implements the DNEA algorithm. In addition to being more computationally efficient, the DNEA R package includes the same features as Filigree, thus making it a powerful tool for the analysis of untargeted metabolomics and lipidomics data. The overall workflow of the package and functions is outlined in **Figure 2.12A**. Briefly, the `createDNEAobject()` function creates an R object from the input data (plain text file containing per-sample unadjusted intensity values and group information). This function also runs some diagnostics on the data to inform the user if <u>feature reduction</u> should be performed before continuing the analysis. The `BICtune()`

function determines the optimal lambda parameter for glasso by computing the Bayesian information criterion (BIC) and liklihood for a range of lambda values, while the `stabilitySelection()` function performs stability selection using the optimal lambda value. `getNetworks()` then computes the final network based on the selection probabilities from the previous step, and `runConsensusCluster()` and `runNetGSA()` perform consensus clustering and differential analysis respectively.

The feature reduction (`reduceFeatures()`) and stability selection coupled with additional subsampling functions are detailed below.

2.4.1 `reduceFeatures()`

There are three options available to perform feature reduction: (i) correlation-based (**Figure 2.12B**), (ii) knowledge-based, and (iii) correlation- and knowledge-based (hybrid). They are indicated as arguments to the reduceFeatures() function. In every case, each collapsed feature is represented by the average intensity of its constituent metabolites. The function returns a new data matrix with the collapsed features and a two-column matrix with the feature group membership of all the input metabolites.

Correlation-based feature reduction: In this method (**Figure 2.12B**), the user supplies a correlation coefficient threshold ($corr\_threshold$; default value is 0.9) above which metabolites should be merged into “collapsed features”. The metabolites must be correlated with each other above this threshold in both experimental groups. The input to the function is a matrix of metabolic expression values. Each row corresponds to an individual sample. The first column corresponds to sample ID while the second column corresponds to condition/group. First, a Pearson’s correlation matrix is computed for each condition separately. The correlation matrices are then clustered using hierarchical clustering (using dissimilarity measure) and the dendrograms are cut at a height of $(1 - corr\_threshold)$. This generates two sets of clusters (or feature groups), one for each experimental condition. The intersection of these sets of feature groups is performed by generating a consensus matrix (block diagonal matrix) encoding the final collapsed features. Further, the consensus matrix is converted to a graph object using the `igraph` R package[186] and connected components are identified. These connected components represent the final set of collapsed features.

Knowledge-based feature reduction: In this method, the user supplies a two-column feature grouping file as an input to the function. The function returns the 'collapsed' data matrix accordingly. No correlations are computed with this method.

Correlation- and knowledge-based (hybrid) feature reduction: This method combines the correlation and knowledge-based feature reduction strategies. Here, the user provides a two-column feature grouping file as well as a correlation coefficient threshold ($corr\_threshold$; default value is 0.9). The function then computes collapsed features based on correlation coefficients (as described above) within each of the user-defined feature groups i.e., only metabolites that are correlated with each other greater than $corr\_threshold$ and belong to the same feature group as defined by the user will get collapsed.

2.4.2 Stability selection with additional subsampling

This method (**Figure 2.12C**) can be used when the sample groups in the data are highly unbalanced. As detailed in Filigree's description, having highly unbalanced groups can lead to unstable partial correlation networks and heavy bias towards the group with greater number of samples. This can mask any potentially interesting biological differences between the groups. In order to circumvent this issue, we implemented a modified stability selection procedure for network estimation. While stability selection[161] itself performs subsampling to recover robust edges in the network, the modified version performs an additional downsampling in the larger group to reduce bias. This modification is indicated by a flag (`subSample = TRUE`) in the `stabilitySelection()` function. It is comprised of the following steps:

i. Determine size of smaller group (n_{min});
ii. At every iteration of the stability selection (default value set to 500 iterations), create new data matrices for the two groups as follows:

   a. For the larger group, randomly sample $(\alpha \times n_{min})$ samples without replacement.
   b. For the smaller group, randomly sample $(\beta \times n_{min})$ samples without replacement. Additionally, in order to maintain some degree of randomness in the smaller group, $(1-\beta) \times n_{min}$ samples are randomly chosen from this and added back.

iii. Fit the training model for the new subsamples of the data at every iteration;

iv. Obtain edge selection probabilities and retain edges with a selection probability of > τ;

v. Use the selection probabilities as weights when estimating the partial correlation networks.

Based on extensive experimentation, the following parameter values were chosen:

i. $\alpha = 1.3$,

ii. $\beta = 0.9$, and

iii. $\tau = 0.9$.

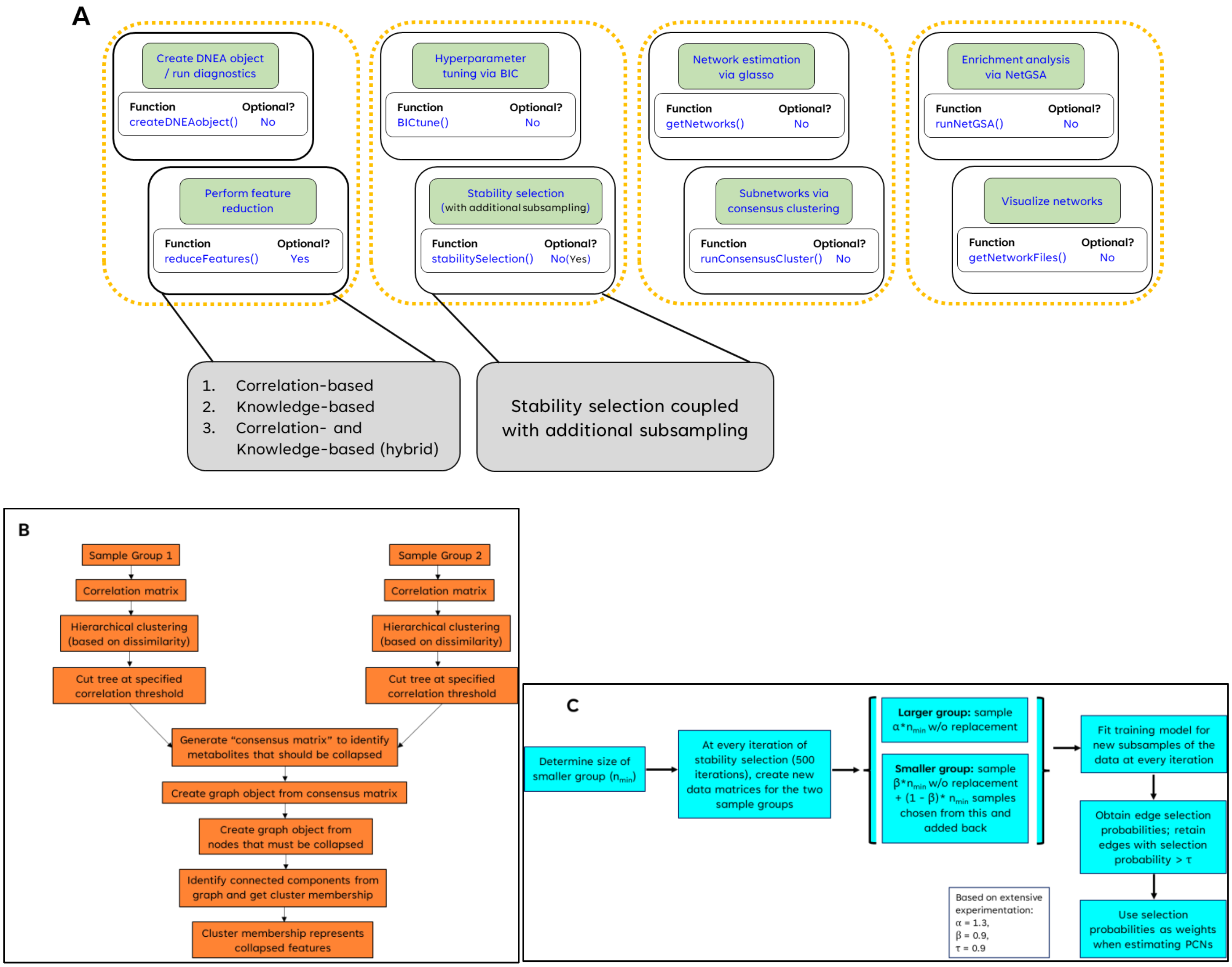

A
Create DNEA object / run diagnostics
Function Optional?
createDNEAobject() No
Perform feature reduction
Function Optional?
reduceFeatures() Yes
Hyperparameter tuning via BIC
Function Optional?
BICtune() No
Stability selection (with additional subsampling)
Function Optional?
stabilitySelection() No(Yes)
Network estimation via glasso
Function Optional?
getNetworks() No
Subnetworks via consensus clustering
Function Optional?
runConsensusCluster() No
Enrichment analysis via NetGSA
Function Optional?
runNetGSA() No
Visualize networks
Function Optional?
getNetworkFiles() No
1. Correlation-based
2. Knowledge-based
3. Correlation- and Knowledge-based (hybrid)
Stability selection coupled with additional subsampling
B
Sample Group 1
Correlation matrix
Hierarchical clustering (based on dissimilarity)
Cut tree at specified correlation threshold
Sample Group 2
Correlation matrix
Hierarchical clustering (based on dissimilarity)
Cut tree at specified correlation threshold
Generate "consensus matrix" to identify metabolites that should be collapsed
Create graph object from consensus matrix
Create graph object from nodes that must be collapsed
Identify connected components from graph and get cluster membership
Cluster membership represents collapsed features
C
Determine size of smaller group (n_{min})
At every iteration of stability selection (500 iterations), create new data matrices for the two sample groups
Larger group: sample $\alpha * n_{min}$ w/o replacement
Smaller group: sample $\beta * n_{min}$ w/o replacement + $(1 - \beta) * n_{min}$ samples chosen from this and added back
Fit training model for new subsamples of the data at every iteration
Obtain edge selection probabilities; retain edges with selection probability > τ
Use selection probabilities as weights when estimating PCNs
Based on extensive experimentation:
$\alpha = 1.3$,
$\beta = 0.9$,
$\tau = 0.9$

Figure 2.12: (**A**) Overall workflow of the DNEA R package. (**B**) Workflow of correlation-based feature reduction. (**C**) Workflow of the stability selection coupled with additional subsampling (for highly imbalanced sample groups) function.

CHAPTER III

Identification of Metabolic Markers of COVID-19 Severity

A manuscript covering this work is in preparation, with myself as the first author.

3.1 Introduction

Since the SARS CoV-2 2019 (COVID-19) virus was first identified, it has produced a global pandemic responsible for hundreds of millions of cases and millions of deaths. While most infected patients recover without requiring hospitalization, some develop severe disease that can lead to respiratory failure, multiple organ failure, and death. The importance of metabolic diseases such as hypertension, diabetes, and obesity on the risk of developing severe disease was recognized early[187–190]. Several hypotheses for this association have been suggested, including a high baseline inflammatory milieu which leads to an increased 'cytokine storm' associated with COVID-19 infections[191], hypertension and potential interaction with angiotensin converting enzyme 2 receptors (ACE2R)[192], mechanical compression of lungs due to intraabdominal fat, and underlying metabolic dysfunction in tissues exacerbated by the enhanced metabolic dysfunction induced by cytokines[193]. However, the connection between the triumvirate - hypertension, obesity, and diabetes with COVID-19 is still being investigated and is a rapidly evolving area of research.

Metabolomics has shown promise as a tool to understand the development of organ failure in COVID-19 and other acute illnesses by uncovering prognostic biomarkers and identifying metabolic alterations that may contribute to worse outcomes[194]. Multiple studies of the metabolome of human serum or plasma have demonstrated abnormalities of metabolism in patients with more severe disease. Other studies have shown differences in blood metabolite concentrations between those with mild and severe disease[195]. In patients who do survive COVID-19, those with metabolic abnormalities are at increased risk for developing chronic problems[196,197].

The goal of this study was to use broad spectrum untargeted metabolic profiling of blood plasma to differentiate healthy controls and patients with mild and severe COVID-19 disease. Patients who had type 2 diabetes (T2D) and higher BMI tended to have more severe disease. We performed differential analysis to identify a pool of potential metabolic markers of disease severity and tested their predictive power using parsimonious random forest models. We found that metabolite-based models performed better than similar models based on known COVID-19 risk factors such as age, BMI, race, gender, and diabetic status metabolite, indicating their potential value for predicting COVID severity.

3.2 Methods

3.2.1 Study Population & Sample Selection

This study was approved by the Institutional Review Board (IRBMED) at the University of Michigan, Ann Arbor, MI, USA. During the initial surge (April 2020) of COVID-19 admissions at a single tertiary care medical center, excess clinical specimens were collected from hospitalized patients. We selected patients who had available plasma samples collected on two different days. In general, the first and last samples available during hospitalization were used (referred to as timepoint 1 and timepoint 2 respectively). COVID-19 patients were separated into mild (n=73) or severe (n=132) groups. The mild group was defined as those individuals who were discharged from the hospital and never required intubation for mechanical ventilation. The severe group was defined as either those who required intubation and mechanical ventilation as part of their care and/or those who died during the hospitalization. In addition, 136 healthy controls were selected from the MGI-MEND (Michigan Genomics Initiative - Metabolism, Endocrinology & Diabetes) and IWMC

(Investigational Weight Management Clinic) cohorts, who were appropriately matched with the COVID-19 patients for age, gender, and race.

Samples were processed and sent to the University of Michigan Central Biorepository for storage. All patients greater than 18 years old at the time of sample collection and had plasma samples from multiple days available for analysis were eligible for inclusion. Patients with any limitations on medical therapy when the samples were collected were excluded.

3.2.2 Sample preparation

Untargeted metabolomics profiling was completed by the Michigan Regional Comprehensive Metabolomics Resource Core (MRC)^2 (www.mrc2.umich.edu), which has extensive experience in production and analysis of metabolomics data. Samples were arranged in a semi-randomized fashion so that control/COVID cases and COVID severity (mild/severe) were evenly distributed across batches. Samples were aliquoted for the untargeted metabolomics assays at the start of the project and pools (batch and global) were created from the individual plasma samples and treated identically to the samples in all subsequent steps. For a single sample batch (approx. 80-96 samples), samples were removed from -80 °C storage and maintained on wet ice throughout the processing steps. To each 50 µL sample, 200 µL of extraction solvent (1:1:1 Methanol:Acetonitrile:Acetone) containing internal standards was added. Samples were vortexed then allowed to incubate overnight at -20°C. Post incubation, the vortex step was repeated, and samples were centrifuged for 10 minutes at 14,000 RPM in 4° C to precipitate protein. 200 µL of supernatant was transferred to an autosampler vial and brought to complete dryness using a nitrogen blower in ambient conditions. Samples and controls were reconstituted with 100 µL of water: methanol (8:2 by volume).

3.2.3 Optimized LC-MS methods

For RPLC-MS, samples were analyzed on an Agilent 1290 Infinity II / 6545 qTOF MS system with the JetStream Ionization (ESI) source (Agilent Technologies, Inc., Santa Clara, CA USA) using the Waters Acquity HSS T3 1.8 µ 50 mM column (Waters Corporation, Milford, MA).

Each sample was analyzed twice, once in positive and once in negative ion mode. Mobile phase A was 100% water with 0.1% formic acid and mobile phase B was 100% methanol with 0.1% formic acid. The gradient for both positive and negative ion modes was as follows: 2% B (0 min), 75% B (20 min), 98%B (22 min), 98%B (30 min), 2% B (30.1 min) was used. The column was then reconditioned for 7 min with 2%B before moving to the next injection The flow rate was 0.46 mL/min and the column temperature was 40°C. The injection volume for positive and negative mode was 5 µL and 8 µL, respectively. Source parameters were: drying gas temperature 350°C, drying gas flow rate 10 L/min, nebulizer pressure 30 psig, sheath gas temp 350°C and flow 11 l/min, and capillary voltage 3500V, with internal reference mass correction.

3.2.4 Metabolite Analysis

Semi-quantitative data for known compounds is obtained by manually integration using Profinder v8.00 (Agilent Technologies, Santa Clara, CA.) Metabolites were identified by matching the retention time (+/- 0.1 min), mass (+/- 10 ppm) and isotope profile (peak height and spacing) to authentic standards.

3.2.5 Statistical and Bioinformatics Data Analysis

All data analyses were performed in R (v 4.1.1) statistical programming language and environment (https://www.R-project.org/).

Multiple linear regression (MLR) was performed on base 10 log-transformed measurements to describe differences in metabolite abundances due to selected covariates (age, gender, race, BMI, and T2D) i.e., number of regression models constructed was based on the number of metabolites tested. The direction of association was determined based on the sign of the regression coefficient. The residuals from the MLR models were taken as metabolite levels that are adjusted for the selected covariates and were used to test for differences between healthy controls and patients with COVID-19 (mild and severe) disease.

Analysis of Variance (ANOVA) and pairwise Student's t-tests were performed to identify differential metabolites between healthy controls and mild and severe COVID-19 patients.

The significance levels (p values) were adjusted for multiple hypothesis testing according to Benjamini and Hochberg[43] at a false discovery rate (FDR) of 5%.

Metabolites selected from differential analysis were utilized to build random forest classification models to compare their capability in classifying COVID-19 severity with that of clinical factors like age, gender, race, BMI, and diabetic status. The samples were split into 70% training (n = 145) and 30% test (n = 60) sets. The training data was used to construct the model and final model performance was validated using the test data (OOB error). Final model classification performance was validated through prediction of class labels for the test set and are reported as the area under the receiver operator characteristic curve.

Partial correlation networks were constructed for 294 metabolites from the healthy controls (n = 136), patients with mild COVID-19 (n = 73), and patients with severe COVID-19 (n = 132) using the Debiased Sparse Partial Correlation (DSPC) algorithm implemented in CorrelationCalculator[109]. DSPC is especially useful to estimate partial correlations in a high-dimensional setting (n << p), under the assumption that the true connectivity among the metabolites is much smaller than the sample size i.e., *sparse*. Significance of the partial correlation between a pair of metabolites i.e., edges in the partial correlation network was defined as an FDR-adjusted p-value < 0.1.

3.3 Results

3.3.1 Study Design and Patient Demographics

Samples were collected from patients admitted to Michigan Medicine during the initial surge of COVID-19. We selected 205 patients who had available plasma samples from at least two different days during hospitalization. The first and last samples available during the hospitalization were used for this analysis. COVID-19 patients were separated into mild or severe. Mild group (n = 73) was defined as those individuals who were discharged from the hospital and never required intubation for mechanical ventilation. Severe group (n = 132) was defined as either those who did require intubation and mechanical ventilation as part of their care and/or those who died during the hospitalization. The study also included 136 control individuals with similar characteristics to the COVID-19 cohort.

Compared to those with mild disease, patients with severe COVID-19 were more likely to have a higher BMI (37.0 ± 21.7 vs. 30.1 ± 9.0, p=0.002) and T2D (47.7% vs. 31.5%, p=0.021) (**Table 3.1**). Patients with severe disease were younger (56 vs. 63 years, p=0.011), and had a greater proportion of male gender (64 vs. 40%, p=0.003). Race did not differ significantly between groups.

While effort was made to select a control cohort with characteristics similar to the COVID-19 cohort, several significant differences were found between controls and each of the COVID-19 severities, potentially due to differences between the mild and severe COVID-19 groups (**Table 3.1**). Compared to controls, those with mild COVID-19 were older (62.6 vs. 56.9 years, p=0.019), had a lower BMI (30.1 vs 32.9 kg/m2, p=0.028), and were more likely to be diabetic (68.5% vs. 45.6%, 0=0.002). Comparing the severe COVID-19 cohort with controls only showed a difference in BMI, with the severe COVID-19 cohort tending to have a higher BMI compared to controls (37.0 vs. 32.9 kg/m2, p=0.042).

Table 3. 1: Control and COVID-19 Population Demographics

	Control (N=136)	Mild (N=73)	Severe (N=132)	p-value
Age (years)				
Mean (SD)	56.8 (14.5)	62.6 (17.9)	56.2 (15.9)	0.0127
Median [Min, Max]	60.0 [23.0, 88.0]	64.0 [23.0, 89.0]	58.0 [20.0, 89.0]	
BMI				
Mean (SD)	32.9 (7.82)	30.1 (9.03)	37.0 (21.7)	0.0068
Median [Min, Max]	32.55 [18.8, 53.8]	27.7 [16.9, 58.0]	33.4 [18.6, 21.8]	
Gender				
Female	66 (48.5%)	44 (60.3%)	49 (37.1%)	0.0049
Male	70 (51.5%)	29 (39.7%)	83 (62.9%)	
Race				
African American	50 (36.8%)	26 (35.6%)	56 (42.4%)	0.5341
Non-African American	86 (63.2%)	47 (64.4%)	76 (57.6%)	
Diabetic status				
Diabetes	74 (54.4%)	23 (31.5%)	63 (47.7%)	0.0093
No diabetes	62 (45.6%)	50 (68.5%)	69 (52.3%)	

*p-values for continuous and categorical variables correspond to ANOVA and Chi-squared tests, respectively

3.3.2 Metabolomics Analysis

Untargeted metabolomics profiling generated a dataset that contained 8599 and 4714 features in the positive and negative ionizations modes respectively. After data reduction using Binner[198], there were 5298 and 3273 features in the positive and negative modes respectively. 294 putatively annotated metabolites were included in the analysis.

3.3.2.1 Effect of clinical covariates on metabolome

Since the metabolome is known to be strongly influenced by various clinical factors, we looked at the association of these metabolites with age, gender, BMI, race, and diabetic status in multiple linear regression models. At p-value < 0.05, 87 metabolites were associated with age, 55 with gender, 23 with BMI, 78 with race and 61 with diabetic status (**Supplementary Table 3.1; Figure 3.1**). Majority of these metabolites were lipids that belonged to the following classes: fatty acyls, glycerophospholipids, sphingolipids, and sterol lipids. Other classes of compounds included organic acids and derivatives, organo-heterocyclic compounds, benzenoids, and nucleic acids.

Additionally, we examined the effect of a commonly used anesthetic substance, propofol on plasma metabolome. In COVID patients, propofol was administered prior to intubation. It must be noted that all patients who received propofol were from the severe group (n= 63). We found that 201 metabolites were associated with propofol administration (p-value < 0.05) (**Supplementary Table 3.1; Figure 3.1**). The majority of these were lipids, including sterols, sphingolipids, glycerolipids, glycerophospholipids and fatty acyls.

To eliminate the influence of age, gender, race, BMI, T2D, and propofol administration, we constructed a linear regression model with these covariates and disease status (Control, mild COVID, and severe COVID). We found that there were 196 significant metabolites (p < 0.05) between control and mild COVID groups, 234 significant metabolites (p < 0.05) between control and severe COVID groups, and 167 significant metabolites (p < 0.05) between mild and severe COVID groups.

To build partial correlation networks, we constructed a linear regression model with age, gender, race, BMI, T2D, and propofol administration (excluding disease status) and used the residuals from the model in subsequent analyses. We performed differential analysis on the same adjusted data and visualized the differential metabolites in the resulting networks.

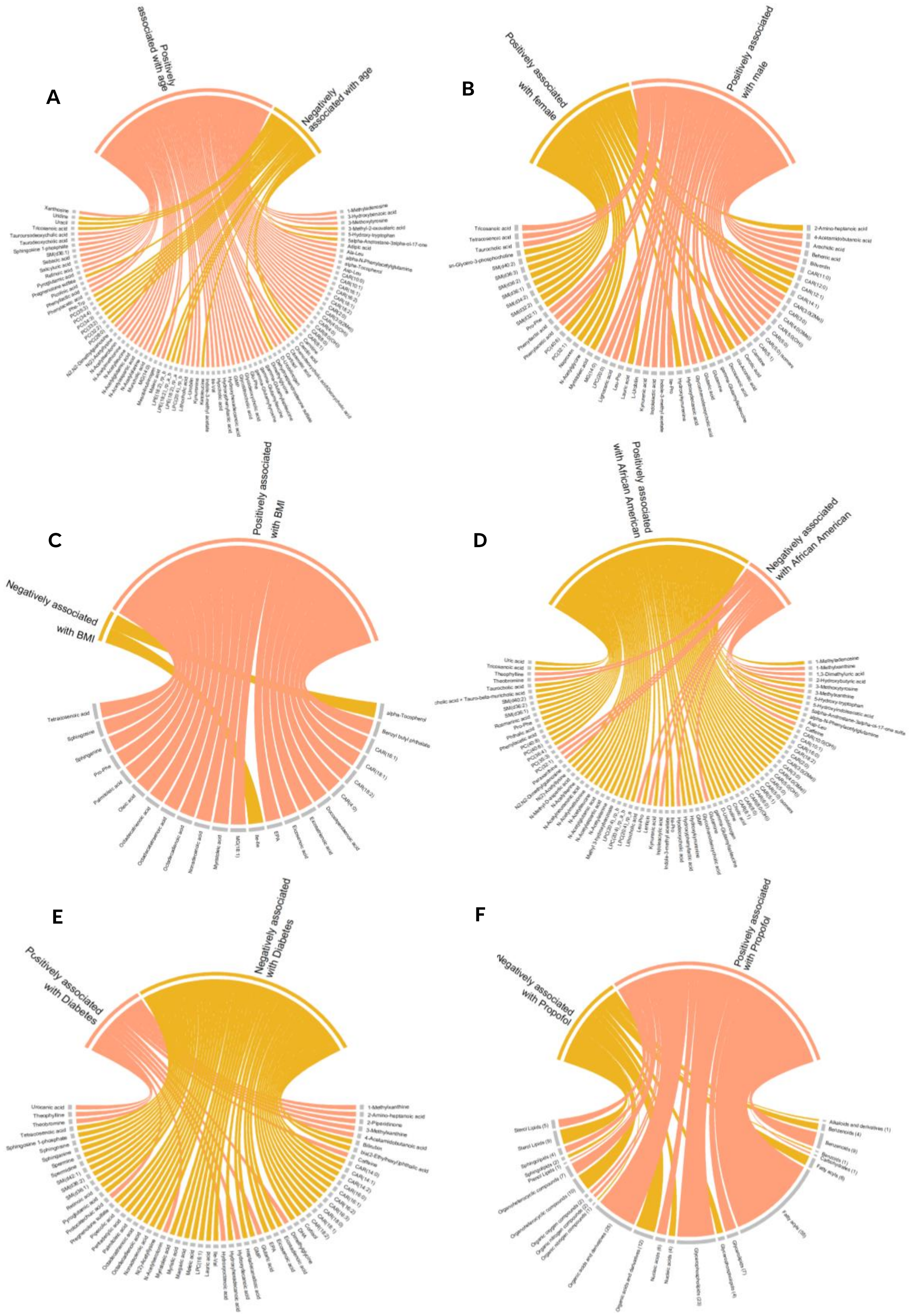

A
Positively associated with age
Negatively associated with age
Xanthosine
Uridine
Uracil
Tricosanoic acid
Tauroursodeoxycholic acid
Taurodeoxycholic acid
Sphingosine 1-phosphate
SM(d36:1)
Sebacic acid
Salicyluric acid
Retinoic acid
Pyroglutamic acid
Pregnenolone sulfate
Picolinic acid
Phenyllactic acid
Phenylacetic acid
Phe-Trp
PC(35:3)
PC(34:4)
PC(34:3)
PC(33:2)
PC(32:2)
PC(28:0)
N2,N2-Dimethylguanosine
N(2)-Acetyllysine
N-Acetylserotonin
N-Acetylneuraminic acid
N-Acetylleucine
N-Acetylglutamic acid
N-Acetylaspartic acid
N-Acetylalanine
Muramic acid
MG(14:0)
Mesobilirubinogen
Malic acid
LPE(18:2)_rp_b
LPE(18:2)_rp_a_b
LPE(18:2)_rp_a
LPC(20:4)_rp_b
Lithocholic acid
L-Urobilin
Kynurenine
Ketoleucine
Indole-3-methyl acetate
Ile-Val
Hyocholic acid
Hydroxyphenyllactic acid
Hydroxyhexadecanoic acid
GMP
Glycoursocholic acid
Glycodeoxycholic acid
Glu-Phe
gamma-Glutamyltyrosine
gamma-Glutamylleucine
Dimethylglycine
D-Urobilinogen
Cortisol
Cortisone
Dehydroepiandrosterone sulfate
Citramalic acid
Chenodeoxycholic acid/Deoxycholic acid
Cholic acid
Carnitine
CAR(8:1)
CAR(8:0)
CAR(6:0)
CAR(5:0(OH))
CAR(4:0)
CAR(4:0(OH))
CAR(3:0(2Me))
CAR(2:0)
CAR(18:2)
CAR(18:1)
CAR(16:2)
CAR(16:1)
CAR(10:1)
CAR(10:0)
Asp-Leu
alpha-Tocopherol
alpha-N-Phenylacetylglutamine
Ala-Leu
Adipic acid
5alpha-Androstane-3alpha-ol-17-one
5-Hydroxy-tryptophan
3-Methyl-2-oxovaleric acid
3-Methoxytyrosine
3-Hydroxybenzoic acid
1-Methyladenosine
B
Positively associated with female
Positively associated with male
Tricosanoic acid
Tetracosenoic acid
Taurocholic acid
sn-Glycero-3-phosphocholine
SM(d40:2)
SM(d36:3)
SM(d36:2)
SM(d36:1)
SM(d34:2)
SM(d32:2)
SM(d32:1)
Pro-Phe
Phenyllactic acid
Phenylacetic acid
PC(40:6)
PC(32:1)
Naproxen
N-Acetylglycine
Myristoleic acid
MG(14:0)
LPC(20:0)
Lignoceric acid
Leu-Pro
Lauric acid
L-Urobilin
Kynurenic acid
Indolelactic acid
Indole-3-methyl acetate
Ile-Pro
Hydroxykynurenine
Hydroxydecanoic acid
Glycochenodeoxycholic acid
Glutaric acid
Glutamine
gamma-Glutamylleucine
Docosenoic acid
cis-Aconitic acid
Citric acid
Carnitine
CAR(5:1)
CAR(5:0) isomers
CAR(5:0)
CAR(5:0(OH))
CAR(4:0(3Me))
CAR(3:0)
CAR(3:0(2Me))
CAR(14:1)
CAR(12:1)
CAR(12:0)
CAR(11:0)
Biliverdin
Behenic acid
Arachidic acid
4-Acetamidobutanoic acid
2-Amino-heptanoic acid
C
Positively associated with BMI
Negatively associated with BMI
Tetracosenoic acid
Sphingosine
Sphinganine
Pro-Phe
Palmitoleic acid
Oleic acid
Octadecatrienoic acid
Octadecatetraenoic acid
Octadecadienoic acid
Nonadecenoic acid
Myristoleic acid
MG(18:1)
Ile-Ile
EPA
Eicosenoic acid
Eicosatrienoic acid
Docosapentaenoic acid
CAR(4:0)
CAR(18:2)
CAR(18:1)
CAR(16:1)
Benzyl butyl phthalate
alpha-Tocopherol
D
Positively associated with African American
Negatively associated with African American
Uric acid
Tricosanoic acid
Theophylline
Theobromine
Taurocholic acid
cholic acid + Tauro-beta-muricholic acid
SM(d40:2)
SM(d36:2)
SM(d36:1)
Rosmarinic acid
Pro-Phe
Phthalic acid
Phenylacetic acid
PC(40:8)
PC(40:6)
PC(36:4)
PC(35:3)
PC(32:1)
Paraxanthine
N2,N2-Dimethylguanosine
N(2)-Acetyllysine
N-Methyl-D-aspartic acid
N-Acetyltaurine
N-Acetylneuraminic acid
N-Acetylmethionine
N-Acetylleucine
N-Acetylglutamic acid
N-Acetylaspartic acid
N-Acetylalanine
Methyl 3-hydroxybenzoate
LPC(20:4)_rp_b
LPC(20:4)_rp_a_b
LPC(20:4)_rp_a
Lithocholic acid
Leu-Pro
Lenticin
Kynurenic acid
Indoleacrylic acid
Indole-3-methyl acetate
Ile-Pro
Hyodeoxycholic acid
Hydroxyphenyllactic acid
Hydroxykynurenine
GMP
Glycochenodeoxycholic acid
Glutamine
gamma-Glutamylleucine
D-Urobilinogen
Choline
Cholic acid
CAR(8:1)
CAR(8:0)
CAR(6:0)
CAR(5:1)
CAR(5:0) isomers
CAR(5:0)
CAR(5:0(OH))
CAR(4:0(3Me))
CAR(3:0)
CAR(3:0(2Me))
CAR(2:0)
CAR(18:2)
CAR(16:0)
CAR(10:1)
CAR(10:0(OH))
Caffeine
Asp-Leu
alpha-N-Phenylacetylglutamine
5alpha-Androstane-3alpha-ol-17-one sulfa
5-Hydroxyindoleacetic acid
5-Hydroxy-tryptophan
3-Methylxanthine
3-Methoxytyrosine
2-Hydroxybutyric acid
1,3-Dimethyluric acid
1-Methylxanthine
1-Methyladenosine
E
Positively associated with Diabetes
Negatively associated with Diabetes
Urocanic acid
Theophylline
Theobromine
Tetracosenoic acid
Sphingosine 1-phosphate
Sphingosine
Sphinganine
Spermine
Spermidine
SM(d42:1)
SM(d36:2)
SM(d36:1)
Retinoic acid
Pyroglutamic acid
Protocatechuic acid
Pregnenolone sulfate
Pipecolic acid
Pentadecylic acid
Palmitoleic acid
Octadecatrienoic acid
Octadecadienoic acid
Nonadecenoic acid
N(2)-Acetyllysine
N-Acetylserotonin
Myristoleic acid
Myristic acid
Margaric acid
Maleic acid
LPC(16:1)
Lauric acid
Ile-Val
Hydroxyoctanoic acid
Hydroxyhexadecanoic acid
Hydroxydecanoic acid
Heptadecanoic acid
GMP
Glutaric acid
EPA
Eicosenoic acid
Eicosadienoic acid
Dimethylglycine
DHA
Cortisol
CAR(18:2)
CAR(18:1)
CAR(18:0)
CAR(16:3)
CAR(16:2)
CAR(16:1)
CAR(16:0)
CAR(14:2)
CAR(14:1)
CAR(14:0)
Caffeine
bis(2-Ethylhexyl)phthalic acid
Bilirubin
4-Acetamidobutanoic acid
3-Methylxanthine
2-Piperidinone
2-Amino-heptanoic acid
1-Methylxanthine
F
Negatively associated with Propofol
Positively associated with Propofol
Sterol Lipids (5)
Sterol Lipids (9)
Sphingolipids (4)
Sphingolipids (2)
Prenol Lipids (1)
Organoheterocyclic compounds (7)
Organoheterocyclic compounds (10)
Organic oxygen compounds (2)
Organic nitrogen compounds (2)
Organic nitrogen compounds (1)
Organic acids and derivatives (25)
Organic acids and derivatives (12)
Nucleic acids (6)
Nucleic acids (4)
Glycerophospholipids (23)
Glycerophospholipids (4)
Glycerolipids (7)
Fatty acyls (35)
Fatty acyls (6)
Carbohydrates (1)
Benzoids (1)
Benzenoids (9)
Benzenoids (4)
Alkaloids and derivatives (1)

Figure 3.1: Chord diagrams illustrating the association of age (**A**), gender (**B**), race (**C**), BMI (**D**), diabetes (**E**), and propofol administration (**F**) with the metabolome in the controls and COVID-19 patients. Metabolites associated with propofol administration (**F**) are grouped into chemical classes for ease of visualization.

3.3.2.2 Differential analysis

We performed a one-way ANOVA, followed by Tukey's HSD post-hoc test to compare the metabolomes of the controls, mild COVID, and severe COVID populations using the first collected blood sample (timepoint 1). We identified 244 significant (FDR < 0.05) metabolites, suggesting that the difference in the levels of these metabolites was due to disease status without any confounding variable (**Table 3.2**). These were primarily fatty acyls (75), organic acids and derivatives (43), glycerophospholipids (33), organo-heterocyclic compounds (23), sterol lipids (15), and benzenoids (15) (**Figure 3.2**).

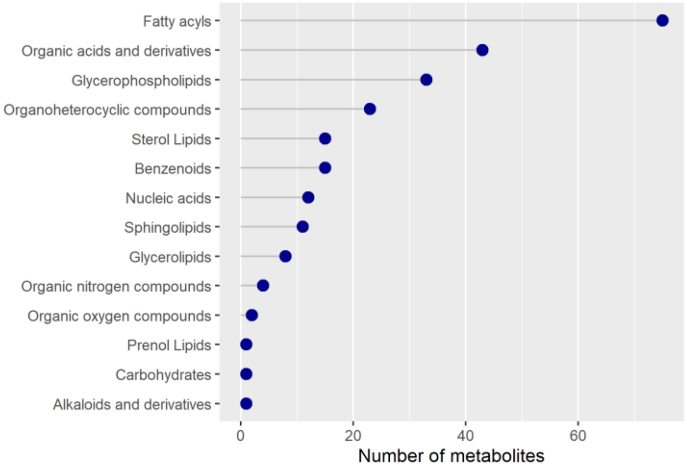


Figure 3.2: Lollipop plot of classes of metabolites that are significantly differential between controls and COVID-19 patients.

Table 3.2: Significantly differential metabolites between controls and COVID-19 patients

Metabolite	F-statistic	p-value	FDR	Tukey's HSD
Pyroglutamic acid	125.43	1.58E-41	3.20E-39	Mild - Control; Severe - Control; Mild - Severe
Sphingosine	124.88	2.17E-41	3.20E-39	Mild - Control; Severe - Control; Mild - Severe
Tetracosenoic acid	116.96	2.22E-39	2.17E-37	Mild - Control; Severe - Control; Severe - Mild
Maleic acid	110.66	9.70E-38	7.13E-36	Mild - Control; Severe - Control; Mild - Severe
CAR(18:1)	97.675	3.04E-34	1.79E-32	Mild - Control; Severe - Control; Severe - Mild
CAR(18:2)	96.56	6.18E-34	3.03E-32	Mild - Control; Severe - Control; Severe - Mild
Sphinganine	96.081	8.39E-34	3.53E-32	Mild - Control; Severe - Control; Severe - Mild
Piperine	95.706	1.07E-33	3.92E-32	Control - Mild; Control - Severe; Mild - Severe
Docosatrienoic acid	80.197	2.94E-29	9.60E-28	Mild - Control; Severe - Control; Severe - Mild
Protocatechuic acid	73.555	2.85E-27	8.38E-26	Control - Mild; Control - Severe; Severe - Mild
MG(18:1)	72.275	6.98E-27	1.87E-25	Mild - Control; Severe - Control; Severe - Mild
Eicosenoic acid	71.322	1.36E-26	3.34E-25	Mild - Control; Severe - Control; Severe - Mild
Arachidic acid	66.83	3.33E-25	7.52E-24	Mild - Control; Severe - Control; Severe - Mild
3-Methylxanthine	65.368	9.53E-25	2.00E-23	Control - Mild; Control - Severe; Mild - Severe
CAR(16:1)	62.281	8.99E-24	1.76E-22	Mild - Control; Severe - Control; Severe - Mild
3-Indolepropionic acid	60.527	3.26E-23	5.99E-22	Control - Mild; Control - Severe; Mild - Severe
Docosenoic acid	58.809	1.16E-22	2.01E-21	Mild - Control; Severe - Control; Severe - Mild
Retinoic acid	58.327	1.67E-22	2.72E-21	Mild - Control; Severe - Control; Mild - Severe
Azelaic acid	57.048	4.33E-22	6.70E-21	Mild - Control; Severe - Control; Mild - Severe
DG(34:1)	56.655	5.81E-22	8.54E-21	Mild - Control; Severe - Control; Severe - Mild
Theophylline	55.838	1.07E-21	1.50E-20	Control - Mild; Control - Severe; Mild - Severe
Phe-Trp	55.082	1.90E-21	2.54E-20	Control - Mild; Control - Severe; Severe - Mild
CAR(16:0)	54.841	2.28E-21	2.91E-20	Mild - Control; Severe - Control; Severe - Mild
Caffeine	54.539	2.86E-21	3.51E-20	Control - Mild; Control - Severe; Mild - Severe
Ile-Val	54.13	3.90E-21	4.59E-20	Mild - Control; Severe - Control; Severe - Mild
Palmitoleic acid	53.837	4.87E-21	5.51E-20	Mild - Control; Severe - Control; Mild - Severe
Theobromine	52.962	9.49E-21	1.03E-19	Control - Mild; Control - Severe; Mild - Severe
Ser-Leu	52.869	1.02E-20	1.07E-19	Mild - Control; Severe - Control; Mild - Severe
Octadecatrienoic acid	51.796	2.31E-20	2.34E-19	Mild - Control; Severe - Control; Severe - Mild
N(2)-Acetyllysine	50.822	4.89E-20	4.71E-19	Mild - Control; Severe - Control; Severe - Mild
Nonadecenoic acid	50.803	4.96E-20	4.71E-19	Mild - Control; Severe - Control; Severe - Mild
1,3-Dimethyluric acid	49.977	9.38E-20	8.62E-19	Control - Mild; Control - Severe; Mild - Severe
Octadecadienoic acid	49.42	1.44E-19	1.29E-18	Mild - Control; Severe - Control; Severe - Mild
Undecanedioic acid	48.152	3.87E-19	3.34E-18	Mild - Control; Severe - Control; Mild - Severe
Eicosadienoic acid	47.025	9.32E-19	7.83E-18	Mild - Control; Severe - Control; Severe - Mild
Paraxanthine	46.772	1.14E-18	9.29E-18	Control - Mild; Control - Severe; Mild - Severe
Margaric acid	45.601	2.86E-18	2.27E-17	Mild - Control; Severe - Control; Severe - Mild
DG(34:2)	45.093	4.26E-18	3.30E-17	Mild - Control; Severe - Control; Severe - Mild
Myristic acid	45.012	4.55E-18	3.43E-17	Mild - Control; Severe - Control; Mild - Severe
Pentadecylic acid	43.7	1.29E-17	9.46E-17	Mild - Control; Severe - Control; Mild - Severe
DG(32:0)	42.829	2.58E-17	1.85E-16	Mild - Control; Severe - Control; Severe - Mild
Behenic acid	41.939	5.25E-17	3.68E-16	Mild - Control; Severe - Control; Severe - Mild
Ala-Leu	41.176	9.70E-17	6.63E-16	Mild - Control; Severe - Control; Mild - Severe
2-Aminooctanoic acid	38.856	6.34E-16	4.24E-15	Control - Mild; Control - Severe; Mild - Severe
N-Acetyl-D-tryptophan	37.802	1.50E-15	9.79E-15	Control - Mild; Control - Severe; Severe - Mild
CAR(4:0(OH))	37.293	2.28E-15	1.45E-14	Mild - Control; Severe - Control; Mild - Severe
gamma-Glutamylmethionine	36.048	6.33E-15	3.93E-14	Control - Mild; Control - Severe; Severe - Mild
DG(18:1_18:1)	36.034	6.41E-15	3.93E-14	Mild - Control; Severe - Control; Severe - Mild

Hydroxydecanoic acid	35.846	7.49E-15	4.48E-14	Control - Mild; Control - Severe; Severe - Mild
Octadecatetraenoic acid	35.824	7.62E-15	4.48E-14	Mild - Control; Severe - Control; Severe - Mild
Leu-Pro	35.749	8.11E-15	4.68E-14	Mild - Control; Severe - Control; Severe - Mild
GMP	35.631	8.94E-15	5.06E-14	Mild - Control; Severe - Control; Severe - Mild
N-Acetylglutamic acid	35.507	9.90E-15	5.49E-14	Mild - Control; Severe - Control; Severe - Mild
PC(34:4)	35.438	1.05E-14	5.71E-14	Control - Mild; Control - Severe; Severe - Mild
Ile-Pro	34.87	1.68E-14	8.97E-14	Mild - Control; Severe - Control; Severe - Mild
1-Methylxanthine	34.692	1.95E-14	1.02E-13	Control - Mild; Control - Severe; Severe - Mild
Glutamine	32.899	8.68E-14	4.48E-13	Control - Mild; Control - Severe; Mild - Severe
3-Hydroxybutyric acid	32.047	1.78E-13	9.00E-13	Mild - Control; Severe - Control; Severe - Mild
5-Hydroxy-tryptophan	31.659	2.46E-13	1.23E-12	Mild - Control; Severe - Control; Severe - Mild
CAR(14:0)	31.491	2.84E-13	1.39E-12	Mild - Control; Severe - Control; Severe - Mild
Tetradecadienoic acid	31.473	2.88E-13	1.39E-12	Mild - Control; Severe - Control; Severe - Mild
CAR(18:0)	30.929	4.56E-13	2.16E-12	Mild - Control; Severe - Control; Severe - Mild
Hyodeoxycholic acid	30.853	4.86E-13	2.27E-12	Control - Mild; Control - Severe; Mild - Severe
N2,N2-Dimethylguanosine	30.702	5.53E-13	2.54E-12	Mild - Control; Severe - Control; Severe - Mild
CAR(5:0(OH))	30.59	6.08E-13	2.75E-12	Mild - Control; Severe - Control; Severe - Mild
Docosatetraenoic acid	30.326	7.61E-13	3.39E-12	Mild - Control; Severe - Control; Severe - Mild
Niacinamide	29.473	1.57E-12	6.89E-12	Mild - Control; Severe - Control; Mild - Severe
Sphingosine 1-phosphate	29.332	1.77E-12	7.63E-12	Mild - Control; Severe - Control; Mild - Severe
DG(36:3)	29.319	1.79E-12	7.63E-12	Mild - Control; Severe - Control; Severe - Mild
CAR(9:0)	29.019	2.31E-12	9.71E-12	Control - Mild; Control - Severe; Severe - Mild
Docosapentaenoic acid	28.758	2.89E-12	1.20E-11	Mild - Control; Severe - Control; Severe - Mild
Acetaminophen	28.486	3.65E-12	1.49E-11	Mild - Control; Severe - Control; Mild - Severe
Stearic acid	28.37	4.03E-12	1.62E-11	Mild - Control; Severe - Control; Severe - Mild
Myristoleic acid	28.319	4.21E-12	1.67E-11	Mild - Control; Severe - Control; Mild - Severe
Deoxyguanosine	28.054	5.29E-12	2.07E-11	Mild - Control; Severe - Control; Severe - Mild
CAR(2:0)	27.904	6.01E-12	2.33E-11	Mild - Control; Severe - Control; Severe - Mild
Phenyllactic acid	27.779	6.69E-12	2.56E-11	Mild - Control; Severe - Control; Severe - Mild
Lauric acid	27.707	7.12E-12	2.69E-11	Mild - Control; Severe - Control; Mild - Severe

Next, we explored the differences between the control group and patients with mild and severe COVID at timepoint 1 using pairwise Student's t-tests and fold change analysis. We found that there were 73 significantly (FDR < 0.05) differential metabolites between mild and severe COVID groups, 192 differential metabolites between controls and mild COVID, and 222 differential metabolites between controls and severe COVID (Supplementary Table 3.2, **Figure 3.3 A-C**). Notably, levels of several fatty acyls were higher in COVID patients, with severe COVID patients having higher levels of these than mild COVID patients. Similarly, levels of organic acids and derivatives are also higher in COVID patients. These results were consistent with the linear regression model that included disease status along with clinical covariates.

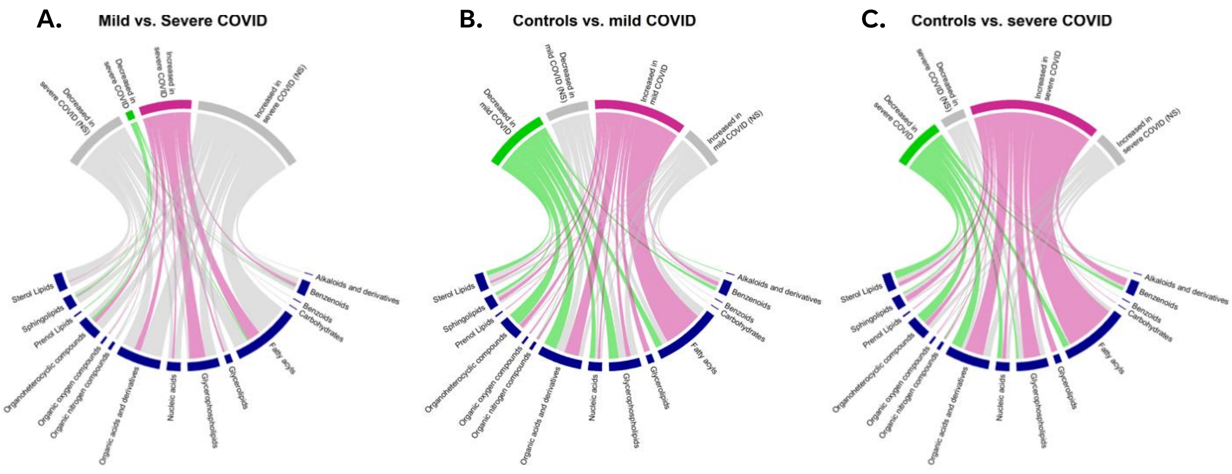


Figure 3.3: Chord diagrams illustrating the differential status of the classes of metabolites based on t-tests performed in mild vs. severe COVID **(A)**, controls vs. mild COVID **(B)** and controls vs. severe COVID **(C)**. The thickness of connections and sectors is proportional to the number of metabolites. Pink color indicates increased levels while green color indicates decreased levels of metabolites in COVID (mild or severe) patients. Gray color indicates metabolites that are not significantly differential.

3.3.2.3 Network Analysis of Metabolomics Data

To examine the relationships among differentiating metabolites and to visualize the metabolic changes we used a data-driven approach that uses partial correlations. The Debiased Sparse Partial Correlation (DSPC) algorithm implemented in the CorrelationCalculator[109] program allows the estimation of partial correlations in a high-dimensional setting (n << p) under the assumption that the true connectivity among the metabolites is much smaller than the sample size i.e., sparse. Under this assumption, DSPC reconstructs a graphical model and provides partial correlation coefficients and *P*-values for every pair of metabolic features in the dataset. Thus, DSPC allows discovering connectivity among large numbers of metabolites using fewer samples.

We constructed a partial correlation network for 294 metabolites based on metabolomics measurements for the control as well as mild and severe COVID groups at timepoint 1. (**Figure 3.4**). The resulting partial correlation network contained 282 metabolites and 422 edges with an FDR-adjusted p-value < 0.05. Two metabolites were excluded due to low correlation with other metabolites in the dataset.

Partial correlation networks have been shown to recapitulate known relationships between metabolites, i.e., biochemically and structurally related metabolites tend to cluster together. In order to visualize the resulting networks we imported them into Cytoscape[127]. Cytoscape allows to create custom visualizations displaying experimental changes in the network context. The levels of fatty acids and DAGs were increased in both mild and severe COVID. The LPEs and carnitines were increased in the severe group. The levels of several purines and bile acids were lower in COVID patients.

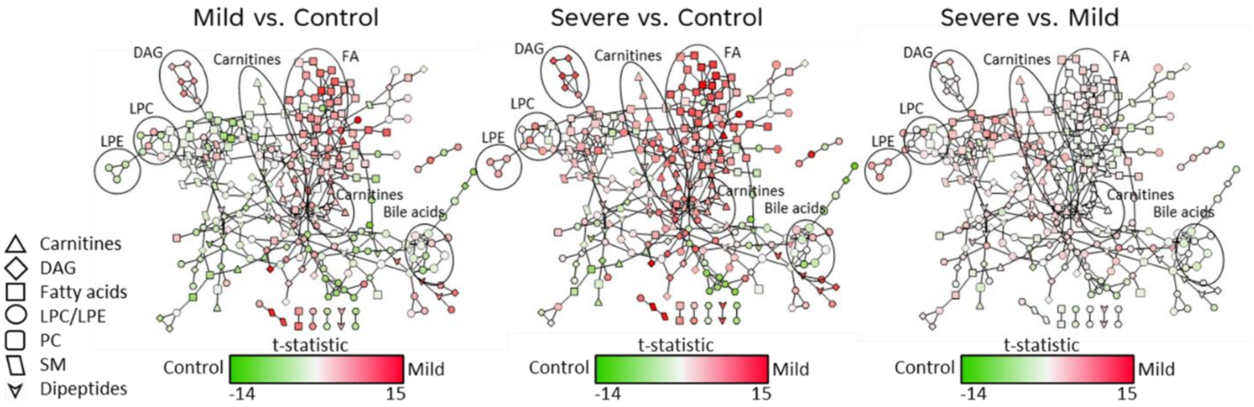


Figure 3.4: Partial correlation network of the plasma metabolome. The nodes in the networks are colored based on the value of the t-statistic in the respective pairwise Student's t-test. Nodes with a bold border represent significantly differential metabolites (FDR < 0.05). The thickness of the edges is based on the adjusted p-value of the partial correlation. Thicker edges represent greater statistical significance.

3.3.2.4 Identification of metabolic markers of COVID severity

To determine the ability of metabolomics to identify patients with severe disease, we constructed parsimonious random forest models to predict severe COVID-19. In order to avoid any potential confounding, we excluded patients who received propofol, leaving 144 patients who were included into this analysis. We compared the predictive performance of traditional COVID-19 risk factors such as age, BMI, race, gender, and diabetic status to the 73 differentially expressed metabolites in discriminating mild and severe COVID at timepoint 1. The samples were divided into training (70%) and test (30%) sets and random forest classifier was trained on the training data and its performance was measured on the test data as the area under the receiver-operator curve (ROC-AUC). We found that the model built with the differential metabolites has a much better performance (AUC-ROC = 0.885 ± 0.054) than the model built with clinical COVID risk factors (0.677 ± 0.084) (**Figure 3.5A**). The top 20 most important variables primarily included short-chain carnitines (C3-C5), long-chain diglycerides, phosphatidylcholines, and lysophosphatidylethanolamines. Interestingly, the dipeptide Phe-Phe also featured in the top 20 metabolites despite having a strong association with propofol administration (**Table 3.3; Figure 3.5B**).

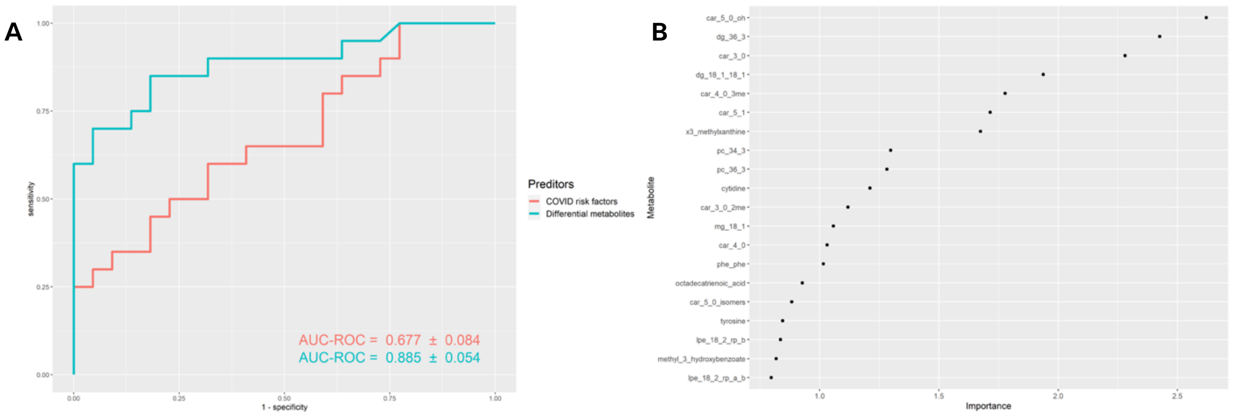


Figure 3.5: Predictive performance of differentially expressed metabolites and clinical COVID risk factors. (**A**) ROC of the random forest models. The orange curve represents the model constructed with COVID risk factors while the blue curve represents the model constructed with the differential metabolites. The area under the receiver-operator curve (AUC-ROC) is given as AUC ± SD. (**B**) Top 20 metabolites ranked by their contribution to classification accuracy, given by the mean decrease in the Gini index.

Table 3.3: Top 20 important metabolites in the Random Forest model ranked by the mean decrease in the Gini index.

Metabolite	MeanDecreaseGini	MeanDecreaseAccuracy
CAR 5:1	2.586	6.661
CAR 5:1 OH	2.332	8.651
MG 18:1	2.239	6.601
DG 18:1/18:1	2.16	6.987
DG 36:3	2.101	6.168
CAR 3:0	2.014	6.69
3-Methyxanthine	1.876	6.185
CAR 4:0 3me	1.301	5.251
CAR 4:0	1.231	4.623
Phe-Phe	1.165	3.563
PC 34:3	1.163	5.73
Arachidic acid	1.084	4.192

CAR 3:0 2me	0.982	2.658
LPE 18:2 (A)	0.955	3.884
5-Methylthioadenosine	0.89	3.918
PC 36:3	0.853	0.783
CAR 5:0 isomers	0.851	3.417
LPE 18:2 (B)	0.805	2.877
Retinoic acid	0.796	1.719
Behenic acid	0.751	3.65

3.3.2.5 Assessing metabolic difference associated with disease progression

We compared the metabolomes of COVID patients between the two timepoints. While there were only 2 significantly differential (FDR < 0.05) metabolites between timepoint 1 and 2 in patients with mild COVID, there were 44 differential metabolites across the two timepoints in patients with severe COVID (**Table 3.4**). The levels of 42 out of the 44 metabolites decreased at timepoint 2.

Table 3.4: Differential metabolites (FDR < 0.05) timepoints 1 and 2 among patients with severe COVID-19.

Metabolite	t-statistic	p-value	FDR
Phenyllactic acid	5.7147	2.96E-08	8.71E-06
Phenylacetic acid	5.1043	6.36E-07	9.34E-05
Phe-Phe	4.9689	1.21E-06	0.000119
Kynurenine	4.662	4.98E-06	0.000366
CAR(4:0(3Me))	4.511	9.72E-06	0.000571
Hydroxyphenyllactic acid	4.3905	1.64E-05	0.000602
Aminobutyric acid	4.4128	1.49E-05	0.000602
3-Hydroxybutyric acid	4.4489	1.27E-05	0.000602
Guanine	4.2461	3.02E-05	0.000985
CAR(5:0) isomers	4.1912	3.79E-05	0.001114
Arachidic acid	4.0549	6.61E-05	0.001766
Behenic acid	3.9754	9.08E-05	0.002224
CAR(5:1)	3.9084	0.000118	0.002672
LPE(18:2)_rp_a	3.7552	0.000213	0.004475
Lenticin	3.6755	0.000287	0.004972
CAR(5:0)	3.6756	0.000287	0.004972
Pyridoxamine	3.6902	0.000272	0.004972
Niacinamide	-3.6278	0.000343	0.005604
LPE(18:2)_rp_a_b	3.5115	0.000524	0.006417
bis(2-Ethylhexyl)phthalic acid	3.5132	0.000521	0.006417
Octadecadienoic acid	3.5207	0.000507	0.006417
Lignoceric acid	3.5259	0.000498	0.006417
Indoleacrylic acid	3.542	0.000469	0.006417

MG(18:1)	3.5699	0.000424	0.006417
Tetradecadienoic acid	3.4366	0.000684	0.008045
gamma-Glutamylmethionine	3.3916	0.000801	0.008726
LPE(18:2)_rp_b	3.399	0.000781	0.008726
Eicosenoic acid	3.2824	0.001168	0.011841
DHA	3.2864	0.001152	0.011841
Mesobilirubinogen	3.1945	0.001571	0.015395
L-Urobilin	3.1311	0.001937	0.017259
LPC(18:2)_rp_a	3.1316	0.001934	0.017259
Docosenoic acid	3.1368	0.001901	0.017259
Hydroxykynurenine	3.1174	0.002026	0.017522
Tyrosine	3.0708	0.002358	0.019804
Taurolithocholic acid	3.0223	0.002755	0.0225
Ile-Val	3.0134	0.002835	0.022526
Leucine/Isoleucine	2.9718	0.003233	0.025016
Hyocholic acid	2.9348	0.003631	0.027375
Isoleucine	2.9191	0.003813	0.028024
Lysine	2.8344	0.004946	0.033815
PC(35:2)	2.8367	0.004912	0.033815
LPC(20:3)_rp_a	2.8488	0.004734	0.033815
2-Deoxy-glucose	2.8042	0.005419	0.036207
Thyroxine	-2.7043	0.007289	0.038012
Hippuric acid	2.7006	0.00737	0.038012
DG(36:3)	2.7007	0.007367	0.038012
Cytidine	2.7027	0.007324	0.038012
Eicosadienoic acid	2.706	0.007253	0.038012
Margaric acid	2.712	0.007128	0.038012
LPC(18:2)_rp_a_b	2.7129	0.007109	0.038012
5-Hydroxy-tryptophan	2.7333	0.006694	0.038012
LPC(16:0)_rp_a	2.7404	0.006555	0.038012
Docosapentaenoic acid	2.7419	0.006526	0.038012
Nonadecenoic acid	2.7431	0.006502	0.038012
Sphingosine	2.7455	0.006456	0.038012
SM(d32:1)	2.7612	0.006164	0.038012
Pyroglutamic acid	2.6496	0.008544	0.043309
2-Hydroxy-3-methylbutyric acid	2.605	0.009707	0.048372
DG(18:1_18:1)	2.5922	0.010067	0.049327

3.4 Discussion

This study investigated changes of plasma metabolome of patients with mild and severe COVID-19, compared to healthy controls. In our patient population the incidence of type 2 diabetes was higher in the severe group compared to the group who had mild COVID-19. Patients with severe COVID-19 had higher BMI. It has been shown that certain chronic comorbidities, such as hypertension, cardiovascular disease, obesity, diabetes, and kidney disease, are highly prevalent in people with COVID-19. While these comorbidities do not appear to increase the risk of developing COVID-19, they are associated with an increased risk of a more severe case of the condition as well as mortality[199]. To account for the influence of age, race, gender, BMI, T2D and propofol administration, we built a multiple linear regression model. We found that 160 metabolites were significantly associated with one or more covariate. We also found that the administration of propofol induces profound metabolic changes especially affecting lipid metabolism (201 out of 292 metabolites were associated with propofol with p-value < 0.05). Adjusting the data for these factors allowed us to focus on metabolic changes associated with disease severity.

We found that several classes of lipids, including fatty acids and acylcarnitines were increased in COVID patients, especially in the severe group. This is consistent with previous findings by Thomas et al. who also found that these were elevated in patients with COVID-19. Further, these elevations were often more pronounced in older patients and those with higher levels of IL-6[200]. Several other studies have made similar observations with regard to acylcarnitines and fatty acid levels with COVID-19[201,202]. These observations may reflect an inability of these patients to mount an adequate metabolic response[203]. Carnitine is vital for moving long-chain fatty acids into the mitochondria to undergo beta-oxidation and dysregulation of this process could cause an increase in the plasma concentrations of these compounds[204]. When acylcarnitines cannot be oxidized in the mitochondria they can be exported from the cell into the circulation[205]. In contrast to acylcarnitines we found that the level of L-carnitine was lower in both COVID groups.

The majority of plasma bile acids were lower in COVID patients than in controls. This could be seen at odds with the common observation that bile acids and other products normally excreted by the liver tend to accumulate in critically ill patients[206]. However, adjusted levels of taurolithocholic acid were elevated in our study and this compound was the only primary bile acid that was associated with COVID-19 severity. The other bile acids that were decreased in both mild and severe patients were secondary bile acids, meaning they were all dehydroxylated by gut bacteria and subsequently resorbed into the blood via the enterohepatic circulation. A similar pattern was recently noted in patients with acute respiratory response distress syndrome in which primary bile acids increased early in the course of the disease but secondary bile acids levels in the serum remain unchanged[207]. Also, plasma concentrations of secondary bile acids may have a direct role in the outcome of patients with COVID-19 since it has been postulated that secondary bile acids, such as chenodeoxycholic and ursodeoxycholic acid, may bind SARS-CoV-2 to angiotensin-converting enzyme 2, preventing it from infecting cells[208–210].

Tetracosenoic acid, also known as nervonic acid, was one of the top 5 differential metabolites both in mild and severe COVID groups. Previous studies have shown tetracosenoic acid to have a protective effect for patients with metabolic disorders[211]. Thyroxine was also lower in both disease groups in our study. This is consistent with other studies, including a metanalysis, which found that low thyroxine levels were associated with hospital mortality[212].

Next, we tested the ability of metabolites to identify patients with more severe disease. We found that the metabolites were much better than patient characteristics at identifying individuals with more severe disease with the performance of the random forest model having an AUC of 0.885. This suggests that metabolic alterations play a significant role in the early response to COVID-19 and support the use of metabolomics to uncover the mechanisms of these diseases. This is in line with other studies that have shown that the metabolome of patients with COVID-19 to be strongly predictively of disease severity[195]. While not addressed in this study, others have also demonstrated that metabolomic profiles may be able to identify patients at risk for developing severe disease when measured prior to infection[213].

Several limitations of this study should be considered when interpreting the results. First, the onset of disease was unknown for this study population. Some patients may have had symptoms for days before being admitted to the hospital while others may have experienced a rapid progression of their symptoms that necessitated hospitalization. This is a common problem in the study of acute infections and related sequelae such as sepsis and septic shock. The inclusion of two timepoints for each COVID-19 subjects allows some identification and features that are important early or late in disease. However, no baseline exists for subjects and the exact date of infection with SARS-CoV-2 or the timing of symptoms is not available for the study population. Second, as is true with any retrospective study, a lack of randomization means that results could be biased by any unmeasured confounds. Third, these subjects were recruited during the early stages of the COVID-19 outbreak. Therefore, vaccines and treatments that are currently in common use were being used. This limits applicability of findings to current patients.

In conclusion, the plasma metabolome of patients with COVID-19 can be used to predict disease severity. Future studies are needed to determine if these relationships hold true with recently developed antiviral treatments and in other similar situations such as bacterial and fungal sepsis.

CHAPTER IV

Body Mass Index Associates with Amyotrophic Lateral Sclerosis Survival and Metabolomic Profiles

The materials presented in this chapter have been accepted for publication as: Stephen A Goutman, Jonathan Boss, **Gayatri Iyer**, Hani Habra, Masha G Savelieff, Alla Karnovsky, Bhramar Mukherjee, Eva L Feldman (2022). "Body mass index associates with amyotrophic lateral sclerosis survival and metabolomic profiles". *Muscle and Nerve*. this Chapter also includes the unpublished data on the association of BMI-related metabolic modules with ALS survival.

4.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a very rare and unpredictable neurological disease that causes progressive loss of muscle function and leads to loss of life. Currently, there are medications that slow progression but there are no lifesaving therapies. ALSdiagnosis is preceded by a pre-symptomatic phase, characterized by initiation of the disease process but lacking pronounced clinical symptoms [214–216]. ALS patients frequently suffer from a rapid decrease in body mass index (BMI) and the rate of loss early in the disease course is a strong prognostic factor[217]. Therefore, BMI loss may reflect an early and pre-symptomatic manifestation of disease. Indeed, individuals with ALS develop BMI loss many years before symptom onset[218]. Additionally, lower BMI earlier in life may both increase ALS risk [218–222] and decrease ALS survival [218,223].

BMI decreases in ALS patients are linked to lower energy intake from dysphagia and higher energy expenditure[224,225], including hypermetabolism, altered glucose and lipid metabolism, and mitochondrial dysfunction[226]. Perturbations in metabolism in ALS are supported by correlations in basic lipid profiles with risk and outcomes. Increased low-density lipoprotein cholesterol (LDL-C) and apolipoprotein B levels years prior raise risk of ALS onset[227] or at diagnosis correlate with longer ALS survival[228]. However, basic lipid profiles do not capture the full spectrum of metabolic changes that occur in the disease. Rather, the metabolome and lipidome, the cumulative profile of all metabolites and lipids, may more comprehensively reflect the metabolic state. Indeed, metabolomics profiles correlate with BMI[229–231] and disease phenotypes, such as cardiometabolic risk[229,230]. Metabolomics signatures may one day be useful in combination with BMI as predictors of disease outcomes[229].

However, the correlation of BMI with metabolomics profile and disease outcomes has not been investigated in ALS. Thus, our goal in this current study was to leverage our case/control study to examine trends in BMI trajectory in ALS versus control participants correlated to survival and metabolomics profile.

4.2 Methods

4.2.1 Participants and Samples

Recruitment and data collection procedures are published[232–235]. Briefly, all patients seen at the Pranger ALS Clinic at University of Michigan with an ALS diagnosis, age > 18 years, and ability to consent in English were asked to participate. Neurologically healthy controls, recruited through population outreach, completed the same procedures. All participants provided oral and written informed consent and the study was approved by the Institutional Review Board. Demographic characteristics and available prior heights and weights from the medical records of the participants were obtained, as were ALS disease characteristics such as Revised El Escorial criteria (rEEC)[236]. Participants were asked to self-report height in feet and inches and weight in pounds 10 years ago, 5 years ago, and at the present time. For ALS participants, present weight was typically equivalent to weight at diagnosis since enrollment occurred shortly after diagnosis. BMI was calculated from height and weight as follows: $weight(kg)/[height(m)]^2$ [237]. ALS participants with an interval of more than 5 years from symptom onset to a diagnosis were not included in the analysis as the goal was to investigate pre-symptomatic differences in BMI. A subset of participants provided plasma for metabolomics analysis, as published[238,239].

4.2.2 Descriptive Analysis

Descriptive statistics were calculated for demographic characteristics including age, sex, onset segment, and disease duration (time from symptom onset to diagnosis). Study population differences were compared between BMI groups by analysis of variance tests and chi-square tests. Lin's concordance correlation coefficient quantified agreement between available self-reported and measured BMIs.

4.2.3 BMI Progression Analysis and Group Assignment

Generalized estimating equations (GEE) with unstructured correlation structure assessed differences in BMI changes for ALS and control participants, while accounting for within-participant correlation between self-reported BMI measurements[240]. The GEE outcome was self-reported BMI, and the covariates were interaction terms between ALS/control status and the three time points adjusted for age and sex at study entry. Differences in average BMI rate of change between ALS and controls were assessed with the Wald test and performed with the R geepack package[241].

After subtracting self-reported BMI 10 years prior to consent from all timepoints, k-means clustering for longitudinal data (kml R package[242]) grouped ALS cases based on their self-reported changes in BMI, for use in ALS survival models. This subtraction step ensured that the k-means procedure clustered exclusively on BMI changes over time, rather than differences in baseline BMI. After considering 2-6 clusters, the selected number of clusters maximized the Calinski and Harabasz criterion[243] a measure of between cluster variation relative to within-cluster variation for longitudinal data[244]. The distance metric used for clustering was Euclidean distance with Gower adjustment[244].

4.2.4 Survival Analysis

Kaplan Meier plots of survival from diagnosis by cluster were produced. Cox proportional hazards models determined associations between cluster groups and ALS survival, defined as the time from diagnosis to death. Associations were adjusted for sex, age, baseline BMI (i.e., 10 years prior), onset segment, diagnosis rEEC, and time from symptom onset to diagnosis. Proportional hazards assumptions were checked using global and individual Schoenfeld tests with graphical assessment of the rescaled Schoenfeld residuals over time. Due to proportional hazards violations in some models, accelerated failure time (AFT) models were constructed.

4.2.5 Sensitivity Analyses

Two sensitivity analyses were performed: (i) As some participants did not provide BMI data during the study period, a sensitivity analysis for missing data was performed with inverse probability weighting and models were rerun using this weighted dataset; (ii) As BMI is an ALS prognostic factor, the participant's reference/baseline BMI (10 years prior to study entry, *i.e.*, 10-year BMI) and clustering trajectory by groupings was captured. BMI was divided by tertiles, and clustering provided three trajectories creating a total of 9 groups, designated as cluster*BMI groups.

4.2.6 Missing Data

To handle missing BMI trajectories, inverse probability weighted complete data analysis was performed for all models described in the methods, since BMI trajectories for almost all participants were either fully observed (381 cases, 266 controls) or completely missing (306 cases, 30 controls). Weights were constructed by modeling the probability of having an observed BMI trajectory with case and control stratified generalized additive models using the R mgcv package[245]. Stratification by case or control status was performed so that Revised El Escorial criteria (rEEC) at study entry, time between symptom onset and diagnosis, and onset segment could be included as covariates in the generalized additive model. Both generalized additive models adjusted for age and sex at study entry. The weights were calculated by inverting the estimated probability of having an observed BMI trajectory obtained from the generalized additive models[246]. The proposed weighting scheme is an alternative strategy to multiple imputation for handling missing data, where individuals are weighted by how likely they are to have a missing BMI trajectory based on their age, sex, and ALS clinical characteristics[246].

4.2.7 Metabolomics Data Analysis

Non-fasting plasma samples from ALS participants were analyzed by Metabolon (Morrisville, NC) and previously published as case-control analyses[238,239]. Samples were run in two separate batches and batch effect was corrected by Z-normalization. Within each batch, each metabolite's values were mean centered and scaled by the standard deviation to produce a normal distribution (*i.e.*, autoscaled). Then, datasets were merged by shared compound IDs, and adjusted by age and sex.

Student's t-tests were performed to identify differential metabolites between the three BMI trajectory groups (decreasing, flat, and increasing). Pearson's correlations were calculated between metabolites and current BMI. In both tests, the statistical significance was determined using a false discovery rate (FDR)-adjusted p-value < 0.05. Lasso regression was performed using the R package glmnet[247] to select metabolites associated with BMI trajectory. Lasso[138] (least absolute shrinkage and selection operator or Lasso) is a $l1$-regularized linear regression model that is used for covariate selection when the number of covariates (in this case, metabolites) is much larger than the available sample size. The $l1$-penalty shrinks the coefficients of uninformative covariates to zero, thereby excluding them. Ten-fold cross-validation was performed to select the tuning parameter that minimized cross-validation error. The final model was generated by re-fitting the lasso model with the selected tuning parameter value.

Next, we constructed partial correlation networks from the metabolic profiles of ALS participants to infer direct metabolic interactions under disease condition. Partial correlation is the conditional dependence between a pair of variables, given all the other variables. This eliminates potentially spurious indirect associations between metabolites. Prior studies have demonstrated that metabolic modules derived from correlation networks contain biochemically and functionally related metabolites[105,126,248]. We utilized the Debiased Sparse Partial Correlation (DSPC) algorithm implemented in the CorrelationCalculator[109] program that allows the estimation of partial correlations in a high-dimensional setting ($n \ll p$), as in the case in this study. The assumption made is that the true connectivity among the metabolites in a biological context is much smaller than the sample size *i.e.*, the connectivity is sparse. Significance of the partial correlation between a pair of metabolites *i.e.*, edges in the partial correlation network was defined as an FDR-adjusted p-value < 0.1.

The partial correlation network was then clustered using a consensus clustering[249] approach to obtain densely connected metabolic modules. Consensus clustering integrates multiple graph-clustering solutions, thereby generating more robust modules. Consensus clustering employs the following seven graph clustering algorithms implemented in the igraph R package (https://igraph.org/): `cluster_edge_betweenness()`, `cluster_fast_greedy()`, `cluster_infomap()`, `cluster_label_prop()`, `cluster_leading_eigen()`, `cluster_louvain()`, `cluster_walktrap`(). The final module assignment is decided based on the consensus of the graph partitions from each of these algorithms.

Next, we tested the association between the metabolic modules and BMI clusters (increasing and decreasing BMI) in group-penalized lasso (group lasso) regression models using the gglasso R package. Group lasso[250] is a special case of lasso regression where covariate-selection is performed on a group-level rather than on individual covariates, under a sparse setting. Here, the grouping structure information is provided by the metabolic module assignment of the metabolites. Mathematically, group lasso solves the following optimization problem:

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \|\beta_l\|_2 \right)$$

Here, y is the response vector (that can be a continuous or binary variable) consisting of N observations, X is the design matrix with dimensions $N\, x\, p$, where p is the number of covariates divided into L groups, p_l is the number of covariates in group l, with β_l being the corresponding vector of beta coefficients. λ is the tuning parameter that controls the degree of sparsity of the beta coefficients. Ten-fold cross-validation was performed to select the tuning parameter λ^* that minimizes cross-validation error. The final model was selected by re-fitting the group lasso model with λ^*.

All analyses were performed using R statistical software version 4.0.2.

4.3 Results

4.3.1 Participants

For those with observed BMI at all three timepoints, ALS participants represented a typical patient population, according to onset age, distribution of segment onset, among other variables. Controls (n=266) were slightly younger than cases (n=381) (61.3 vs. 64.9 years p < 0.001) (**Table 4.1**). Two ALS participants with an uncertain onset segment and one control with a BMI greater than 100 kg/m^2 labeled as an outlier were removed from subsequent analysis. Demographics for ALS and control participants with missing data are detailed in **Table 4.2**. The analysis results for missing data for BMI trajectories in ALS versus control participants were similar to the analysis of participants with complete data. The differences in BMI at -5 years versus 0 years was 1.75 kg/m^2 for ALS cases (95% CI: 1.32 kg/m^2 to 2.19 kg/m^2; p = 3x10^{-15}) and 0.00 kg/m^2 for controls (95% CI: -0.39 kg/m^2 to 0.38 kg/m^2; p = 1) for the analysis of missing data.

Table 4.1: Participant Demographics

Covariate	Overall (n=647)	ALS cases (n=381)	Controls (n=266)	P-Value
Age at survey consent (years)	63.3 (56.5-69.9)	64.9 (57.6-71.4)	61.3 (55.2-68.2)	<0.001
Sex				0.143
Female	317 (49.0)	177 (46.5)	140 (52.6)	
Male	330 (51.0)	204 (53.5)	126 (47.4)	
Last contact event				NA
Death		251 (64.9)	NA	
Censored		130 (34.1)	NA	
Original and/or Revised El Escorial criteria				NA

Possible/Suspected	53 (13.9)	NA	
Probable, LS	104 (27.3)	NA	
Probable	123 (32.3)	NA	
Definite	101 (26.5)	NA	
Onset segment			NA
Bulbar	113 (29.7)	NA	
Cervical	126 (33.1)	NA	
Lumbar	142 (37.3)	NA	
Time between symptom onset and diagnosis (years)	1.01 (0.64-1.66)	NA	NA

For continuous variables, median (25th – 75th percentile); for categorical variables, N (%). P-values for continuous and categorical variables correspond to analysis of variance tests and chi-squared tests, respectively.
ALS, amyotrophic lateral sclerosis; LS, laboratory supported; NA, not applicable.

Table 4.2: Participant Demographics with Missing BMI Data

Covariate	**Overall (n=336)**	**ALS cases (n=306)**	**Controls (n=30)**	**P-Value**
Age at entry (years)	61.7 (53.7-69.7)	62.1 (54.3-70.8)	59.4 (51.1-62.6)	0.008
Sex				0.140
Female	164 (48.8)	145 (47.4)	19 (63.3)	
Male	172 (51.2)	161 (52.6)	11 (36.7)	
Last contact event				NA
Death		223 (72.9)	NA	
Censored		83 (27.1)	NA	
Revised El Escorial criteria				NA
Possible/Suspected		31 (10.1)	NA	
Probable, LS		69 (22.5)	NA	
Probable		98 (32.0)	NA	
Definite		108 (35.3)	NA	
Onset segment				NA
Bulbar		97 (31.7)	NA	
Cervical		98 (32.0)	NA	
Lumbar		111 (36.3)	NA	
Time between symptom onset and diagnosis (years)		0.99 (0.64-1.54)	NA	NA

BMI, body mass index; LS, laboratory supported; NA, not applicable.

4.3.2 BMI Trends in Cases Versus Controls

The Lin's concordance correlation coefficient examined whether self-reported BMI was similar to measured BMI abstracted from medical records. Abstracted BMI was available at the -5 and 0 timepoints. The Lin's concordance correlation coefficient for the self-reported BMI at -5 years was 0.952 and at enrollment was 0.966, indicating participants with previously measured BMIs accurately recalled their weights, therefore minimizing recall bias. Thus, self-reported weights were used due to availability of a larger sample size. Lin's concordance correlation coefficient showed consistency between self-reported and measured BMI values. The mean BMI (±SD, number of observations (n)) for ALS cases at -10, -5, and 0 years was: 27.3 kg/ m^2 (±5.19, n=373), 28.0 kg/m^2 (±5.35, n=377), and 26.3 kg/m^2 (±4.84, n=381), respectively. For controls these were 26.5 kg/m^2 (±4.96, n=265), 27.6 kg/m^2 (±5.51, n=266), and 27.6 kg/m^2 (±5.36, n=266), respectively. ALS and control participants reported BMI increases in the 10- to 5-year period prior to study entry (**Figure 4.1A**). Unlike controls, however, ALS cases had an overall BMI decrease in the 5-year prior to study entry time window. The age- and sex-adjusted GEE model showed average ALS BMI change from -5 to 0 years was 1.75 kg/ m^2 (95% CI: 1.35 kg/m^2 to 2.16 kg/m^2; $p < 1x10^{-17}$) but was only 0.02 kg/ m^2 for controls (95% CI: -0.35 kg/ m^2 to 0.40 kg/ m^2; p = 0.9). Thus, ALS participants report BMI loss occurring 5 years before diagnosis/study entry, while control participants had no significant BMI change during the same timeframe. The kml algorithm applied to the ALS participant BMI trajectories generates 20 random starting conditions to ensure that the clustering results are robust to various initial algorithmic configurations. Of the 20 random starting conditions and across the different numbers of candidate clusters, the maximal value of the Calinski and Harabasz criterion corresponded to one of the three cluster partitions. Therefore, the analysis proceeded with three clusters. The three BMI trajectory clusters can be qualitatively described as: participants with an overall decrease in BMI (decrease group), participants with an overall slight decrease in BMI (mild decrease group), and participants with an overall increase in BMI (increase group) (**Figure 4.1B**, **Table 4.3**).

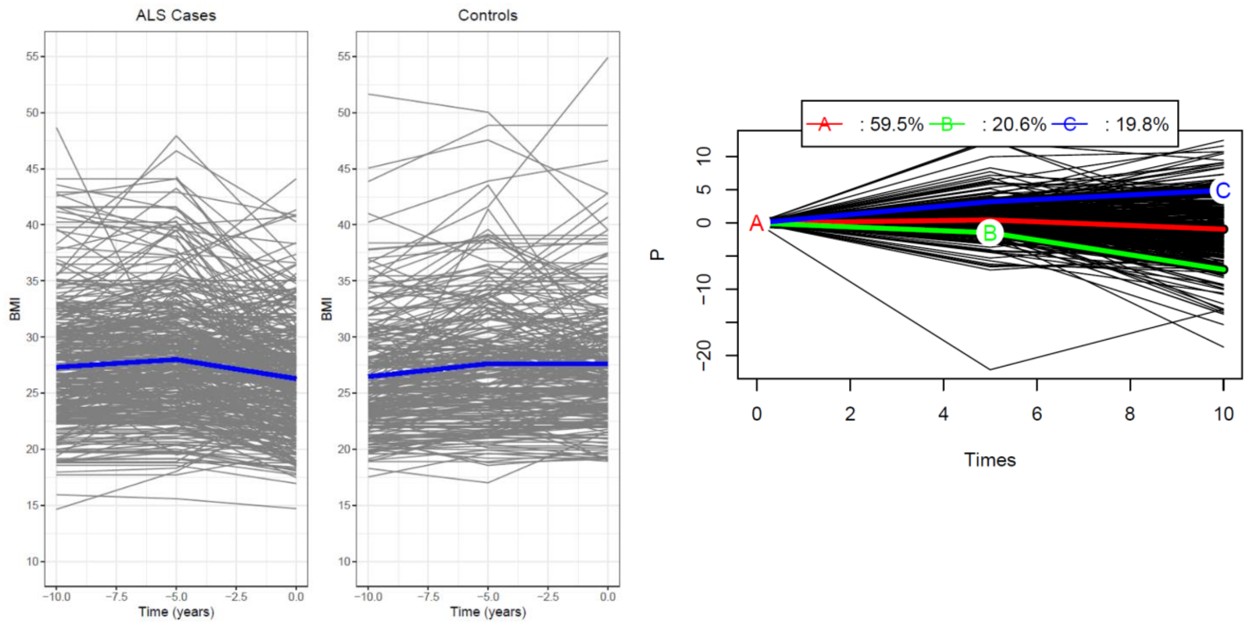


Figure 4.1 (A) BMIs for ALS and Control Participants. Spaghetti plots of BMI calculated from self-reported height and weight, for ALS and control participants at 10 years prior, 5 years prior, and at study entry. Blue line indicates the mean BMI. **(B) ALS BMI Clusters at -10, -5, and 0 Years.** Longitudinal BMI trajectory cluster for ALS participants shows three groups labeled as "increase" (19.8% of participants, blue line, C), "mild decrease" (59.5% of participants, red line, A), and "decrease" (20.6% of participants, green line, B). Individual BMI trajectories are shown by spaghetti plot (black lines). Y-axis shows difference of BMI at 10 years prior, 5 years prior, and 0 years prior compared to BMI at 10 years prior (reference).

Table 4.3: Participant Demographics by Cluster

	BMI Trajectory		
	Decrease (n=77)	**Mild decrease (n=222)**	**Increase (n=74)**
Age at entry (years)	67.9 (62.4-71.6)	65.0 (57.7-72.0)	58.3 (53.1-66.6)
BMI at entry	25.3 (21.6-27.8)	25.1 (21.9-27.7)	30.7 (27.4-34.0)
BMI 5 years before entry	30.3 (26.6-34.5)	26.2 (23.5-29.1)	28.1 (25.9-32.0)
BMI 10 Years before entry	31.3 (28.3-36.5)	25.7 (23.4-28.2)	25.2 (23.0-29.3)
Follow-up time (years)	1.25 (0.75-1.84)	1.87 (1.06-2.97)	1.62 (1.29-2.67)
Last contact event			
Death	55 (71.4)	141 (63.5)	49 (66.2)
Censored	22 (28.6)	81 (36.5)	25 (33.8)
Sex			
Female	35 (45.5)	93 (41.9)	44 (59.5)
Male	42 (54.5)	129 (58.1)	30 (40.5)
El Escorial Criteria			
Possible/Suspected	14 (18.2)	31 (14.0)	8 (10.8)
Probable	26 (33.8)	74 (33.3)	22 (29.7)
Probable, LS	16 (20.8)	61 (27.5)	25 (33.8)
Definite	21 (27.3)	56 (25.2)	19 (25.7)
Onset segment			

Bulbar	25 (32.5)	67 (30.2)	18 (24.3)
Cervical	26 (33.8)	80 (36.0)	17 (23.0)
Lumbar	26 (33.8)	75 (33.8)	39 (52.7)
Time between symptom onset and diagnosis (years)	1.03 (0.60-1.75)	1.00 (0.63-1.62)	1.03 (0.73-1.53)
Number of participants with metabolomics profiles	37	133	37

BMI, body mass index; LS, laboratory supported.

4.3.3 Survival Analysis

Unadjusted Kaplan-Meier survival analysis showed decreased absolute median survival times for the decrease BMI cluster (**Figure 4.2A**). Some Cox models violated proportional hazards by Schoenfeld residuals, so AFT models were constructed. After adjusting for age, sex, baseline BMI (i.e., 10 years prior), onset segment, rEEC, and time from symptom onset to diagnosis, participants in the decrease BMI cluster had a 27.1% shorter survival (95% CI: -42.6% to -7.3%; p = 0.010) versus the mild decrease group (**Figure 4.2B**, **Table 4.4**). Results were similar in missing BMI data sensitivity analyses with the accelerated failure time (AFT) models for participants with missing data (**Figure 4.3**, **Table 4.5**) showed a 25.4% reduction in survival for participants in the decrease BMI group (95% CI: -37.6% to -10.8%, p = 0.001).

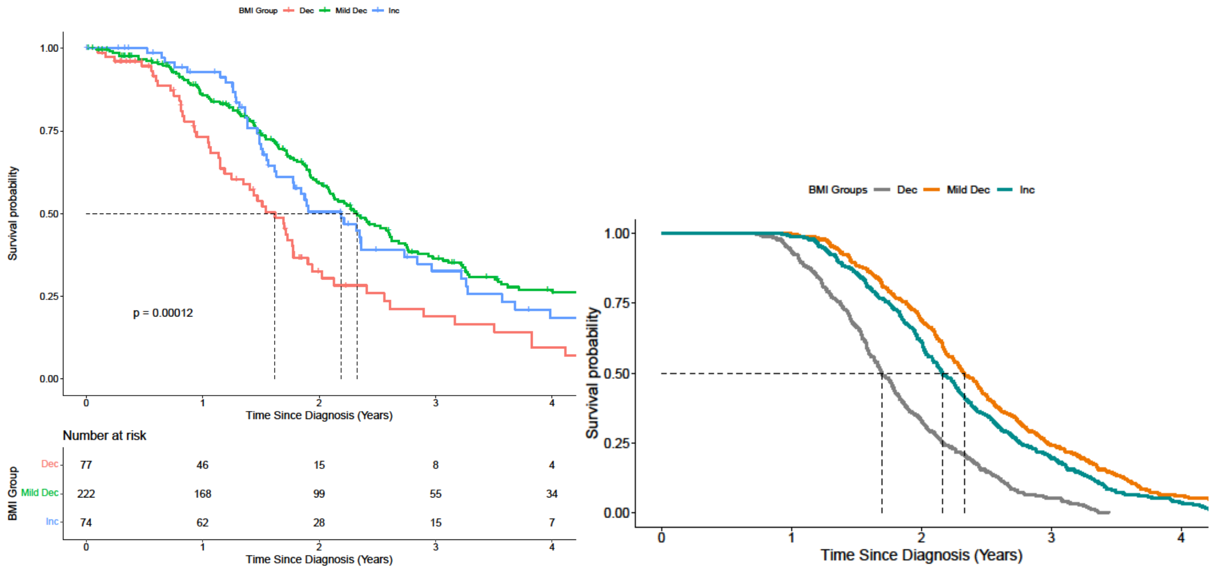


Figure 4.2: (A) Unadjusted Kaplan-Meier Survival Plots for BMI Cluster Groups. Kaplan-Meier survival plots for the body mass index (BMI) cluster groups. Median survival for the decrease group (red line) is 1.70 years. Median survival for the mild decrease group (green line) is 2.33 years. Median survival for the increase group (blue line) is 2.16 years. Difference in survival among all groups is significant (p = 0.00012). Difference in survival between decrease and increase groups is also significant (p = 0.0052). **(B) Accelerated Failure Time Model Plots.** Covariate adjusted survival curves corresponding to the unweighted accelerated failure time model with BMI cluster groups. The estimated median survival time is 1.7 years for the decrease BMI group, 2.33 years for the mild decrease BMI group, and 2.16 years for the increase BMI group.

Table 4.4: Accelerated Failure Time Model

	Percent Change in Survival	**LCL**	**UCL**	**P-Value**
Age at entry (years)	-1.0	-1.9	-0.2	0.016
Symptom onset to diagnosis (log years)	17.3	3.3	33.2	0.014
Baseline BMI	-1.0	-2.7	0.8	0.278
Decrease BMI trajectory	-27.1	-42.6	-7.3	0.010
Increase BMI trajectory	-7.1	-25.2	15.5	0.509
Male	0.1	-16.1	19.4	0.994
Cervical onset	41.0	13.0	76.0	0.002
Lumbar onset	21.3	-1.4	49.3	0.068
rEEC Possible/Suspected	88.3	41.9	149.7	0.000
rEEC Probable	23.4	-0.7	53.3	0.058
rEEC Probable, laboratory supported	61.6	28.5	103.1	0.000

BMI, body mass index; LCL, lower confidence limit; rEEC, revised El Escorial criteria;
UCL, upper confidence limit.

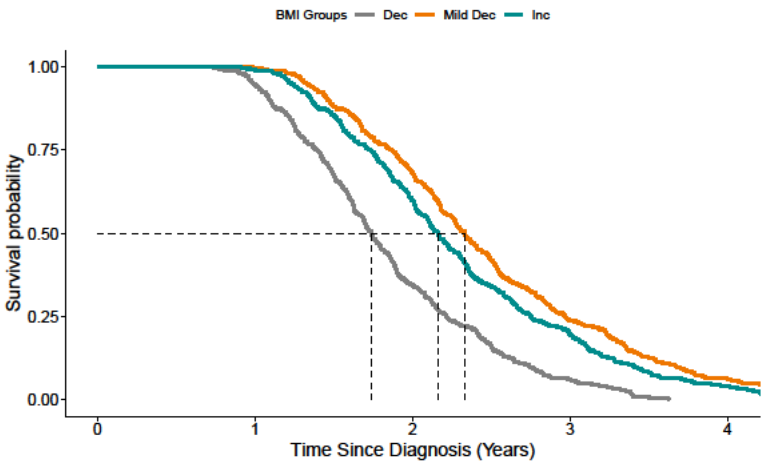


Figure 4.3: Accelerated Failure Time Model Plots for Incomplete Data Analysis. Covariate adjusted survival curves corresponding to the inverse probability weighted accelerated failure time model with the body mass index (BMI) cluster groups. The estimated median survival time for the decrease BMI group is 1.74 years, 2.16 years for the mild decrease BMI group, and 2.33 years for the increase BMI group.

Table 4.5: Accelerated Failure Time Model Sensitivity Analysis for Missing Data

	Percent Change in Survival	LCL	UCL	P-Value
Age at entry (years)	-1.0	-1.5	-0.4	0.001
Symptom onset to diagnosis (log years)	17.0	6.6	28.4	0.001
Baseline BMI	-0.9	-2.2	0.4	0.157
Decrease BMI trajectory	-25.4	-37.6	-10.8	0.001
Increase BMI trajectory	-7.3	-20.8	8.6	0.347
Male	-2.9	-14.7	10.5	0.653
Cervical onset	45.5	23.7	71.0	0.000
Lumbar onset	19.0	2.3	38.5	0.024
rEEC Possible/Suspected	92.1	55.1	137.9	0.000
rEEC Probable	23.2	5.6	43.8	0.008
rEEC Probable, laboratory supported	62.3	37.5	91.5	0.000

BMI, body mass index; LCL, lower confidence limit; rEEC, revised El Escorial criteria; UCL, upper confidence limit.

4.3.4 Metabolic modules associated with BMI trajectories.

Metabolomic differences by BMI cluster (decrease, mild decrease, increase) were investigated for the 207 participants with available previously published untargeted metabolomics[238,239]. The final curated dataset included 640 metabolites from plasma collected near the time of diagnosis. Differential analysis revealed no significant metabolites between decrease vs. mild decrease and increase vs. mild decrease groups. Only two metabolites (1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0) and behenoyl dihydrosphingomyelin (d18:0/22:0)) were significantly differential between the decrease and increase BMI groups. However, lasso regression did select nine metabolites that associated with the increase (odds ratio > 1) or decrease (odds ratio < 1) BMI trajectory groups (**Table 4.6**).

Table 4.6: Metabolites Associated with BMI Trajectory Groups from Lasso Regression Model

Metabolite	Odds Ratio	BMI Cluster
Behenoyl dihydrosphingomyelin (d18:0/22:0)	1.432	Increase
1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	0.719	Decrease
Glycocholate	1.182	Increase
N6-carboxymethyllysine	1.166	Increase
N-acetylglycine	0.891	Decrease
S-methylcysteine sulfoxide	0.949	Decrease
Myristoyl-linoleoyl-glycerol (14:0/18:2)	1.049	Increase
Undecanedioate (C11-DC)	1.035	Increase
Allantoin	1.026	Increase

The partial correlation network was constructed using recently published data from 349 ALS participants[239], of whom 207 were also in this analysis. The utilization of additional samples generated a more informative network since partial correlation methods are sensitive to sample size. The resulting partial correlation network contained 600 metabolites connected by 887 edges (FDR-adjusted p < 0.1), of which 31 had a negative partial correlation coefficient. Consensus clustering identified 26 metabolic modules spanning 555 highly connected metabolites. The remaining 45 metabolites did not cluster due to weak correlations leading to poor connectivity. Metabolic module size ranged from 5 to 66 metabolites.

Group lasso selected eight modules containing 152 metabolites, which associated with the decrease and increase BMI clusters (**Figure 4.4**, **Table 4.7**, **Table 4.8**), with odds ratios (OR) ranging from 0.92 to 1.1 (**Table 4.9**). The largest module 1 (47 metabolites) included ceramides and sphingomyelins, of which 36 had OR > 1, indicating associations with the increase BMI cluster. The second largest module 2 (30 metabolites) included primary and secondary bile acid metabolites, taurine and its derivatives, AMP, ADP, and sterols. Primary bile acids associated with the increase BMI cluster (OR > 1), while most secondary bile acids and taurine metabolites associated with the decrease BMI cluster (OR < 1). Module 3 (22 metabolites) primarily contained amino acid and nucleotide metabolites, half of which associated with the decrease BMI cluster. Module 4 (15 metabolites) was composed of plasmalogens, lyso-plasmalogens, and phosphatidylcholines, 11 of which associated with the decrease BMI cluster. Module 5 (13 metabolites) had mostly acyl carnitines, acyl amino acids, and some other amino acid metabolites, which mostly associated with the decrease BMI cluster. The remaining smaller module 6 (13 metabolites; sugar and nucleotide metabolites, xenobiotics, amino-sugar), module 7 and module 8 (6 metabolites each; xenobiotics, cofactors, vitamins, modified amino acids) contained various metabolites.

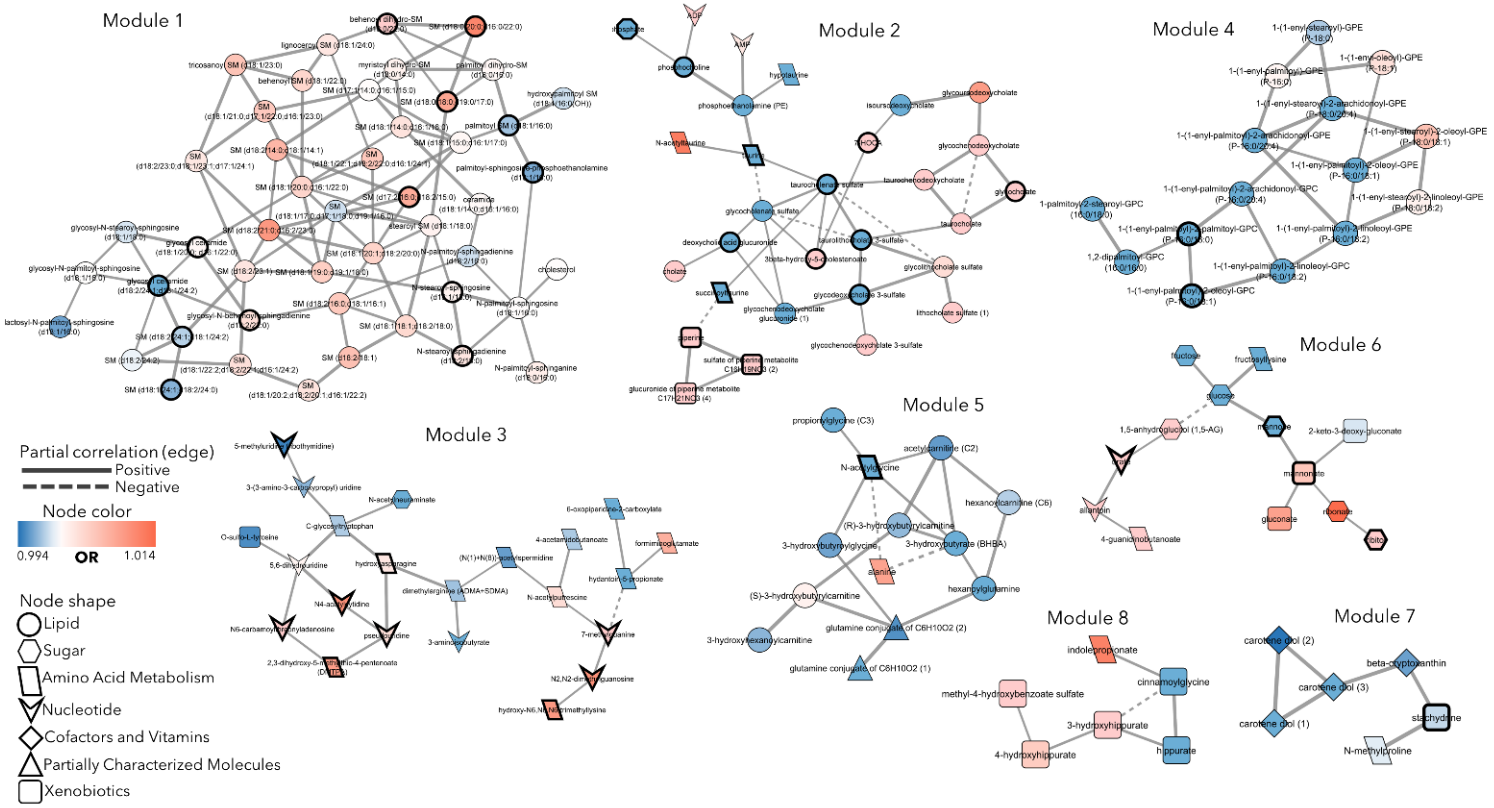


Figure 4.4: Metabolic Modules Associated with BMI Trajectory. Eight metabolic modules containing 152 total metabolites associated with BMI trajectory in group lasso regression models. Node color indicates odds ratio (OR) from group lasso; OR>1 indicates association with the increase BMI cluster (red node), OR<1 indicates association with the decrease BMI cluster (blue node). Nodes with a bold border significantly correlate with current BMI (FDR < 0.05). Node shape indicates the sub-pathway a metabolite belongs to. Solid edge between metabolites indicates positive partial correlation coefficient, dashed edge indicates negative partial correlation coefficient.

Table 4.7: Metabolic Modules Associated with BMI Trajectory Groups from Group Lasso Regression Model

Metabolic module	Number of nodes (metabolites)	Number of edges	Average degree[1]	Metabolic pathways
1	47	88	3.76	Ceramides, Sphingomyelins
2	30	41	1.367	Bile Acid metabolism, Amino Acid and Purine metabolism
3	22	23	2.09	Amino Acid, Nucleotide metabolism
4	15	21	2.8	Plasmalogens, Lyso-plasmalogens, Phosphatidylcholines
5	13	18	2.77	Fatty Acid metabolism (Acyl carnitines, Acyl Amino Acids)
6	13	12	1.85	Carbohydrate, Amino Acid, Nucleotide metabolism

7	6	6	2	Vitamin A metabolism, Amino Acid metabolism
8	6	6	2	Benzoate metabolism, Amino acid metabolism

[1] Average degree represents the average number of connections each node (metabolite) makes within the module and indicates the network/module density.

Table 4.8: Number of Metabolites in Increase and Decrease BMI Cluster by Module and Pathway

Module	Super-pathway	Sub-pathway	Decrease BMI	Increase BMI
Module 1	Lipid	Ceramide Pes	1	0
Module 1	Lipid	Ceramides	1	4
Module 1	Lipid	Dihydroceramides	0	1
Module 1	Lipid	Dihydrosphingomyelins	0	5
Module 1	Lipid	Hexosylceramides (HCER)	2	3
Module 1	Lipid	Lactosylceramides (LCER)	1	0
Module 1	Lipid	Sphingomyelins	6	22
Module 1	Lipid	Sterol	0	1
Module 2	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	3	1
Module 2	Energy	Oxidative Phosphorylation	1	0
Module 2	Lipid	Phospholipid Metabolism	2	0
Module 2	Lipid	Primary Bile Acid Metabolism	1	6
Module 2	Lipid	Secondary Bile Acid Metabolism	3	6
Module 2	Lipid	Sterol	0	2
Module 2	Nucleotide	Purine Metabolism, Adenine containing	0	2
Module 2	Xenobiotics	Food Component/Plant	0	3
Module 3	Amino Acid	Alanine and Aspartate Metabolism	0	1
Module 3	Amino Acid	Histidine Metabolism	1	1
Module 3	Amino Acid	Lysine Metabolism	1	1
Module 3	Amino Acid	Methionine, Cysteine, SAM and Taurine Metabolism	0	1
Module 3	Amino Acid	Polyamine Metabolism	2	1
Module 3	Amino Acid	Tryptophan Metabolism	1	0
Module 3	Amino Acid	Urea cycle; Arginine and Proline Metabolism	1	0
Module 3	Carbohydrate	Aminosugar Metabolism	1	0
Module 3	Nucleotide	Purine Metabolism, Adenine containing	0	3
Module 3	Nucleotide	Pyrimidine Metabolism, Cytidine containing	0	1
Module 3	Nucleotide	Pyrimidine Metabolism, Thymine containing	1	0
Module 3	Nucleotide	Pyrimidine Metabolism, Uracil containing	2	2
Module 3	Xenobiotics	Chemical	1	0
Module 4	Lipid	Lyso-plasmalogen	1	2
Module 4	Lipid	Phosphatidylcholine (PC)	2	0

Module 4	Lipid	Plasmalogen	8	2
Module 5	Amino Acid	Alanine and Aspartate Metabolism	0	1
Module 5	Amino Acid	Glycine, Serine and Threonine Metabolism	1	0
Module 5	Lipid	Fatty Acid Metabolism (Acyl Carnitine, Hydroxy)	2	1
Module 5	Lipid	Fatty Acid Metabolism (Acyl Carnitine, Medium Chain)	1	0
Module 5	Lipid	Fatty Acid Metabolism (Acyl Carnitine, Short Chain)	1	0
Module 5	Lipid	Fatty Acid Metabolism (Acyl Glutamine)	1	0
Module 5	Lipid	Fatty Acid Metabolism (Acyl Glycine)	1	0
Module 5	Lipid	Fatty Acid Metabolism (also BCAA Metabolism)	1	0
Module 5	Lipid	Ketone Bodies	1	0
Module 5	Partially Characterized Molecules	Partially Characterized Molecules	2	0
Module 6	Amino Acid	Guanidino and Acetamido Metabolism	0	1
Module 6	Amino Acid	Lysine Metabolism	1	0
Module 6	Carbohydrate	Fructose, Mannose and Galactose Metabolism	2	0
Module 6	Carbohydrate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	1	1
Module 6	Carbohydrate	Pentose Metabolism	0	2
Module 6	Nucleotide	Purine Metabolism, (Hypo)Xanthine/Inosine containing	0	2
Module 6	Xenobiotics	Food Component/Plant	1	2
Module 7	Amino Acid	Urea cycle; Arginine and Proline Metabolism	1	0
Module 7	Cofactors and Vitamins	Vitamin A Metabolism	4	0
Module 7	Xenobiotics	Food Component/Plant	1	0
Module 8	Amino Acid	Tryptophan Metabolism	0	1
Module 8	Xenobiotics	Benzoate Metabolism	1	3
Module 8	Xenobiotics	Food Component/Plant	1	0

Table 4.9: Odds Ratios (OR) from Group Lasso Regression for the Metabolic Modules Associated with BMI Trajectory

Metabolite	Module	Group lasso OR	BMI cluster group
glycosyl ceramide (d18:2/24:1; d18:1/24:2)	1	0.9942	decrease
lactosyl-N-palmitoyl-sphingosine (d18:1/16:0)	1	0.9963	decrease
palmitoyl-sphingosine-phosphoethanolamine (d18:1/16:0)	1	0.9966	decrease

sphingomyelin (d18:1/24:1; d18:2/24:0)	1	0.997	decrease
palmitoyl sphingomyelin (d18:1/16:0)	1	0.9976	decrease
sphingomyelin (d18:2/24:1; d18:1/24:2)	1	0.9982	decrease
hydroxypalmitoyl sphingomyelin (d18:1/16:0(OH))	1	0.9987	decrease
sphingomyelin (d18:1/17:0; d17:1/18:0; d19:1/16:0)	1	0.9988	decrease
N-palmitoyl-sphingadienine (d18:2/16:0)	1	0.9988	decrease
glycosyl-N-stearoyl-sphingosine (d18:1/18:0)	1	0.999	decrease
sphingomyelin (d18:2/24:2)	1	0.9995	decrease
glycosyl-N-palmitoyl-sphingosine (d18:1/16:0)	1	1.0001	increase
cholesterol	1	1.0002	increase
N-palmitoyl-sphingosine (d18:1/16:0)	1	1.0003	increase
sphingomyelin (d17:1/14:0; d16:1/15:0)	1	1.0003	increase
ceramide (d18:1/14:0; d16:1/16:0)	1	1.0006	increase
glycosyl ceramide (d18:1/20:0; d16:1/22:0)	1	1.0007	increase
palmitoyl dihydrosphingomyelin (d18:0/16:0)	1	1.0008	increase
sphingomyelin (d17:1/16:0; d18:1/15:0; d16:1/17:0)	1	1.0008	increase
N-stearoyl-sphingosine (d18:1/18:0)	1	1.0012	increase
stearoyl sphingomyelin (d18:1/18:0)	1	1.0012	increase
myristoyl dihydrosphingomyelin (d18:0/14:0)	1	1.0013	increase
N-palmitoyl-sphinganine (d18:0/16:0)	1	1.0016	increase
sphingomyelin (d18:2/23:0; d18:1/23:1; d17:1/24:1)	1	1.0023	increase
lignoceroyl sphingomyelin (d18:1/24:0)	1	1.0024	increase
sphingomyelin (d18:2/23:1)	1	1.0025	increase
N-stearoyl-sphingadienine (d18:2/18:0)	1	1.0025	increase
sphingomyelin (d18:1/20:2; d18:2/20:1; d16:1/22:2)	1	1.0027	increase
sphingomyelin (d18:1/14:0; d16:1/16:0)	1	1.0027	increase
sphingomyelin (d18:1/22:2; d18:2/22:1; d16:1/24:2)	1	1.0031	increase
glycosyl-N-behenoyl-sphingadienine (d18:2/22:0)	1	1.0031	increase
sphingomyelin (d18:1/18:1; d18:2/18:0)	1	1.0037	increase
behenoyl sphingomyelin (d18:1/22:0)	1	1.004	increase
sphingomyelin (d18:1/20:0; d16:1/22:0)	1	1.0041	increase
sphingomyelin (d18:1/20:1; d18:2/20:0)	1	1.0042	increase
sphingomyelin (d18:1/19:0; d19:1/18:0)	1	1.0047	increase
sphingomyelin (d18:2/16:0; d18:1/16:1)	1	1.0052	increase
sphingomyelin (d18:1/21:0; d17:1/22:0; d16:1/23:0)	1	1.0052	increase
sphingomyelin (d18:1/22:1; d18:2/22:0; d16:1/24:1)	1	1.0053	increase
tricosanoyl sphingomyelin (d18:1/23:0)	1	1.0057	increase
sphingomyelin (d18:2/18:1)	1	1.0061	increase
sphingomyelin (d18:2/14:0; d18:1/14:1)	1	1.007	increase
sphingomyelin (d17:2/16:0; d18:2/15:0)	1	1.008	increase
sphingomyelin (d18:0/18:0; d19:0/17:0)	1	1.0086	increase

sphingomyelin (d18:2/21:0; d16:2/23:0)	1	1.01	increase
sphingomyelin (d18:0/20:0; d16:0/22:0)	1	1.0116	increase
behenoyl dihydrosphingomyelin (d18:0/22:0)	1	1.0144	increase
phosphate	2	0.9409	decrease
succinoyltaurine	2	0.9419	decrease
glycocholenate sulfate	2	0.9511	decrease
taurocholenate sulfate	2	0.9681	decrease
deoxycholic acid glucuronide	2	0.9693	decrease
hypotaurine	2	0.9767	decrease
glycodeoxycholate 3-sulfate	2	0.9779	decrease
glycochenodeoxycholate glucuronide (1)	2	0.979	decrease
phosphoethanolamine (PE)	2	0.9804	decrease
taurine	2	0.984	decrease
phosphocholine	2	0.9891	decrease
taurolithocholate 3-sulfate	2	0.9925	decrease
isoursodeoxycholate	2	0.9942	decrease
AMP	2	1.0023	increase
glycolithocholate sulfate	2	1.0033	increase
glycoursodeoxycholate	2	1.0096	increase
N-acetyltaurine	2	1.0126	increase
lithocholate sulfate (1)	2	1.0154	increase
glycochenodeoxycholate 3-sulfate	2	1.0216	increase
glucuronide of piperine metabolite C17H21NO3 (4)	2	1.0295	increase
glycochenodeoxycholate	2	1.0323	increase
3beta-hydroxy-5-cholestenoate	2	1.0323	increase
ADP	2	1.0343	increase
piperine	2	1.0369	increase
cholate	2	1.0438	increase
taurochenodeoxycholate	2	1.0451	increase
sulfate of piperine metabolite C16H19NO3 (2)	2	1.0452	increase
7-HOCA	2	1.0663	increase
taurocholate	2	1.0699	increase
glycocholate	2	1.0954	increase
3-aminoisobutyrate	3	0.9848	decrease
6-oxopiperidine-2-carboxylate	3	0.9863	decrease
hydantoin-5-propionate	3	0.9868	decrease
N-acetylneuraminate	3	0.9926	decrease
5-methyluridine (ribothymidine)	3	0.995	decrease
O-sulfo-L-tyrosine	3	0.9954	decrease
(N(1) + N(8))-acetylspermidine	3	0.9961	decrease
3-(3-amino-3-carboxypropyl)uridine	3	0.9971	decrease

dimethylarginine (ADMA + SDMA)	3	0.9976	decrease
4-acetamidobutanoate	3	0.9979	decrease
C-glycosyltryptophan	3	0.998	decrease
5;6-dihydrouridine	3	1.0011	increase
hydroxyasparagine	3	1.0023	increase
pseudouridine	3	1.0032	increase
N-acetylputrescine	3	1.0032	increase
formiminoglutamate	3	1.0076	increase
2;3-dihydroxy-5-methylthio-4-pentenoate (DMTPA)	3	1.0094	increase
hydroxy-N6;N6;N6-trimethyllysine	3	1.0099	increase
N4-acetylcytidine	3	1.0106	increase
N2;N2-dimethylguanosine	3	1.0106	increase
N6-carbamoylthreonyladenosine	3	1.0149	increase
7-methylguanine	3	1.0167	increase
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)	4	0.9216	decrease
1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)	4	0.923	decrease
1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)	4	0.9565	decrease
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)	4	0.957	decrease
1;2-dipalmitoyl-GPC (16:0/16:0)	4	0.9689	decrease
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	4	0.9719	decrease
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (P-16:0/20:4)	4	0.9779	decrease
1-(1-enyl-palmitoyl)-2-oleoyl-GPE (P-16:0/18:1)	4	0.9872	decrease
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)	4	0.991	decrease
1-(1-enyl-stearoyl)-2-arachidonoyl-GPE (P-18:0/20:4)	4	0.9928	decrease
1-(1-enyl-stearoyl)-GPE (P-18:0)	4	0.9979	decrease
1-(1-enyl-palmitoyl)-GPE (P-16:0)	4	1.001	increase
1-(1-enyl-stearoyl)-2-linoleoyl-GPE (P-18:0/18:2)	4	1.0014	increase
1-(1-enyl-oleoyl)-GPE (P-18:1)	4	1.003	increase
1-(1-enyl-stearoyl)-2-oleoyl-GPE (P-18:0/18:1)	4	1.0059	increase
N-acetylglycine	5	0.9827	decrease
hexanoylglutamine	5	0.9888	decrease
3-hydroxybutyrate (BHBA)	5	0.9903	decrease
propionylglycine (C3)	5	0.9912	decrease
glutamine conjugate of C6H10O2 (1)	5	0.9932	decrease
glutamine conjugate of C6H10O2 (2)	5	0.9956	decrease
acetylcarnitine (C2)	5	0.996	decrease
3-hydroxybutyroylglycine	5	0.9963	decrease
3-hydroxyhexanoylcarnitine (1)	5	0.997	decrease
(R)-3-hydroxybutyrylcarnitine	5	0.9972	decrease
hexanoylcarnitine (C6)	5	0.9981	decrease
(S)-3-hydroxybutyrylcarnitine	5	1.0014	increase

alanine	5	1.0089	increase
mannose	6	0.9811	decrease
fructose	6	0.9856	decrease
fructosyllysine	6	0.9913	decrease
glucose	6	0.994	decrease
2-keto-3-deoxy-gluconate	6	0.999	decrease
mannonate	6	1.0054	increase
gluconate	6	1.009	increase
ribonate	6	1.0142	increase
1;5-anhydroglucitol (1;5-AG)	6	1.0242	increase
4-guanidinobutanoate	6	1.0257	increase
ribitol	6	1.0283	increase
urate	6	1.035	increase
allantoin	6	1.056	increase
carotene diol (1)	7	0.9936	decrease
carotene diol (3)	7	0.994	decrease
carotene diol (2)	7	0.9946	decrease
beta-cryptoxanthin	7	0.9962	decrease
stachydrine	7	0.9985	decrease
N-methylproline	7	0.9993	decrease
cinnamoylglycine	8	0.9216	decrease
hippurate	8	0.9289	decrease
4-hydroxyhippurate	8	1.0049	increase
indolepropionate	8	1.0121	increase
methyl-4-hydroxybenzoate sulfate	8	1.0283	increase
3-hydroxyhippurate	8	1.0695	Increase

Next, we looked at the correlation between the 152 metabolites selected by the group lasso model with BMI at the time of ALS diagnosis (**Table 4.10**). 65 metabolites were significantly correlated with BMI (p < 0.05). However, the magnitude of the correlation coefficients was relatively low ($|corr| \sim 0.3$) (**Figure 4.4**). We also tested the association of these 152 metabolites with ALS survival in a Cox Proportional Hazards model. 31 metabolites were found to be significantly associated with ALS survival (p < 0.05) (**Figure 4.5**, **Table 4.11**). Of the 70 metabolites associated with the decrease BMI cluster from the group lasso model, 7 metabolites (N-palmitoyl-sphingadienine (d18:2/16:0), Hippurate, 3-(3-amino-3-carboxypropyl) uridine, succinoyltaurine, phosphocholine, N-methylproline and isoursodeoxycholate) were also significantly associated with poorer ALS survival (HR > 1). Conversely, of the 82 metabolites associated with the increase BMI cluster, 4 metabolites (1-(1-enyl-stearoyl)-2-oleoyl-GPE (P-18:0/18:1), 1-(1-enyl-oleoyl)-GPE (P-18:1), glycoursodeoxycholate and N4-acetylcytidine) were significantly associated with better ALS survival (HR < 1).

Taken together, these results suggest that the metabolic profiles of ALS patients at the time of ALS diagnosis are reflective of the change in BMI over a period of 10 years and have potential to predict survival among ALS patients.

Table 4.10: Correlation of BMI at the time of ALS diagnosis with 152 metabolites associated with BMI trajectory

Metabolite	Correlation Coefficient	p-value
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*	-0.31879	1.10E-09
urate	0.311685	2.67E-09
N4-acetylcytidine	0.259687	8.72E-07
sphingomyelin (d18:0/18:0, d19:0/17:0)*	0.252995	1.69E-06
glycosyl ceramide (d18:2/24:1, d18:1/24:2)*	-0.24827	2.66E-06
glucuronide of piperine metabolite C17H21NO3 (4)*	0.238718	6.51E-06
taurocholenate sulfate*	-0.23794	7.00E-06
behenoyl dihydrosphingomyelin (d18:0/22:0)*	0.228318	1.65E-05
piperine	0.221056	3.09E-05
glycosyl ceramide (d18:1/20:0, d16:1/22:0)*	-0.21462	5.29E-05
ribitol	0.207173	9.67E-05
sphingomyelin (d18:1/24:1, d18:2/24:0)*	-0.20665	0.000101
phosphocholine	-0.20496	0.000115
taurolithocholate 3-sulfate	-0.20378	0.000126
mannose	0.202698	0.000137

N-acetylglycine	-0.19643	0.000222
N2,N2-dimethylguanosine	0.191606	0.000318
palmitoyl-sphingosine-phosphoethanolamine (d18:1/16:0)	-0.18857	0.000397
succinoyltaurine	-0.1876	0.000426
palmitoyl sphingomyelin (d18:1/16:0)	-0.18635	0.000466
mannonate*	0.181915	0.000638
7-HOCA	0.178599	0.000804
taurine	-0.17851	0.000809
2,3-dihydroxy-5-methylthio-4-pentenoate (DMTPA)*	0.177993	0.000838
sphingomyelin (d18:0/20:0, d16:0/22:0)*	0.17501	0.001027
hydroxyasparagine	0.174774	0.001043
hydroxy-N6,N6,N6-trimethyllysine*	0.172488	0.001216
N6-carbamoylthreonyladenosine	0.172126	0.001246
3beta-hydroxy-5-cholestenoate	-0.16742	0.001698
phosphate	-0.16462	0.002033
sulfate of piperine metabolite C16H19NO3 (2)*	0.16462	0.002033
glycodeoxycholate 3-sulfate	-0.16423	0.002084
7-methylguanine	0.159502	0.002806
sphingomyelin (d17:2/16:0, d18:2/15:0)*	0.155557	0.003575
glycosyl-N-behenoyl-sphingadienine (d18:2/22:0)*	-0.15539	0.003612
stachydrine	0.153009	0.004169
N-stearoyl-sphingadienine (d18:2/18:0)*	0.152802	0.004221
N-stearoyl-sphingosine (d18:1/18:0)*	0.151331	0.004607
5-methyluridine (ribothymidine)	-0.14561	0.006431
glycocholate	0.145365	0.006521
pseudouridine	0.144025	0.007039
sphingomyelin (d18:2/24:1, d18:1/24:2)*	-0.14254	0.007654
1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)*	-0.14179	0.007985
deoxycholic acid glucuronide	0.140525	0.008567
glycosyl-N-palmitoyl-sphingosine (d18:1/16:0)	-0.13759	0.010069
hydroxypalmitoyl sphingomyelin (d18:1/16:0(OH))	-0.13596	0.011
sphingomyelin (d18:1/18:1, d18:2/18:0)	0.135138	0.0115
sphingomyelin (d18:1/20:2, d18:2/20:1, d16:1/22:2)*	0.133715	0.012409
1,2-dipalmitoyl-GPC (16:0/16:0)	-0.13061	0.014615
propionylglycine (C3)	-0.12844	0.016363
1-palmitoyl-2-stearoyl-GPC (16:0/18:0)	-0.12404	0.020452
sphingomyelin (d18:2/14:0, d18:1/14:1)*	0.1213	0.023432
sphingomyelin (d18:2/18:1)*	0.119013	0.026197
cholate	0.11733	0.028407
3-hydroxyhippurate	0.116567	0.02946
(R)-3-hydroxybutyrylcarnitine	0.115493	0.031001
1-(1-enyl-oleoyl)-GPE (P-18:1)*	-0.11414	0.033041
N-palmitoyl-sphingadienine (d18:2/16:0)*	0.112602	0.03549
6-oxopiperidine-2-carboxylate	-0.11248	0.035692
lactosyl-N-palmitoyl-sphingosine (d18:1/16:0)	-0.11076	0.038623
1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*	-0.10988	0.040219
ceramide (d18:1/14:0, d16:1/16:0)*	0.107616	0.044531
sphingomyelin (d18:2/21:0, d16:2/23:0)*	0.106	0.047847
glycosyl-N-stearoyl-sphingosine (d18:1/18:0)	-0.10513	0.049721
glutamine conjugate of C6H10O2 (1)*	0.105007	0.049987
N-methylproline	0.103106	0.054304
phosphoethanolamine (PE)	-0.10282	0.054978
sphingomyelin (d18:2/16:0, d18:1/16:1)*	0.102358	0.056085

alanine	0.098618	0.065737
5,6-dihydrouridine	0.098054	0.067303
stearoyl sphingomyelin (d18:1/18:0)	0.095389	0.075128
hypotaurine	-0.09362	0.080711
1-(1-enyl-stearoyl)-2-arachidonoyl-GPE (P-18:0/20:4)*	0.092376	0.084849
indolepropionate	-0.09008	0.092905
hexanoylcarnitine (C6)	0.088642	0.098273
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (P-16:0/20:4)*	0.087952	0.100929
glycolithocholate sulfate*	-0.0878	0.101535
glycochenodeoxycholate	0.087428	0.102984
ribonate	0.086694	0.105919
N-palmitoyl-sphinganine (d18:0/16:0)	0.085147	0.112319
cinnamoylglycine	-0.08492	0.113299
1-(1-enyl-palmitoyl)-2-oleoyl-GPE (P-16:0/18:1)*	-0.07948	0.13839
ADP	-0.07921	0.13973
(S)-3-hydroxybutyrylcarnitine	0.078271	0.144504
glucose	0.074395	0.165521
allantoin	0.074267	0.166249
sphingomyelin (d18:1/22:2, d18:2/22:1, d16:1/24:2)*	0.073722	0.169389
1-(1-enyl-stearoyl)-GPE (P-18:0)*	-0.07251	0.176535
(N(1) + N(8))-acetylspermidine	-0.06956	0.194842
lignoceroyl sphingomyelin (d18:1/24:0)	-0.0688	0.199785
beta-cryptoxanthin	-0.06873	0.200215
fructosyllysine	0.0677	0.207074
N-acetylputrescine	-0.06692	0.21236
palmitoyl dihydrosphingomyelin (d18:0/16:0)*	-0.0652	0.224351
3-hydroxybutyrate (BHBA)	-0.06457	0.228886
sphingomyelin (d18:2/23:0, d18:1/23:1, d17:1/24:1)*	-0.06354	0.236406
3-(3-amino-3-carboxypropyl)uridine*	0.063182	0.239084
sphingomyelin (d18:1/19:0, d19:1/18:0)*	0.06303	0.240219
isoursodeoxycholate	0.060565	0.259137
4-hydroxyhippurate	0.060516	0.259529
4-guanidinobutanoate	0.060396	0.260473
3-aminoisobutyrate	-0.05844	0.276231
C-glycosyltryptophan	0.057148	0.287034
hexanoylglutamine	-0.05524	0.303438
1-(1-enyl-palmitoyl)-GPE (P-16:0)*	-0.05423	0.312432
1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)*	0.052717	0.326107
taurocholate	0.050666	0.34531
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*	-0.04884	0.363001
carotene diol (2)	-0.04691	0.38229
taurochenodeoxycholate	0.046029	0.391297
sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0)	-0.04545	0.397259
3-hydroxyhexanoylcarnitine (1)	0.045117	0.400765
lithocholate sulfate (1)	0.044483	0.407417
carotene diol (3)	0.043217	0.420907
formiminoglutamate	0.042029	0.433801
sphingomyelin (d18:1/22:1, d18:2/22:0, d16:1/24:1)*	0.041785	0.436486
3-hydroxybutyroylglycine	-0.04115	0.443483
2-keto-3-deoxy-gluconate	0.040143	0.454731
sphingomyelin (d17:1/14:0, d16:1/15:0)*	-0.03944	0.462681
fructose	-0.03923	0.465102
gluconate	0.039072	0.466864

1-(1-enyl-stearoyl)-2-oleoyl-GPE (P-18:0/18:1)	-0.03863	0.471923
glutamine conjugate of C6H10O2 (2)*	0.037194	0.48857
sphingomyelin (d18:2/23:1)*	0.03674	0.493894
behenoyl sphingomyelin (d18:1/22:0)*	0.035662	0.506665
tricosanoyl sphingomyelin (d18:1/23:0)*	0.033714	0.530174
hippurate	0.032705	0.542562
1,5-anhydroglucitol (1,5-AG)	0.032492	0.545184
dimethylarginine (ADMA + SDMA)	-0.03224	0.548285
myristoyl dihydrosphingomyelin (d18:0/14:0)*	-0.03029	0.572723
sphingomyelin (d18:1/21:0, d17:1/22:0, d16:1/23:0)*	0.029253	0.585996
glycochenodeoxycholate glucuronide (1)	-0.02575	0.631631
1-(1-enyl-palmitoyl)-2-linoleoyl-GPE (P-16:0/18:2)*	0.024542	0.647742
glycocholenate sulfate*	-0.0232	0.665844
acetylcarnitine (C2)	0.021302	0.691682
sphingomyelin (d18:1/14:0, d16:1/16:0)*	0.020148	0.707602
sphingomyelin (d18:1/20:1, d18:2/20:0)*	-0.02013	0.707881
N-acetyltaurine	-0.0201	0.708302
sphingomyelin (d18:2/24:2)*	0.019032	0.723115
AMP	-0.01823	0.734282
glycochenodeoxycholate 3-sulfate	-0.01797	0.737928
hydantoin-5-propionate	-0.01753	0.7442
N-palmitoyl-sphingosine (d18:1/16:0)	-0.01738	0.746328
sphingomyelin (d18:1/20:0, d16:1/22:0)*	-0.01392	0.795471
N-acetylneuraminate	-0.01389	0.795933
carotene diol (1)	-0.01313	0.806895
4-acetamidobutanoate	-0.01187	0.825156
1-(1-enyl-stearoyl)-2-linoleoyl-GPE (P-18:0/18:2)*	-0.01055	0.844316
glycoursodeoxycholate	0.009603	0.85812
methyl-4-hydroxybenzoate sulfate	0.008669	0.871804
cholesterol	0.003682	0.94536
O-sulfo-L-tyrosine	0.002645	0.96073

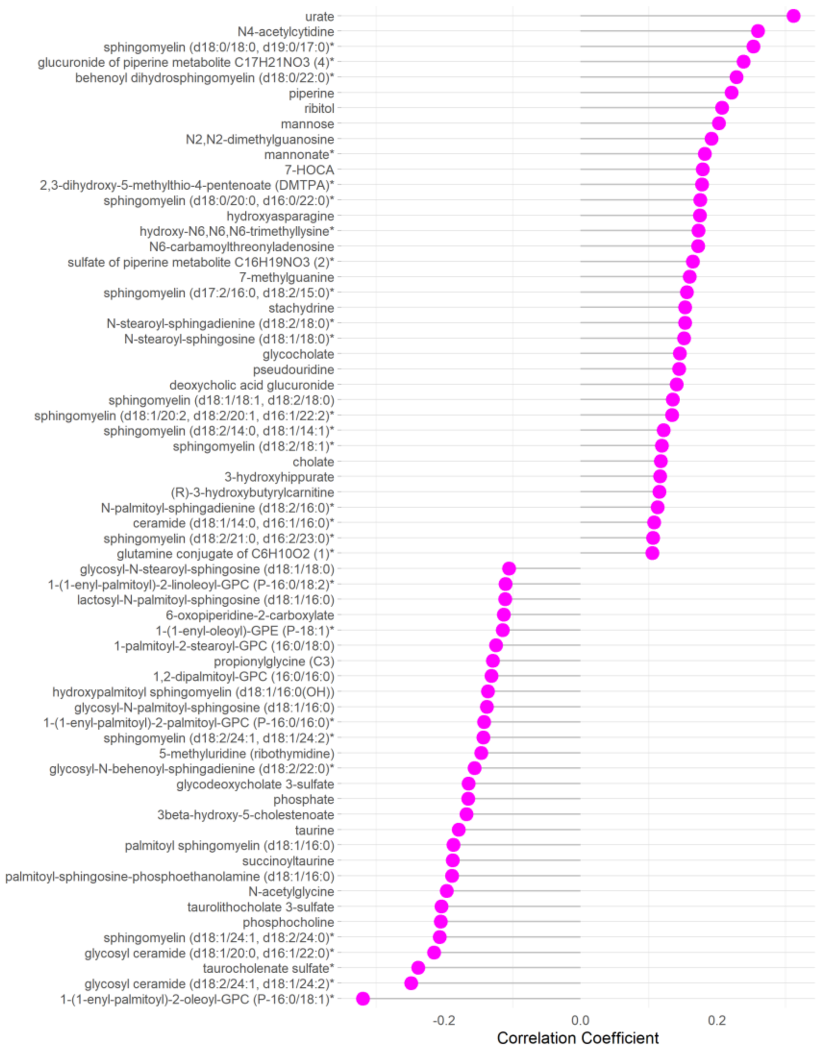


Figure 4.5: Lollipop plot of the metabolites significantly ($p < 0.05$) correlated with BMI at the time of ALS diagnosis.

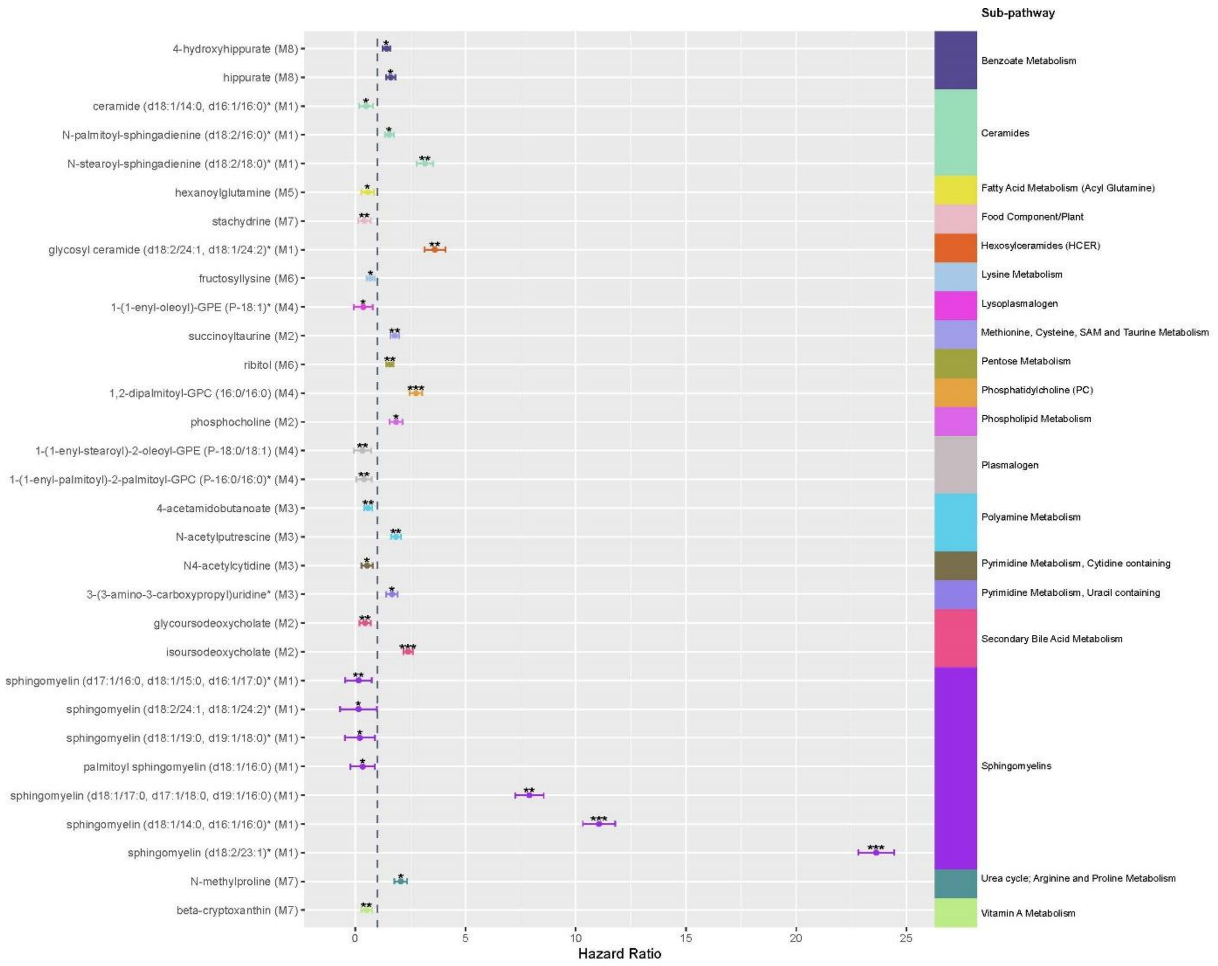


Figure 4.6: Hazard Ratios for metabolites significantly associated with ALS survival in Cox proportional hazards models. Metabolites associated with BMI trajectory were tested for their association with ALS survival. Module assignment for each metabolite is indicated in parentheses after the metabolite name. * Significant at $p < 0.05$; ** significant at $p < 0.01$; *** significant at $p < 0.001$.

4.4 Discussion

This study adds to the growing body of evidence that pre-symptomatic BMI loss is linked to ALS risk and survival. We show that ALS participants are characterized by significant BMI loss five years, but not 10 years, prior to study entry versus control participants. A decrease in BMI trajectory was associated with shorter survival in ALS, which also correlated with a distinct metabolomic profile. Our study also suggests that BMI loss may occur during the pre-symptomatic phase of ALS leading up to diagnosis. Several other studies have similarly shown BMI decrease preceding ALS diagnosis, out to 10 years prior to onset[218] and even within the decades preceding ALS[219,222]. Although we found BMI trajectories differed over the 10-year window, we found that absolute BMI did not vary between ALS and control participants 10 years before study entry when participants would have had a mean age of 54.9 (ALS) and 51.3 (controls) years. In contrast, other studies report that lower mid-to-late life BMI increases ALS risk[221,222,251], although one study reported ALS survival depends on BMI change, not on BMI before or at diagnosis[217]. Another recent study suggests that BMI in ALS patients diverges from controls 10 years prior to disease onset[252].

Next, we found that that ALS participants with a 10-year decrease BMI trend had shorter survival. Our results are consistent with several studies demonstrating that a drop in BMI prior to ALS diagnosis correlates with poorer survival[217,218,222,253].In particular, analysis of the Piemonte and Valle d'Aosta Register for ALS found that BMI loss at diagnosis was more prognostic of survival than BMI either before or at diagnosis[217]. However, since there is literature that BMI is an ALS risk factor[221,222,251], we conducted sensitivity analyses to assess the interaction of baseline BMI with BMI trajectory. We found that normal baseline BMI lengthened survival in the decrease BMI trajectory group, whereas obese baseline BMI shortened survival in the increase BMI trajectory group. Baseline BMI only marginally influenced survival in the mild decrease BMI trajectory group. Interestingly, the European Prospective Investigation into Cancer and Nutrition study also showed that obese females had shorter survival that did not reach significance[223], whereas the Piemonte and Valle d'Aosta Register found no impact of BMI on survival[217].

The reasons for survival differences by BMI or BMI change in ALS are not known. However, the prevailing theories are related to impaired energy homeostasis[224], with lowered energy intake fighting against higher energy expenditure. Dysphagia is a frequent cause of reduced energy intake, however in ALS BMI loss also occurs independent of dysphagia[217,253] indicating the presence of significantly elevated energy expenditure. Indeed, hypermetabolism is more frequent in ALS versus control participants and correlates inversely with survival[225]. Resting energy expenditure may additionally interact with BMI and fat mass to influence survival in ALS[254,255].

In the current study, we employed data driven network analysis to identify highly interconnected metabolic modules and assessed their correlation with BMI trajectory groups. The largest of these, module 1, contained ceramides (13 species) and sphingomyelins (33 species). The latter were primarily associated with the increase BMI group. We and others previously found that sphingomyelins also differ in analyses of ALS versus control participant plasma[238,239,256–259]. Further, one recent study reported that higher sphingomyelin levels may correlate with faster disease progression[259]. Sphingomyelins are a large class of lipids that have structural roles in cell membranes and lipid rafts, and, through hydrolysis to ceramides, with signaling activity, e.g., pro-apoptotic, excitotoxic, neurotoxic[260,261]. Impaired sphingomyelin metabolism may be an integral factor in ALS as supported by investigations of genetic models[262]. Of the 47 metabolites in module 1, only 13 significantly correlated with BMI at diagnosis, suggesting associations of the remaining 34 metabolites with BMI trajectory may be related to the ALS disease process.

The second largest module 2 mostly contained primary and secondary bile acids, which generally associated with the increase BMI trajectory, in addition to metabolites of methionine, cysteine, S-adenosyl methionine, and taurine metabolism and oxidative phosphorylation. Nearly half of the metabolites in this module also significantly correlated with diagnosis BMI (13 species). Bile acids play important roles in nutrient absorption, regulation of cholesterol metabolism, and systemic energy expenditure[263], so the correlation with BMI trajectory herein is unsurprising. Interestingly, although not present in the module, two bile acids ursodeoxycholic and its taurine derivative tauroursodeoxycholic acid (taurursodiol) have shown some efficacy in ALS clinical trials[264–267].

Module 3 contained modified amino acids and nucleotide derivatives spanning 22 species evenly split between the decrease and increase BMI groups, of which 9 significantly correlated with diagnosis BMI. Module 4 contained several bioactive lipids, plasmalogens (10 species), lyso-plasmalogens (3 species), and phosphatidylcholines (2 species), which mostly associated with the decrease BMI group, i.e., poorer survival. Only two species were significantly linked to diagnosis BMI. We[239] and others[256,259,268,269] have previously shown phosphatidylcholines differentiate ALS from control participants, in particular, phosphatidylcholine 36:4[259,268].

Modules 5 and 6 comprised candidates related to energy metabolism. Module 5 contained four short-chain acyl-carnitines, intermediates of, which all save one correlated with the decrease BMI group. We previously reported acyl-carnitines, along with free fatty acids, contributed to the discrimination between ALS versus control participants[238,239], which we attributed to either dysfunctional or at capacity β-oxidation[270]. Modules 6, 7 and 8 contained few metabolites equally divided in their correlation with either the decrease or increase BMI trajectory group, suggesting ALS status may be a stronger determinant of these metabolites than BMI trajectory.

Overall, across some modules, e.g., module 5, there were more metabolites from various biochemical pathways relating to energy utilization (e.g. fatty acid β-oxidation) that are more discerning of ALS versus control participants than of BMI trajectories. These findings suggest that ALS status is an important determinant of energy metabolism. One possibility is that metabolites correlate with fat mass loss in ALS patients[271], an idea supported by studies where ALS polygenic risk associates with body fat percentage in addition to BMI[272,273]. Interestingly, neither creatine nor creatinine were among the metabolites correlating with BMI change or diagnosis BMI, indicating weight changes may be more pronounced for fat mass than muscle mass. However, lacking body composition measures, we could not evaluate this possibility in this study.

This study had limitations. Participants self-reported weight, potentially incurring recall bias; however, Lin's concordance correlation coefficient was high for participants with available weight, indicating good recall. Our study did not query weight at frequent intervals, so we cannot determine if BMI loss in ALS participants was linear in the 5 years prior to study entry or more pronounced closer to diagnosis. It is also possible we failed to detect an onset in BMI changes between the 10-to-5-year window before diagnosis due to the lack of granular BMI information. Next, we only asked participants to report current height, and use this for BMI calculations at all timepoints. However, such changes in height over the life course are not anticipated to cause bias in statistical models[274]. We also did not collect a dietary or physical activity survey for this analysis. Additionally, our metabolomics analysis was untargeted, and thus did not measure all metabolites in every relevant biochemical pathway. While BMI analysis was longitudinal, metabolomics analysis was cross-sectional. Plasma samples for untargeted metabolomics were non-fasted for ethical reasons, as noted in our prior publications[238,239].

In summary, we found that ALS participants have distinct BMI trajectories versus controls, with the most significant BMI drop occurring within 5 years before diagnosis. ALS participants with normal baseline BMI and decrease BMI trajectory, or baseline obese BMI and increase BMI trajectory have shorter survival. BMI trajectories correlate with metabolic changes, especially with sphingomyelins and bile acids.

CHAPTER V

Conclusions and Future Perspectives

5.1 General conclusions

Experimental design in metabolomics commonly involves assessing metabolite levels in two or more disease or experimental conditions. Metabolomics data acquired from such experiments are amenable to univariate analysis, followed by pathway mapping and enrichment analysis. While overall this approach has proven to be extremely useful, there are certain limitations. First, univariate analysis assesses differences in individual metabolite levels but does not account for the interactions between them. Second, application of pathway mapping and enrichment analysis is hampered by the low coverage of metabolites in biological pathway databases. This is particularly true for lipids and secondary metabolites. The problem is compounded by the relatively small number of known metabolites measured in most experiments which limits both statistical significance and overall reliability of the analyses. In this dissertation, I presented computational approaches to overcome these limitations and gain deeper insights into high dimensional metabolomics and lipidomics data.

In Chapter 2 of this dissertation, I presented *Filigree*, a new computational approach and tool that provides an alternative to traditional pathway-centric approaches. *Filigree* constructs partial correlation networks among metabolites directly from experimental measurements. In lieu of knowledge-based metabolic pathways, *Filigree* generates topology-based sets (subnetworks) comprised of biochemically and structurally related metabolites. *Filigree* then assesses changes in both the level of these metabolite sets and, importantly, the degree of interaction among the metabolites and how these interactions are disrupted by disease, thus providing a systems level view of the data. We made *Filigree* more robust by developing mathematical approaches to allow for severely limited sample sizes or grossly imbalanced experimental groups in the data. We analyzed previously published studies assessing the metabolome in the context of metabolic disorders (type 1 and type 2 diabetes) and the interplay between maternal and infant lipidome during pregnancy. We observed strong differential connectivity in metabolite networks in T1D and T2D. We were also able to demonstrate the influence of maternal lipidome on infant birthweight. With these analyses, we showed that topology-based enrichment methods are more powerful than traditional enrichment testing. *Filigree* therefore provides a clear advantage and is a powerful hypothesis-generating tool. In the final section of this chapter, I detail my contributions to the development of the DNEA R package. Specifically, I conceptualized and wrote the functions for feature aggregation and stability selection coupled with additional subsampling. Feature aggregation is crucial in a high dimensional setting i.e., when the available sample size is much smaller than the number of measured metabolic features. By collapsing highly correlated or chemically related metabolites, this approach effectively reduces the feature space and enables recovering a robust network. Further, when the number of samples in one experimental group is much larger than those in the other, the edges in the resulting network are heavily biased towards the larger group. Stability selection coupled with additional subsampling probes the larger group and allows us to recover a more balanced set of edges in the network.

In Chapter 3 of this dissertation, I described the changes in the plasma metabolome associated with COVID-19 disease severity. We established the association of the plasma metabolome with patient characteristics such as age, gender, race, BMI, and diabetes. As expected, a substantial portion of the metabolome was influenced by these clinical variables. In addition, we also found a strong influence of anesthetic administration (propofol) on the plasma metabolome of COVID-19 patients. Differential analysis revealed substantial changes between healthy controls and COVID-19 patients with mild and severe disease. In particular, levels of fatty acids and acylcarnitines were elevated in COVID-19 patients, with patients in the severe group having higher levels than those in the mild group. Levels of several bile acids as well as the hormone Thyroxine were lowered in COVID-19 patients. Further, the metabolites performed much better in discriminating disease severity compared to clinical characteristics that are traditionally used to ascertain disease predisposition. Several acylcarnitines, diacylglycerols, and phosphocholines contributed the most to model performance, recapitulating as well as augmenting some of the known metabolic markers of COVID-19 severity. In conclusion, we demonstrated that the plasma metabolome can be used to predict COVID-19 severity.

In Chapter 4 of this dissertation, I described the association of data-driven metabolic modules with BMI trajectory and survival in patients with Amyotrophic Lateral Sclerosis (ALS). ALS patients showed a significant BMI loss 5 to 10 years prior to diagnosis. This decrease in BMI over time correlated with poorer survival. We constructed data-driven partial correlation networks from the metabolic profiles of these patients and clustered them to obtain interconnected metabolic modules. We found that 8 of these modules associated with BMI trajectory, primarily those modules containing sphingomyelins and bile acids. Notably, assessing the associations of individual metabolites with BMI trajectory did not yield any significant result. However, when we look at modules of chemically and functionally related metabolites, we discover a lot more associations, demonstrating that subtle and nuanced associations can be identified when we look at groups of correlated metabolites as opposed to individual metabolites. Additionally, we found that a subset of the metabolites associated with BMI trajectory was also associated with ALS survival.

In conclusion, the body of work in this dissertation highlights the importance of data-driven analysis in the field of metabolomics. Further, this work also underscores the advantages of building data-driven metabolic networks in lieu of knowledge-based pathways to obtain biologically relevant information from metabolomics data. This work overcomes challenges associated with knowledge-based analysis and offers suitable alternatives through the computational tools developed and employed in analyzing a variety of metabolomics data types.

5.2 Future perspectives

5.2.1 Data-driven network analysis

One of the most important considerations in application of data-driven analysis methods for metabolomics (or any other omics data) is the number of samples vs. the number of variables. A larger sample size provides higher degrees of freedom that increases the power of the analysis. One way to boost the power of data-driven network analysis methods is to incorporate prior knowledge of metabolite relationships (*Composite networks*). These relationships can come from pathway databases such as KEGG, BioCyc, MetaCyc, and Reactome or from chemical ontologies such as ClassyFire[275]. Incorporation of prior knowledge into data-driven metabolic networks can potentially increase their robustness and provide additional biological context.

Another potential improvement that could boost that interpretability of data-driven metabolic modules generated by our method is computing a summary measure for each module. Borrowing the concept of "module eigengene" from WGCNA method[111], we can compute a "module eigenmetabolite" and utilize this measure for downstream association analysis. These module-specific eigenmetabolites can also be used to compare metabolic modules across datasets i.e., for the meta-analysis of partial correlation networks.

A natural extension to data-driven network analysis of metabolomics data is performing data-driven multi-omics integration. Exploring relationships between key metabolic changes and alterations in gene expression, for example, can provide additional levels of information and help build biological insights from experimental data. A key challenge remains that the number of features that can be included in data-driven integration tend to be limited by the number of available samples. Therefore, data reduction (or feature selection) becomes a crucial step that requires rigorous exploration.

Data-driven networks can be applied to sufficiently large longitudinal metabolomics data to assess topological changes over time, especially within modules containing metabolites of interest associated with a specific phenotype. Temporal changes in the relationship among metabolites can potentially inform us of the underlying metabolic rewiring taking place over time and how that could affect their associations with external traits.

5.2.2 Metabolic markers of COVID-19 severity

Our analysis revealed that several lipids (fatty acids, acylcarnitines) had the potential to discriminate COVID-19 patients based on disease severity. It would therefore be interesting to explore the plasma lipidome of these patients to elaborate on some of our findings and identify more nuanced changes in lipid profiles/pathways leading to more severe disease.

Our analysis also focused on a set of 294 putatively annotated metabolites from the untargeted metabolomics data. It would be worthwhile to investigate the unannotated portion of the dataset as well, using the same analysis pipeline, to gain deeper insights.

COVID-19 has been studied extensively and it is clear that its etiology is highly complex. Changes in the metabolome associated with severity of disease is likely a reflection of orchestrated changes in epigenome, transcriptome and proteome. An integrative approach is therefore required to get a wholistic understanding of the perturbations and better rationalize some of our findings.

Finally, while our classification models revealed some interesting lipids are markers of disease severity, they will need to be validated experimentally in *in vitro* and *in vivo* COVID-19 models, for any translational applications.

5.2.3 Association of BMI trajectory with ALS survival and metabolic modules

The goal of a typical clinical metabolomics study is to identify predictive marker(s) of the disease under study. While it would be very important and useful to be able to utilize patient metabolic profiles to predict future ALS before onset, the current study design does not permit us to explore this avenue. The metabolomics data was collected from ALS patients at the time of diagnosis, and this remains one of the biggest challenges of this study. Obtaining samples from patients 5 or 10 years prior to ALS diagnosis is also remarkably challenging and will require a prospective cohort study design.

ALS manifests as a complex and heterogeneous disease arising from a combination of genetic susceptibility, environmental exposures, as well as metabolic events like hypermetabolism and mitochondrial dysfunction. Therefore, alterations in the metabolome of ALS patients can likely be a cause or effect. It follows then that the association between the metabolic modules and BMI trajectory in ALS patients that we observe is likely more convoluted than a direct association. Teasing out these relationships will require collection of several other data types like gene expression, protein expression, and environmental exposures, and integrating them to be able to formulate a disease risk score.

Finally, we can explore ALS patients' stratification based on their metabolic profiles i.e., *metabotypes* and correlate them with the BMI trajectory stratification to gain a better understanding of the complex interplay between the change in BMI and metabolome over time in ALS patients.

5.2.4 Final thoughts

Partial correlation networks offer plenty of advantages for untargeted metabolomics data. With the development of increasingly sensitive analytical platforms, the proportion of high confidence annotations in these datasets is also increasing. Incorporation of knowledge-based metabolic pathways, requiring well-annotated metabolites, into data-driven partial correlation networks will significantly increase the power and interpretability of these networks. Likewise, the integration of metabolomics data with gene and protein expression data, specifically in a data-driven manner, can further augment the biological findings. Finally, the availability of longitudinal metabolomics data would aid in understanding the change in cellular mechanisms across time and how this change differs in the disease under study.

BIBLIOGRAPHY

(1) Liu, J. Y.; Wellen, K. E. Advances into Understanding Metabolites as Signaling Molecules in Cancer Progression. *Curr. Opin. Cell Biol.* **2020**, *63*, 144–153. https://doi.org/10.1016/j.ceb.2020.01.013.

(2) Martínez-Reyes, I.; Chandel, N. S. Mitochondrial TCA Cycle Metabolites Control Physiology and Disease. *Nat. Commun.* **2020**, *11* (1), 102. https://doi.org/10.1038/s41467-019-13668-3.

(3) Zhang, Z.; Tang, H.; Chen, P.; Xie, H.; Tao, Y. Demystifying the Manipulation of Host Immunity, Metabolism, and Extraintestinal Tumors by the Gut Microbiome. *Signal Transduct. Target. Ther.* **2019**, *4* (1), 41. https://doi.org/10.1038/s41392-019-0074-5.

(4) Li, X.; Egervari, G.; Wang, Y.; Berger, S. L.; Lu, Z. Regulation of Chromatin and Gene Expression by Metabolic Enzymes and Metabolites. *Nat. Rev. Mol. Cell Biol.* **2018**, *19* (9), 563–578. https://doi.org/10.1038/s41580-018-0029-7.

(5) Liu, J.; Harada, B. T.; He, C. Regulation of Gene Expression by N-Methyladenosine in Cancer. *Trends Cell Biol.* **2019**, *29* (6), 487–499. https://doi.org/10.1016/j.tcb.2019.02.008.

(6) Haws, S. A.; Leech, C. M.; Denu, J. M. Metabolism and the Epigenome: A Dynamic Relationship. *Trends Biochem. Sci.* **2020**, *45* (9), 731–747. https://doi.org/10.1016/j.tibs.2020.04.002.

(7) Steuer, A. E.; Brockbals, L.; Kraemer, T. Metabolomic Strategies in Biomarker Research-New Approach for Indirect Identification of Drug Consumption and Sample Manipulation in Clinical and Forensic Toxicology? *Front. Chem.* **2019**, *7*, 319. https://doi.org/10.3389/fchem.2019.00319.

(8) Carneiro, G.; Radcenco, A. L.; Evaristo, J.; Monnerat, G. Novel Strategies for Clinical Investigation and Biomarker Discovery: A Guide to Applied Metabolomics. *Horm. Mol. Biol. Clin. Investig.* **2019**, *38* (3), /j/hmbci.2019.38.issue-3/hmbci-2018-0045/hmbci-2018-0045.xml. https://doi.org/10.1515/hmbci-2018-0045.

(9) Ishikawa, S.; Sugimoto, M.; Kitabatake, K.; Tu, M.; Sugano, A.; Yamamori, I.; Iba, A.; Yusa, K.; Kaneko, M.; Ota, S.; Hiwatari, K.; Enomoto, A.; Masaru, T.; Iino, M. Effect of Timing of Collection of Salivary Metabolomic Biomarkers on Oral Cancer Detection. *Amino Acids* **2017**, *49* (4), 761–770. https://doi.org/10.1007/s00726-017-2378-5.

(10) Lou, S.; Balluff, B.; Cleven, A. H. G.; Bovée, J. V. M. G.; McDonnell, L. A. Prognostic Metabolite Biomarkers for Soft Tissue Sarcomas Discovered by Mass Spectrometry Imaging. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (2), 376–383. https://doi.org/10.1007/s13361-016-1544-4.

(11) Rodrigues, D.; Monteiro, M.; Jerónimo, C.; Henrique, R.; Belo, L.; Bastos, M. de L.; Guedes de Pinho, P.; Carvalho, M. Renal Cell Carcinoma: A Critical Analysis of Metabolomic Biomarkers Emerging from Current Model Systems. *Transl. Res.* **2017**, *180*, 1–11. https://doi.org/10.1016/j.trsl.2016.07.018.

(12) Ko, D.; Riles, E. M.; Marcos, E. G.; Magnani, J. W.; Lubitz, S. A.; Lin, H.; Long, M. T.; Schnabel, R. B.; McManus, D. D.; Ellinor, P. T.; Ramachandran, V. S.; Wang, T. J.; Gerszten, R. E.; Benjamin, E. J.; Yin, X.; Rienstra, M. Metabolomic Profiling in Relation to New-Onset Atrial Fibrillation (from the Framingham Heart Study). *Am. J. Cardiol.* **2016**, *118* (10), 1493–1496. https://doi.org/10.1016/j.amjcard.2016.08.010.

(13) Würtz, P.; Havulinna, A. S.; Soininen, P.; Tynkkynen, T.; Prieto-Merino, D.; Tillin, T.; Ghorbani, A.; Artati, A.; Wang, Q.; Tiainen, M. Metabolite Profiling and Cardiovascular Event Risk: A Prospective Study of 3 Population-Based Cohorts. *Circulation* **2015**, *131*, 774–785.

(14) Afshinnia, F.; Rajendiran, T. M.; Karnovsky, A.; Soni, T.; Wang, X.; Xie, D.; Yang, W.; Shafi, T.; Weir, M. R.; He, J. Lipidomic Signature of Progression of Chronic Kidney Disease in the Chronic Renal Insufficiency Cohort. *Kidney Int Rep* **2016**, *1*, 256–268.

(15) Elmariah, S.; Farrell, L. A.; Daher, M.; Shi, X.; Keyes, M. J.; Cain, C. H.; Pomerantsev, E.; Vlahakes, G. J.; Inglessis, I.; Passeri, J. J. Metabolite Profiles Predict Acute Kidney Injury and Mortality in Patients Undergoing Transcatheter Aortic Valve Replacement. *J Am Heart Assoc* **2016**, *5*, 002712.

(16) Paige, M.; Burdick, M. D.; Kim, S.; Xu, J.; Lee, J. K.; Michael Shim, Y. Pilot Analysis of the Plasma Metabolite Profiles Associated with Emphysematous Chronic Obstructive Pulmonary Disease Phenotype. *Biochem. Biophys. Res. Commun.* **2011**, *413* (4), 588–593. https://doi.org/10.1016/j.bbrc.2011.09.006.

(17) Sysi-Aho, M.; Ermolov, A.; Gopalacharyulu, P. V.; Tripathi, A.; Seppänen-Laakso, T.; Maukonen, J.; Mattila, I.; Ruohonen, S. T.; Vähätalo, L.; Yetukuri, L. Metabolic Regulation in Progression to Autoimmune Diabetes. *PLoS Comput Biol* **2011**, *7*, 1002257.

(18) Galderisi, A.; Pirillo, P.; Moret, V.; Stocchero, M.; Gucciardi, A.; Perilongo, G.; Moretti, C.; Monciotti, C.; Giordano, G.; Baraldi, E. Metabolomics Reveals New Metabolic Perturbations in Children with Type 1 Diabetes. *Pediatr Diabetes* **2018**, *19*, 59–67.

(19) Wang, T. J.; Larson, M. G.; Vasan, R. S.; Cheng, S.; Rhee, E. P.; McCabe, E.; Lewis, G. D.; Fox, C. S.; Jacques, P. F.; Fernandez, C.; O'Donnell, C. J.; Carr, S. A.; Mootha, V. K.; Florez, J. C.; Souza, A.; Melander, O.; Clish, C. B.; Gerszten, R. E. Metabolite Profiles and the Risk of Developing Diabetes. *Nat. Med.* **2011**, *17* (4), 448–453. https://doi.org/10.1038/nm.2307.

(20) Cheng, S.; Rhee, E. P.; Larson, M. G.; Lewis, G. D.; McCabe, E. L.; Shen, D.; Palma, M. J.; Roberts, L. D.; Dejam, A.; Souza, A. L. Metabolite Profiling Identifies Pathways Associated with Metabolic Risk in Humans. *Circulation* **2012**, *125*, 2222–2231.

(21) Mass Spectrometry-Based Hair Metabolomics for Biomarker Discovery. *Mass Spectrom. Lett.* **2022**, *13* (1), 2–10. https://doi.org/10.5478/MSL.2022.13.1.2.

(22) Arn, P. H. Newborn Screening: Current Status. *Health Aff. Proj. Hope* **2007**, *26* (2), 559–566. https://doi.org/10.1377/hlthaff.26.2.559.

(23) Kashani, K.; Rosner, M. H.; Ostermann, M. Creatinine: From Physiology to Clinical Application. *Eur. J. Intern. Med.* **2020**, *72*, 9–14. https://doi.org/10.1016/j.ejim.2019.10.025.

(24) Targher, G.; Byrne, C. D. Circulating Markers of Liver Function and Cardiovascular Disease Risk. *Arterioscler. Thromb. Vasc. Biol.* **2015**, *35* (11), 2290–2296. https://doi.org/10.1161/ATVBAHA.115.305235.

(25) Sreekumar, A.; Poisson, L. M.; Rajendiran, T. M.; Khan, A. P.; Cao, Q.; Yu, J.; Laxman, B.; Mehra, R.; Lonigro, R. J.; Li, Y.; Nyati, M. K.; Ahsan, A.; Kalyana-Sundaram, S.; Han, B.; Cao, X.; Byun, J.; Omenn, G. S.; Ghosh, D.; Pennathur, S.; Alexander, D. C.; Berger, A.; Shuster, J. R.; Wei, J. T.; Varambally, S.; Beecher, C.; Chinnaiyan, A. M. Metabolomic Profiles Delineate Potential Role for Sarcosine in Prostate Cancer Progression. *Nature* **2009**, *457* (7231), 910–914. https://doi.org/10.1038/nature07762.

(26) Koeth, R. A.; Wang, Z.; Levison, B. S.; Buffa, J. A.; Org, E.; Sheehy, B. T.; Britt, E. B.; Fu, X.; Wu, Y.; Li, L.; Smith, J. D.; DiDonato, J. A.; Chen, J.; Li, H.; Wu, G. D.; Lewis, J. D.; Warrier, M.; Brown, J. M.; Krauss, R. M.; Tang, W. H. W.; Bushman, F. D.; Lusis, A. J.; Hazen, S. L. Intestinal Microbiota Metabolism of L-Carnitine, a Nutrient in Red Meat, Promotes Atherosclerosis. *Nat. Med.* **2013**, *19* (5), 576–585. https://doi.org/10.1038/nm.3145.

(27) Mayers, J. R.; Wu, C.; Clish, C. B.; Kraft, P.; Torrence, M. E.; Fiske, B. P.; Yuan, C.; Bao, Y.; Townsend, M. K.; Tworoger, S. S.; Davidson, S. M.; Papagiannakopoulos, T.; Yang, A.; Dayton, T. L.; Ogino, S.; Stampfer, M. J.; Giovannucci, E. L.; Qian, Z. R.; Rubinson, D. A.; Ma, J.; Sesso, H. D.; Gaziano, J. M.; Cochrane, B. B.; Liu, S.; Wactawski-Wende, J.; Manson, J. E.; Pollak, M. N.; Kimmelman, A. C.; Souza, A.; Pierce, K.; Wang, T. J.; Gerszten, R. E.; Fuchs, C. S.; Vander Heiden, M. G.; Wolpin, B. M. Elevation of Circulating Branched-Chain Amino Acids Is an Early Event in Human Pancreatic Adenocarcinoma Development. *Nat. Med.* **2014**, *20* (10), 1193–1198. https://doi.org/10.1038/nm.3686.

(28) Emwas, A.-H. M.; Salek, R. M.; Griffin, J. L.; Merzaban, J. NMR-Based Metabolomics in Human Disease Diagnosis: Applications, Limitations, and Recommendations. *Metabolomics* **2013**, *9* (5), 1048–1072. https://doi.org/10.1007/s11306-013-0524-y.

(29) López-Hernández, Y.; Monárrez-Espino, J.; Oostdam, A.-S. H.; Delgado, J. E. C.; Zhang, L.; Zheng, J.; Valdez, J. J. O.; Mandal, R.; González, F. de L. O.; Moreno, J. C. B.; Trejo-Medinilla, F. M.; López, J. A.; Moreno, J. A. E.; Wishart, D. S. Targeted Metabolomics Identifies High Performing Diagnostic and Prognostic Biomarkers for COVID-19. *Sci. Rep.* **2021**, *11* (1), 14732. https://doi.org/10.1038/s41598-021-94171-y.

(30) Aderemi, A. V.; Ayeleso, A. O.; Oyedapo, O. O.; Mukwevho, E. Metabolomics: A Scoping Review of Its Role as a Tool for Disease Biomarker Discovery in Selected Non-Communicable Diseases. *Metabolites* **2021**, *11* (7), 418. https://doi.org/10.3390/metabo11070418.

(31) Castelli, F. A.; Rosati, G.; Moguet, C.; Fuentes, C.; Marrugo-Ramírez, J.; Lefebvre, T.; Volland, H.; Merkoçi, A.; Simon, S.; Fenaille, F.; Junot, C. Metabolomics for Personalized Medicine: The Input of Analytical Chemistry from Biomarker Discovery to Point-of-Care Tests. *Anal. Bioanal. Chem.* **2022**, *414* (2), 759–789. https://doi.org/10.1007/s00216-021-03586-z.

(32) Long, N. P.; Yoon, S. J.; Anh, N. H.; Nghi, T. D.; Lim, D. K.; Hong, Y. J.; Hong, S.-S.; Kwon, S. W. A Systematic Review on Metabolomics-Based Diagnostic Biomarker Discovery and Validation in Pancreatic Cancer. *Metabolomics* **2018**, *14* (8), 109. https://doi.org/10.1007/s11306-018-1404-2.

(33) Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: Beyond Biomarkers and towards Mechanisms. *Nat. Rev. Mol. Cell Biol.* **2016**, *17* (7), 451–459. https://doi.org/10.1038/nrm.2016.25.

(34) Barnes, S.; Benton, H. P.; Casazza, K.; Cooper, S. J.; Cui, X.; Du, X.; Engler, J.; Kabarowski, J. H.; Li, S.; Pathmasiri, W.; Prasain, J. K.; Renfrow, M. B.; Tiwari, H. K. Training in Metabolomics Research. I. Designing the Experiment, Collecting and Extracting Samples and Generating Metabolomics Data. *J. Mass Spectrom. JMS* **2016**, *51* (7), 461–475. https://doi.org/10.1002/jms.3782.

(35) Vuckovic, D. Sample Preparation in Global Metabolomics of Biological Fluids and Tissues. In *Proteomic and Metabolomic Approaches to Biomarker Discovery*; Elsevier, 2020; pp 53–83. https://doi.org/10.1016/B978-0-12-818607-7.00004-9.

(36) Fiehn, O. Metabolomics by Gas Chromatography–Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr. Protoc. Mol. Biol.* **2016**, *114* (1). https://doi.org/10.1002/0471142727.mb3004s114.

(37) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779–787. https://doi.org/10.1021/ac051437y.

(38) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics* **2010**, *11* (1), 395. https://doi.org/10.1186/1471-2105-11-395.

(39) Pang, Z.; Chong, J.; Zhou, G.; de Lima Morais, D. A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P.-É.; Li, S.; Xia, J. MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* **2021**, *49* (W1), W388–W396. https://doi.org/10.1093/nar/gkab382.

(40) Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M. A Lipidome Atlas in MS-DIAL 4. *Nat. Biotechnol.* **2020**, *38* (10), 1159–1163. https://doi.org/10.1038/s41587-020-0531-2.

(41) Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine. In *Current Protocols in Bioinformatics*; Baxevanis, A. D., Petsko, G. A., Stein, L. D., Stormo, G. D., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012; p bi1411s37. https://doi.org/10.1002/0471250953.bi1411s37.

(42) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem.* **2017**, *89* (17), 8689–8695. https://doi.org/10.1021/acs.analchem.7b01069.

(43) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57* (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

(44) Bujak, R.; Struck-Lewicka, W.; Markuszewski, M. J.; Kaliszan, R. Metabolomics for Laboratory Diagnostics. *J. Pharm. Biomed. Anal.* **2015**, *113*, 108–120. https://doi.org/10.1016/j.jpba.2014.12.017.

(45) Xi, B.; Gu, H.; Baniasadi, H.; Raftery, D. Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data. In *Mass Spectrometry in Metabolomics*; Raftery, D., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2014; Vol. 1198, pp 333–353. https://doi.org/10.1007/978-1-4939-1258-2_22.

(46) Principal Component Analysis for Special Types of Data. In *Principal Component Analysis*; Springer Series in Statistics; Springer-Verlag: New York, 2002; pp 338–372. https://doi.org/10.1007/0-387-22440-8_13.

(47) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*, 6th ed.; Pearson Prentice Hall: Upper Saddle River, N.J, 2007.

(48) Johnson, S. C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32* (3), 241–254. https://doi.org/10.1007/BF02289588.

(49) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28* (1), 100. https://doi.org/10.2307/2346830.

(50) Bezdek, J. C.; Ehrlich, R.; Full, W. FCM: The Fuzzy c-Means Clustering Algorithm. *Comput. Geosci.* **1984**, *10* (2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

(51) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17* (3), 166–173. https://doi.org/10.1002/cem.785.

(52) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis – a Marriage of Convenience or a Shotgun Wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

(53) Want, E.; Masson, P. Processing and Analysis of GC/LC-MS-Based Metabolomics Data. In *Metabolic Profiling*; Metz, T. O., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 708, pp 277–298. https://doi.org/10.1007/978-1-61737-985-7_17.

(54) Mehmood, T.; Martens, H.; Sæbø, S.; Warringer, J.; Snipen, L. A Partial Least Squares Based Algorithm for Parsimonious Variable Selection. *Algorithms Mol. Biol.* **2011**, *6* (1), 27. https://doi.org/10.1186/1748-7188-6-27.

(55) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer New York: New York, NY, 2000. https://doi.org/10.1007/978-1-4757-3264-1.

(56) Xu, Y.; Zomer, S.; Brereton, R. G. Support Vector Machines: A Recent Method for Classification in Chemometrics. *Crit. Rev. Anal. Chem.* **2006**, *36* (3–4), 177–188. https://doi.org/10.1080/10408340600969486.

(57) Breiman, L. [No Title Found]. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(58) Liland, K. H. Multivariate Methods in Metabolomics – from Pre-Processing to Dimension Reduction and Statistical Analysis. *TrAC Trends Anal. Chem.* **2011**, *30* (6), 827–841. https://doi.org/10.1016/j.trac.2011.02.007.

(59) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13* (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

(60) Park, M. Y.; Hastie, T. Penalized Logistic Regression for Detecting Gene Interactions. *Biostatistics* **2008**, *9* (1), 30–50. https://doi.org/10.1093/biostatistics/kxm010.

(61) Kuo, T.-C.; Tian, T.-F.; Tseng, Y. J. 3Omics: A Web-Based Systems Biology Tool for Analysis, Integration and Visualization of Human Transcriptomic, Proteomic and Metabolomic Data. *BMC Syst. Biol.* **2013**, *7*, 64. https://doi.org/10.1186/1752-0509-7-64.

(62) Paley, S. M.; Karp, P. D. The Pathway Tools Cellular Overview Diagram and Omics Viewer. *Nucleic Acids Res.* **2006**, *34* (13), 3771–3778. https://doi.org/10.1093/nar/gkl334.

(63) García-Alcalde, F.; García-López, F.; Dopazo, J.; Conesa, A. Paintomics: A Web Based Tool for the Joint Visualization of Transcriptomics and Metabolomics Data. *Bioinforma. Oxf. Engl.* **2011**, *27* (1), 137–139. https://doi.org/10.1093/bioinformatics/btq594.

(64) Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V. G.; Scardoni, G.; Laudanna, C.; Sartor, M. A.; Stringer, K. A.; Jagadish, H. V.; Burant, C.; Athey, B.; Omenn, G. S. Metscape 2 Bioinformatics Tool for the Analysis and Visualization of Metabolomics and Gene Expression Data. *Bioinforma. Oxf. Engl.* **2012**, *28* (3), 373–380. https://doi.org/10.1093/bioinformatics/btr661.

(65) Cavalcante, R. G.; Patil, S.; Weymouth, T. E.; Bendinskas, K. G.; Karnovsky, A.; Sartor, M. A. ConceptMetab: Exploring Relationships among Metabolite Sets to Identify Links among Biomedical Concepts. *Bioinforma. Oxf. Engl.* **2016**, *32* (10), 1536–1543. https://doi.org/10.1093/bioinformatics/btw016.

(66) López-Ibáñez, J.; Pazos, F.; Chagoyen, M. MBROLE 2.0-Functional Enrichment of Chemical Compounds. *Nucleic Acids Res.* **2016**, *44* (W1), W201-204. https://doi.org/10.1093/nar/gkw253.

(67) Xia, J.; Wishart, D. S. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.* **2016**, *55*, 14.10.1-14.10.91. https://doi.org/10.1002/cpbi.11.

(68) Chagoyen, M.; Pazos, F. Tools for the Functional Interpretation of Metabolomic Experiments. *Brief. Bioinform.* **2013**, *14* (6), 737–744. https://doi.org/10.1093/bib/bbs055.

(69) Khatri, P.; Sirota, M.; Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput. Biol.* **2012**, *8* (2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375.

(70) Wieder, C.; Frainay, C.; Poupin, N.; Rodríguez-Mier, P.; Vinson, F.; Cooke, J.; Lai, R. P.; Bundy, J. G.; Jourdan, F.; Ebbels, T. Pathway Analysis in Metabolomics: Recommendations for the Use of over-Representation Analysis. *PLOS Comput. Biol.* **2021**, *17* (9), e1009105. https://doi.org/10.1371/journal.pcbi.1009105.

(71) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res.* **2009**, *37* (1), 1–13. https://doi.org/10.1093/nar/gkn923.

(72) Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27–30. https://doi.org/10.1093/nar/28.1.27.

(73) Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; Paley, S. M.; Subhraveti, P. The BioCyc Collection of Microbial Genomes and Metabolic Pathways. *Brief. Bioinform.* **2019**, *20* (4), 1085–1093. https://doi.org/10.1093/bib/bbx085.

(74) Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; Loney, F.; May, B.; Milacic, M.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Weiser, J.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2019**, gkz1031. https://doi.org/10.1093/nar/gkz1031.

(75) Frolkis, A.; Knox, C.; Lim, E.; Jewison, T.; Law, V.; Hau, D. D.; Liu, P.; Gautam, B.; Ly, S.; Guo, A. C.; Xia, J.; Liang, Y.; Shrivastava, S.; Wishart, D. S. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D480–D487. https://doi.org/10.1093/nar/gkp1002.

(76) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci.* **2005**, *102* (43), 15545–15550. https://doi.org/10.1073/pnas.0506580102.

(77) Barry, W. T.; Nobel, A. B.; Wright, F. A. Significance Analysis of Functional Categories in Gene Expression Studies: A Structured Permutation Approach. *Bioinformatics* **2005**, *21* (9), 1943–1949. https://doi.org/10.1093/bioinformatics/bti260.

(78) Efron, B.; Tibshirani, R. On Testing the Significance of Sets of Genes. *Ann. Appl. Stat.* **2007**, *1* (1). https://doi.org/10.1214/07-AOAS101.

(79) Jiang, Z.; Gentleman, R. Extensions to Gene Set Enrichment. *Bioinformatics* **2007**, *23* (3), 306–313. https://doi.org/10.1093/bioinformatics/btl599.

(80) Xia, J.; Wishart, D. S. MSEA: A Web-Based Tool to Identify Biologically Meaningful Patterns in Quantitative Metabolomic Data. *Nucleic Acids Res.* **2010**, *38* (Web Server), W71–W77. https://doi.org/10.1093/nar/gkq329.

(81) Chagoyen, M.; Pazos, F. MBRole: Enrichment Analysis of Metabolomic Data. *Bioinformatics* **2011**, *27* (5), 730–731. https://doi.org/10.1093/bioinformatics/btr001.

(82) Kankainen, M.; Gopalacharyulu, P.; Holm, L.; Orešič, M. MPEA—Metabolite Pathway Enrichment Analysis. *Bioinformatics* **2011**, *27* (13), 1878–1879. https://doi.org/10.1093/bioinformatics/btr278.

(83) Kamburov, A.; Cavill, R.; Ebbels, T. M. D.; Herwig, R.; Keun, H. C. Integrated Pathway-Level Analysis of Transcriptomics and Metabolomics Data with IMPaLA. *Bioinformatics* **2011**, *27* (20), 2917–2918. https://doi.org/10.1093/bioinformatics/btr499.

(84) Caspi, R.; Billington, R.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Midford, P. E.; Ong, W. K.; Paley, S.; Subhraveti, P.; Karp, P. D. The MetaCyc Database of Metabolic Pathways and Enzymes - a 2019 Update. *Nucleic Acids Res.* **2020**, *48* (D1), D445–D453. https://doi.org/10.1093/nar/gkz862.

(85) Aggio, R. B. M. Pathway Activity Profiling (PAPi): A Tool for Metabolic Pathway Analysis. In *Yeast Metabolic Engineering*; Mapelli, V., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2014; Vol. 1152, pp 233–250. https://doi.org/10.1007/978-1-4939-0563-8_14.

(86) Kutmon, M.; van Iersel, M. P.; Bohler, A.; Kelder, T.; Nunes, N.; Pico, A. R.; Evelo, C. T. PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLOS Comput. Biol.* **2015**, *11* (2), e1004085. https://doi.org/10.1371/journal.pcbi.1004085.

(87) Barupal, D. K.; Haldiya, P. K.; Wohlgemuth, G.; Kind, T.; Kothari, S. L.; Pinkerton, K. E.; Fiehn, O. MetaMapp: Mapping and Visualizing Metabolomic Data by Integrating Information from Biochemical Pathways and Chemical and Mass Spectral Similarity. *BMC Bioinformatics* **2012**, *13* (1), 99. https://doi.org/10.1186/1471-2105-13-99.

(88) Cottret, L.; Wildridge, D.; Vinson, F.; Barrett, M. P.; Charles, H.; Sagot, M.-F.; Jourdan, F. MetExplore: A Web Server to Link Metabolomic Experiments and Genome-Scale Metabolic Networks. *Nucleic Acids Res.* **2010**, *38* (Web Server), W132–W137. https://doi.org/10.1093/nar/gkq312.

(89) Amara, A.; Frainay, C.; Jourdan, F.; Naake, T.; Neumann, S.; Novoa-del-Toro, E. M.; Salek, R. M.; Salzer, L.; Scharfenberg, S.; Witting, M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front. Mol. Biosci.* **2022**, *9*, 841373. https://doi.org/10.3389/fmolb.2022.841373.

(90) Robinson, J. L.; Kocabaş, P.; Wang, H.; Cholley, P.-E.; Cook, D.; Nilsson, A.; Anton, M.; Ferreira, R.; Domenzain, I.; Billa, V.; Limeta, A.; Hedin, A.; Gustafsson, J.; Kerkhoven, E. J.; Svensson, L. T.; Palsson, B. O.; Mardinoglu, A.; Hansson, L.; Uhlén, M.; Nielsen, J. An Atlas of Human Metabolism. *Sci. Signal.* **2020**, *13* (624), eaaz1482. https://doi.org/10.1126/scisignal.aaz1482.

(91) Lacroix, V.; Cottret, L.; Thebault, P.; Sagot, M.-F. An Introduction to Metabolic Networks and Their Structural Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2008**, *5* (4), 594–617. https://doi.org/10.1109/TCBB.2008.79.

(92) Liggi, S.; Griffin, J. L. Metabolomics Applied to Diabetes−lessons from Human Population Studies. *Int. J. Biochem. Cell Biol.* **2017**, *93*, 136–147. https://doi.org/10.1016/j.biocel.2017.10.011.

(93) del Mar Amador, M.; Colsch, B.; Lamari, F.; Jardel, C.; Ichou, F.; Rastetter, A.; Sedel, F.; Jourdan, F.; Frainay, C.; Wevers, R. A.; Roze, E.; Depienne, C.; Junot, C.; Mochel, F. Targeted versus Untargeted Omics — the CAFSA Story. *J. Inherit. Metab. Dis.* **2018**, *41* (3), 447–456. https://doi.org/10.1007/s10545-017-0134-3.

(94) Faust, K.; Dupont, P.; Callut, J.; van Helden, J. Pathway Discovery in Metabolic Networks by Subgraph Extraction. *Bioinformatics* **2010**, *26* (9), 1211–1218. https://doi.org/10.1093/bioinformatics/btq105.

(95) Bánky, D.; Iván, G.; Grolmusz, V. Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLoS ONE* **2013**, *8* (1), e54204. https://doi.org/10.1371/journal.pone.0054204.

(96) Frainay, C.; Aros, S.; Chazalviel, M.; Garcia, T.; Vinson, F.; Weiss, N.; Colsch, B.; Sedel, F.; Thabut, D.; Junot, C.; Jourdan, F. MetaboRank: Network-Based Recommendation System to Interpret and Enrich Metabolomics Results. *Bioinformatics* **2019**, *35* (2), 274–283. https://doi.org/10.1093/bioinformatics/bty577.

(97) Thiele, I.; Vlassis, N.; Fleming, R. M. T. FastGapFill: Efficient Gap Filling in Metabolic Networks. *Bioinformatics* **2014**, *30* (17), 2529–2531. https://doi.org/10.1093/bioinformatics/btu321.

(98) Pan, S.; Reed, J. L. Advances in Gap-Filling Genome-Scale Metabolic Models and Model-Driven Experiments Lead to Novel Metabolic Discoveries. *Curr. Opin. Biotechnol.* **2018**, *51*, 103–108. https://doi.org/10.1016/j.copbio.2017.12.012.

(99) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcantara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2007**, *36* (Database), D344–D350. https://doi.org/10.1093/nar/gkm791.

(100) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.;

Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25* (1), 25–29. https://doi.org/10.1038/75556.

(101) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. Mass Spectral Molecular Networking of Living Microbial Colonies. *Proc. Natl. Acad. Sci.* **2012**, *109* (26). https://doi.org/10.1073/pnas.1203689109.

(102) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34* (8), 828–837. https://doi.org/10.1038/nbt.3597.

(103) Olivon, F.; Elie, N.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D. MetGem Software for the Generation of Molecular Networks Based on the T-SNE Algorithm. *Anal. Chem.* **2018**, *90* (23), 13900–13908. https://doi.org/10.1021/acs.analchem.8b03099.

(104) Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; van der Hooft, J. J. J. Spec2Vec: Improved Mass Spectral Similarity Scoring through Learning of Structural Relationships. *PLOS Comput. Biol.* **2021**, *17* (2), e1008724. https://doi.org/10.1371/journal.pcbi.1008724.

(105) 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* (1), 21. https://doi.org/10.1186/1752-0509-5-21.

(106) Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data*; Springer Series in Statistics; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-20192-9.

(107) Friedman, J.; Hastie, T.; Tibshirani, R. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* **2008**, *9*, 432–441.

(108) Meinshausen, N.; Bühlmann, P. High-Dimensional Graphs and Variable Selection with the Lasso. *Ann. Stat.* **2006**, *34* (3). https://doi.org/10.1214/009053606000000281.

(109) Basu, S.; Duren, W.; Evans, C. R.; Burant, C. F.; Michailidis, G.; Karnovsky, A. Sparse Network Modeling and Metscape-Based Visualization Methods for the Analysis of Large-Scale Metabolomics Data. *Bioinformatics* **2017**, btx012. https://doi.org/10.1093/bioinformatics/btx012.

(110) Zhang, B.; Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4* (1). https://doi.org/10.2202/1544-6115.1128.

(111) Langfelder, P.; Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* **2008**, *9* (1), 559. https://doi.org/10.1186/1471-2105-9-559.

(112) Osterhoff, M.; Frahnow, T.; Seltmann, A.; Mosig, A.; Neunübel, K.; Sales, S.; Sampaio, J.; Hornemann, S.; Kruse, M.; Pfeiffer, A. Identification of Gene-Networks Associated with Specific Lipid Metabolites by Weighted Gene Co-Expression Network Analysis (WGCNA). *Exp. Clin. Endocrinol. Diabetes* **2014**, *122* (03), s-0034-1372115. https://doi.org/10.1055/s-0034-1372115.

(113) Vernocchi, P.; Gili, T.; Conte, F.; Del Chierico, F.; Conta, G.; Miccheli, A.; Botticelli, A.; Paci, P.; Caldarelli, G.; Nuti, M.; Marchetti, P.; Putignani, L. Network Analysis of Gut Microbiome and Metabolome to Discover Microbiota-Linked Biomarkers in Patients Affected by Non-Small Cell Lung Cancer. *Int. J. Mol. Sci.* **2020**, *21* (22), 8730. https://doi.org/10.3390/ijms21228730.

(114) Petersen, C.; Dai, D. L. Y.; Boutin, R. C. T.; Sbihi, H.; Sears, M. R.; Moraes, T. J.; Becker, A. B.; Azad, M. B.; Mandhane, P. J.; Subbarao, P.; Turvey, S. E.; Finlay, B. B. A Rich Meconium Metabolome in Human Infants Is Associated with Early-Life Gut Microbiota Composition and Reduced Allergic Sensitization. *Cell Rep. Med.* **2021**, *2* (5), 100260. https://doi.org/10.1016/j.xcrm.2021.100260.

(115) Wu, J.; Ye, Y.; Quan, J.; Ding, R.; Wang, X.; Zhuang, Z.; Zhou, S.; Geng, Q.; Xu, C.; Hong, L.; Xu, Z.; Zheng, E.; Cai, G.; Wu, Z.; Yang, J. Using Nontargeted LC-MS Metabolomics to Identify the Association of Biomarkers in Pig Feces with Feed Efficiency. *Porc. Health Manag.* **2021**, *7* (1), 39. https://doi.org/10.1186/s40813-021-00219-w.

(116) DiLeo, M. V.; Strahan, G. D.; den Bakker, M.; Hoekenga, O. A. Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLoS ONE* **2011**, *6* (10), e26683. https://doi.org/10.1371/journal.pone.0026683.

(117) German, J. B.; Zivkovic, A. M.; Dallas, D. C.; Smilowitz, J. T. Nutrigenomics and Personalized Diets: What Will They Mean for Food? Annu. *Rev Food Sci Technol* **2001**, *2*, 97–123.

(118) McKay, A. J.; Mathers, C. J. Diet Induced Epigenetic Changes and Their Implications for Health. *Acta Physiol* **2011**, *202*, 103–118.

(119) Conterno, L.; Fava, F.; Viola, R.; Tuohy, K. M. Obesity and the Gut Microbiota: Does up-Regulating Colonic Fermentation Protect against Obesity and Metabolic Disease? Genes Nutr, 2011, *6*, 241–260.

(120) Wild, C. P. Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiol Biomark Prev* **2005**, *14*, 1847–1850.

(121) Llorach, R.; Garcia-Aloy, M.; Tulipani, S.; Vazquez-Fresno, R.; Andres-Lacueva, C. Nutrimetabolomic Strategies to Develop New Biomarkers of Intake and Health Effects. *J Agric Food Chem* **2012**, *60*, 8797–8808.

(122) Cairns, R. A.; Harris, I. S.; Mak, T. W. Regulation of Cancer Cell Metabolism. *Nat Rev Cancer* **2011**, *11*, 85–95.

(123) Ko, D.; Riles, E. M.; Marcos, E. G.; Magnani, J. W.; Lubitz, S. A.; Lin, H.; Long, M. T.; Schnabel, R. B.; McManus, D. D.; Ellinor, P. T. Metabolomic Profiling in Relation to New-Onset Atrial Fibrillation (from the Framingham Heart Study. *Am J Cardiol* **2016**, *118*, 1493–1496.

(124) Gardinassi, L. G.; Xia, J.; Safo, S. E.; Li, S. Bioinformatics Tools for the Interpretation of Metabolomics Data. *Curr Pharmacol Rep* **2017**, *3*, 374–383.

(125) Hollywood, K.; Brison, D. R.; Goodacre, R. Metabolomics: Current Technologies and Future Trends. *Proteomics* **2016**, *6*, 4716–4723.

(126) Ma, J.; Karnovsky, A.; Afshinnia, F.; Wigginton, J.; Rader, D. J.; Natarajan, L.; Sharma, K.; Porter, A. C.; Rahman, M.; He, J.; Hamm, L.; Shafi, T.; Gipson, D.; Gadegbeku, C.; Feldman, H.; Michailidis, G.; Pennathur, S. Differential Network Enrichment Analysis Reveals Novel Lipid Pathways in Chronic Kidney Disease. *Bioinformatics* **2019**, *35* (18), 3441–3452. https://doi.org/10.1093/bioinformatics/btz114.

(127) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **2003**, *13*, 2498–2504.

(128) Guo, J.; Levina, E.; Michailidis, G.; Zhu, J. Joint Estimation of Multiple Graphical Models. *Biometrika* **2011**, *98*, 1–15.

(129) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res* **2018**, *46*, 608–617.

(130) Kind, T.; Liu, K. H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. LipidBlast in Silico Tandem Mass Spectrometry Database for Lipid Identification. *Nat Methods* **2013**, *10*, 755–758.

(131) Fahrmann, J.; Grapov, D.; Yang, J.; Hammock, B.; Fiehn, O.; Bell, G. I.; Hara, M. Systemic Alterations in the Metabolome of Diabetic NOD Mice Delineate Increased Oxidative Stress Accompanied by Reduced Inflammation and Hypertriglyceremia. *Am J Physiol Endocrinol Metab* **2015**, *308*, 978–989.

(132) Grapov, D.; Fahrmann, J.; Hwang, J.; Poudel, A.; Jo, J.; Periwal, V.; Fiehn, O.; Hara, M. Diabetes Associated Metabolomic Perturbations in NOD Mice. *Metabolomics* **2015**, *11*, 425–437.

(133) Kannel, W. B.; McGee, D. L. Diabetes and Cardiovascular Disease: The Framingham Study. *JAMA* **1979**, *241*, 2035–2038.

(134) Merino, J.; Leong, A.; Liu, C. T.; Porneala, B.; Walford, G. A.; Grotthuss, M.; Wang, T. J.; Flannick, J.; Dupuis, J.; Levy, D. Metabolomics Insights into Early Type 2 Diabetes Pathogenesis and Detection in Individuals with Normal Fasting Glucose. *Diabetologia* **2018**, *61*, 1315–1324.

(135) LaBarre, J. L.; Puttabyatappa, M.; Song, P. X.; Goodrich, J. M.; Zhou, L.; Rajendiran, T. M.; Soni, T.; Domino, S. E.; Treadwell, M. C.; Dolinoy, D. C. Maternal Lipid Levels across Pregnancy Impact the Umbilical Cord Blood Lipidome and Infant Birth Weight. *Sci Rep* **2020**, *10*, 1–15.

(136) Yuan, M.; Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *J R Stat Soc Ser B Stat Methodol* **2006**, *68*, 49–67.

(137) Yang, Y.; Zou, H. G. G. L. A. S. S. O. Group Lasso Penalized Learning Using a Unified BMD Algorithm. *R Package Version* **2013**, *1*.

(138) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58* (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

(139) Shojaie, A.; Michailidis, G. Analysis of Gene Sets Based on the Underlying Regulatory Network. *J Comput Biol* **2009**, *16*, 407–426.

(140) Shojaie, A.; Michailidis, G. Network Enrichment Analysis in Complex Experiments. *Stat Appl Genet Mol Biol* **2010**, 9.

(141) Maritim, A. C.; Sanders, A.; Watkins Iii, J. B. Diabetes, Oxidative Stress, and Antioxidants: A Review. *J Biochem Mol Toxicol* **2003**, *17*, 24–38.

(142) Rabinovitch, A. L. E. X.; Suarez-Pinzon, W. L.; Strynadka, K.; Lakey, J. R.; Rajotte, R. V. Human Pancreatic Islet Beta-Cell Destruction by Cytokines Involves Oxygen Free Radicals and Aldehyde Production. *J Clin Endocrinol Metab* **1996**, *81*, 3197–3202.

(143) Hayes, J. D.; McLELLAN, L. I. Glutathione and Glutathione-Dependent Enzymes Represent a Co-Ordinately Regulated Defence against Oxidative Stress. *Free Radic Res* **1999**, *31*, 273–300.

(144) Murakami, K.; Takahito, K.; Ohtsuka, Y.; Fujiwara, Y.; Shimada, M.; Kawakami, Y. Impairment of Glutathione Metabolism in Erythrocytes from Patients with Diabetes Mellitus. *Metabolism* **1989**, *38*, 753–758.

(145) Samiec, P. S.; Drews-Botsch, C.; Flagg, E. W.; Kurtz, J. C.; Sternberg, P., Jr.; Reed, R. L.; Jones, D. P. Glutathione in Human Plasma: Decline in Association with Aging, Age-Related Macular Degeneration, and Diabetes. *Free Radic Biol Med* **1998**, *24*, 699–704.

(146) Darmaun, D.; Smith, S. D.; Sweeten, S.; Sager, B. K.; Welch, S.; Mauras, N. Evidence for Accelerated Rates of Glutathione Utilization and Glutathione Depletion in Adolescents with Poorly Controlled Type 1 Diabetes. *Diabetes* **2005**, *54*, 190–196.

(147) Dincer, Y.; Akcay, T.; Alademir, Z.; Ilkova, H. Effect of Oxidative Stress on Glutathione Pathway in Red Blood Cells from Patients with Insulin-Dependent Diabetes Mellitus. *Metab Clin Exp* **2002**, *51*, 1360–1362.

(148) Dotan, I.; Shechter, I. Thiol-Disulfide-Dependent Interconversion of Active and Latent Forms of Rat Hepatic 3-Hydroxy-3-Methylglutaryl-Coenzyme A Reductase. *Biochim Biophys Acta BBA Lipids Lipid Metab* **1982**, *713*, 427–434.

(149) Roitelman, J.; Shechter, I. Regulation of Rat Liver 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase. Evidence for Thiol-Dependent Allosteric Modulation of Enzyme Activity. *J Biol Chem* **1984**, *259*, 870–877.

(150) Cappel, R. E.; Gilbert, H. F. Thiol/Disulfide Exchange between 3-Hydroxy-3-Methylglutaryl-CoA Reductase and Glutathione. A Thermodynamically Facile Dithiol Oxidation. *J Biol Chem* **1988**, *263*, 12204–12212.

(151) Gustafsson, J.; Carlsson, B.; Larsson, A. Cholesterol Synthesis in Patients with Glutathione Deficiency. *Eur J Clin Investig* **1990**, *20*, 470–474.

(152) Sample, C. E.; Ness, G. C. Regulation of the Activity of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase by Insulin. *Biochem Biophys Res Commun* **1986**, *137*, 201–207.

(153) Konorev, E. A.; Hogg, N.; Kalyanaraman, B. Rapid and Irreversible Inhibition of Creatine Kinase by Peroxynitrite. *FEBS Lett* **1998**, *427*, 171–174.

(154) Jiang, Z.; Kohzuki, M.; Harada, T.; Sato, T. Glutathione Suppresses Increase of Serum Creatine Kinase in Experimental Hypoglycemia. *Diabetes Res Clin Pr.* **2007**, *77*, 357–362.

(155) Horecker, B. L.; Land, K.; Takagi, Y. I. S. on M. Physiology and Clinical Use of Pentoses and Pentitols, 1969.

(156) Chukwuma, C. I.; Islam, M. S. Xylitol Improves Anti-Oxidative Defense System in Serum, Liver, Heart, Kidney and Pancreas of Normal and Type 2 Diabetes Model of Rats. *Acta Pol Pharm* **2017**, *74*, 817–826.

(157) Burant, C. F.; Flink, S.; DePaoli, A. M.; Chen, J.; Lee, W. S.; Hediger, M. A.; Buse, J. B.; Chang, E. B. Small Intestine Hexose Transport in Experimental Diabetes. Increased Transporter MRNA and Protein Expression in Enterocytes. *J Clin Investig* **1994**, *93*, 578–585.

(158) Vaarala, O. Leaking Gut in Type 1 Diabetes. *Curr Opin Gastroenterol* **2008**, *24*, 701–706.

(159) Wołoszyn-Durkiewicz, A.; Myśliwiec, M. The Prognostic Value of Inflammatory and Vascular Endothelial Dysfunction Biomarkers in Microvascular and Macrovascular Complications in Type 1 Diabetes. *Pediatr Endocrinol Diabetes Metab* **2019**, *25*, 28–35.

(160) Wang, T. J.; Wollert, K. C.; Larson, M. G.; Coglianese, E.; McCabe, E. L.; Cheng, S.; Ho, J. E.; Fradley, M. G.; Ghorbani, A.; Xanthakis, V. Prognostic Utility of Novel Biomarkers of Cardiovascular Stress: The Framingham Heart Study. *Circulation* **2012**, *126*, 1596–1604.

(161) Meinshausen, N.; Bühlmann, P. Stability Selection. *J R Stat Soc Ser B Stat Methodol* **2010**, *72*, 417–473.

(162) Yu, E.; Papandreou, C.; Ruiz-Canela, M.; Guasch-Ferre, M.; Clish, C. B.; Dennis, C.; Liang, L.; Corella, D.; Fitó, M.; Razquin, C. Association of Tryptophan Metabolites with Incident Type 2 Diabetes in the PREDIMED Trial: A Case–Cohort Study. *Clin Chem* **2018**, *64*, 1211–1220.

(163) Rebnord, E. W.; Strand, E.; Midttun, Ø.; Svingen, G. F.; Christensen, M. H.; Ueland, P. M.; Mellgren, G.; Njølstad, P. R.; Tell, G. S.; Nygård, O. K. The Kynurenine: Tryptophan Ratio as a Predictor of Incident Type 2 Diabetes Mellitus in Individuals with Coronary Artery Disease. *Diabetologia* **2017**, *60*, 1712–1721.

(164) Wang, T. J.; Ngo, D.; Psychogios, N.; Dejam, A.; Larson, M. G.; Vasan, R. S.; Ghorbani, A.; O'Sullivan, J.; Cheng, S.; Rhee, E. P. 2-Aminoadipic Acid Is a Biomarker for Diabetes Risk. *J Clin Investig* **2013**, *123*, 4309–4317.

(165) Kushiyama, A.; Nakatsu, Y.; Matsunaga, Y.; Yamamotoya, T.; Mori, K.; Ueda, K.; Inoue, Y.; Sakoda, H.; Fujishiro, M.; Ono, H. Role of Uric Acid Metabolism-Related Inflammation in the Pathogenesis of Metabolic Syndrome Components Such as Atherosclerosis and Nonalcoholic Steatohepatitis. *Mediat Inflamm* **2016**, 1–15.

(166) Cicero, A. F. G.; Fogacci, F.; Giovannini, M.; Grandi, E.; Rosticci, M.; D'Addato, S.; Borghi, C. Serum Uric Acid Predicts Incident Metabolic Syndrome in the Elderly in an Analysis of the Brisighella Heart Study. *Sci Rep* **2018**, *8*, 1–6.

(167) Patti, M. E.; Corvera, S. The Role of Mitochondria in the Pathogenesis of Type 2 Diabetes. *Endocr Rev* **2010**, *31*, 364–395.

(168) Miselli, M. A.; Dalla Nora, E.; Passaro, A.; Tomasi, F.; Zuliani, G. Plasma Triglycerides Predict Ten-Years All-Cause Mortality in Outpatients with Type 2 Diabetes Mellitus: A Longitudinal Observational Study. *Cardiovasc Diabetol* **2014**, *13*, 135.

(169) Zhao, J.; Zhang, Y.; Wei, F.; Song, J.; Cao, Z.; Chen, C.; Zhang, K.; Feng, S.; Wang, Y.; Li, W. D. Triglyceride Is an Independent Predictor of Type 2 Diabetes among Middle-Aged and Older Adults: A Prospective Study with 8-Year Follow-Ups in Two Cohorts. *J Transl Med* **2019**, *17*, 403.

(170) Bennion, L. J.; Grundy, S. M. Effects of diabetes mellitus on cholesterol metabolism in man. N. *Engl J Med* **1977**, *296*, 1365–1371.

(171) Staels, B.; Kuipers, F. Bile Acid Sequestrants and the Treatment of Type 2 Diabetes Mellitus. *Drugs* **2007**, *67*, 1383–1392.

(172) Lefebvre, P.; Cariou, B.; Lien, F.; Kuipers, F.; Staels, B. Role of Bile Acids and Bile Acid Receptors in Metabolic Regulation. *Physiol Rev* **2009**, *89*, 147–191.

(173) Suhre, K.; Meisinger, C.; Döring, A.; Altmaier, E.; Belcredi, P.; Gieger, C.; Chang, D.; Milburn, M. V.; Gall, W. E.; Weinberger, K. M. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS ONE* **2010**, *5*, 13953.

(174) Prawitt, J.; Caron, S.; Staels, B. Bile Acid Metabolism and the Pathogenesis of Type 2 Diabetes. *Curr Diabetes Rep* **2011**, *11*, 160.

(175) Guiastrennec, B.; Sonne, D. P.; Bergstrand, M.; Vilsbøll, T.; Knop, F. K.; Karlsson, M. O. Model-Based Prediction of Plasma Concentration and Enterohepatic Circulation of Total Bile Acids in Humans. *CPT Pharmacomet Syst Pharmacol* **2018**, *7*, 603–612.

(176) Newgard, C. B.; An, J.; Bain, J. R.; Muehlbauer, M. J.; Stevens, R. D.; Lien, L. F.; Haqq, A. M.; Shah, S. H.; Arlotto, M.; Slentz, C. A. A Branched-Chain Amino Acid-Related Metabolic Signature That Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cell Metab* **2009**, *9*, 311–326.

(177) Pettitt, D. J.; Jovanovic, L. Birth Weight as a Predictor of Type 2 Diabetes Mellitus: The U-Shaped Curve. *Curr Diabetes Rep* **2001**, *1*, 78–81.

(178) Fenton, T. R.; Kim, J. H. A Systematic Review and Meta-Analysis to Revise the Fenton Growth Chart for Preterm Infants. *BMC Pediatr* **2013**, *13*, 59.

(179) Lu, Y. P.; Reichetzeder, C.; Prehn, C.; Yin, L. H.; Yun, C.; Zeng, S.; Chu, C.; Adamski, J.; Hocher, B. Cord Blood Lysophosphatidylcholine 16: 1 Is Positively Associated with Birth Weight. *Cell Physiol Biochem* **2018**, *45*, 614–624.

(180) Maeba, R.; Nishimukai, M.; Sakasegawa, S. I.; Sugimori, D.; Hara, H. Plasma/Serum Plasmalogens: Methods of Analysis and Clinical Significance. In *Advances in Clinical Chemistry*; Amsterdam, The Netherlands: Elsevier, 2015; Vol. 70, pp 31–94.

(181) Brenseke, B.; Prater, M. R.; Bahamonde, J.; Gutierrez, J. C. Current Thoughts on Maternal Nutrition and Fetal Programming of the Metabolic Syndrome. *J Pregnancy* **2013**, 1–13.

(182) Sonagra, A. D.; Biradar, S. M.; Dattatreya, K.; Jayaprakash Murthy, D. S. Normal Pregnancy-a State of Insulin Resistance. *J Clin Diagn Res JCDR* **2014**, *8*, 01–03.

(183) Haggarty, P.; Page, K.; Abramovich, D. R.; Ashton, J.; Brown, D. Long-Chain Polyunsaturated Fatty Acid Transport across the Perfused Human Placenta. *Placenta* **1997**, *18*, 635–642.

(184) Martínez-Victoria, E.; Yago, M. D. Omega 3 Polyunsaturated Fatty Acids and Body Weight. *Br J Nutr* **2012**, *107*, 107–116.

(185) Prieto-Sánchez, M. T.; Ruiz-Palacios, M.; Blanco-Carnero, J. E.; Pagan, A.; Hellmuth, C.; Uhl, O.; Peissner, W.; Ruiz-Alcaraz, A. J.; Parrilla, J. J.; Koletzko, B. Placental MFSD2a Transporter Is Related to Decreased DHA in Cord Blood of Women with Treated Gestational Diabetes. *Clin Nutr* **2017**, *36*, 513–521.

(186) Csárdi, Gábor; Nepusz, Tamás; Horvát, Szabolcs; Traag, Vincent; Zanini, Fabio; Noom, Daniel. Igraph, 2022. https://doi.org/10.5281/ZENODO.3630268.

(187) Clerkin, K. J.; Fried, J. A.; Raikhelkar, J.; Sayer, G.; Griffin, J. M.; Masoumi, A.; Jain, S. S.; Burkhoff, D.; Kumaraiah, D.; Rabbani, L.; Schwartz, A.; Uriel, N. COVID-19 and Cardiovascular Disease. *Circulation* **2020**, *141* (20), 1648–1655. https://doi.org/10.1161/CIRCULATIONAHA.120.046941.

(188) Guan, W.; Ni, Z.; Hu, Y.; Liang, W.; Ou, C.; He, J.; Liu, L.; Shan, H.; Lei, C.; Hui, D. S. C.; Du, B.; Li, L.; Zeng, G.; Yuen, K.-Y.; Chen, R.; Tang, C.; Wang, T.; Chen, P.; Xiang, J.; Li, S.; Wang, J.; Liang, Z.; Peng, Y.; Wei, L.; Liu, Y.; Hu, Y.; Peng, P.; Wang, J.; Liu, J.; Chen, Z.; Li, G.; Zheng, Z.; Qiu, S.; Luo, J.; Ye, C.; Zhu, S.; Zhong, N. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **2020**, *382* (18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032.

(189) Yang, X.; Yu, Y.; Xu, J.; Shu, H.; Xia, J.; Liu, H.; Wu, Y.; Zhang, L.; Yu, Z.; Fang, M.; Yu, T.; Wang, Y.; Pan, S.; Zou, X.; Yuan, S.; Shang, Y. Clinical Course and Outcomes of Critically Ill Patients with SARS-CoV-2 Pneumonia in Wuhan, China: A Single-Centered, Retrospective, Observational Study. *Lancet Respir. Med.* **2020**, *8* (5), 475–481. https://doi.org/10.1016/S2213-2600(20)30079-5.

(190) Zhou, F.; Yu, T.; Du, R.; Fan, G.; Liu, Y.; Liu, Z.; Xiang, J.; Wang, Y.; Song, B.; Gu, X.; Guan, L.; Wei, Y.; Li, H.; Wu, X.; Xu, J.; Tu, S.; Zhang, Y.; Chen, H.; Cao, B. Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: A Retrospective Cohort Study. *The Lancet* **2020**, *395* (10229), 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3.

(191) Mehta, P.; McAuley, D. F.; Brown, M.; Sanchez, E.; Tattersall, R. S.; Manson, J. J. COVID-19: Consider Cytokine Storm Syndromes and Immunosuppression. *The Lancet* **2020**, *395* (10229), 1033–1034. https://doi.org/10.1016/S0140-6736(20)30628-0.

(192) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pöhlmann, S. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181* (2), 271-280.e8. https://doi.org/10.1016/j.cell.2020.02.052.

(193) De Jong, A.; Chanques, G.; Jaber, S. Mechanical Ventilation in Obese ICU Patients: From Intubation to Extubation. *Crit. Care* **2017**, *21* (1), 63. https://doi.org/10.1186/s13054-017-1641-1.

(194) Maile, M. D.; Standiford, T. J.; Engoren, M. C.; Stringer, K. A.; Jewell, E. S.; Rajendiran, T. M.; Soni, T.; Burant, C. F. Associations of the Plasma Lipidome with Mortality in the Acute Respiratory Distress Syndrome: A Longitudinal Cohort Study. *Respir. Res.* **2018**, *19* (1), 60. https://doi.org/10.1186/s12931-018-0758-3.

(195) Shen, B.; Yi, X.; Sun, Y.; Bi, X.; Du, J.; Zhang, C.; Quan, S.; Zhang, F.; Sun, R.; Qian, L.; Ge, W.; Liu, W.; Liang, S.; Chen, H.; Zhang, Y.; Li, J.; Xu, J.; He, Z.; Chen, B.; Wang, J.; Yan, H.; Zheng, Y.; Wang, D.; Zhu, J.; Kong, Z.; Kang, Z.; Liang, X.; Ding, X.; Ruan, G.; Xiang, N.; Cai, X.; Gao, H.; Li, L.; Li, S.; Xiao, Q.; Lu, T.; Zhu, Y.; Liu, H.; Chen, H.; Guo, T. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **2020**, *182* (1), 59-72.e15. https://doi.org/10.1016/j.cell.2020.05.032.

(196) Wu, Q.; Zhou, L.; Sun, X.; Yan, Z.; Hu, C.; Wu, J.; Xu, L.; Li, X.; Liu, H.; Yin, P.; Li, K.; Zhao, J.; Li, Y.; Wang, X.; Li, Y.; Zhang, Q.; Xu, G.; Chen, H. Altered Lipid Metabolism in Recovered SARS Patients Twelve Years after Infection. *Sci. Rep.* **2017**, *7* (1), 9110. https://doi.org/10.1038/s41598-017-09536-z.

(197) Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2. *Nat. Microbiol.* **2020**, *5* (4), 536–544. https://doi.org/10.1038/s41564-020-0695-z.

(198) Kachman, M.; Habra, H.; Duren, W.; Wigginton, J.; Sajjakulnukit, P.; Michailidis, G.; Burant, C.; Karnovsky, A. Deep Annotation of Untargeted LC-MS Metabolomics Data with *Binner*. *Bioinformatics* **2020**, *36* (6), 1801–1806. https://doi.org/10.1093/bioinformatics/btz798.

(199) Singh, A. K.; Gillies, C. L.; Singh, R.; Singh, A.; Chudasama, Y.; Coles, B.; Seidu, S.; Zaccardi, F.; Davies, M. J.; Khunti, K. Prevalence of Co-morbidities and Their Association with Mortality in Patients with COVID -19: A Systematic Review and Meta-analysis. *Diabetes Obes. Metab.* **2020**, *22* (10), 1915–1924. https://doi.org/10.1111/dom.14124.

(200) Thomas, T.; Stefanoni, D.; Reisz, J. A.; Nemkov, T.; Bertolone, L.; Francis, R. O.; Hudson, K. E.; Zimring, J. C.; Hansen, K. C.; Hod, E. A.; Spitalnik, S. L.; D'Alessandro, A. COVID-19 Infection Alters Kynurenine and Fatty Acid Metabolism, Correlating with IL-6 Levels and Renal Status. *JCI Insight* **2020**, *5* (14), e140327. https://doi.org/10.1172/jci.insight.140327.

(201) Barberis, E.; Timo, S.; Amede, E.; Vanella, V. V.; Puricelli, C.; Cappellano, G.; Raineri, D.; Cittone, M. G.; Rizzi, E.; Pedrinelli, A. R.; Vassia, V.; Casciaro, F. G.; Priora, S.; Nerici, I.; Galbiati, A.; Hayden, E.; Falasca, M.; Vaschetto, R.; Sainaghi, P. P.; Dianzani, U.; Rolla, R.; Chiocchetti, A.; Baldanzi, G.; Marengo, E.; Manfredi, M. Large-Scale Plasma Analysis Revealed New Mechanisms and Molecules Associated with the Host Response to SARS-CoV-2. *Int. J. Mol. Sci.* **2020**, *21* (22), 8623. https://doi.org/10.3390/ijms21228623.

(202) Valdés, A.; Moreno, L. O.; Rello, S. R.; Orduña, A.; Bernardo, D.; Cifuentes, A. Metabolomics Study of COVID-19 Patients in Four Different Clinical Stages. *Sci. Rep.* **2022**, *12* (1), 1650. https://doi.org/10.1038/s41598-022-05667-0.

(203) Kerner, J.; Hoppel, C. Fatty Acid Import into Mitochondria. *Biochim. Biophys. Acta BBA - Mol. Cell Biol. Lipids* **2000**, *1486* (1), 1–17. https://doi.org/10.1016/S1388-1981(00)00044-5.

(204) Reuter, S. E.; Evans, A. M. Carnitine and Acylcarnitines: Pharmacokinetic, Pharmacological and Clinical Aspects. *Clin. Pharmacokinet.* **2012**, *51* (9), 553–572. https://doi.org/10.1007/BF03261931.

(205) Rinaldo, P.; Cowan, T. M.; Matern, D. Acylcarnitine Profile Analysis. *Genet. Med.* **2008**, *10* (2), 151–156. https://doi.org/10.1097/GIM.0b013e3181614289.

(206) Jenniskens, M.; Langouche, L.; Vanwijngaerden, Y.-M.; Mesotten, D.; Van den Berghe, G. Cholestatic Liver (Dys)Function during Sepsis and Other Critical Illnesses. *Intensive Care Med.* **2016**, *42* (1), 16–27. https://doi.org/10.1007/s00134-015-4054-0.

(207) Harnisch, L.-O.; Mihaylov, D.; Bein, T.; Apfelbacher, C.; Kiehntopf, M.; Bauer, M.; Moerer, O.; Quintel, M. Determination of Individual Bile Acids in Acute Respiratory Distress Syndrome Reveals a Specific Pattern of Primary and Secondary Bile Acids and a Shift to the Acidic Pathway as an Adaptive Response to the Critical Condition. *Clin. Chem. Lab. Med. CCLM* **2022**, *60* (6), 891–900. https://doi.org/10.1515/cclm-2021-1176.

(208) Carino, A.; Moraca, F.; Fiorillo, B.; Marchianò, S.; Sepe, V.; Biagioli, M.; Finamore, C.; Bozza, S.; Francisci, D.; Distrutti, E.; Catalanotti, B.; Zampella, A.; Fiorucci, S. Hijacking SARS-CoV-2/ACE2 Receptor Interaction by Natural and Semi-Synthetic Steroidal Agents Acting on Functional Pockets on the Receptor Binding Domain. *Front. Chem.* **2020**, *8*, 572885. https://doi.org/10.3389/fchem.2020.572885.

(209) Kumar, Y.; Yadav, R.; Bhatia, A. Can Natural Detergent Properties of Bile Acids Be Used Beneficially in Tackling Coronavirus Disease-19? *Future Virol.* **2020**, *15* (12), 779–782. https://doi.org/10.2217/fvl-2020-0210.

(210) Poochi, S. P.; Easwaran, M.; Balasubramanian, B.; Anbuselvam, M.; Meyyazhagan, A.; Park, S.; Bhotla, H. K.; Anbuselvam, J.; Arumugam, V. A.; Keshavarao, S.; Kanniyappan, G. V.; Pappusamy, M.; Kaul, T. Employing Bioactive Compounds Derived from *Ipomoea Obscura* (L.) to Evaluate Potential Inhibitor for SARS-CoV-2 Main Protease and ACE2 Protein. *Food Front.* **2020**, *1* (2), 168–179. https://doi.org/10.1002/fft2.29.

(211) Oda, E.; Hatada, K.; Kimura, J.; Aizawa, Y.; Thanikachalam, P. V.; Watanabe, K. Relationships Between Serum Unsaturated Fatty Acids and Coronary Risk Factors Negative Relations Between Nervonic Acid and Obesity-Related Risk Factors: **Negative Relations Between Nervonic Acid and Obesity-Related Risk Factors**. *Int. Heart. J.* **2005**, *46* (6), 975–985. https://doi.org/10.1536/ihj.46.975.

(212) Chen, Y.; Li, X.; Dai, Y.; Zhang, J. The Association Between COVID-19 and Thyroxine Levels: A Meta-Analysis. *Front. Endocrinol.* **2022**, *12*, 779692. https://doi.org/10.3389/fendo.2021.779692.

(213) Julkunen, H.; Cichońska, A.; Slagboom, P. E.; Würtz, P.; Nightingale Health UK Biobank Initiative. Metabolic Biomarker Profiling for Identification of Susceptibility to Severe Pneumonia and COVID-19 in the General Population. *eLife* **2021**, *10*, e63033. https://doi.org/10.7554/eLife.63033.

(214) Goutman, S. A.; Hardiman, O.; Al-Chalabi, A.; Chió, A.; Savelieff, M. G.; Kiernan, M. C.; Feldman, E. L. Recent Advances in the Diagnosis and Prognosis of Amyotrophic Lateral Sclerosis. *Lancet Neurol.* **2022**, *21* (5), 480–493. https://doi.org/10.1016/S1474-4422(21)00465-8.

(215) Goutman, S. A.; Hardiman, O.; Al-Chalabi, A.; Chió, A.; Savelieff, M. G.; Kiernan, M. C.; Feldman, E. L. Emerging Insights into the Complex Genetics and Pathophysiology of Amyotrophic Lateral Sclerosis. *Lancet Neurol.* **2022**, *21* (5), 465–479. https://doi.org/10.1016/S1474-4422(21)00414-2.

(216) Benatar, M.; Turner, M. R.; Wuu, J. Defining Pre-Symptomatic Amyotrophic Lateral Sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* **2019**, *20* (5–6), 303–309. https://doi.org/10.1080/21678421.2019.1587634.

(217) Moglia, C.; Calvo, A.; Grassano, M.; Canosa, A.; Manera, U.; D’Ovidio, F.; Bombaci, A.; Bersano, E.; Mazzini, L.; Mora, G.; Chiò, A. Early Weight Loss in Amyotrophic Lateral Sclerosis: Outcome Relevance and

Clinical Correlates in a Population-Based Cohort. *J. Neurol. Neurosurg. Psychiatry* **2019**, *90* (6), 666–673. https://doi.org/10.1136/jnnp-2018-319611.

(218) Peter, R. S.; Rosenbohm, A.; Dupuis, L.; Brehme, T.; Kassubek, J.; Rothenbacher, D.; Nagel, G.; Ludolph, A. C. Life Course Body Mass Index and Risk and Prognosis of Amyotrophic Lateral Sclerosis: Results from the ALS Registry Swabia. *Eur. J. Epidemiol.* **2017**, *32* (10), 901–908. https://doi.org/10.1007/s10654-017-0318-z.

(219) Nakken, O.; Meyer, H. E.; Stigum, H.; Holmøy, T. High BMI Is Associated with Low ALS Risk: A Population-Based Study. *Neurology* **2019**, *93* (5), e424–e432. https://doi.org/10.1212/WNL.0000000000007861.

(220) O'Reilly, É. J.; Wang, H.; Weisskopf, M. G.; Fitzgerald, K. C.; Falcone, G.; McCullough, M. L.; Thun, M.; Park, Y.; Kolonel, L. N.; Ascherio, A. Premorbid Body Mass Index and Risk of Amyotrophic Lateral Sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* **2013**, *14* (3), 205–211. https://doi.org/10.3109/21678421.2012.735240.

(221) O'Reilly, É. J.; Wang, M.; Adami, H.-O.; Alonso, A.; Bernstein, L.; van den Brandt, P.; Buring, J.; Daugherty, S.; Deapen, D.; Freedman, D. M.; English, D. R.; Giles, G. G.; Håkansson, N.; Kurth, T.; Schairer, C.; Weiderpass, E.; Wolk, A.; Smith-Warner, S. A. Prediagnostic Body Size and Risk of Amyotrophic Lateral Sclerosis Death in 10 Studies. *Amyotroph. Lateral Scler. Front. Degener.* **2018**, *19* (5–6), 396–406. https://doi.org/10.1080/21678421.2018.1452944.

(222) Mariosa, D.; Beard, J. D.; Umbach, D. M.; Bellocco, R.; Keller, J.; Peters, T. L.; Allen, K. D.; Ye, W.; Sandler, D. P.; Schmidt, S.; Fang, F.; Kamel, F. Body Mass Index and Amyotrophic Lateral Sclerosis: A Study of US Military Veterans. *Am. J. Epidemiol.* **2017**, *185* (5), 362–371. https://doi.org/10.1093/aje/kww140.

(223) Gallo, V.; Wark, P. A.; Jenab, M.; Pearce, N.; Brayne, C.; Vermeulen, R.; Andersen, P. M.; Hallmans, G.; Kyrozis, A.; Vanacore, N.; Vahdaninia, M.; Grote, V.; Kaaks, R.; Mattiello, A.; Bueno-de-Mesquita, H. B.; Peeters, P. H.; Travis, R. C.; Petersson, J.; Hansson, O.; Arriola, L.; Jimenez-Martin, J.-M.; Tjonneland, A.; Halkjaer, J.; Agnoli, C.; Sacerdote, C.; Bonet, C.; Trichopoulou, A.; Gavrila, D.; Overvad, K.; Weiderpass, E.; Palli, D.; Quiros, J. R.; Tumino, R.; Khaw, K.-T.; Wareham, N.; Barricante-Gurrea, A.; Fedirko, V.; Ferrari, P.; Clavel-Chapelon, F.; Boutron-Ruault, M.-C.; Boeing, H.; Vigl, M.; Middleton, L.; Riboli, E.; Vineis, P. Prediagnostic Body Fat and Risk of Death from Amyotrophic Lateral Sclerosis: The EPIC Cohort. *Neurology* **2013**, *80* (9), 829–838. https://doi.org/10.1212/WNL.0b013e3182840689.

(224) Ioannides, Z. A.; Ngo, S. T.; Henderson, R. D.; McCombe, P. A.; Steyn, F. J. Altered Metabolic Homeostasis in Amyotrophic Lateral Sclerosis: Mechanisms of Energy Imbalance and Contribution to Disease Progression. *Neurodegener. Dis.* **2016**, *16* (5–6), 382–397. https://doi.org/10.1159/000446502.

(225) Steyn, F. J.; Ioannides, Z. A.; van Eijk, R. P. A.; Heggie, S.; Thorpe, K. A.; Ceslis, A.; Heshmat, S.; Henders, A. K.; Wray, N. R.; van den Berg, L. H.; Henderson, R. D.; McCombe, P. A.; Ngo, S. T. Hypermetabolism in ALS Is Associated with Greater Functional Decline and Shorter Survival. *J. Neurol. Neurosurg. Psychiatry* **2018**, *89* (10), 1016–1023. https://doi.org/10.1136/jnnp-2017-317887.

(226) Dupuis, L.; Pradat, P.-F.; Ludolph, A. C.; Loeffler, J.-P. Energy Metabolism in Amyotrophic Lateral Sclerosis. *Lancet Neurol.* **2011**, *10* (1), 75–82. https://doi.org/10.1016/S1474-4422(10)70224-6.

(227) Mariosa, D.; Hammar, N.; Malmström, H.; Ingre, C.; Jungner, I.; Ye, W.; Fang, F.; Walldius, G. Blood Biomarkers of Carbohydrate, Lipid, and Apolipoprotein Metabolisms and Risk of Amyotrophic Lateral Sclerosis: A More than 20-Year Follow-up of the Swedish AMORIS Cohort: Blood Biomarkers of Energy Metabolism and ALS Risk. *Ann. Neurol.* **2017**, *81* (5), 718–728. https://doi.org/10.1002/ana.24936.

(228) Ingre, C.; Chen, L.; Zhan, Y.; Termorshuizen, J.; Yin, L.; Fang, F. Lipids, Apolipoproteins, and Prognosis of Amyotrophic Lateral Sclerosis. *Neurology* **2020**, *94* (17), e1835–e1844. https://doi.org/10.1212/WNL.0000000000009322.

(229) Cirulli, E. T.; Guo, L.; Leon Swisher, C.; Shah, N.; Huang, L.; Napier, L. A.; Kirkness, E. F.; Spector, T. D.; Caskey, C. T.; Thorens, B.; Venter, J. C.; Telenti, A. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metab.* **2019**, *29* (2), 488-500.e2. https://doi.org/10.1016/j.cmet.2018.09.022.

(230) Ho, J. E.; Larson, M. G.; Ghorbani, A.; Cheng, S.; Chen, M.-H.; Keyes, M.; Rhee, E. P.; Clish, C. B.; Vasan, R. S.; Gerszten, R. E.; Wang, T. J. Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. *PLOS ONE* **2016**, *11* (2), e0148361. https://doi.org/10.1371/journal.pone.0148361.

(231) Kraus, W. E.; Pieper, C. F.; Huffman, K. M.; Thompson, D. K.; Kraus, V. B.; Morey, M. C.; Cohen, H. J.; Ravussin, E.; Redman, L. M.; Bain, J. R.; Stevens, R. D.; Newgard, C. B. Association of Plasma Small-Molecule Intermediate Metabolites With Age and Body Mass Index Across Six Diverse Study Populations. *J. Gerontol. A. Biol. Sci. Med. Sci.* **2016**, *71* (11), 1507–1513. https://doi.org/10.1093/gerona/glw031.

(232) Goutman, S. A.; Boss, J.; Patterson, A.; Mukherjee, B.; Batterman, S.; Feldman, E. L. High Plasma Concentrations of Organic Pollutants Negatively Impact Survival in Amyotrophic Lateral Sclerosis. *J. Neurol. Neurosurg. Psychiatry* **2019**, *90* (8), 907–912. https://doi.org/10.1136/jnnp-2018-319785.

(233) Su, F.-C.; Goutman, S. A.; Chernyak, S.; Mukherjee, B.; Callaghan, B. C.; Batterman, S.; Feldman, E. L. Association of Environmental Toxins With Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **2016**, *73* (7), 803. https://doi.org/10.1001/jamaneurol.2016.0594.

(234) Yu, Y.; Su, F.-C.; Callaghan, B. C.; Goutman, S. A.; Batterman, S. A.; Feldman, E. L. Environmental Risk Factors and Amyotrophic Lateral Sclerosis (ALS): A Case-Control Study of ALS in Michigan. *PLoS ONE* **2014**, *9* (6), e101186. https://doi.org/10.1371/journal.pone.0101186.

(235) Goutman, S. A.; Boss, J.; Godwin, C.; Mukherjee, B.; Feldman, E. L.; Batterman, S. A. Associations of Self-Reported Occupational Exposures and Settings to ALS: A Case–Control Study. *Int. Arch. Occup. Environ. Health* **2022**, *95* (7), 1567–1586. https://doi.org/10.1007/s00420-022-01874-4.

(236) Brooks, B. R.; Miller, R. G.; Swash, M.; Munsat, T. L. El Escorial Revisited: Revised Criteria for the Diagnosis of Amyotrophic Lateral Sclerosis. *Amyotroph. Lateral Scler. Other Motor Neuron Disord.* **2000**, *1* (5), 293–299. https://doi.org/10.1080/146608200300079536.

(237) Keys, A.; Fidanza, F.; Karvonen, M. J.; Kimura, N.; Taylor, H. L. Indices of Relative Weight and Obesity. *J. Chronic Dis.* **1972**, *25* (6–7), 329–343. https://doi.org/10.1016/0021-9681(72)90027-6.

(238) Goutman, S. A.; Boss, J.; Guo, K.; Alakwaa, F. M.; Patterson, A.; Kim, S.; Savelieff, M. G.; Hur, J.; Feldman, E. L. Untargeted Metabolomics Yields Insight into ALS Disease Mechanisms. *J. Neurol. Neurosurg. Psychiatry* **2020**, *91* (12), 1329–1338. https://doi.org/10.1136/jnnp-2020-323611.

(239) Goutman, S. A.; Guo, K.; Savelieff, M. G.; Patterson, A.; Sakowski, S. A.; Habra, H.; Karnovsky, A.; Hur, J.; Feldman, E. L. Metabolomics Identifies Shared Lipid Pathways in Independent Amyotrophic Lateral Sclerosis Cohorts. *Brain* **2022**, awac025. https://doi.org/10.1093/brain/awac025.

(240) Liang, K.-Y.; Zeger, S. L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **1986**, *73* (1), 13–22. https://doi.org/10.1093/biomet/73.1.13.

(241) Halekoh, U.; Højsgaard, S.; Yan, J. The *R* Package **Geepack** for Generalized Estimating Equations. *J. Stat. Softw.* **2006**, *15* (2). https://doi.org/10.18637/jss.v015.i02.

(242) Genolini, C.; Alacoque, X.; Sentenac, M.; Arnaud, C. **Kml** and **Kml3d** : *R* Packages to Cluster Longitudinal Data. *J. Stat. Softw.* **2015**, *65* (4). https://doi.org/10.18637/jss.v065.i04.

(243) Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat. - Theory Methods* **1974**, *3* (1), 1–27. https://doi.org/10.1080/03610927408827101.

(244) Gower, J. C. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika* **1966**, *53* (3–4), 325–338. https://doi.org/10.1093/biomet/53.3-4.325.

(245) Wood, S. N. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models: Estimation of Semiparametric Generalized Linear Models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2011**, *73* (1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.

(246) Little, R. J. A.; Rubin, D. B. *Statistical Analysis with Missing Data*, Second edition.; Wiley: Hoboken, 2010.

(247) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33* (1), 1–22.

(248) Iyer, G. R.; Wigginton, J.; Duren, W.; LaBarre, J. L.; Brandenburg, M.; Burant, C.; Michailidis, G.; Karnovsky, A. Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations. *Metabolites* **2020**, *10* (12), E479. https://doi.org/10.3390/metabo10120479.

(249) Lancichinetti, A.; Fortunato, S. Consensus Clustering in Complex Networks. *Sci. Rep.* **2012**, *2* (1), 336. https://doi.org/10.1038/srep00336.

(250) Friedman, J.; Hastie, T.; Tibshirani, R. A Note on the Group Lasso and a Sparse Group Lasso. **2010**. https://doi.org/10.48550/ARXIV.1001.0736.

(251) Huisman, M. H. B.; Seelen, M.; van Doormaal, P. T. C.; de Jong, S. W.; de Vries, J. H. M.; van der Kooi, A. J.; de Visser, M.; Schelhaas, H. J.; van den Berg, L. H.; Veldink, J. H. Effect of Presymptomatic Body Mass Index and Consumption of Fat and Alcohol on Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **2015**, *72* (10), 1155. https://doi.org/10.1001/jamaneurol.2015.1584.

(252) Westeneng, H.-J.; van Veenhuijzen, K.; van der Spek, R. A.; Peters, S.; Visser, A. E.; van Rheenen, W.; Veldink, J. H.; van den Berg, L. H. Associations between Lifestyle and Amyotrophic Lateral Sclerosis Stratified by C9orf72 Genotype: A Longitudinal, Population-Based, Case-Control Study. *Lancet Neurol.* **2021**, *20* (5), 373–384. https://doi.org/10.1016/S1474-4422(21)00042-9.

(253) Janse van Mantgem, M. R.; van Eijk, R. P. A.; van der Burgh, H. K.; Tan, H. H. G.; Westeneng, H.-J.; van Es, M. A.; Veldink, J. H.; van den Berg, L. H. Prognostic Value of Weight Loss in Patients with Amyotrophic Lateral Sclerosis: A Population-Based Study. *J. Neurol. Neurosurg. Psychiatry* **2020**, *91* (8), 867–875. https://doi.org/10.1136/jnnp-2020-322909.

(254) Nakamura, R.; Kurihara, M.; Ogawa, N.; Kitamura, A.; Yamakawa, I.; Bamba, S.; Sanada, M.; Sasaki, M.; Urushitani, M. Prognostic Prediction by Hypermetabolism Varies Depending on the Nutritional Status in Early Amyotrophic Lateral Sclerosis. *Sci. Rep.* **2021**, *11* (1), 17943. https://doi.org/10.1038/s41598-021-97196-5.

(255) Jésus, P.; Fayemendy, P.; Nicol, M.; Lautrette, G.; Sourisseau, H.; Preux, P.-M.; Desport, J.-C.; Marin, B.; Couratier, P. Hypermetabolism Is a Deleterious Prognostic Factor in Patients with Amyotrophic Lateral Sclerosis. *Eur. J. Neurol.* **2018**, *25* (1), 97–104. https://doi.org/10.1111/ene.13468.

(256) Bjornevik, K.; Zhang, Z.; O'Reilly, É. J.; Berry, J. D.; Clish, C. B.; Deik, A.; Jeanfavre, S.; Kato, I.; Kelly, R. S.; Kolonel, L. N.; Liang, L.; Marchand, L. L.; McCullough, M. L.; Paganoni, S.; Pierce, K. A.; Schwarzschild, M. A.; Shadyab, A. H.; Wactawski-Wende, J.; Wang, D. D.; Wang, Y.; Manson, J. E.; Ascherio, A. Prediagnostic Plasma Metabolomics and the Risk of Amyotrophic Lateral Sclerosis. *Neurology* **2019**, 10.1212/WNL.0000000000007401. https://doi.org/10.1212/WNL.0000000000007401.

(257) Lawton, K. A.; Brown, M. V.; Alexander, D.; Li, Z.; Wulff, J. E.; Lawson, R.; Jaffa, M.; Milburn, M. V.; Ryals, J. A.; Bowser, R.; Cudkowicz, M. E.; Berry, J. D.; On behalf of the Northeast ALS Consortium. Plasma Metabolomic Biomarker Panel to Distinguish Patients with Amyotrophic Lateral Sclerosis from Disease Mimics. *Amyotroph. Lateral Scler. Front. Degener.* **2014**, *15* (5–6), 362–370. https://doi.org/10.3109/21678421.2014.908311.

(258) Lawton, K. A.; Cudkowicz, M. E.; Brown, M. V.; Alexander, D.; Caffrey, R.; Wulff, J. E.; Bowser, R.; Lawson, R.; Jaffa, M.; Milburn, M. V.; Ryals, J. A.; Berry, J. D. Biochemical Alterations Associated with ALS. *Amyotroph. Lateral Scler.* **2012**, *13* (1), 110–118. https://doi.org/10.3109/17482968.2011.619197.

(259) Sol, J.; Jové, M.; Povedano, M.; Sproviero, W.; Domínguez, R.; Piñol-Ripoll, G.; Romero-Guevara, R.; Hye, A.; Al-Chalabi, A.; Torres, P.; Andres-Benito, P.; Area-Gómez, E.; Pamplona, R.; Ferrer, I.; Ayala, V.; Portero-Otín, M. Lipidomic Traits of Plasma and Cerebrospinal Fluid in Amyotrophic Lateral Sclerosis Correlate with Disease Progression. *Brain Commun.* **2021**, *3* (3), fcab143. https://doi.org/10.1093/braincomms/fcab143.

(260) Cutler, R. G.; Pedersen, W. A.; Camandola, S.; Rothstein, J. D.; Mattson, M. P. Evidence That Accumulation of Ceramides and Cholesterol Esters Mediates Oxidative Stress-Induced Death of Motor Neurons in Amyotrophic Lateral Sclerosis. *Ann. Neurol.* **2002**, *52* (4), 448–457. https://doi.org/10.1002/ana.10312.

(261) Wang, G.; Bieberich, E. Sphingolipids in Neurodegeneration (with Focus on Ceramide and S1P). *Adv. Biol. Regul.* **2018**, *70*, 51–64. https://doi.org/10.1016/j.jbior.2018.09.013.

(262) Mohassel, P.; Donkervoort, S.; Lone, M. A.; Nalls, M.; Gable, K.; Gupta, S. D.; Foley, A. R.; Hu, Y.; Saute, J. A. M.; Moreira, A. L.; Kok, F.; Introna, A.; Logroscino, G.; Grunseich, C.; Nickolls, A. R.; Pourshafie, N.; Neuhaus, S. B.; Saade, D.; Gangfuß, A.; Kölbel, H.; Piccus, Z.; Le Pichon, C. E.; Fiorillo, C.; Ly, C. V.; Töpf, A.; Brady, L.; Specht, S.; Zidell, A.; Pedro, H.; Mittelmann, E.; Thomas, F. P.; Chao, K. R.; Konersman, C. G.; Cho, M. T.; Brandt, T.; Straub, V.; Connolly, A. M.; Schara, U.; Roos, A.; Tarnopolsky, M.; Höke, A.; Brown, R. H.; Lee, C.-H.; Hornemann, T.; Dunn, T. M.; Bönnemann, C. G. Childhood Amyotrophic Lateral Sclerosis Caused by Excess Sphingolipid Synthesis. *Nat. Med.* **2021**, *27* (7), 1197–1204. https://doi.org/10.1038/s41591-021-01346-1.

(263) Di Ciaula, A.; Garruti, G.; Lunardi Baccetto, R.; Molina-Molina, E.; Bonfrate, L.; Wang, D. Q.-H.; Portincasa, P. Bile Acid Physiology. *Ann. Hepatol.* **2017**, *16*, S4–S14. https://doi.org/10.5604/01.3001.0010.5493.

(264) Paganoni, S.; Macklin, E. A.; Hendrix, S.; Berry, J. D.; Elliott, M. A.; Maiser, S.; Karam, C.; Caress, J. B.; Owegi, M. A.; Quick, A.; Wymer, J.; Goutman, S. A.; Heitzman, D.; Heiman-Patterson, T.; Jackson, C. E.; Quinn, C.; Rothstein, J. D.; Kasarskis, E. J.; Katz, J.; Jenkins, L.; Ladha, S.; Miller, T. M.; Scelsa, S. N.; Vu, T. H.; Fournier, C. N.; Glass, J. D.; Johnson, K. M.; Swenson, A.; Goyal, N. A.; Pattee, G. L.; Andres, P. L.; Babu, S.; Chase, M.; Dagostino, D.; Dickson, S. P.; Ellison, N.; Hall, M.; Hendrix, K.; Kittle, G.; McGovern, M.; Ostrow, J.; Pothier, L.; Randall, R.; Shefner, J. M.; Sherman, A. V.; Tustison, E.; Vigneswaran, P.; Walker, J.; Yu, H.; Chan, J.; Wittes, J.; Cohen, J.; Klee, J.; Leslie, K.; Tanzi, R. E.; Gilbert, W.; Yeramian, P. D.; Schoenfeld, D.; Cudkowicz, M. E. Trial of Sodium Phenylbutyrate–Taurursodiol for Amyotrophic Lateral Sclerosis. *N. Engl. J. Med.* **2020**, *383* (10), 919–930. https://doi.org/10.1056/NEJMoa1916945.

(265) Paganoni, S.; Hendrix, S.; Dickson, S. P.; Knowlton, N.; Macklin, E. A.; Berry, J. D.; Elliott, M. A.; Maiser, S.; Karam, C.; Caress, J. B.; Owegi, M. A.; Quick, A.; Wymer, J.; Goutman, S. A.; Heitzman, D.; Heiman-Patterson, T. D.; Jackson, C. E.; Quinn, C.; Rothstein, J. D.; Kasarskis, E. J.; Katz, J.; Jenkins, L.; Ladha, S.; Miller, T. M.; Scelsa, S. N.; Vu, T. H.; Fournier, C. N.; Glass, J. D.; Johnson, K. M.; Swenson, A.; Goyal, N. A.; Pattee, G. L.; Andres, P. L.; Babu, S.; Chase, M.; Dagostino, D.; Hall, M.; Kittle, G.; Eydinov, M.; McGovern, M.; Ostrow, J.; Pothier, L.; Randall, R.; Shefner, J. M.; Sherman, A. V.; St Pierre, M. E.; Tustison, E.; Vigneswaran, P.; Walker, J.; Yu, H.; Chan, J.; Wittes, J.; Yu, Z.; Cohen, J.; Klee, J.; Leslie, K.; Tanzi, R. E.; Gilbert, W.; Yeramian, P. D.; Schoenfeld, D.; Cudkowicz, M. E. Long-term Survival of Participants in the CENTAUR Trial of Sodium Phenylbutyrate-taurursodiol in AMYOTROPHIC LATERAL SCLEROSIS. *Muscle Nerve* **2021**, *63* (1), 31–39. https://doi.org/10.1002/mus.27091.

(266) Parry, G. J.; Rodrigues, C. M. P.; Aranha, M. M.; Hilbert, S. J.; Davey, C.; Kelkar, P.; Low, W. C.; Steer, C. J. Safety, Tolerability, and Cerebrospinal Fluid Penetration of Ursodeoxycholic Acid in Patients With Amyotrophic Lateral Sclerosis. *Clin. Neuropharmacol.* **2010**, *33* (1), 17–21. https://doi.org/10.1097/WNF.0b013e3181c47569.

(267) Min, J.-H.; Hong, Y.-H.; Sung, J.-J.; Kim, S.-M.; Lee, J. B.; Lee, K.-W. Oral Solubilized Ursodeoxycholic Acid Therapy in Amyotrophic Lateral Sclerosis: A Randomized Cross-Over Trial. *J. Korean Med. Sci.* **2012**, *27* (2), 200. https://doi.org/10.3346/jkms.2012.27.2.200.

(268) Blasco, H.; Veyrat-Durebex, C.; Bocca, C.; Patin, F.; Vourc'h, P.; Kouassi Nzoughet, J.; Lenaers, G.; Andres, C. R.; Simard, G.; Corcia, P.; Reynier, P. Lipidomics Reveals Cerebrospinal-Fluid Signatures of ALS. *Sci. Rep.* **2017**, *7* (1), 17652. https://doi.org/10.1038/s41598-017-17389-9.

(269) Chang, K.-H.; Lin, C.-N.; Chen, C.-M.; Lyu, R.-K.; Chu, C.-C.; Liao, M.-F.; Huang, C.-C.; Chang, H.-S.; Ro, L.-S.; Kuo, H.-C. Altered Metabolic Profiles of the Plasma of Patients with Amyotrophic Lateral Sclerosis. *Biomedicines* **2021**, *9* (12), 1944. https://doi.org/10.3390/biomedicines9121944.

(270) van Eunen, K.; Simons, S. M. J.; Gerding, A.; Bleeker, A.; den Besten, G.; Touw, C. M. L.; Houten, S. M.; Groen, B. K.; Krab, K.; Reijngoud, D.-J.; Bakker, B. M. Biochemical Competition Makes Fatty-Acid β-Oxidation Vulnerable to Substrate Overload. *PLoS Comput. Biol.* **2013**, *9* (8), e1003186. https://doi.org/10.1371/journal.pcbi.1003186.

(271) Lee, I.; Kazamel, M.; McPherson, T.; McAdam, J.; Bamman, M.; Amara, A.; Smith, D. L.; King, P. H. Fat Mass Loss Correlates with Faster Disease Progression in Amyotrophic Lateral Sclerosis Patients: Exploring the Utility of Dual-Energy x-Ray Absorptiometry in a Prospective Study. *PLOS ONE* **2021**, *16* (5), e0251087. https://doi.org/10.1371/journal.pone.0251087.

(272) Li, C.; Ou, R.; Wei, Q.; Shang, H. Shared Genetic Links between Amyotrophic Lateral Sclerosis and Obesity-Related Traits: A Genome-Wide Association Study. *Neurobiol. Aging* **2021**, *102*, 211.e1-211.e9. https://doi.org/10.1016/j.neurobiolaging.2021.01.023.

(273) Zhang, L.; Tang, L.; Huang, T.; Fan, D. Life Course Adiposity and Amyotrophic Lateral Sclerosis: A Mendelian Randomization Study. *Ann. Neurol.* **2020**, *87* (3), 434–441. https://doi.org/10.1002/ana.25671.

(274) Fernihough, A.; McGovern, M. E. Physical Stature Decline and the Health Status of the Elderly Population in England. *Econ. Hum. Biol.* **2015**, *16*, 30–44. https://doi.org/10.1016/j.ehb.2013.12.010.

(275) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J. Cheminformatics* **2016**, *8* (1), 61. https://doi.org/10.1186/s13321-016-0174-y.