A Metabolomic Modeling Approach for Functional Microbiome Analysis

Jasmine Chong

Institute of Parasitology

McGill University, Montreal, Canada

April 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Table of Contents

Abstract **5**
Abrégé **6**
Acknowledgements **7**
Contribution to Original Knowledge **8**
Contribution of Authors **10**
List of Tables **12**
List of Figures **13**
List of Abbreviations **15**
Chapter 1. Introduction and Literature Review **17**
1.1 Background 17
1.2 Overview of the human gut microbiome 18
1.3 Composition and structure of the gut microbiome 19
1.4 Plasticity of the gut microbiome 20
1.5 Dysbiosis: moving beyond associations towards causation 22
1.6 Inferring the functional potential of the gut microbiome 22
1.7 The gut metabolome as a functional readout of microbial activity 24
1.8 Application of metabolomics towards characterizing the microbiome 25
1.9 Deciphering the Metabolic Function of the Gut Microbiome 28
1.10 Future Perspectives 29
1.11 Rationale and objectives 30
Preamble to Chapter 2 ..... **33**
Chapter 2. Improving Biological Insights from Metabolomics Data **34**
Abstract 35
Introduction 36
Overview of the MetaboAnalyst 4.0 framework 38
MetaboAnalystR and improved transparency/reproducibility 41
MetaboAnalyst's knowledgebase update 44
New module #1: Mummichog and MS peaks-to-pathways 45

New module #2: Meta-analysis of metabolomics data 48
New module #3: Network explorer 49
Other feature updates 52
Implementation 52
Comparison with other tools 53
Conclusions 55
Preamble to Chapter 3 57
Chapter 3. From Raw Spectra to Biological Insights 59
Abstract 60
Introduction 61
Results 64
Benchmark Case Study 66
IBD Case Study 67
Discussion 72
Conclusion 73
Materials and Methods 74
Spectral Processing 74
Prediction of Pathway Activities 75
Benchmark Case Studies 76
Acknowledgements 78
Author Contributions 78
Conflicts of Interest 78
Supplemental Information: MetaboAnalystR 2.0 79
Preamble to Chapter 4 85
Chapter 4. Enhancing Biological Insights from Paired Microbiome and Metabolomics Data 86
Abstract 87
Background 88
Methods 92
Genome-scale metabolic models 94
Matching to genome-scale metabolic models 95

Creation of a community metabolic network 95
Available pathway libraries 96
Predicted pathway activity 97
Data visualization 98
Case Study 98
Results 99
Crohn's Disease Case Study 100
Discussion 107
Conclusion 111
Supplementary Materials: Metabolome Coverage of AGORA Versus CarveMe Metabolic Models 112
Comparison of Super Chemical Class 112
Comparison of Main Chemical Class 115
Comparison of Sub Chemical Class 117
Summary 120
Supplementary Materials: Crohn's Disease Case Study 121
Chapter 5. Discussion 129
Chapter 6. Conclusion 134
References 137

Abstract

The gut microbiome is a complex biological system that impacts many aspects of human health. While several studies have identified long lists of microbes implicated in disease, why they are associated with differential host phenotypes remains unclear. Metabolomics can complement sequencing-based approaches by providing a snapshot of host-microbial co-metabolism, however, its use in the field of microbiomics is still in its infancy. The objectives of my project are therefore to (i) to become proficient in metabolomics data processing and analysis and translate this knowledge into the development of bioinformatics tools for metabolomics data, (ii) to improve biological insights obtained from untargeted metabolomics data in an open-source and transparent matter, and (iii) to implement a novel bioinformatic framework to integrate untargeted metabolomics data and taxonomic microbial data to model changes in microbial metabolism. Ultimately, this framework will permit researchers to understand metabolic mechanisms of the gut microbiome and aide the design of novel therapeutics.

Abrégé

Le microbiome intestinal est un système biologique complexe qui a un impact sur de nombreux aspects de la santé humaine. Bien que plusieurs études aient identifié de longues listes de microbes impliqués dans les maladies, la raison pour laquelle ils sont associés à des phénotypes d'hôtes différentiels reste incertaine. La métabolomique peut compléter les approches basées sur le séquençage en fournissant un apérçu du co-métabolisme hôte-microbien, cependant, son utilisation dans le domaine de la microbiomique en est encore à ses debuts. L'un des principaux objectifs de la recherche sur le microbiome est de définir la fonction d'un microbe et son impact sur l'hôte. Les objectifs de mon projet sont donc de (i) acquérir une compréhension globale de la métabolomique de bout en bout et d'analyser des données microbiologiques, (ii) déterminer les microbes clés liés à d'importants changements métabolomiques, et (iii) mettre en œuvre une méthode basée sur un réseau métabolique pour prédire les fonctions du microbiome. Finalement, cette structure permettra aux chercheurs de comprendre les mécanismes métaboliques du microbiome intestinal et aidera à la conception de nouvelles thérapies.

Acknowledgements

I would first like to sincerely thank my supervisor Dr. Jianguo Xia for his continuous support, wisdom, and guidance throughout my entire PhD adventure. It was through his motivating talks, expertise across so many fields, opportunities bestowed upon me, and financial support that this body of work was made possible. Thank you for helping me grow as a researcher. I also gratefully acknowledge the support from my committee members of inspiring women in science, Dr. Jennifer Ronholm and Dr. Marilyn Scott. Next, I would like to mention our collaborators Dr. Shuzhao Li and Dr. David Wishart. They were always willing to help, from our requests for data or meetings to delve deep into metabolomics. I would also like to thank all my fellow Xia Lab members for being a wonderful and kind group of people who support one another and share good food and laughs. I would like to shout out Dr. Othman Soufan, Yannan Fan, Achal Dhariwal, and Guangyan Zhou for showing me the ropes and embracing me with open arms in the beginning of my PhD. Without their help at the start, I would only be a fraction of the scientist I am today. I am also thankful to all my friends and staff at the Institute for being so warm and helpful. Thank you to my partner, my sisters, my parents, my in-laws, and my friends for believing in me and encouraging me. Your support has been a source of motivation through hard times and I could not have done this without you all. Finally, I am grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Alexander Graham Bell Doctoral Scholarship and McGill University for the GREAT award.

Contribution to Original Knowledge

From the field of metabolomics to microbiomics, I have made the following novel contributions to research throughout my PhD:

1 *Developed MetaboAnalyst 4.0, a comprehensive web application for metabolomic data analysis, interpretation, and integration with other omics data.* This update, as detailed in Chapter 2, aimed to enhance the metabolomics data workflow through (1) support for pathway activity prediction from untargeted metabolomics data, (2) creation of a new module for metabolomic biomarker meta-analysis, and (3) implementation of a knowledge-based network analysis and visualization module for multi-omics data integration.

2 *Enhanced MetaboAnalystR, to support end-to-end analysis of untargeted metabolomics data.* Processing and interpreting untargeted metabolomics datasets are a key challenge in current computational metabolomics. With the development of MetaboAnalystR 2.0, I have enhanced the support for comprehensive LC-MS data processing, statistical analysis, and functional interpretation. This was through the seamless integration of popular R packages for peak annotation and processing, as well as implementing high-performance algorithms for prediction of pathway activities directly from unannotated peaks.

3 *Created microMum, which aims to fill the gap of integrating paired untargeted metabolomic and taxonomic sequencing data to investigate microbiome function.* Despite its popularity, integrating multiple omics profiles obtained from the same samples is still a difficult task.

Further, how to gain functional insights from untargeted metabolomics and taxonomic microbial signatures has yet to be addressed. With microMum, microbiome researchers are now able to use taxonomic microbial signatures to build a metabolomic model, predict changes in microbial metabolism, and form hypotheses whether such changes contribute to disease pathogenesis.

Contribution of Authors

This thesis is written by Jasmine Chong and comprises of three original scholarly manuscripts. Jasmine Chong is the primary author on all three manuscripts. The following authors contributed to one or more of the manuscripts in this thesis:

Jasmine Chong (J.C.)

Jianguo Xia (J.X.)

Othman Soufan (O.S.)

Carin Li (C.I.)

Iurie Caraus (I.C.)

Shuzhao Li (S.L.)

Guillaume Bourque (G.B.)

David Wishart (D.W.)

Mai Yamamoto (MY)

Yao Lu (Y.L.)

Tanisha Shiri (T.S.)

The contributions of the authors are as follows:

Manuscript 1: J.C. and J.X. prepared the manuscript. J.C, O.S, and I.C contributed to the development, testing and implementation of MetaboAnalyst 4.0. C.L. performed data extraction and curation. S.L. provided consultations regarding algorithm development. J.X. oversaw the project planning and execution. All authors critically reviewed the draft manuscript and evaluated the webtool.

Manuscript 2: J.C., M.Y. and J.X. prepared the manuscript. J.C. and M.Y. contributed to the development of algorithms and performing the formal analysis. M.Y. performed the data curation. J.X. supervised the project. All authors read and approved the final manuscript.

Manuscript 3: J.C. created the framework for the tool including writing R functions, setting up the SQLite database, finding case-studies, and implementing the R package. J.C. wrote the manuscript and managed the project planning. Y.L. and T.S. performed data extraction and curation. Y.L. and T.S. also helped refine the methodology and visualizations within the framework. Y.L., T.S., and J.X. revised and edited the manuscript. J.X. provided conceptual advice and supervised the project.

List of Tables

Table 1. Comparison of MetaboAnalyst with other tools. The table compares the main features of MetaboAnalyst (versions 1.0 - 4.0) with other web-based or web-enabled tools. Symbols used for feature evaluations with "√" for present, "-" for absent, and "+" for a more quantitative assessment (more "+" indicate better support). 54
Table 2. Comparison of MetaboAnalystR 2.0 with XCMS Online. This table compares the peak identification and quantification accuracies using the benchmark dataset between MetaboAnalystR 2.0 and the original manuscript using XCMS Online. 66
Table 3. Top metabolic pathway alterations using MetaboAnalystR 2.0. The table shows the top five metabolic pathways identified between patients with pediatric Crohn's disease (n=24) as compared to healthy controls (n=24), using the mummichog algorithm (*PerformMummichog*) and GSEA (*PerformGSEA*) implemented in MetaboAnalystR 2.0. 70
Table 4. Top 11 altered metabolic pathways using microMum. The top 11 predicted pathway activity of untargeted metabolomics data showing the results of the community metabolic model (CMN) permutation. 105

List of Figures

Figure 1. Overview of MetaboAnalyst 4.0 framework. The current modules can be organized into four general categories: (i) exploratory statistical analysis, (ii) functional analysis, (iii) data integration & systems biology and (iv) data processing & utility functions. 40
Figure 2. Summary of new features introduced in MetaboAnalyst 4.0. (A) An illustration showing the R command history panel and the companion MetaboAnalystR package will allow users to easily reproduce their analyses. In this case, the R command history captures all R commands leading to the generation the PLS-DA 2D score plot, which can then be reproduced in MetaboAnalystR using identical R commands (except the file path parameter for user input). (B) A zoomed-in view of the KEGG metabolic network showing the potential metabolite hits predicted from the mummichog algorithm. Clicking on a highlighted node will display all possible matched adduct forms of the corresponding compound. (C) An interactive Venn diagram showing the results froma biomarkermeta-analysis. Clicking on an area will show the corresponding hits. (D)An example of the metabolite-gene-disease interaction network created based on user input..... 43
Figure 3. A typical metabolomics data analysis workflow. The workflow starts from the raw data input and flows to data preprocessing, data processing, and data analysis. 62
Figure 4. OPLS-DA Score plot. The OPLS-DA score plot based on the 4113 features from the stool metabolome of 24 pediatric Crohn's disease patients and 24 healthy children, with a R2 of 0.912 and Q2 of 0.424..... 68
Figure 5. Scatter plot integrating mummichog and GSEA pathway analysis results. On the x-axis are negative log raw GSEA p-values, and on the y-axis are negative log raw mummichog p-values. The size and color of the circles correspond to the value of their transformed combined p-values (ascending, from white to dark red). The blue and pink quadrants represent pathways that were significant using a single pathway analysis algorithm (blue = mummichog, pink = GSEA), and the purple quadrant contains significantly perturbed pathways identified using both algorithms. From the plot, bile acid metabolism and vitamin D3 metabolism show congruence as the top enriched pathways..... 69
Figure 6. microMum workflow. The first step begins with matching the user's taxonomic microbial signature to the internal database of genome-scale metabolic models (GEMs). The matched GEMs are then combined to create a unique community metabolic model. Meanwhile, feature selection is performed on the untargeted metabolomics data. Differentially abundant peaks are then used as input for putative compound annotation. Putatively annotated compounds are overlayed onto the community metabolic network and pathway enrichment is performed. Further exploration of pathway activity results can be used to identify key organisms who contribute to enriched metabolic functions. 93
Figure 7. Lollipop chart of altered metabolic pathways using microMum. The Lollipop chart shows the enrichment of KEGG metabolic pathways in AGORA (yellow lines) and CarveMe

(blue lines) matched microbes implicated in Crohn's Disease. The x-axis is the -log 10 p-value (scaled from 0-5), and the y-axis are the pathways. The square colored boxes on the left of the plot shows the metabolic hierarchy of that pathway to aid interpretation. The length of the lines corresponds to the -log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits). Lines with an asterisk at the end are pathways that were also observed to be significantly altered using real metabolomics data. 103
Figure 8. microMum Bubble chart of Secondary Bile Acid Biosynthesis. The bubble chart shows the predicted presence of metabolites belonging to the secondary bile acid biosynthesis pathway across all AGORA matched microbes implicated in Crohn's Disease and the human genome-scale metabolic network..... 106

List of Abbreviations

NSERC	Natural Sciences and Engineering Research Council of Canada
T1D	Type 1 diabetes
FMT	Fecal Microbiota Transplantation
IBS	Irritable bowel syndrome
IBD	Inflammatory bowel disease
T2D	Type-2 diabetes
CNS	Central nervous system
MWAS	Microbiome-wide association studies
BCAAs	Branched-chain amino acids
ASD	Autism spectrum disorder
4EPS	4-ethylphenylsulfate
CHF	Chronic heart failure
HMDB	Human Metabolome Database
TMAO	Trimethylamine-N-oxide
LPS	Lipopolysaccharide
KEGG	Kyoto Encyclopedia of Genes and Genomes
MSEA	Metabolite set enrichment analysis
MetPA	Metabolic pathway analysis
MetATT	Advanced two-factor and time-series analyses
MS	Mass spectrometry
SNP	Single nucleotide polymorphism
NMR	Nuclear magnetic resonance
GC-MS	Gas chromatography - mass spectrometry
LC-MS	Liquid chromatography - mass spectrometry
PLS	Partial least square
OPLS	Orthogonal PLS
CIHR	Canadian Institutes of Health Research
CRC	Canada Research Chairs
ORA	Over-representation analysis
GSEA	Gene-set enrichment analysis
EICs	Extracted ion chromatograms
TIC	Total Ion Chromatogram
BPC	Base Peak Chromatogram
iHMP	Integrative Human Microbiome Project Consortium
IQR	Inter quantile range

CVD	Cardiovascular disease
CRC	Colorectal cancer
NAFLD	Non-alcoholic fatty liver disease
SCFAs	Short chain fatty acids
GEMs	Genome-scale metabolic models
CIDs	PubChem identifiers
PCA	Principal components analysis
RA	Rheumatoid arthritis
CMNs	Community metabolic networks
PCST	Prize-collecting Steiner tree

Chapter 1. Introduction and Literature Review

1.1 Background

Owing to advances in high-throughput sequencing technologies, the past decade has seen a wealth of research highlighting the importance of the human microbiome towards human health, development, disease susceptibility, and behavior (Integrative, 2014). Here, the term 'microbiome' refers to the set of microorganisms inhabiting a shared environment, as well as their genomic content and metabolic products. These complex assemblages of microorganisms inhabiting the body's mucosal surfaces and cavities have evolved with their human hosts over millennia to form a variety of relationships, including mutualistic, symbiotic and even parasitic (Foster, Schluter, Coyte, & Rakoff-Nahoum, 2017; Mazmanian, Round, & Kasper, 2008). Humans are rapidly colonized by microbes at birth, and their microbiota show great interpersonal variation due to several factors such as age, lifestyle, environment, and genetics (Bäckhed et al., 2015; Fierer, Hamady, Lauber, & Knight, 2008; Goodrich et al., 2016; Human Microbiome Project, 2012; Yatsunenko et al., 2012). Further, it has been estimated that humans are home to over 5000 different genera of bacteria, with a 1:1 ratio of human to bacterial cells (Rojo et al., 2017; Sender, Fuchs, & Milo, 2016). Understanding the complex interplay between gut microbiota and its human host is paramount to uncover the role the microbiome plays in maintaining host homeostasis.

1.2 Overview of the human gut microbiome

The human intestinal tract is home to the greatest density and diversity of microbes, containing archea, bacteria, fungi, viruses, and eukaryotes (Y. K. Lee & Mazmanian, 2010). Bacteria are the most predominant microorganisms in the human gut, consisting of approximately 1000 species largely belonging to the phyla *Firmicutes* and *Bacteriodetes* (Human Microbiome Project, 2012; Qin et al., 2010). The gut microbiome is intrinsically involved in maintaining homoeostasis and performs several vital functions that are otherwise inaccessible to its human host. For instance, the gut microbiota metabolize host-indigestible polysaccharides and synthesize essential vitamins (Thursby & Juge, 2017); maintain the integrity of the intestinal epithelial and mucosal barriers (Natividad & Verdu, 2013); provide protection against opportunistic pathogens; and shape the development of the host immune system (Belkaid & Hand, 2014). Its reach also extends well beyond the gastrointestinal (GI) tract, affecting host processes such as bone homeostasis (Sjögren et al., 2012), adiposity (Bäckhed et al., 2004), brain function and behavior (Cryan & Dinan, 2012). Moreover, the gut microbiome encompasses an immense toolbox of biochemical and metabolic capacities, containing 150-fold more genes than the entire human genome (Human Microbiome Project, 2012; Ursell et al., 2014). Due to these rich functionalities that greatly expand that of its host, the gut microbiome is considered a virtual organ within the human body (Baquero & Nombela, 2012; O'Hara & Shanahan, 2006; Ursell et al., 2014). Despite its importance, microbiome research towards understanding mechanisms underlying the gut microbiota’s influence on human health and disease is still in its infancy.

1.3 Composition and structure of the gut microbiome

Taxonomic profiling of the gut microbiome permits researchers to pinpoint species as biomarkers of disease and gain insight into ecological processes driving its composition. Numerous large-scale studies have demonstrated that the human gut microbiome is highly personalized and relatively stable after infancy, with considerable intra-individual taxonomic diversity even amongst healthy individuals (Flores et al., 2014; Human Microbiome Project, 2012; Kostic et al., 2015; Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012; Zhernakova et al., 2016). Remarkably, it has been shown that twins can share only 50% or less of their microbial genera (Shafquat, Joice, Simmons, & Huttenhower, 2014; Peter J Turnbaugh et al., 2009). While high taxonomic diversity is considered an indicator of a 'healthy' gut microbiome, reduced taxonomic diversity has been associated with poor health outcomes such as frailty (Jackson et al., 2016), type 1 diabetes (T1D) (Kostic et al., 2015), irritable bowel syndrome (IBS) (Zhernakova et al., 2016), and obesity (Peter J Turnbaugh et al., 2009). Several studies have also highlighted the potential for microbial biomarkers as non-invasive targets for early detection of various diseases including Inflammatory Bowel Disease (IBD) (Rooks et al., 2014), hepatocellular carcinoma (Ren et al., 2018), and T1D (Davis-Richardson et al., 2014). More recently, a large-scale population study attributed increased abundance in *Enterobacteriaceae* with a greater mortality risk (Salosensaari et al., 2021). Overall, taxonomic profiling can provide some insight into how the microbiome changes in association with phenotype.

Despite substantial taxonomic variability between individuals, metagenomics has revealed that in the absence of disease, humans share a functional core microbiome (Human Microbiome

Project, 2012; Lozupone et al., 2012; Peter J Turnbaugh et al., 2009). This set of conserved functions consists of (i) housekeeping functions required for universal maintenance of life such as transcription and translation (Human Microbiome Project, 2012), (ii) shared human-associated microbial processes such as transportation of small molecules and biosynthesis of host-required compounds (i.e. amino acids, lipids), and finally (iii) functions specific to the gut including carbohydrate processing and vitamin biosynthesis (Shafquat et al., 2014). Functional redundancy amongst different taxa residing within the same shared environment may have evolved to maintain gut homeostasis permitting their colonization (Backhed et al., 2012; Levy, Thaiss, & Elinav, 2016). For instance, functional redundancy would permit a microbial community to withstand perturbations from environmental impacts such as antibiotics (resistance), or to return to equilibrium following stress-induced changes (resilience) (Backhed et al., 2012). Tian et al. recently demonstrated that high-functional redundancy in recipient's gut microbiota pre-fecal microbiota transplantation (FMT) increases resilience, thus reducing the transplantation efficacy (Tian et al., 2020). In comparison, recipient's gut microbiota with low functional redundancy prior to FMT had a better chance of returning to a healthy state following transplantation. Ultimately, gut microbiome function is important to maintain a healthy host state and alterations in this function likely play an important role in disease progression and treatment.

1.4 Plasticity of the gut microbiome

While specific mechanisms may be unclear, several environmental and genetic factors are known to shape the gut microbiome (Candela, Biagi, Maccaferri, Turroni, & Brigidi, 2012; Maurice, Haiser, & Turnbaugh, 2013; Quercia et al., 2014). A large-scale twin study of over

1000 twin pairs demonstrated the genetic heritability of ~83 taxa (1% of the overall gut microbiome composition) that were stable across repeat samplings (Goodrich et al., 2016). Compared to environmental factors, the effect of host genetics on shaping the gut microbiome is relatively modest (Kurilshikov, Wijmenga, Fu, & Zhernakova, 2017). A recent study of >1000 individuals with divergent genetic backgrounds yet living in similar environments demonstrated that it was non-genetic factors such as household sharing, diet, and age that shaped the composition of the gut microbiome, explaining ~20% of observed variance (Rothschild et al., 2018). In comparison, genetic ancestry was not significantly associated with gut microbiome composition. Of the environmental factors, diet is known to be one of the most important modifiers of the gut microbiome (Singh et al., 2017; Zmora, Suez, & Elinav, 2019). A landmark study of gnobiotic mice demonstrated that transitioning from a plant-rich, low-fat diet to a 'Western' high fat, high-sugar diet can alter the microbial community structure and function within a single day (P. J. Turnbaugh et al., 2009). Another study comparing the fecal microbiome of children from Burkina Faso and Italy demonstrated clear differences in their microbial community compositions attributed to their diet, including a decrease in microbial richness for Italians and an increase in short-chain fatty acids (SCFAs) in Africans (De Filippo et al., 2010). Significant efforts are now being put forward to leverage such immense plasticity to manipulate the microbiome through interventions such as antibiotics, probiotics, fecal microbiota and phage therapy to restore imbalances and improve host health (David et al., 2014; Kau, Ahern, Griffin, Goodman, & Gordon, 2011; Kuntz & Gilbert, 2017; S. S. Li et al., 2016; Mirzaei & Maurice, 2017; Quigley & Gajula, 2020).

1.5 Dysbiosis: moving beyond associations towards causation

Perturbations of the gut microbiome, known as 'dysbiosis', are characterized as changes to its composition and/or functions that negatively affect the host's health. Dysbiosis of the gut microbiome has been linked to a wide-array of diseases from GI-related disorders including IBD (Lewis et al., 2015), type-2 diabetes (T2D) (X. Li, Watanabe, & Kimura, 2017), and obesity (Carding, Verbeke, Vipond, Corfe, & Owen, 2015), to central nervous system (CNS) related disorders such as Alzheimer's (Vogt et al., 2017), autism (Strati et al., 2017), and depression (Luna & Foster, 2015). Notably, microbiome-wide association studies (MWAS) have provided a wealth of information linking various microbiome features to disease (Gilbert et al., 2016). For instance, a recent large-scale investigation of ~3400 individuals identified >1500 significant associations between 102 bacterial genera and 142 host factors (Manor et al., 2020). While microbiome studies have provided long lists of implicated microbes and/or genes, their functional role within the gut microbiome is largely unknown (Schmidt, Raes, & Bork, 2018; Surana & Kasper, 2017). Specifically, how these microbes impact host physiology is required to refine such lists into experimentally validated causal features and ultimately translate this knowledge into actionable therapies.

1.6 Inferring the functional potential of the gut microbiome

The majority of studies highlighting the functional importance of the gut microbiome have focused on gene-based approaches, extrapolating from relative abundances of species and their genes to enriched functions and pathways in a microbial community (Abubucker et al.,

2012; M. G. Langille et al., 2013). These methods, however, are inherently limited in their ability to identify which functions are conferred by specific microbes, and are incapable of directly measuring functional activity (Heintz-Buschart & Wilmes, 2018). Functional profiling of the microbiome from 16S rRNA sequencing/metagenomes requires mapping reads to annotated genes, proteins or genomes. The accuracy of functional predictions thereby depends on the accuracy of *a priori* functional annotations. DNA-based techniques usually do not differentiate between microbes that are dead or alive, confounding interpretations of processes important to host-microbiome interactions (Bajaj et al., 2018; Blazewicz, Barnard, Daly, & Firestone, 2013; Cangelosi & Meschke, 2014).

A functional microbiome is a product of its expressed genes, reflected upstream of the functional hierarchy by transcripts, proteins, and metabolites. Functional profiling of the microbiome using DNA-based omics assumes that all functionally predicted genes are expressed equally - a flawed notion that has been disproven with other omics technologies including metatranscriptomics and proteomics (Franzosa et al., 2014; Verberkmoes et al., 2009). Moreover, DNA-based techniques provide a taxonomic classification of a microbial community, yet taxonomic resolution of 16S rRNA gene sequencing is at best to species-level (Drewes et al., 2017), and strain-level taxonomic resolution of metagenomics is just emerging (Nayfach & Pollard, 2016). The lack of high-resolution characterization is important as there can exist great functional differences at the strain level of identical microbial species, with major consequences to host's health (Noecker, McNally, Eng, & Borenstein, 2017; Rosen & Palm, 2017). For instance, some strains of *Propionibacterium acnes* are shown to be enriched in acne, while others are associated with healthy skin (Fitz-Gibbon et al., 2013). Therefore, inferring the functional

capacity of the microbiome is predicated upon precise and accurate mapping of the microbiome. Despite its shortcomings, DNA-based strategies have yielded important insight into the composition and functional potential of the microbiome, aptly answering “who is there?” and “what can they do?”.

1.7 The gut metabolome as a functional readout of microbial activity

Many biological processes that occur within the gut and are important to host-microbial interactions far exceed the taxonomic and genomic contributions of the gut microbiome (Noecker et al., 2017). From an evolutionary standpoint, humans and their native gut microbiota have formed a mutually beneficial relationship where gut microbes contribute to the production of biologically active small molecules, termed microbial-derived metabolites, that are believed to enhance host fitness (Nicholson et al., 2012). These metabolites serve as signaling molecules, mediating vital host-microbial interactions through a dynamic crosstalk involving numerous molecular pathways. The host and its resident gut microbiota work together to coproduce metabolites through nutrient and xenobiotic metabolism (Nicholson et al., 2012). These small molecule products greatly impact biological processes with important consequences to human health including digestion, immune system development and regulation, inflammation, and neurodevelopment (Hsiao et al., 2013; Sharon et al., 2014). For instance, *Prevotella copri* and *Bacteroides vulgatus* were recently shown to influence circulating levels of branched-chain amino acids (BCAAs), likely inducing insulin resistance in humans (Pedersen et al., 2016). Another study demonstrated that treatment of *B. fragilis* in mice displaying features of autism spectrum disorder (ASD) corrected anxiety-like behavior, attributed to reductions in the levels of

the microbial-derived metabolite 4-ethylphenylsulfate (4EPS) (Hsiao et al., 2013). The gut metabolome therefore reflects the interplay between the host and its microbiota, providing a functional readout of gut microbiome activity (Marcobal et al., 2013; Zierer et al., 2018).

Perturbations in cellular processes are rapid and leave metabolic fingerprints, representing the physiological state of the host and its microbiota (Gilbert et al., 2016; Zierer et al., 2018). These metabolites are the start and end products of various microbial-mediated processes, enabling systems-level insight of the gut microbiome. Metabolites are also the universal language of the microbiome, providing a detailed functional assessment of host-microbiome-disease interactions. As mentioned above, different assemblages of the gut microbiota can alter the metabolome. A recent large-scale study has demonstrated that the observed variance of the fecal metabolome can be significantly explained by the gut microbiome structure (86). Therefore systematically linking changes in metabolomic profiles to compositional shifts in the microbiome has great potential to deliver cutting-edge functional understandings of a complex ecosystem (Noecker et al., 2016; Noecker et al., 2017). Furthermore, by unearthing changes in functional activity and attributing it to a microbe, targeted therapeutics can be created to regain functional homeostasis. The gut metabolome thus complements DNA-based approaches, enabling researchers to gain mechanistic insights into functional processes underlying differential phenotypic states.

1.8 Application of metabolomics towards characterizing the microbiome

Metabolomics is the comprehensive quantification and analysis of small molecules (<1500 Da) within a biological system. The central dogma of metabolomics is that one's

metabolic profile closely represents his or hers health status, reflecting not only genetic influences but impacts from their lifestyle, environment, and native microbiota (Beger et al., 2016; D. S. Wishart, 2016). For this reason, it has been designated as the "link between genotype and phenotype" (Fiehn, 2002). There exist two main metabolomic approaches, untargeted and/or targeted. While untargeted methods aim to measure the global set of metabolites within a sample, targeted metabolomics aim to quantify a predefined set of metabolites (C. H. Johnson, J. Ivanisevic, & G. Siuzdak, 2016). A wide range of sample types can be analyzed, including common biospecimens such as urine, tissues, blood, and stool. In recent years, metabolomic approaches are becoming increasingly popular in microbiome studies, particularly (i) to characterize diseases/disorders or (ii) to investigate the impact of dietary and xenobiotic interventions on the host's metabolic profile (Ryan et al., 2017; Sanguinetti et al., 2018; F. Wang et al., 2018). For instance, Cui et al. performed an integrative metagenomic and metabolomics approach to evaluate the role of the gut microbiota in chronic heart failure (CHF). They correlated distinctive changes in the gut microbiota composition, namely *Faecalibacterium prausnitzii* and *Ruminococcus gnavus*, with differential fecal and plasma metabolic profiles between healthy controls and CHF patients, highlighting the significant impact of gut microbiota dysbiosis on human health (Cui et al., 2018). Marrying other molecular layers of the microbiome with the metabolome permits not only the understanding of mechanisms underlying host-microbial activity, but the exploration of genetic, epigenetic, and environmental impacts on host's health (Rojo et al., 2017).

Obtaining and characterizing the metabolic profile of a sample is not without its challenges. Despite advances in high-throughput mass spectrometry technologies, only a small

fraction of metabolites is measurable by current technologies. The human metabolome has been estimated to consist of 1-3 million compounds, however current targeted metabolomic approaches can only detect about 300-700 metabolites (Uppal et al., 2016). Poor coverage of the human metabolome (<1%) is due in part to the inherent nature of these compounds - they can exist in undetectable levels, they have very high turnover rates, and are vulnerable to chemical modifications (Nielsen, 2017). Attributing the origin of metabolites to specific microbes or the host is also incredibly challenging as core metabolic processes are universally conserved, therefore important small molecules are structurally indistinguishable between species (Newsom & McCall, 2018). Finally, metabolite identification remains a significant challenge in the field of metabolomics (Vinaixa et al., 2016). Biological interpretation and contextualization of metabolites predicates on their proper identification, requiring researchers to manually compare the exact mass (m/z) or processed spectra against comprehensive MS-based spectral databases. Manual identification can often lead to a number of false-positives (Tobias Kind & Oliver Fiehn, 2006), as well metabolite databases can be improperly curated, incomplete, and not available to all researchers (Vinaixa et al., 2016). Furthermore, only 5-10% of all quantifiable compounds can be identified across MS-spectra databases (e.g. Human Metabolome Database (HMDB) and MassBank), requiring major efforts from researchers to query multiple databases to confirm their metabolite identities (Vinaixa et al., 2016). Novel computational and technological innovations are therefore required to facilitate and improve metabolite identification and interpretation. Notwithstanding these limitations, metabolomics provides researchers with an unprecedented and real-time approach to quantify systems-level alterations in the microbiome, reflecting community-wide shifts in functional activity.

1.9 Deciphering the Metabolic Function of the Gut Microbiome

Pinpointing precise modifications of the gut microbiome responsible for phenotypic differences between healthy and diseased individuals is incredibly challenging. While taxonomic profiling can identify differentially enriched bacteria, alone it is ineffective to infer what features of these bacteria have important beneficial or detrimental impacts on its host. These features could be the production of microbial-derived metabolites such as trimethylamine-N-oxide (TMAO), which has been linked to adverse cardiac events and chronic kidney disease (D. Li, Kirsop, & Tang, 2015; Tang et al., 2015), or to their lipopolysaccharide (LPS) outer coating, which have immunoinhibitory effects on the host (d'Hennezel, Abubucker, Murphy, & Cullen, 2017; Zhao, Cong, Jaber, & Lukiw, 2017). Consequently, microbiome function not only includes metabolic activities performed by native gut bacteria, but also their inherent characteristics. Moreover, assigning microbiome function to specific microbes is vital for translational insights (Louca, Parfrey, & Doebeli, 2016; Manor & Borenstein, 2017). Bauer et al. (Bauer, Laczny, Magnusdottir, Wilmes, & Thiele, 2015) demonstrated that metabolic function of microbial strains can greatly differ as predicted by their phylogeny, stressing the importance of linking known metabolic functions to the microbes performing them. Creation of microbial functional profiles, encompassing both *who* the microbes are and *what they do* is required to create targeted therapeutics aiming to manipulate microbiome functionality.

Importantly, downstream interpretation of quantified or predicted causative metabolites and genes relies on proper contextualization. Here, the main approach is to perform pathway prediction utilizing existing reference pathway databases including the Kyoto Encyclopedia of

Genes and Genomes (KEGG) (Minoru Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017) and MetaCyc (R. Caspi et al., 2014; Ron Caspi et al., 2020). These databases hold well-characterized pathways known to play significant roles in host physiology, though the definition of these pathways is incredibly subjective. For instance, the average MetaCyc pathway contains 4.37 reactions, yet the average KEGG pathway contains 28.84 reactions (Altman, Travers, Kothari, Caspi, & Karp, 2013). Moreover, roughly 50% of reactions are shared between the two databases, once again highlighting the necessity for users to utilize multiple knowledgebases to avoid bias. Limiting knowledge-based contextualization of metabolites and genes to only pre-defined functional categories can result in missing specific functions that play key roles in disease. However, inferring novel metabolic pathways is computationally burdensome, requiring users to follow "rules" such as atom-mapping (Blum & Kohlbacher, 2008) and structural transformation (Moriya et al., 2010). For instance, a recent graph-theory based algorithm identified > 1000 pathways from glucose to pyruvate, capturing the well-known glycolysis pathway in addition to alternate pathways that are poorly studied yet important to maintain functional homeostasis (Ravikrishnan, Nasre, & Raman, 2018). Assigning functional units (e.g. metabolites, genes, and reactions) to pathways to gain insight into a given microbiome's metabolism is a difficult yet vital task to facilitate interpretation.

1.10 Future Perspectives

Despite the abundance of available data characterizing the gut microbiome, there is a lack of true understanding of its dynamics, function, and interactions with its host. Omics have been widely used to profile this complex system at multiple levels, though the resulting big data

challenges have significantly limited researchers to make important discoveries and translational applications. A deep understanding of current knowledge and programming skills are required to use novel ‘omics approaches, which represent significant barriers for their wider applications. Easy-to-use bioinformatics tools, such as one to integrate metabolomics and microbial sequencing data, are needed to address this gap. Ultimately, obtaining metabolic perspectives of the gut microbiome can shine a light on underlying mechanisms that greatly impact human health and inform the rational design of novel treatments.

1.11 Rationale and objectives

Microbiome association studies typically result in lists of implicated microbes that require further curation to determine how they impact their host. DNA-based functional profiling is commonly used to gain mechanistic insights but is a measure of functional potential rather than true activity. Conversely, metabolomics provides a snapshot of host-microbial interactions and better represents a complex ecosystem such as the microbiome by reflecting both genetic and environmental inputs. However, the use of metabolomics data is complicated by the fact that only a small fraction of metabolic features can be annotated. Further, the use of metabolomics in microbiome studies is not yet commonplace. It is thus hypothesized genome-scale metabolic models, which are network reconstructions of an organism’s metabolism, can be used to predict alterations in microbial metabolism. The primary objectives of my project are therefore:

1. **To become proficient in metabolomics data processing and analysis and translate this knowledge into the development of bioinformatics tools for metabolomics data.**

Processing and interpreting complex metabolomics datasets is challenging for novice researchers or those without bioinformatics training. I therefore first aim to become proficient in analyzing big metabolomics data by collaborating with other researchers in their data analysis and performing a literature review of existing analytical methods. I then aim to enhance the MetaboAnalyst platform (Chong et al., 2018), which is a freely accessible web-based tool for metabolomics data analysis.

2. **To improve biological insights obtained from untargeted metabolomics data in an open-source and transparent matter.**

Untargeted metabolomics based on high-resolution LC-MS is increasingly employed in large-scale omics studies. However, processing these complex metabolomics datasets is a key challenge in current computational metabolomics. Previous implementations had limited support for raw spectra processing and peak annotation. I thus aim to address two important gaps, 1) raw spectral processing and 2) functional interpretation directly from MS peaks.

3. **To implement a novel bioinformatic framework to integrate paired untargeted metabolomics and taxonomic microbial signatures to predict changes in microbial metabolism.**

A key goal of microbiome research is to define a microbe's function and its impact on the host. It is well known that microbial metabolism can influence a host’s phenotype. However, the integration of untargeted metabolomics and taxonomic sequencing data has yet to be tackled. To overcome this, I will create microMum, which leverages genome-scale metabolic networks to

combine the different data and obtain interpretable functional insights.

Preamble to Chapter 2

Rapid advances in analytical chemistry, mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy have dramatically increased the size and speed of which metabolomics data can be obtained. However, the inability of researchers to extract meaningful biological insights from these increasingly large and complex datasets has now become a major roadblock in current metabolomics research and applications. Furthermore, properly processing complex metabolomics data can be challenging for bench scientists and clinicians with minimal bioinformatic skills. Oftentimes users are required to use multiple tools to perform their analyses, necessitating they have coding skills to utilize the tools and format the data properly for each tool. Moreover, some popular tools are proprietary software that require paid licenses. These obstacles can be overwhelming and restricts users to basic statistics and visualizations. Therefore, to empower the field of metabolomics, the first objective of my thesis was to develop user-friendly and easily accessible bioinformatic tools to bridge the gap between metabolomics data generation and biological insights. The manuscript in Chapter 2 is the first of nine publications I have published in metabolomics, ranging from R packages, investigations of flesh quality of carp, a meta-analysis of COVID-19 metabolomics data, and web-based tools.

Chapter 2. Improving Biological Insights from Metabolomics Data

MetaboAnalyst 4.0 – Towards More Transparent and Integrative Metabolomic Analysis

[1]Jasmine Chong, [1]Othman Soufan, [2]Carin Li, [1]Iurie Caraus, [3]Shuzhao Li, [4,5]Guillaume Bourque, [2,6]David Wishart, [1,7]Jianguo Xia

[1]Institute of Parasitology, and [7]Department of Animal Science, McGill University, Montreal, Québec, Canada

[2]Department of Biological Sciences, and [6]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[3]Department of Medicine, Emory University School of Medicine, Atlanta, USA

[4]Department of Human Genetics, and [5]Canadian Center for Computational Genomics, McGill University, Montreal, Québec, Canada

Published in *Nucleic Acids Research* (Chong et al., 2018).

Abstract

We present a new update to MetaboAnalyst (version 4.0) for comprehensive metabolomic data analysis, interpretation, and integration with other omics data. Since the last major update in 2015, MetaboAnalyst has continued to evolve based on user feedback and technological advancements in the field. For this year's update, four new key features have been added to MetaboAnalyst 4.0, including: 1) Real-time R command tracking and display coupled with the release of a companion MetaboAnalystR package; 2) a new module for pathway prediction from untargeted mass spectral data using the *mummichog* algorithm; 3) a Biomarker Meta-analysis module for robust biomarker identification through the combination of multiple metabolomic datasets; and 4) a Network Explorer module for integrative analysis of metabolomics, metagenomics, and/or transcriptomics data. The user interface of MetaboAnalyst 4.0 has been reengineered to provide a more modern look and feel that gives more space and flexibility to introduce new functions. The underlying knowledgebases (compound libraries, metabolite sets, metabolite-SNP associations, and metabolic pathways) have also been updated using the latest data from the Human Metabolome Database (HMDB). A Docker image of MetaboAnalyst is also now available to facilitate local download and installation of MetaboAnalyst. MetaboAnalyst 4.0 is freely available at http://metaboanalyst.ca.

Introduction

MetaboAnalyst is a comprehensive web server designed to help users easily perform metabolomic data analysis, visualization, and functional interpretation. It was first introduced in 2009 with a single module for metabolomic data processing and statistical analysis (Jianguo Xia, Nick Psychogios, Nelson Young, & David S Wishart, 2009). Since then, it has been continuously updated to meet the evolving needs of the metabolomics research community. Version 2.0, which was released in 2012 (Jianguo Xia, Rupasri Mandal, Igor V Sinelnikov, David Broadhurst, & David S Wishart, 2012), incorporated three new modules for metabolite set enrichment analysis (MSEA (Jianguo Xia & David S Wishart, 2010b)), metabolic pathway analysis (MetPA (Jianguo Xia & David S Wishart, 2010a)), as well as advanced two-factor and time-series analyses (MetATT (Jianguo Xia, Sinelnikov, & Wishart, 2011)). Version 3.0, which was released in 2015, added support for biomarker analysis, power analysis, and joint pathway analysis (i.e. integrating genes/proteins and metabolites), coupled with a major upgrade of the underlying web framework (Jianguo Xia, Igor V Sinelnikov, Beomsoo Han, & David S Wishart, 2015).

With each iteration, MetaboAnalyst has grown more popular. To better handle the growing user traffic, MetaboAnalyst has recently been migrated to a Google cloud server for improved performance and accessibility. According to Google Analytics, over the past 12 months, MetaboAnalyst has processed >1.8 million jobs submitted from ~60,000 users. For instance, MetaboAnalyst has been used to elucidate metabolic differences in breast cancer of African-American and Caucasian women (Tayyari et al., 2018), to identify highly predictive biomarkers of ketosis in dairy cows (G. Zhang et al., 2017), to understand alterations in the intestinal

metabolome during enteric infections (Reynolds et al., 2017), as well as to study many other complex biological processes and diseases (Arts et al., 2016; Cox et al., 2016; Paglia et al., 2016). Based on our citation analysis for 2017, MetaboAnalyst has been used in at least 1/4 of all metabolomics publications for that year, attesting to its status as one of the preferred tools for metabolomic data analysis.

However, the field of metabolomics continues to evolve and it is important that MetaboAnalyst also evolves to keep current with the field and its growing user base. For this year's update, MetaboAnalyst has been substantially upgraded to enhance its user interface, to improve reproducibility/transparency, to support batch processing, to provide improved pathway interpretation from untargeted mass spectrometry (MS) data, to support meta-analysis and multi-omics analysis, to expand its underlying knowledgebase, and to support more facile local installations. In particular, the key features of this year's update include:

- A companion R package (MetaboAnalystR) and an accompanying R-command history panel to permit more transparent and reproducible analysis;
- An expanded set of metabolite, pathway and metabolite-disease/SNP (single nucleotide polymorphism) association knowledgebases to support more accurate and comprehensive functional analysis and interpretation;
- A new module based on the *mummichog* (S. Li et al., 2013b) algorithm for pathway activity prediction from untargeted metabolomics data;
- A new module to support metabolomic biomarker meta-analysis;
- A new module to support multi-omics data integration through knowledge-based network analysis and visualization;

- Other important updates including direct links to online tools for nuclear magnetic resonance (NMR), gas chromatography - mass spectrometry (GC-MS) and liquid chromatography - mass spectrometry (LC-MS) spectral analysis; as well as offering a Docker image for facile download and installation of MetaboAnalyst on local computers.

These changes and updates are all contained in MetaboAnalyst 4.0, which is freely available at http://www.metaboanalyst.ca. For each new module, we have added frequently asked questions (FAQs) and additional functions for more comprehensive analysis report generation. A more detailed description of each of these updates and changes in MetaboAnalyst 4.0 is given below.

Overview of the MetaboAnalyst 4.0 framework

MetaboAnalyst's user interface has been upgraded to provide a more modern "look and feel" that maintains the same easy-to-use modular analytical pipeline. To facilitate navigation, all functions are now organized into 12 analytical modules, which can be arranged into four general categories: 1) exploratory statistical analysis, 2) functional analysis, 3) data integration and systems biology, and 4) data processing & utility functions (**Figure 1**). The exploratory statistical analysis category (general statistics, biomarker analysis, two-factor/time-series analysis, and power analysis) can accept data from either targeted or untargeted metabolomics data sets. The functional analysis category has been expanded to include a new module on pathway activity prediction from MS data, in addition to the two existing modules for metabolite set enrichment analysis and pathway analysis of targeted metabolomics data. The data integration and systems biology category now includes three new modules (biomarker meta-analysis, joint-pathway

analysis, and network explorer). Finally, the data processing and other utilities category contains common data processing tools such as compound ID conversion, batch effect correction, as well as links to several web-based tools for spectra analysis such as Bayesil for automated NMR spectral annotation (Ravanbakhsh et al., 2015), GC-AutoFit and XCMS Online for LC/MS spectral processing (Huan et al., 2017).

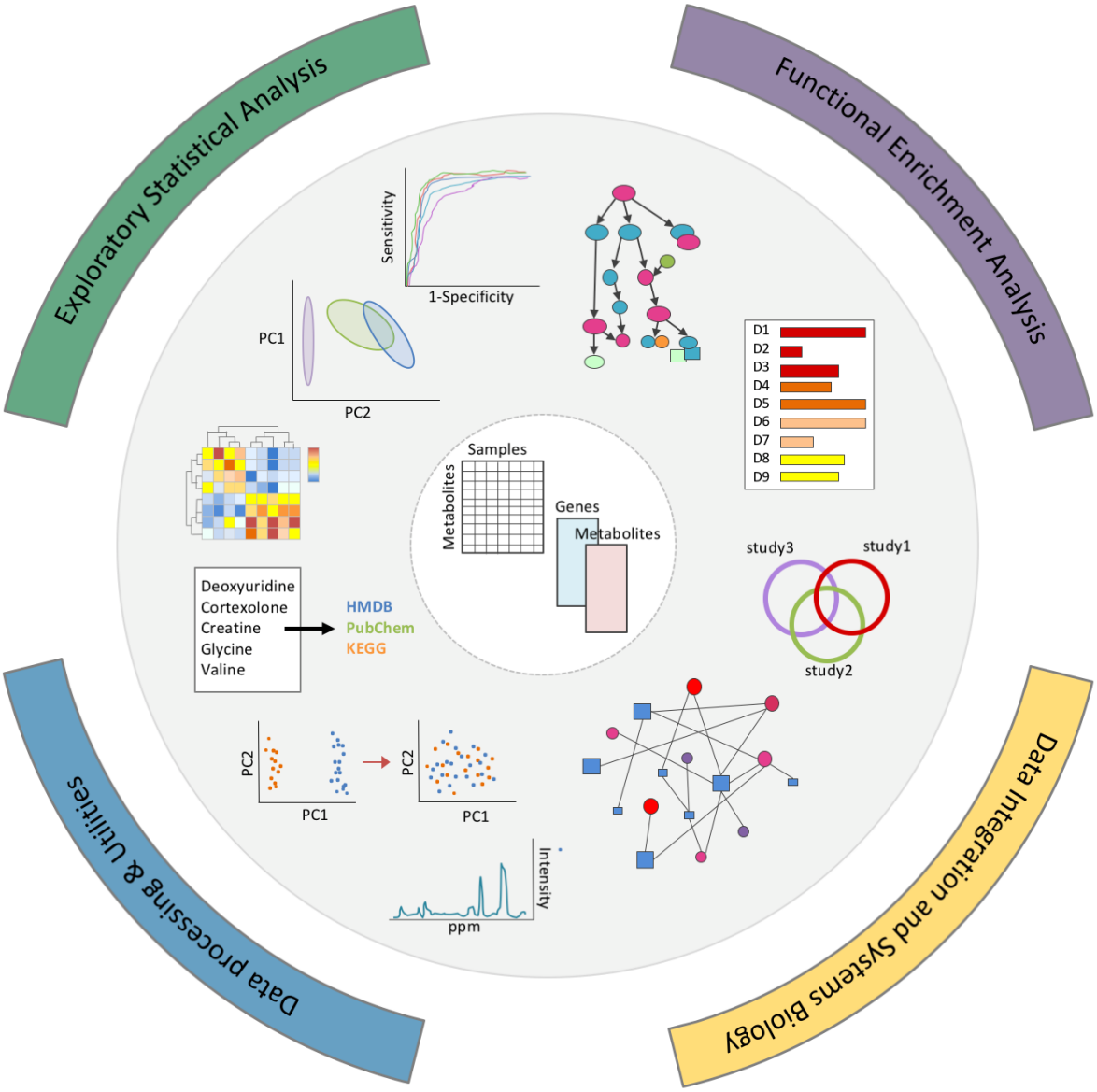


Figure 1. Overview of MetaboAnalyst 4.0 framework. The current modules can be organized into four general categories: (i) exploratory statistical analysis, (ii) functional analysis, (iii) data integration & systems biology and (iv) data processing & utility functions.

MetaboAnalystR and improved transparency/reproducibility

Thanks to continuing technological advancements in the field along with very helpful user feedback, many small updates and feature enhancements have taken place over the years to make MetaboAnalyst faster, more intuitive, and more robust. A potential downside associated with this continuous evolution is that it could lead to long-term reproducibility issues due to small changes in the interface or default parameter settings. While this flexibility is one feature that has made MetaboAnalyst so appealing, it has also made it inherently challenging to fully capture all steps required for reproducible analysis in the future. One possible way to alleviate this issue is to host multiple snapshots of the tool created at different time points. However, the maintenance costs associated with such an approach would be prohibitive. Another approach is to improve MetaboAnalyst's transparency throughout the analysis process. Because most of MetaboAnalyst's analytical tools are based on R functions, it would be much more efficient to capture the workflow using R commands (together with user-selected parameters). Furthermore, many computationally advanced users of MetaboAnalyst have requested improved support for its R functions in order to tailor their analysis to their data and to perform more extensive batch data processing.

To address both of these needs (i.e. greater support for transparency and batch analysis), we have developed the MetaboAnalystR package. This is a companion R-package that permits users to "see" and save the R code that MetaboAnalyst is running in real-time, which can then be used locally to reproduce their analytical workflow. MetaboAnalystR is designed to support more transparent, reproducible yet flexible analysis of metabolomic data within MetaboAnalyst. The R code between MetaboAnalystR and the web server has been extensively modified to ensure that

they are fully interchangeable and have identical functionalities across either platform. During each session of data analysis, these R commands are displayed on the right side of each page in the "R command history" sidebar, and each command appears sequentially based on when the command was executed (**Figure 2A**). MetaboAnalyst also stores the entire R command history as an executable R script that can be downloaded following the completion of each module. This script contains all user-selected parameters and selected tests. We believe that revealing the R code behind MetaboAnalyst improves transparency and allows users to track each step of their analysis in a form (R script) that can be easily shared and reproduced either on the web or locally using the MetaboAnalystR package. Beginner R users will be able to quickly learn the basics of MetaboAnalystR by copying the commands generated via their web-based analysis directly into R and reproduce their analyses; while advanced R users will be easily able to incorporate MetaboAnalystR package into their analytical workflows or customize the code to suit their needs. We believe that the MetaboAnalystR feature not only captures the workflow for better reproducibility, but also offers greater flexibility for more refined analysis and batch processing.

Obviously, with any major update to a resource like MetaboAnalyst, there is also some concern about the reproducibility (or return-accessibility) of data analyses performed using earlier versions of the server. For instance, due to updates to the underlying metabolite set libraries, the ranks and p-values of the top hits would change for the same input data. To help alleviate this issue, the previous version of MetaboAnalyst (version 3.0) will still be maintained, as long as there is sufficient interest and user traffic.

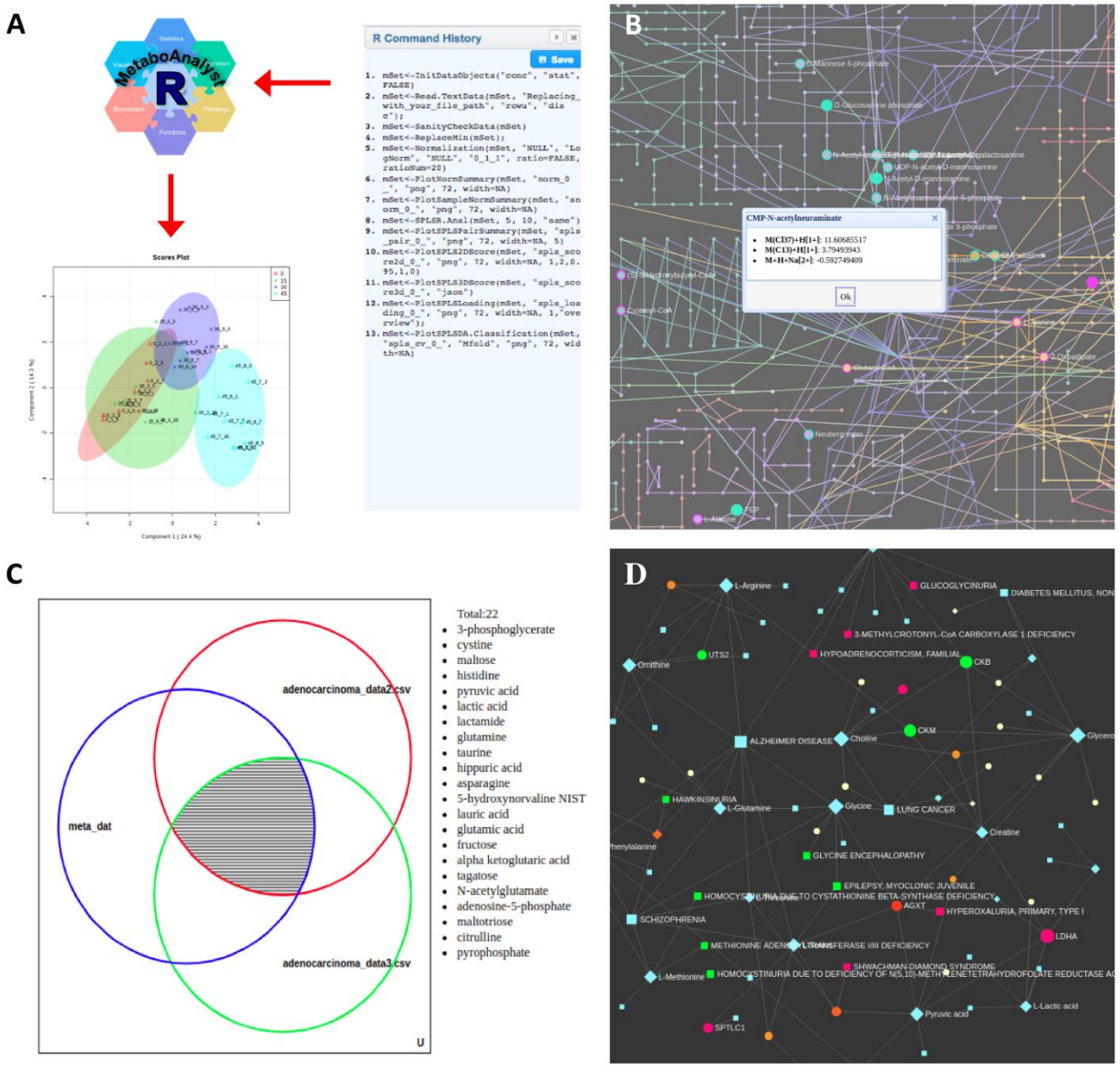


Figure 2. Summary of new features introduced in MetaboAnalyst 4.0. (A) An illustration showing the R command history panel and the companion MetaboAnalystR package will allow users to easily reproduce their analyses. In this case, the R command history captures all R commands leading to the generation the PLS-DA 2D score plot, which can then be reproduced in MetaboAnalystR using identical R commands (except the file path parameter for user input). (B) A zoomed-in view of the KEGG metabolic network showing the potential metabolite hits predicted from the mummichog algorithm. Clicking on a highlighted node will display all possible matched adduct forms of the corresponding compound. (C) An interactive Venn diagram showing the results froma biomarkermeta-analysis. Clicking on an area will show the corresponding hits. (D)An example of the metabolite-gene-disease interaction network created based on user input.

MetaboAnalyst's knowledgebase update

Considerable effort has been put in to update many of MetaboAnalyst's knowledgebases. This has been done to address potential issues such as the decline in analysis quality due to a lack of updated annotations (Marco-Ramell et al., 2018; Wadi, Meyer, Weiser, Stein, & Reimand, 2016). The most noteworthy updates are to the underlying compound databases, to the pathway datasets used for pathway analysis, and to the metabolite sets used for metabolite set enrichment (MSEA). The aim of these updates is to provide users more accurate and far deeper biological insights to help interpret their metabolomic data.

Compound database. MetaboAnalyst performs in-house mapping of common compound names to a wide-variety of database identifiers including KEGG (Minoru Kanehisa et al., 2017), HMDB (Wishart et al., 2017b; Wishart et al., 2012), ChEBI (Hastings et al., 2012), METLIN (Smith et al., 2005), and PubChem (Sunghwan Kim et al., 2015) prior to performing any selected functional analysis. This knowledgebase has been updated with HMDB Version 4.0 (Wishart et al., 2012), including updates of HMDB identifiers and links to other databases. As a result, MetaboAnalyst's compound database has been expanded to ~19 000 compounds. They represent the core subset of HMDB compounds (~114 100) with more detailed annotations relevant for downstream functional analysis.

Metabolite sets and pathway libraries. MetaboAnalyst's metabolite set libraries are primarily used by its metabolite set enrichment analysis module (MSEA). Many MetaboAnalyst users utilize the MSEA module to provide appropriate functional analysis and biological context

to their uploaded metabolomic data. Six metabolite sets, and one newly created metabolite set, were updated using HMDB version 4.0 (Wishart et al., 2012). The updated metabolite sets include disease sets in blood (344 diseases, increased from 330 diseases), cerebral spinal fluid (166 diseases, increased from 108 diseases), and urine (384 diseases, increased from 290 diseases), as well as location-based metabolite sets (73 organs, biofluids and tissues, increased from 57 organs, biofluids and tissues), pathway-based metabolite sets (147 metabolic pathways, increased from 80 metabolic pathways), single nucleotide polymorphisms (SNPs) metabolite set (4598 SNPs, increased from 4501 SNPs), and a new metabolite set consisting of drug-related pathways (461 pathways). It should be noted that these metabolite sets were derived from human-only data. We are currently updating the Pathway Analysis module to support interactive visual analysis of the extensive list of pathways from SMPDB (Jewison et al., 2013).

New module #1: Mummichog and MS peaks-to-pathways

High-throughput analysis and functional interpretation of untargeted MS-based (mass spectrometry-based) metabolomics data continues to be a major bottleneck in metabolomics. Conventional MS-based procedures typically include peak identification, spectral deconvolution, and peak annotation. A number of excellent methods have been developed to deal with the first two tasks (Lommen & Kools, 2012; Tomáš Pluskal, Castillo, Villar-Briones, & Orešič, 2010), which typically yield a list of "clean" MS peaks. Peak annotations are then performed manually by searching through a variety of spectral or compound databases. This process can often generate a number of false positives, due to redundancies in masses or the lack of unique MS spectral signatures for many small compounds (T. Kind & O. Fiehn, 2006; Kind & Fiehn, 2007). High-

resolution MS instruments are increasingly used to reduce these false hits analytically. Computationally, a promising approach is to shift the unit of analysis from individual compounds to individual pathways (or any groups of functionally related compounds which collectively produce more distinctive spectral footprints) - a concept similar to the widely used gene set enrichment analysis or GSEA (Subramanian et al., 2005). The *mummichog* algorithm was an elegant and efficient implementation of this concept that enables direct prediction of pathway activities from high-resolution MS peaks, without performing accurate peak annotation upfront (S. Li et al., 2013b). Currently, the algorithm lacks a graphic interface, thus limiting the access to many bench researchers. Due to its popularity and repeated user requests, we added a new module (called "MS Peaks-to-Pathways") in MetaboAnalyst to support *mummichog*-based MS peak analysis through user-friendly interface. We re-implemented the *mummichog* (version 1.0.10) algorithm in R to be consistent with MetaboAnalyst workflow and the aforementioned strategy of reproducibility. The knowledge-base for this module consists of five genome-scale metabolic models obtained from the original Python implementation which have either been manually curated or downloaded from BioCyc, as well as an expanded library of 21 organisms derived from KEGG metabolic pathways. The inclusion of SMPDB pathways for other model organisms will occur in the next few months (Jewison et al., 2013). While compound identification is generally de-emphasized in *mummichog*, the *post hoc* analysis of the matched compounds is critical for downstream validation and interpretation. To address these needs, we implemented a KEGG style global metabolic network to allow users to visualize the global peak matching patterns as well as to interactively zoom into a particular candidate compound to examine all of its matched isotopic or adduct forms.

To use this module, users must upload a table containing three columns - m/z features, p-values, and statistical scores (e.g. t-scores or fold-change values). If these values have not yet been calculated, users can use MetaboAnalyst's exploratory statistical analysis module to upload their raw m/z peak tables and perform their statistical analysis of choice, then upload these results into the *mummichog* module. Users also need to specify the mass accuracy, the ion mode (positive or negative), and the p-value cutoff to delineate between significantly enriched and non-significantly enriched m/z features. Following data upload, users must select an organism (library) from which to perform the untargeted pathway analysis. For a detailed explanation of the pathway analysis, please refer to MetaboAnalyst's web server FAQs, or to Li et al. 2013 (S. Li et al., 2013b).

The output of the *mummichog* module consists of a table of results containing ranked pathways that are enriched in the user-uploaded data. The table includes the total number of hits, their raw p-values (Fisher's exact test or Hypergeometric), their EASE score, and the p-value modeled on user data using a Gamma distribution. Users can click the "View" link to view the detailed hits for each pathway. A comprehensive table containing the compound matching information for all user-uploaded m/z features is also available for download. Importantly, all of this information (pathways, compounds, and matched hits) can be intuitively explored within the KEGG global metabolic network (**Figure 2B**). The page consists of three sections: 1) a top toolbar containing different menus to control various visualizations, 2) a left-hand section showing the pathway analysis results, and 3) a central view for interactive visual exploration of the metabolic network. Users can scroll their mouse to zoom in and out of the network view. Clicking on a pathway name will highlight all of its compounds within the network. Double-clicking a node will

show all the matching details for the corresponding compound as shown in the dialog (**Figure 2B)**. The current view can be downloaded as either a PNG or SVG file.

New module #2: Meta-analysis of metabolomics data

Biomarker identification continues to be an important area of research in metabolomics (Caroline H Johnson, Julijana Ivanisevic, & Gary Siuzdak, 2016). However, a major challenge in many metabolomics-based biomarker discovery efforts is the validation of potential metabolic markers (Hanash, Pitteri, & Faca, 2008). Questions have been raised about biomarker consistency and robustness across different metabolomics studies conducted on the same disease. As a result, the importance of external validation to improve statistical power for biomarker validation has been increasingly emphasized (Goveia et al., 2016; Tzoulaki, Ebbels, Valdes, Elliott, & Ioannidis, 2014). To address this issue of biomarker validation and reproducibility, there is growing interest among researchers to combine multiple published metabolomics datasets collected under similar conditions. The idea is that this approach would reduce study bias to enable more robust biomarker identification. This practice is often referred to as biomarker "meta-analysis" (Tseng, Ghosh, & Feingold, 2012). When executed properly, biomarker meta-analysis can leverage the collective power of multiple independent studies to overcome potential biases and small effect sizes associated with individual datasets. This can significantly improve the precision in identifying true patterns within the data (Haidich, 2010; Patti, Yanes, & Siuzdak, 2012; Walsh, Hu, Batt, & Santos, 2015). However, user-friendly tools dedicated to support biomarker meta-analysis of metabolomic data are currently lacking (Cambiaghi, Ferrario, & Masseroli, 2016). To address this issue, we have implemented a new module in MetaboAnalyst 4.0 called "Biomarker Meta-analysis". The

primary goal of the Biomarker Meta-analysis module is to provide a user-friendly tool for the integration of individual metabolomics studies to support the identification of robust biomarkers. The main steps for using this Biomarker Meta-analysis module are as follows:

i. Prior to uploading the data, the user should clean all datasets to ensure consistency amongst feature names (compound IDs, spectral bins, or peaks) as well as consistency in the class labels (two groups only) across all included studies;

ii. Once the data is cleaned and uploaded, the user can perform standard data processing, normalization, and differential analysis for each individual data set;

iii. Once each individual data set has been processed via step (ii) above, meta-analysis can be performed using one of several statistical options: a) combining p-values, b) vote counting, or c) direct merging of data for very similar datasets (Walsh et al., 2015);

iv. After step iii has been completed, the result table containing summary statistics for all significant features is then displayed. Users can then click to view a boxplot summary of any feature across the different datasets;

v. After completion, users can explore the results in an interactive Venn diagram to view the shared features among all possible combinations of the datasets. An example is shown in **Figure 2C**.

New module #3: Network explorer

Metabolomics is increasingly being used with other omics platforms such as transcriptomics, proteomics, and metagenomics to study complex diseases and to gain functional insights into microbial communities. However, integrating multiple omics data and interpreting

these results at a systems level has become a significant challenge (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015). A commonly used strategy is to analyze each set of omics data individually using tools and methods already developed for each field, and then piece together the "big picture" using individual lists of significant features (metabolites, genes, proteins, *etc.*). In particular, biological networks are a very intuitive and flexible vehicle to convey our knowledge at a systems level. For instance, known relationships between genes, metabolites, and diseases can be easily represented as knowledge-based networks. By harnessing the power of networks and *a priori* biological knowledge, these lists of significant features can be co-projected onto the networks to reveal important links between them, as well as their associations with diseases or other interesting phenotypes. Such a comprehensive knowledgebase that connects metabolites with other molecular entities or phenotypes of interest, coupled with support for interactive network visualization, will be an essential asset to help address current data integration challenges (Cambiaghi et al., 2016; Charitou, Bryan, & Lynn, 2016). In MetaboAnalyst 4.0, this is addressed in the Network Explorer module. The aim of this module is to provide users with an easy-to-use tool that permits the mapping of their metabolites and/or genes (including KEGG orthologs or KOs) onto different types of molecular interaction networks. This network visualization can then be used to gain novel insights or assist users with the development of new hypotheses.

This new Network Explorer analysis module complements MetaboAnalyst's joint-Pathway Analysis module by allowing the identification of connections that cross pathway boundaries (e.g. metabolite-disease interactions) as well as enabling a more global view of the pathways which may not be obvious when examined individually. The Network Explorer module currently supports five types of biological networks including the KEGG global metabolic network, a gene-

metabolite interaction network, a metabolite-disease interaction network, a metabolite-metabolite interaction network, and a metabolite-gene-disease interaction network. The last four networks are created based on information gathered from HMDB and STITCH databases (Yao et al., 2015), and are applicable to human studies only.

Users can upload either a list of metabolites, a list of genes, or both. For the metabolite list, MetaboAnalyst 4.0 currently accepts compound names, HMDB IDs, or KEGG compound IDs as metabolite identifiers. For the gene/protein list, Entrez IDs, ENSEMBL IDs, official gene symbols, or KEGG orthologs are currently supported. The uploaded list of metabolites and/or genes/proteins is then mapped using MetaboAnalyst's internal databases. Following this step, users can select which of the five networks to begin to visually explore their data. On the network visualization page, users can use their mouse or touchpad to zoom in and out, highlight, drag and drop nodes (except the KEGG global metabolic network), or click on a node/edge for further details. Users can also perform functional enrichment analysis and then highlight those metabolites, genes or proteins involved in functions of interest on the network. The background color of the network and the colors of the nodes and edges can also be customized. An example of the output from MetaboAnalyst's Network Explorer module is shown in **Figure 2D**. Each generated network can then be exported as an SVG or PNG image for publication purposes. We believe that the integration of interactive network exploration, network topological analysis, and functional enrichment analysis will provide users with more informative views and richer contextual information to facilitate the generation of testable hypotheses.

Other feature updates

There have been many small updates based on user suggestions that have accumulated over the past three years. For instance, in the biomarker analysis module, many users indicated that they wanted to be able to select features that give information complementary to biomarkers they already selected. We have therefore added feature similarity information using the cluster membership from k-means analysis to enable this feature selection. Based on additional user feedback, we have also added two variants of the popular partial least square (PLS) methods, including orthogonal PLS (OPLS) and sparse PLS (SPLS) for improved data interpretation and more robust statistical analysis (Lê Cao, Boitard, & Besse, 2011; Thevenot, Roux, Xu, Ezan, & Junot, 2015). For the two-way analysis of variance (ANOVA), we have added support for both type I and type III ANOVA, as well as additional analysis options for different experimental designs. While it is well known that MetaboAnalyst only has limited support for raw spectra profiling, we have attempted to remedy this by adding a "Spectral Analysis" feature to point users to several easy-to-use, web-based tools that are freely available for raw spectra processing, analysis, and annotation. It currently contains links to Bayesil (Ravanbakhsh et al., 2015), GC-AutoFit, and XCMS Online for NMR, GC-MS, and MS spectra processing, respectively.

Implementation

MetaboAnalyst 4.0 was implemented based on the PrimeFaces (v6.1) component library (http://primefaces.org/) and R (version 3.4.3). The interactive network visualization was implemented using the sigma.js JavaScript library (http://sigmajs.org). The entire system is hosted

on a Google Cloud server with 32GB of RAM and eight virtual CPUs with 2.6 GHz each. The server is capable of dealing with 5000~8000 data analysis jobs submitted from ~1000 users on a daily basis. For those who wish to use MetaboAnalyst 4.0 locally, we have provided the options to download the .war file or the MetaboAnalyst Docker image. Detailed instructions for download and installation of the Docker image are provided on the "Resources" page of the web server. The MetaboAnalystR package is available from the GitHub (https://github.com/xia-lab/MetaboAnalystR)

Comparison with other tools

Several web-based as well as several web-enabled tools for metabolomic data analysis have been developed over recent years, including XCMS Online (Tautenhahn, Patti, Rinehart, & Siuzdak, 2012), Workflow4Metabolomics (Giacomoni et al., 2014), Galaxy-M (Davidson, Weber, Liu, Sharma-Oates, & Viant, 2016), and Metabox (Wanichthanarak, Fan, Grapov, Barupal, & Fiehn, 2017). Detailed comparisons between these tools and MetaboAnalyst 4.0, as well as its previous versions are shown in **Table 1**. Based on this table is evident that MetaboAnalyst offers the most comprehensive support for statistical analysis, functional interpretation and integration with other omics data. It is also evident that MetaboAnalyst supports real time interactive data analysis in way that no other tool currently can. While MetaboAnalyst has been limited in its built-in support for raw spectral processing and annotation, the new "Spectral Analysis" module help address this shortcoming. Certainly, raw LC-MS spectra processing and analysis has been a major strength of XCMS online, Galaxy-M, and Workflow4Metabolomics, and these tools continue to be the "go-to" resources for LC-MS data analysis. Overall, the primary strength of MetaboAnalyst

is in its downstream data analysis, just as it is with Metabox. Indeed, the design of Metabox is similar to MetaboAnalyst in that it primarily accepts preprocessed metabolomics data for various statistical computing, functional analysis, and network-based integration. However, as noted in **Table 1**, no public server is currently available for Metabox and researchers must install it locally in order to use this tool.

Table 1. Comparison of MetaboAnalyst with other tools. The table compares the main features of MetaboAnalyst (versions 1.0 - 4.0) with other web-based or web-enabled tools. Symbols used for feature evaluations with "√" for present, "-" for absent, and "+" for a more quantitative assessment (more "+" indicate better support).

Tool name	**MetaboAnalyst**				**XCMS online**	**Galaxy-M**	**W4M**	**Metabox**
	4.0	**3.0**	**2.0**	**1.0**				
Data processing								
Raw spectra	++	+	+	+	+++	+++	+++	-
Data filtering	√	√	√	-	√	√	-	-
Missing-value	√	√	√	√	-	√	-	-
Normalization	+++	+++	++	++	-	++	++	++
Statistical analysis								
Univariate	+++	+++	+++	++	+	++	++	++
Multivariate	+++	++	++	++	++	+	+++	+
Clustering	+++	+++	++	++	+	-	+	+
Classification	++	++	++	++	-	-	-	-
Power analysis	√	√	-	-	-	-	-	√
Biomarker analysis	√	√	-	-	-	-	-	-
Functional analysis								
Enrichment	+++	++	++	-	-	-	-	+
Pathway analysis	+++	++	++	-	√	-	-	√
Mummichog	++	-	-	-	+	-	-	-
Data integration and systems biology								

Joint pathway analysis	√	√	-	-	√	-	-	-
Knowledge - based network analysis	√	-	-	-	-	-	-	√
Correlation-based network analysis	-	-	-	-	-	-	-	√
Biomarker meta-analysis	++	-	-	-	+	-	-	-

- XCMS Online: https://xcmsonline.scripps.edu
- Galaxy-M: https://github.com/Viant-Metabolomics/Galaxy-M
- Workflow4Metabolomics (W4M): http://workflow4metabolomics.org/
- Metabox: http://kwanjeeraw.github.io/metabox/

Conclusions

Perhaps the most visible change to MetaboAnalyst is its newly designed web interface, which allows new features to be more easily "plugged in". It also gives more space to permit interactive exploration of large networks as well as to display R command history during a standard data analysis. We believe the latter feature, in combination with the release of the MetaboAnalystR package, will greatly improve reproducibility and transparency during metabolomics data analysis. Many advanced MetaboAnalyst users have felt constrained by the analysis boundaries defined by its web interface and have asked for a more flexible workflow design and batch processing capabilities. The MetaboAnalystR package addresses these limitations. Users can now create a workflow (R command history) through the web interface, customize the workflow by changing the order of the commands or their parameters, and finally execute the workflow in batch mode using the R package. For those researchers who are already familiar with R programming, it is also

possible to directly modify MetaboAnalyst's R code to suit their needs. Another major focus of this MetaboAnalyst update is the addition of new modules to support further data integration (biomarker meta-analysis and multi-omics analysis), as well as functional analysis for high-resolution untargeted MS (*mummichog*). These additions were made in response to frequent user requests and growing trends seen in metabolomic data analysis practices. Finally, to ensure that the biological interpretation of metabolomic data remains as current and insightful as possible, all of MetaboAnalyst's underlying knowledgebases have been updated. These updates will allow metabolomics researchers to move beyond simply re-iterating common textbook interpretations of metabolism and give them much more useful insights into complex and relevant biological processes that are ultimately driven by metabolites. Overall, we believe these updates will allow MetaboAnalyst to remain at the cutting edge of computational metabolomics and systems biology, and that it will continue to enable new discoveries and greater insights for a growing number of metabolomics researchers.

Funding

The authors would like to acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), Genome Canada, and the Canada Research Chairs (CRC) Program.

Conflict of interest

None declared.

Preamble to Chapter 3

High-throughput analysis and functional interpretation of untargeted mass spectrometry (MS)-based metabolomics data continues to be a major bottleneck in current metabolomics research. Following peak picking and alignment, conventional procedures typically require annotation of peak lists to named metabolites prior to downstream analysis. This process of peak identification is known to be very time consuming and error prone. MS peak lists are defined by a combination of a mass-to-charge ratio (m/z) and retention time. Despite the high mass accuracy of modern instruments, it is not unusual for several metabolites to match a single MS peak. It is also possible that completely unknown metabolites may have identical m/z values to known metabolites. The use of tandem MS (or MS/MS) experiments, which provides additional information beyond m/z values or retention time, can be used to distinguish, or identify some compounds. However, MS/MS experiments adds additional time and cost, therefore most untargeted metabolomics studies still rely on high resolution MS data.

One promising approach to metabolite identification in untargeted high-resolution MS metabolomics studies is to leverage the collective power of metabolic pathways to help resolve the ambiguity of metabolite annotations. In particular, the *mummichog* algorithm bypasses the bottleneck of metabolite identification prior to pathway analysis by leveraging *a priori* pathway and network knowledge to directly infer biological activity based on MS peaks (S. Li et al., 2013b). The underlying assumption in *mummichog* is that if a list of significantly enriched features truly reflects biological activity, the representation of these true metabolites would be enriched on localized structures such as pathways, while false matches would be distributed at random. To use

the algorithm, users must first provide a list of m/z features and a p-value for each feature, hereby referred to as $\mathbf{L_{ref}}$. Two lists will be further drawn from $\mathbf{L_{ref}}$. One list is called $\mathbf{L_{sig}}$, which contains only the significant m/z features and the other list is called $\mathbf{L_{perm}}$, which is a list of randomly drawn m/z features from $\mathbf{L_{ref}}$, but is the same length as $\mathbf{L_{sig}}$. The main steps are as follows:

1. A list of randomly drawn m/z features are drawn from $\mathbf{L_{ref}}$ to create $\mathbf{L_{perm}}$. These m/z features are mapped to potential metabolites, considering all possible isotopic configurations and adducts (e.g. M+H[+] and M-H[-]).
2. The list of potential compounds is then mapped to the user's selected library of pathways and a p-value is calculated per pathway.
3. Steps 1 and 2 are repeated many times to compute the null distribution of p-values (modelled as a gamma distribution).
4. $\mathbf{L_{sig}}$ is then mapped to potential metabolites for pathway enrichment analysis, and the resulting p-values (Fisher's or Hypergeometric, and EASE scores) per pathway are calculated and adjusted for the null distribution.

The following chapter highlights one of several advancements I have made to this algorithm. Other enhancements include the inclusion of retention time to increase the accuracy of putative compound identification (Pang, Chong, Li, & Xia, 2020) and to enable meta-analysis of untargeted metabolomics data (Pang et al., 2021).

Chapter 3. From Raw Spectra to Biological Insights

MetaboAnalystR 2.0: From Raw Spectra to Biological Insights

[1]Jasmine Chong, [1]Mai Yamamoto, and [1,2,3,4]Jianguo Xia

[1]Institute of Parasitology, and [2]Department of Animal Science, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada

[3]Department of Microbiology & Immunology, Montreal, Quebec, Canada

[4]Department of Human Genetics, Montreal, Quebec, Canada

Published in *Metabolites* (Chong, Yamamoto, & Xia, 2019).

Abstract

Global metabolomics based on high-resolution liquid chromatography mass spectrometry (LC-MS) has been increasingly employed in recent large-scale multi-omics studies. Processing and interpretation of complex metabolomics datasets have become a key challenge in current computational metabolomics. Here we introduce MetaboAnalystR 2.0 for comprehensive LC-MS data processing, statistical analysis, and functional interpretation. Compared to the previous version, this new release seamlessly integrates XCMS and CAMERA to support raw spectral processing and peak annotation. Additionally, it features high-performance implementations of mummichog and GSEA approaches for prediction of pathway activities. The application and utility of the MetaboAnalystR 2.0 workflow were demonstrated using a synthetic benchmark dataset and a clinical dataset. In summary, MetaboAnalystR 2.0 offers a unified and flexible workflow that enables end-to-end analysis of LC-MS metabolomics data within the open-source R environment.

Introduction

Metabolomics is the comprehensive study of all small molecule metabolites (<1500 Da) detected within a biological system. An individual's metabolic profile represents the functional product of interactions among genetics, lifestyle, environment, diet, and native microbiota, which closely reflects his or her health status (Beger et al., 2016; David S Wishart, 2016). The metabolome thus serves as the link between genotype and phenotype, and metabolomics plays a critical role in the development and implementation of precision medicine (Fiehn, 2002; Caroline H Johnson, Julijana Ivanisevic, & Gary Siuzdak, 2016).

There are two general approaches in conducting metabolomics. Targeted metabolomics aim to study a predefined set of metabolites, requiring familiarity with the system (Caroline H Johnson, Julijana Ivanisevic, & Gary Siuzdak, 2016). Untargeted metabolomics, also known as global metabolomics, aim to measure the global set of metabolites within a sample without *a prioi* knowledge of the system. A typical metabolomics analysis workflow involves three main steps: raw data processing, statistical analysis, and functional interpretation (**Figure 1**). Global metabolomics requires more sensitive analytics platforms to achieve comprehensive measurement. High-resolution liquid chromatography-mass spectrometry (LC-MS) systems is currently the main workhorse for global metabolomics. The platform often generates thousands of signals, including true biological signals from metabolites, their adducts, fragments, and isotopes, as well as noise signals from contaminants and artifacts (Nash & Dunn, 2018). Computational tools able to significantly reduce noise in MS spectra are crucial for more meaningful downstream analyses (Uppal et al., 2016).

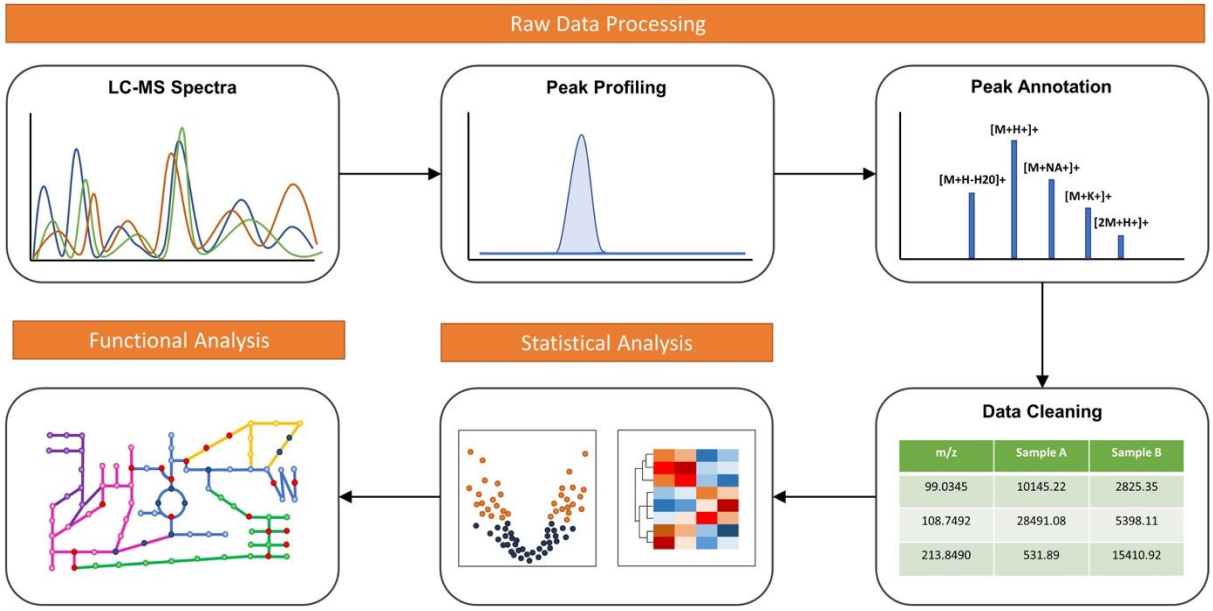


Figure 3. A typical metabolomics data analysis workflow. The workflow starts from the raw data input and flows to data preprocessing, data processing, and data analysis.

There are several powerful computational workflows including commercial tools such as Mass Profiler (Agilent Technologies) and Compound Discoverer (Thermo Scientific), cloud-based software such as XCMS Online (Forsberg et al., 2018) and Workflow4Metabolomics (Giacomoni et al., 2014), desktop software such as MZmine2 (T. Pluskal, Castillo, Villar-Briones, & Oresic, 2010), MS-DIAL (Tsugawa et al., 2015), and Open-MS (Rost et al., 2016), and finally R packages such as MAIT (Fernández-Albert, Llorach, Andrés-Lacueva, & Perera, 2014) and metaX (Wen, Mei, Zeng, & Liu, 2017). Most of these software focus on addressing one of the two main tasks - spectral processing and/or statistical analysis. Consequently, users must often learn several tools to meet their data analysis needs. Due to compatibility issues, it is often necessary to write scripts to convert outputs from one tool to the next.

Tools for functional interpretation of global metabolomics data is in general lacking or poorly addressed (Gardinassi, Xia, Safo, & Li, 2017; J. Xia, 2017). A prerequisite for metabolomics data interpretation is metabolite identification, thereby permitting the contextualization of annotated peaks in metabolic pathways and their integration with other omics data. However, even with high mass accuracy afforded by the current high-resolution MS platforms, it is often impossible to uniquely annotate a given peak based on its mass alone (Kind & Fiehn, 2007). Researchers usually need to manually search compound databases and then perform further experimental validations such as tandem MS. Novel bioinformatics tools are urgently needed to enable researchers to gain biological insights with a minimum amount of manual efforts. To get around this issue, a key concept is to shift the unit of analysis from individual compounds to individual pathways or a group of functionally related compounds (i.e. metabolite sets (J. Xia & D. S. Wishart, 2010)). The general assumption is that the collective behavior of a group is more robust against a certain degree of random errors of individuals. The *mummichog* algorithm is the first implementation of this concept to infer pathway activities from a ranked MS peaks (S. Li et al., 2013b). The original algorithm implements an over-representation analysis (ORA) method to evaluate pathway-level enrichment based on significant features. An alternative approach is the Gene Set Enrichment Analysis (GSEA) method, which is widely used to test enriched functions from ranked gene lists (Subramanian et al., 2005). Unlike ORA, GSEA considers the overall ranks of features without using a significance cutoff. It can detect subtle and consistent changes which can be missed from using ORA methods. Despite its widespread applications in gene expression profiling, it has not yet been used for global metabolomics.

MetaboAnalyst is one of the most widely used tools for statistical and functional analysis of metabolomics data (Chong et al., 2018; J. Xia, R. Mandal, I. V. Sinelnikov, D. Broadhurst, & D. S. Wishart, 2012; J. Xia, N. Psychogios, N. Young, & D. S. Wishart, 2009; J. Xia, I. V. Sinelnikov, B. Han, & D. S. Wishart, 2015). It was initially designed for targeted metabolomics, and subsequent releases gradually introduced many statistical methods applicable to both targeted and untargeted metabolomics. Due to its web-based implementation, there is very limited support for raw spectra processing and peak annotation. The most recent update (version 4.0) was released with a companion R package, MetaboAnalystR (v1.0), to help tackle issues associated with workflow customization, reproducibility, and handling large datasets (Chong & Xia, 2018).

Here we present MetaboAnalystR (v2.0) to address the two important gaps left in its previous version: 1) raw spectral processing - we have implemented comprehensive support for raw LC-MS spectral data processing including peak picking, peak alignment and peak annotations; and 2) functional interpretation directly from m/z peaks - in addition to an efficient implementation of the mummichog algorithm (S. Li et al., 2013b), we have added a new method to support pathway activity prediction based on the well-established GSEA algorithm (Subramanian et al., 2005). We showcase the performance of these new functions through two case studies.

Results

MetaboAnalystR 2.0 consists of a series of flexible R functions that can take a variety of user-supplied data and parameters to perform end-to-end metabolomics data analysis. The source code is freely available at the GitHub repository (https://github.com/xia-lab/MetaboAnalystR).

Detailed instructions, tutorials, troubleshooting tips, example data, and analyses discussed in this paper are also available in this repository.

To demonstrate the utility of MetaboAnalystR 2.0 workflow, we present the results from the two case studies: (*i*) a synthetic benchmark dataset to evaluate the raw MS spectra processing functions, with a focus on its peak detection and quantification performance; and (*ii*) a clinical pediatric inflammatory bowel disease (IBD) dataset to showcase the overall workflow, with a focus on its capacity to provide biological insights. The first case study consists of eight mzML files (~150 MB each) that were preprocessed using MetaboAnalystR 2.0 on an Ubuntu 18.04 desktop computer with 24 GB RAM and an 8 core i7 processor, using two cores in parallel. The total running time was 18 minutes. The clinical case study consists of 48 mzML files (~75 MB each) that were preprocessed using MetaboAnalystR 2.0 on an Ubuntu 16.04 server with 128 GB RAM and a 32 core i7 processor, using 15 cores in parallel. The total running time was 40 minutes. All R scripts to perform the entire metabolomics data analysis pipeline are available from the MetaboAnalystR GitHub repository under the section "Case Studies" (https://github.com/xia-lab/MetaboAnalystR). The accompanying vignette ("The MetaboAnalystR 2.0 Workflow") provides a step-by-step tutorial to demonstrate to users how to use MetaboAnalystR 2.0 to perform an end-to-end metabolomics data analysis on a subset of 12 of the 48 clinical IBD samples. This tutorial was created on a Dell XPS 15 9570 laptop dual-booted with Ubuntu 16.04 and 16 GB of ram. The total running time of the tutorial was 14 minutes, averaging ~1.25 minutes per sample, using 6 cores in parallel and 10.5 GB of ram.

Benchmark Case Study

We first demonstrate the accuracy of the raw data preprocessing module using a benchmark dataset comprised of a mixture of 1100 known compounds ranging in size from 100 to 1300 Da (Z. Li et al., 2018). The original study used a targeted analysis to obtain their benchmark feature list, which we used as the ground truth to evaluate our workflow. As shown in **Table 1**, the original study detected 35,215 peaks using XCMS Online, with 820 classified as true features. Using the same data preprocessing parameters as published, MetaboAnalystR 2.0 detected 21,013 peaks from the benchmark data. Among them, 732 matched the true features based on m/z and retention time (10 ppm and 0.3 min RT tolerance).

Table 2. Comparison of MetaboAnalystR 2.0 with XCMS Online. This table compares the peak identification and quantification accuracies using the benchmark dataset between MetaboAnalystR 2.0 and the original manuscript using XCMS Online.

	Methods	**Features detected**	**True Features**		
			Total	**Accurately quantified**	**Discriminating**
Li et al. 2018 (Z. Li et al., 2018)	Targeted	-	836	836	-
	Untargeted (XCMS Online)	35215	820	731	45
MetaboAnalystR 2.0	Untargeted	21013	732	632	45

Next, we compared the number of accurately quantified true features using MetaboAnalystR 2.0 to those from the original manuscript using XCMS Online (**Table 1**). Features were accurately quantified if their fold changes had a <20% relative error as compared to the benchmarked data. MetaboAnalystR 2.0 accurately quantified 632 features and identified 45 truly discriminating features.

IBD Case Study

The 48 fecal samples were obtained from 24 pediatric Crohn's Disease (CD) patients and 24 pediatric healthy controls (**Table S1**). Our workflow detected 8187 features which reduced to 6930 features after filtering out isotopes and missing features within >50% of samples. After exclusion of low-variance features, a total of 4113 features were statistically analyzed using the standard MetaboAnalystR functions.

Mann-Whitney U test and fold change analysis detected 59 features that were significantly different between CD and healthy controls. Differences between CD and healthy controls were evaluated using PCA, PLS-DA, and OPLS-DA. The PCA showed an overlapping of clusters along the first two components, with CD exhibiting a wider data distribution (**Figure S1**). This indicates an overall similarity of the metabolic profiles between CD and healthy controls but larger heterogeneity within CD patients. The PLS-DA score plot showed a clear separation between the two groups (**Figure S2**). Ten-fold cross validation of two PLS-DA components gave an R2 of 0.912 and Q2 of 0.424 (**Figure S3**). The OPLS-DA score plot shows a clear separation between CD and healthy controls (**Figure 2**). The R2Y and Q2Y values from the OPLS-DA were 0.501 and 0.272 respectively, indicating a moderate goodness of fit but poor predictive ability. To further

evaluate the model, we performed permutation tests (n=1000). The R2Y and Q2Y values from the OPLS-DA permutation tests were 0.979 ($p = 0.026$) and 0.522 ($p < 0.001$). Altogether, a clear distinction between the metabolome of CD and healthy controls was observed.

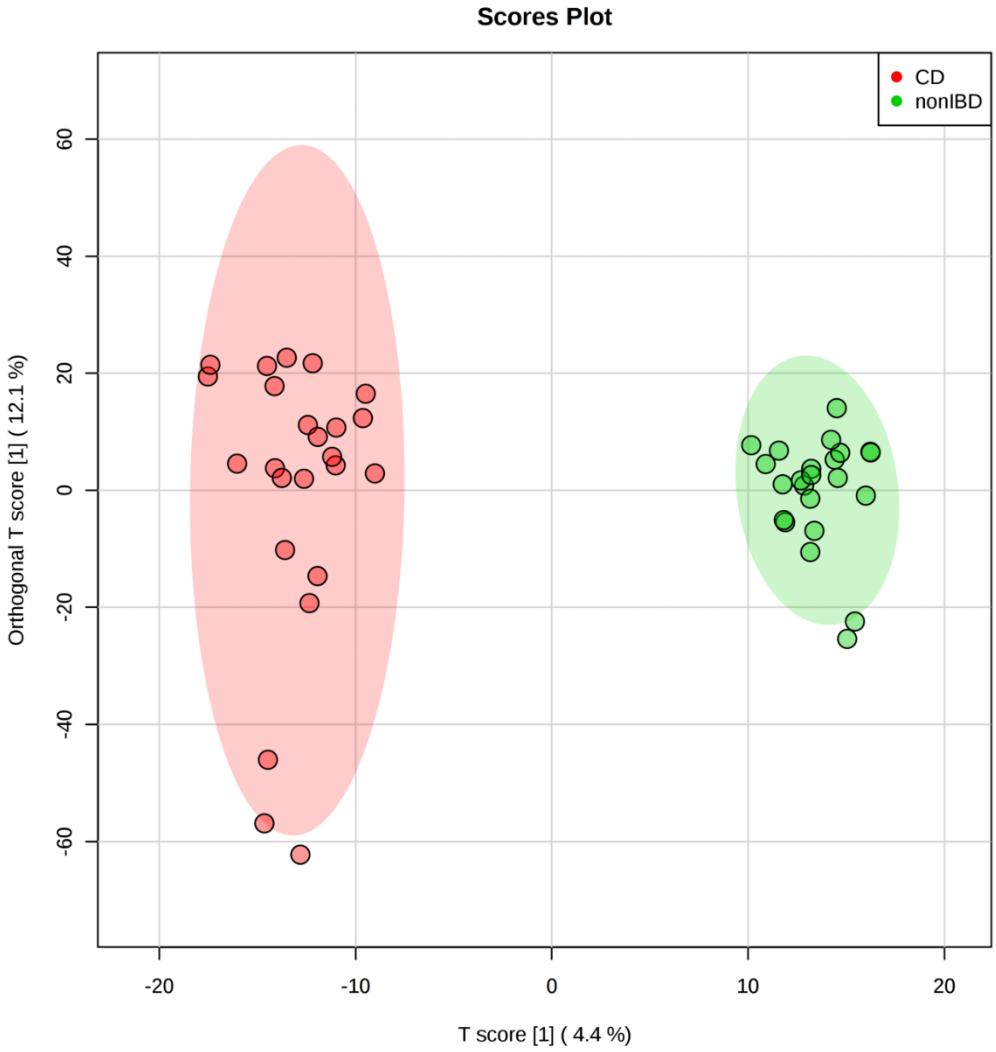


Figure 4. OPLS-DA Score plot. The OPLS-DA score plot based on the 4113 features from the stool metabolome of 24 pediatric Crohn's disease patients and 24 healthy children, with a R2 of 0.912 and Q2 of 0.424.

To gain potential biological insights from the global metabolomics data, we applied both mummichog and GSEA algorithms and integrated their results (**Figure 5**). Mummichog suggested that differentially abundant features between CD and healthy patients were associated with perturbations in bile acid biosynthesis and fatty acid activation, as well as vitamin E, fatty acid, and vitamin D3 metabolism. The GSEA algorithm also identified alterations in bile acid biosynthesis. Moreover, it identified differences in androgen and estrogen biosynthesis and metabolism, squalene and cholesterol biosynthesis, biopterin metabolism, and butanoate metabolism. More details of the top enriched pathways from both methods are given in **Table 2**.

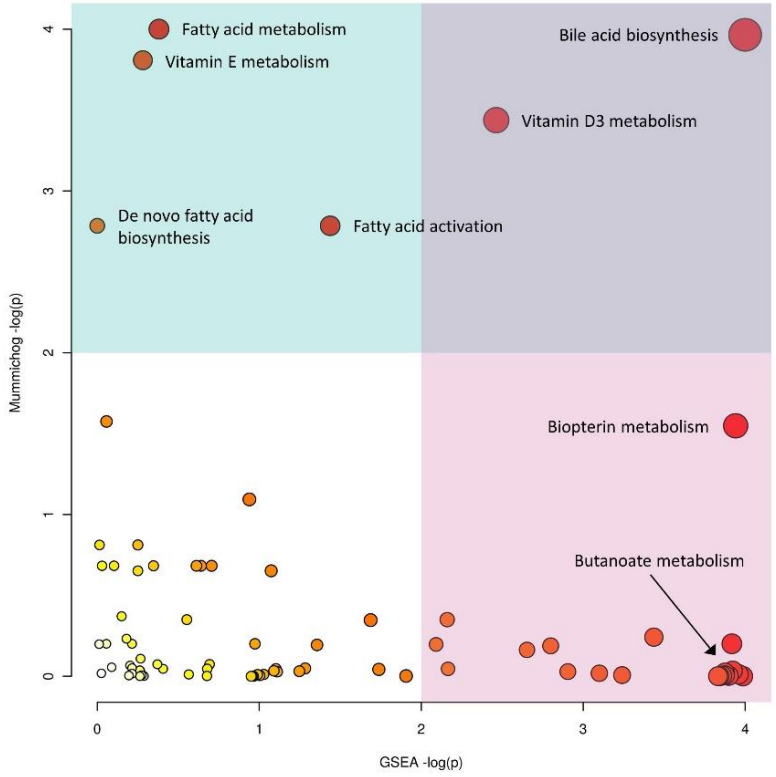


Figure 5. Scatter plot integrating mummichog and GSEA pathway analysis results. On the x-axis are negative log raw GSEA p-values, and on the y-axis are negative log raw mummichog p-values. The size and color of the circles correspond to the value of their transformed combined

p-values (ascending, from white to dark red). The blue and pink quadrants represent pathways that were significant using a single pathway analysis algorithm (blue = mummichog, pink = GSEA), and the purple quadrant contains significantly perturbed pathways identified using both algorithms. From the plot, bile acid metabolism and vitamin D3 metabolism show congruence as the top enriched pathways.

Table 3. Top metabolic pathway alterations using MetaboAnalystR 2.0. The table shows the top five metabolic pathways identified between patients with pediatric Crohn's disease (n=24) as compared to healthy controls (n=24), using the mummichog algorithm (*PerformMummichog*) and GSEA (*PerformGSEA*) implemented in MetaboAnalystR 2.0.

Mummichog			**GSEA**		
Pathway Name	Compound Hits*	P-Value	Pathway Name	Compound Hits	P-Value
Bile acid biosynthesis	29/52	0.00282	Bile acid biosynthesis	52	0.001761
Vitamin E metabolism	20/33	0.00356	Androgen and estrogen biosynthesis and metabolism	10	0.01465
Fatty Acid Metabolism	9/11	0.00268	Squalene and cholesterol biosynthesis	7	0.02214
Vitamin D3 metabolism	8/10	0.00616	Biopterin metabolism	14	0.07806
Fatty acid activation	10/15	0.01620	Butanoate metabolism	11	0.08318

* The mummichog compound hits represent the number of significant compound hits divided by the total number of compound hits per pathway.

Interestingly, the GSEA algorithm identified Butanoate metabolism as a significantly enriched pathway (q = 0.01988), whereas the mummichog algorithm did not (q = 0.43279). The mummichog algorithm only utilizes significantly different m/z features, therefore only three features in Butanoate metabolism were used to calculate a pathway enrichment score. On the other hand, GSEA utilized all 20 compound hits (corresponding to 38 m/z features) to calculate enrichment **(Figure S4)**. Of these features, 145.0496 m/z was putatively annotated as (S)-2-Aceto-2-hydroxybutanoate (a deprotonated ion), as was 205.0710 m/z (a formic acid adduct). Furthermore, 124.0392 m/z corresponded to 2-Butynoate. This demonstrates the ability of GSEA to pick up on subtle changes, such as perturbations in Butanoate metabolism, and the utility of using both algorithms to gain biological insight.

We further examined the 17 features that overlap between the putatively annotated features in the pathway analysis and the important features found in univariate statistical analysis. Notably, 431.3164 m/z was putatively annotated as a deprotonated ion of 3-β, 7-α-dihydroxy-5-cholestenoate (C17336) based on its correspondence to the exact mass of C17336 from the KEGG database (Minoru Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2011). This compound is found in the primary bile acid pathway. Additionally, the same mass also corresponds to a deprotonated ion of 23S, 25, 26-trihydroxyvitamin D3 (CE2202). Exact identification of this feature requires further experiments, which is beyond the scope of this manuscript. In addition to this compound, five additional compounds out of the 17 have been previously found as stool metabolites in the context of IBD (Franzosa et al., 2019). Representative extracted ion chromatograms (EICs), boxplots and corresponding information, such as m/z, retention time and p-values are highlighted in supplemental information (**Figure S5**).

Discussion

In this paper, we have described the new functions introduced in MetaboAnalystR 2.0 to support global metabolomics data analysis, covering raw LC-MS spectra processing to generation of biological insights. These functions were showcased through two case studies. For the benchmark dataset, despite applying the same parameters used by Li et al. (Z. Li et al., 2018), we were unable to reproduce the identification and quantification accuracy obtained by the original authors using XCMS Online. Their setup detected >14000 (68%) more features compared to those obtained using our pipeline. We tried several options, including the suggested parameters for a HPLC or UPLC coupled with a Q Exactive HF mass spectrometer. We posit this incongruity arose because the authors did not specify the exact peak width used, which is a critical parameter for peak picking. Additionally, the data conversion step from .RAW to mzML used in our workflow may have resulted in a slight difference in the input data when compared to the data conversion used in XCMS Online. It is also important to note that our workflow integrated the latest version of XCMS (version 3.4.4), which have introduced many new functionalities and updates in existing functions. Overall, our preprocessing workflow performed well, executing peak picking, annotation, and filtering on the eight benchmark samples in less than twenty minutes.

For the IBD case study, we observed a clear separation in the metabolic profiles between pediatric CD patients and healthy controls using either PLS-DA or OPLS-DA. Furthermore, our analysis highlighted several metabolic pathways associated with CD, without performing accurate metabolite identification. For instance, alterations in bile acid biosynthesis is well known among CD patients as a result of inflammation in the terminal ileum, which is the critical site of bile acid

absorption (Duboc et al., 2013; Hofmann & Hagey, 2008). Combining the results of pathway analysis and statistical analysis also putatively identified some promising metabolic features that could be used to as potential biomarkers. In addition to bile acids, vitamin D has been shown to play an immunomodulatory role in IBD pathogenesis (Limketkai, Mullin, Limsui, & Parian, 2017). Taken together, this use case demonstrates the ease of which MetaboAnalystR 2.0 can be utilized to gain mechanistic insights and generate hypotheses for future experimental validation.

Conclusion

The previous version (v1.0) of MetaboAnalystR features comprehensive normalization and statistical methods inherited from the MetaboAnalyst web server. The version 2.0 seamlessly integrates XCMS and CAMERA to support raw MS spectral processing and peak annotation, it also contains efficient implementations of mummichog and GSEA methods for prediction of pathway activities. The performance of this workflow was evaluated on a published benchmark dataset as well as a recent clinical study on IBD. The MetaboAnalystR package is maintained in conjunction with the cloud-based MetaboAnalyst web application and is under continuous development based on the community feedback. Our next focus is to support isotope-resolved metabolomics as well as development of a Galaxy-based platform for raw data processing (Afgan et al., 2018).

Materials and Methods

Spectral Processing

Three main wrapper functions have been implemented for metabolomics data processing based on XCMS (version 3.4.4) and CAMERA (version 1.38.1) (Benton, Want, & Ebbels, 2010; Kuhl, Tautenhahn, Bottcher, Larson, & Neumann, 2011; Tautenhahn, Boettcher, & Neumann, 2008) including: (*i*) the *ImportRawMSData* function for reading in raw data files, (*ii*) the *PerformPeakProfiling* function for peak picking and alignment, and (*iii*) the *PerformPeakAnnotation* function for peak annotation. These functions are described below in further detail.

The *ImportRawMSData* function reads in raw MS data files and saves it as an *OnDiskMSnExp* object. To avoid potential memory issues on a user's desktop/laptop, the function will limit the number of cores used to half of the available number of cores. The function outputs two plots - the Total Ion Chromatogram (TIC) which provides an overview of the entire spectra, and the Base Peak Chromatogram (BPC) which is a cleaner profile of the spectra based on the most abundant signals. These plots are useful to inform the setting of parameters downstream. For users who wish to view a peak of interest, an Extracted Ion Chromatogram (EIC) can be generated using the *PlotEIC* function.

The *PerformPeakProfiling* function is a wrapper of several XCMS R functions that performs peak detection, alignment, and grouping in a single step. The resulting peaks are outputted as a *XCMSnExp* object. The function also generates two diagnostic plots including a

retention time adjustment map, and a PCA plot showing the overall sample clustering prior to data cleaning and statistical analysis. Users can specify several parameters such as the mass accuracy, peak width, and retention time range using the *SetPeakParam* function to optimize the peak picking function. A detailed table of suggested parameters for common LC-MS platforms is provided in **Table S2**.

The *PerformPeakAnnotation* function annotates isotope and adduct peaks using the CAMERA package (Kuhl et al., 2011). CAMERA matches m/z features to potential isotopes and adducts based on molecular mass using a dynamic rule set. It does not utilize any structural databases to perform annotation. It outputs the result as a CSV file ("annotated_peaklist.csv") and saves the annotated peaks as an *xsAnnotate* object. Finally, the peak list is formatted to the correct structure for MetaboAnalystR and filtered based upon user's specifications using the *FormatPeakList* function. This function permits the filtering of adducts (i.e. removal of all adducts except for $[M+H]^+$/$[M-H]^-$) and filtering of isotopes (i.e. removal of all isotopes except for monoisotopic peaks). The goal of filtering peaks is to remove degenerative signals and reduce the file size.

Prediction of Pathway Activities

Several metabolic databases are supported at the moment including KEGG (Minoru Kanehisa et al., 2011), BioCyc (P. D. Karp et al., 2017), etc. The main mummichog algorithm is available in the *PerformMummichog* function. Users need to specify a pre-defined cutoff based on either t-statistics or fold changes. The *PerformGSEA* function contains the GSEA implementation based on the high-performance *fgsea* R package (Sergushichev, 2016).

Benchmark Case Studies

The benchmark data created by Li et al. 2018 (Z. Li et al., 2018) is comprised of two standard mixtures (A and B) consisting of 1100 known compounds, with four replicates per mixture. The Google Drive link to the raw data is https://drive.google.com/drive/folders/1PRDIvihGFgkmErp2fWe41UR2Qs2VY_5G?usp=sharing_eip&ts=5b8ab35f. For this manuscript, we selected the dataset that was generated from a Q Exactive HF mass spectrometry (Thermo Fisher Scientific) in positive ion mode, coupled with a Dionex UltiMate 3000 HPLC equipped with a ZORBAX Eclipse Plus C18 column (Agilent Technologies). Parameters for our workflow were selected based on the default values provided for HPLC-Q Exactive Orbitrap data on XCMS Online (mass error: 5 ppm and peak width: 10-60 seconds).

The second dataset consists of pediatric IBD stool samples obtained from the Integrative Human Microbiome Project Consortium (iHMP) (Consortium, 2014). The original study included samples longitudinally collected from IBD patients and non-IBD controls over 50 weeks. The link to the raw metabolomics data is https://ibdmdb.org/tunnel/public/summary.html, under the subheadings HMP2, Metabolites, 2017.23. For our evaluation purpose, we collected samples that meet the following criteria for the diseased group: (*i*) age between 6 and 19 and (*ii*) diagnosed as Crohn's disease. Samples obtained at the earliest clinical visit of each patient who met criteria (*i*) and (*ii*) were included in our study. For the healthy control, samples of non-IBD individuals between age 6 and 19 collected during their first and second clinical visits were included. The dataset was generated from a Q-Exactive Plus orbitrap mass spectrometer (Thermo Fisher

Scientific) in negative ion mode, coupled with a Nexera X2-U-HPLC system (Shimadzu Scientific Instruments) equipped with an ACQUITY BEH C18 column (Waters).

All raw data in .RAW format were converted into .mzML format using ProteoWizard 3.0 MSConvert (Holman, Tabb, & Mallick, 2014) with parameters summarized in the supplemental materials (**Table S3**). Following the spectral processing described earlier, data cleaning and statistical analysis of the clinical data was performed on the clinical data using various functions within MetaboAnalystR. Firstly, missing value imputation was performed by replacing them with half of the minimum value found for each feature. Features containing more than 50% missing values across all samples were removed. Features with nearly constant values across samples were also filtered out based on the inter quantile range (IQR), which removed approximately 25% of total features. Subsequently, value of each feature was normalized with the median value of all features per sample to account for variable water content of stool samples. Finally, generalized log-transformation and auto-scaling were applied to data prior to multivariate statistical analysis. For univariate analysis, non-parametric methods (i.e. Mann-Whitney U test and fold change calculation) were applied to untransformed data to avoid false positives due to data manipulation (Di Guida et al., 2016). A minimum fold change >2 and <0.5, and a false discovery rate (FDR) adjusted p-value of 0.05 were used as cut-off values. To infer pathway activities, we applied both mummichog and GSEA to predict pathway activities. The human BiGG and Edinburgh Model (hsa_mfn) library was selected as the knowledge base, with the p-value cutoff set to 0.05 and the instrumentation accuracy set to 5 ppm.

Acknowledgements

This work is supported by Genome Canada and Genome Quebec. J.C. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). J.X. is supported by the Canada Research Chairs (CRC) Program. We gratefully acknowledge the maintainers of XCMS and CAMERA, Drs. Steffen Neumann and Johannes Rainer for their valuable contribution to the metabolomics community.

Author Contributions

Conceptualization, Jianguo Xia; Data curation, Mai Yamamoto; Formal analysis, Jasmine Chong and Mai Yamamoto; Funding acquisition, Jianguo Xia; Methodology, Jasmine Chong, Mai Yamamoto and Jianguo Xia; Supervision, Jianguo Xia; Writing – original draft, Jasmine Chong and Mai Yamamoto; Writing – review & editing, Jianguo Xia.

Conflicts of Interest

The authors declare no conflicts of interests.

Supplemental Information: MetaboAnalystR 2.0

Table S1. Characteristics of pediatric IBD patients and healthy controls included in this study.

	CD	Healthy
n	24	24
Female gender (n)	9	15
Median age, years (range)	14.5 (8-19)	11 (6-17)

Table S2. Suggested peak picking parameters for commonly used LC-MS platforms.

		SetPeakParam()		
Vendor	**Instrument**	**ppm**	**min_pkw**	**max_pkw**
Agilent	HPLC/Q-TOF	30	10	60
Agilent	HPLC/UHD Q-TOF	15	10	60
Agilent	HILIC_HPLC/UHD Q-TOF neg[1]	15	10	120
Bruker	HPLC/Q-TOF neg[1]	10	10	60
Bruker	HPLC/Q-TOF pos[1]	10	5	20
ABSciex	UPLC/TripleTOF	15	5	20
Waters	UPLC/HRMS	15	2	25
Waters	HPLC/TOF	30	10	60
Thermo	UPLC/Q-Exactive	5	5	20
Thermo	HPLC/Orbitrap	3	10	60

[1] neg = negative ion mode, pos = positive ion mode

Parameters are extracted from XCMS Online default settings (PMID 29494574)

Table S3. Parameters used to convert .RAW files to mzML format on ProteoWizard MSConvert.

Filter category	Parameter
Peak picking	Vendor msLevel = 1 -
Threshold Peak Filter	Absolute 1000 most-intense
Subset	msLevel 1 – 1

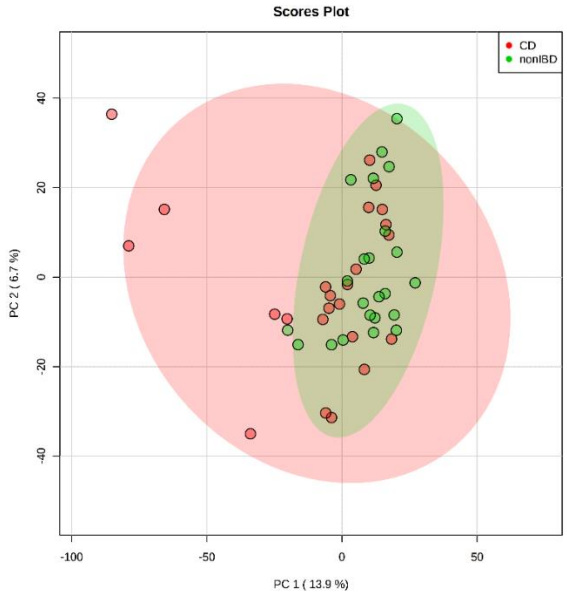


Figure S1. PCA plot of pediatric IBD stool metabolome. Data including 4113 features were median-normalized, log-transformed, and auto-scaled.

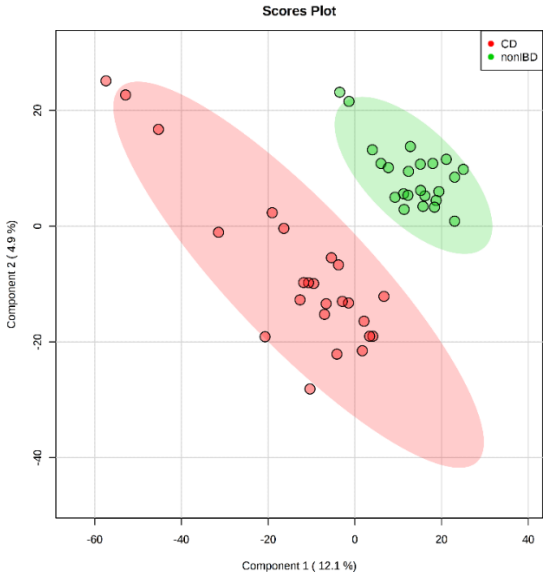


Figure S2. PLS-DA plot of pediatric IBD stool metabolome. Data including 4113 features were median-normalized, log-transformed and auto-scaled.

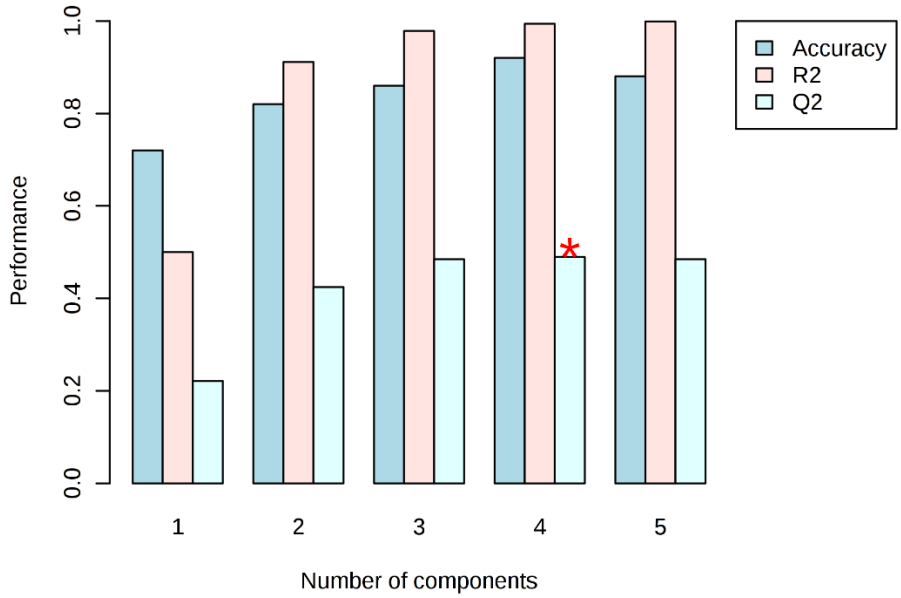


Figure S3. 10-fold cross validation of PLS-DA model (Figure S3) generated from the pediatric IBD stool metabolome data.

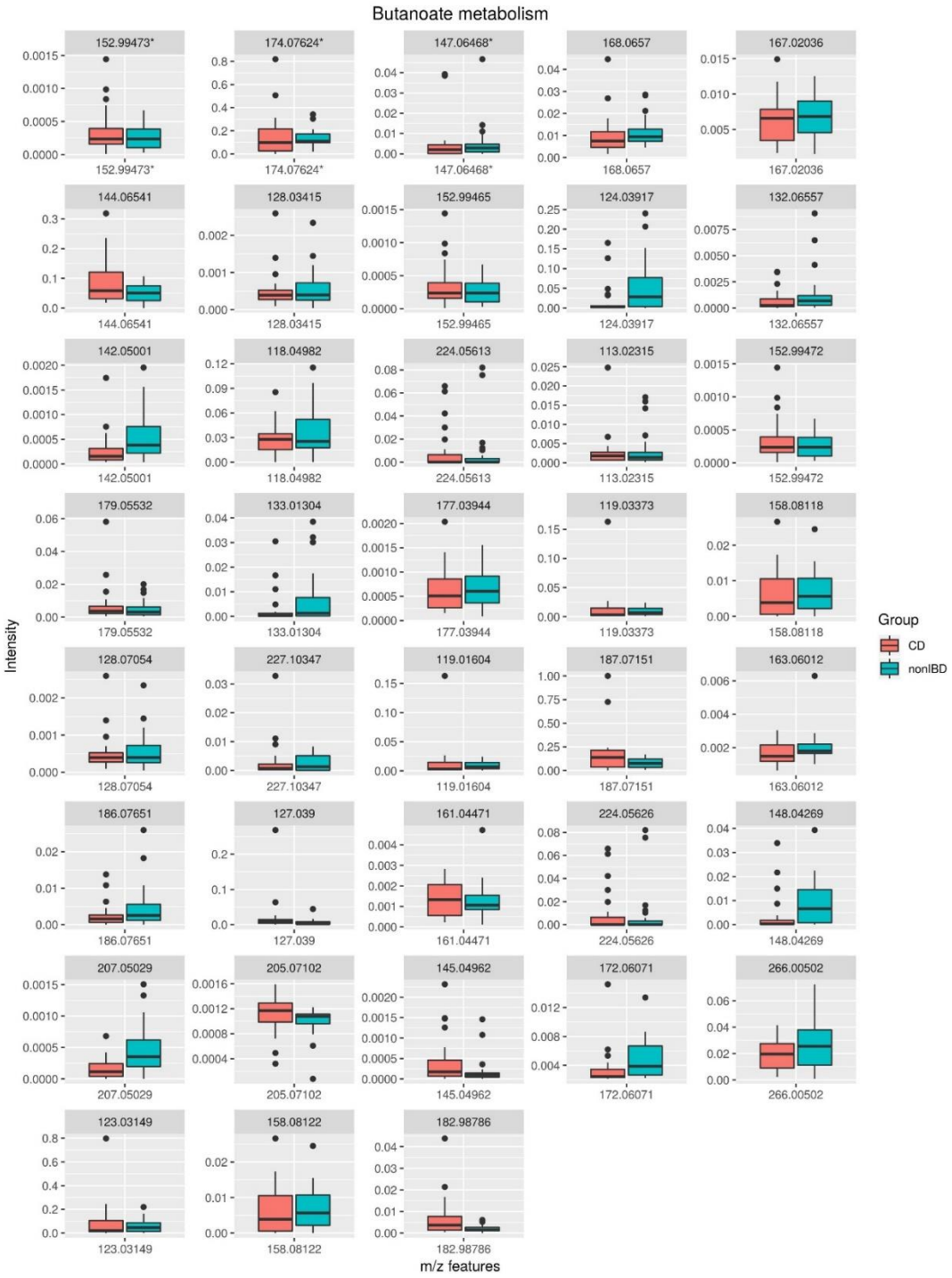
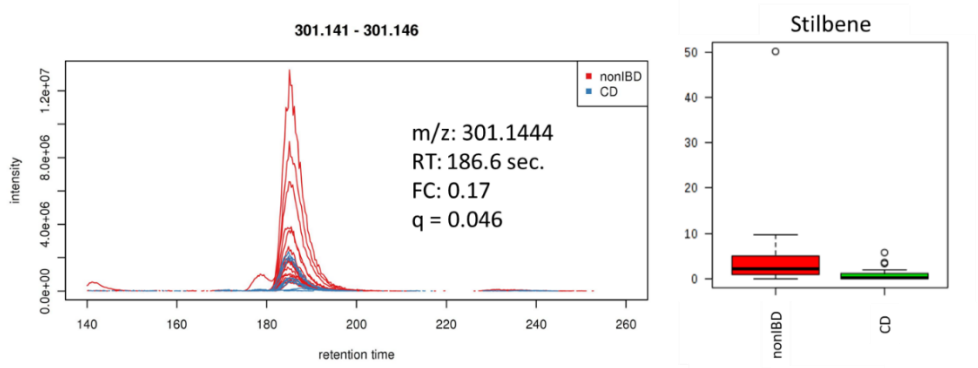
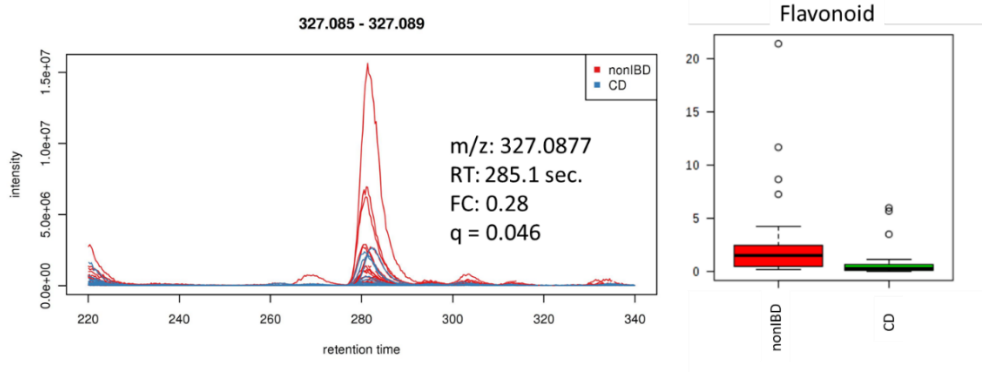
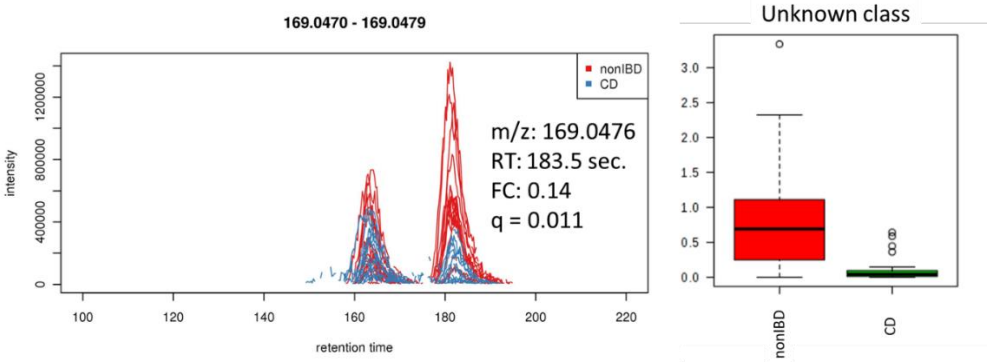
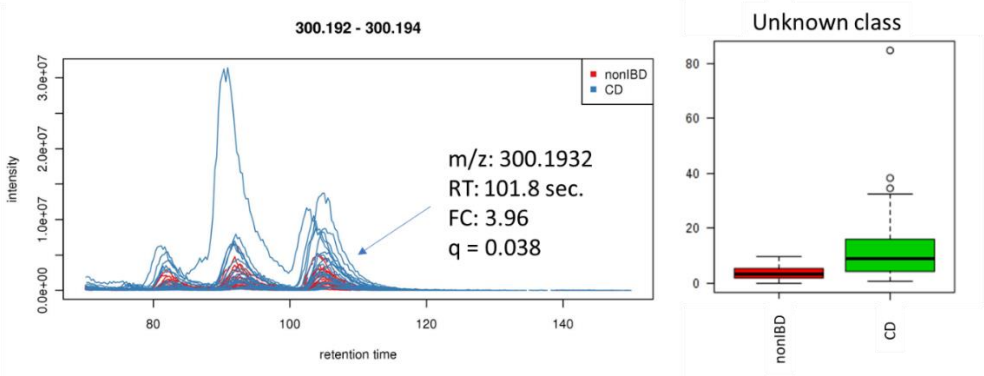


Figure S4. Boxplots of m/z features used for functional interpretation. The m/z features with an asterisk were used by the mummichog algorithm, while all m/z features were used by the GSEA algorithm. The red boxes represent Crohn's disease pediatric patients, and the blue boxes represent healthy patients.


301.141 - 301.146
nonIBD
CD
m/z: 301.1444
RT: 186.6 sec.
FC: 0.17
q = 0.046
intensity
retention time
Stilbene
nonIBD
CD


327.085 - 327.089
nonIBD
CD
m/z: 327.0877
RT: 285.1 sec.
FC: 0.28
q = 0.046
intensity
retention time
Flavonoid
nonIBD
CD


169.0470 - 169.0479
nonIBD
CD
m/z: 169.0476
RT: 183.5 sec.
FC: 0.14
q = 0.011
intensity
retention time
Unknown class
nonIBD
CD


300.192 - 300.194
nonIBD
CD
m/z: 300.1932
RT: 101.8 sec.
FC: 3.96
q = 0.038
intensity
retention time
Unknown class
nonIBD
CD

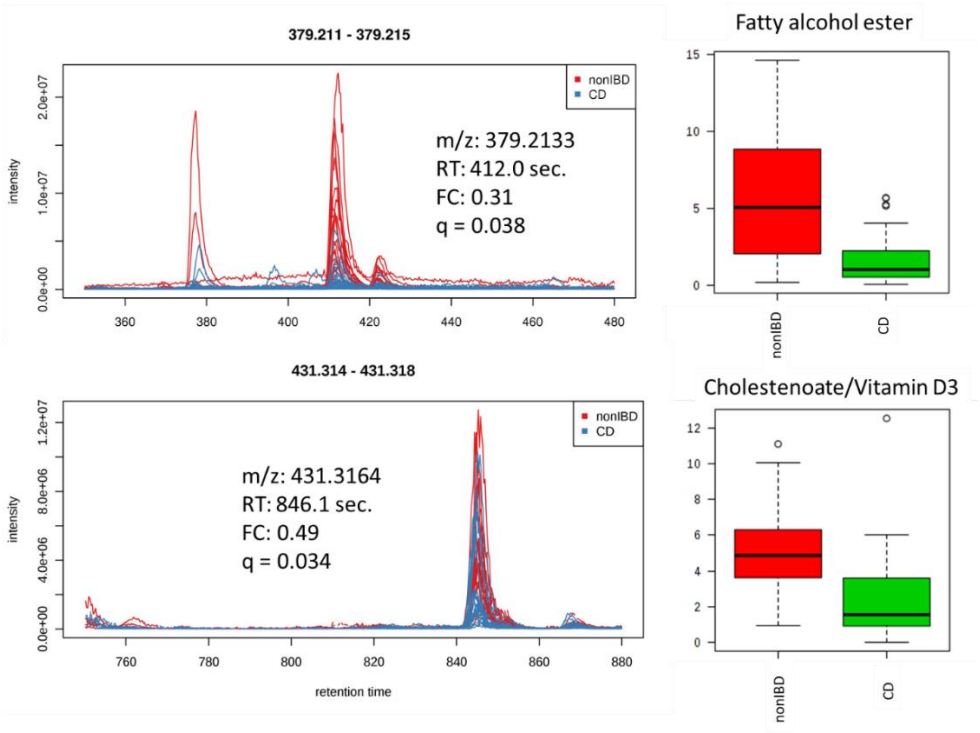


Figure S5. Representative EICs and boxplots of compounds differentially excreted in stool samples of healthy children and pediatric CD patients based on pathway analysis and Mann-Whitney U test (FDR adjusted p-value < 0.05). The m/z of all compounds highlighted above exactly matches with the m/z of compounds detected in the previously published study on adult IBD patients (PMID: 30531976). Putative IDs of each compound assigned in this study are shown above each boxplot.

Preamble to Chapter 4

MicrobiomeAnalyst was the first project I worked on upon starting my PhD and is a web-tool for comprehensive analysis of the microbiome. This work, as well as numerous collaborations with microbiome researchers, had provided me with a solid foundation in the current state of microbiome research. Using this knowledge, I further built upon this platform, streamlining the underlying R code, updating the internal knowledgebases, enhancing visualizations, adding algorithms, and creating a companion R package which supports processing 16S rRNA sequencing data. These updates were published as a well-documented protocol (Chong, Liu, Zhou, & Xia, 2020).

Following my work in the fields of metabolomics and microbiomics, my goal was to marry these two worlds (Chong & Xia, 2017). From metabolomics I understood that metabolites reflect not just one's environment and genetics, but also the interactions between a host and its microbiota. With microbiomics, I consistently saw that while the goal of most studies was to attribute the microbiota to a disease, they would fall short at just characterizing a list of bacteria associated with disease. However, why were these microbes associated with that phenotype? From an evolutionary perspective, an organism can be found in an environment where it can thrive. Therefore, a bacteria's metabolism must play a significant role in their presence. Connecting these thoughts together, I created the microMum workflow, which will be presented in Chapter 4.

Chapter 4. Enhancing Biological Insights from Paired Microbiome and Metabolomics Data

microMum: Enhancing biological insights from paired microbial and untargeted metabolomics data

[1]Jasmine Chong, [1]Yao Lu, [1]Tanisha Shiri and [1,2,3,4]Jianguo Xia

[1]Institute of Parasitology, and [2]Department of Animal Science, McGill University, Ste-Anne-de-Bellevue, Quebec, Canada

[3]Department of Microbiology & Immunology, Montreal, Quebec, Canada

[4]Department of Human Genetics, Montreal, Quebec, Canada

.

Abstract

Metabolism is fundamental to how an organism interacts with its community and its surroundings. Within the gut, the host and its resident microbiota have evolved over millennia to maintain homeostasis. Therefore, deciphering host-microbe and microbe-microbe interactions at the metabolic level is crucial to understanding the gut microbiome. Metabolomics is considered the most promising 'omic technology to reveal deep mechanistic insights into the gut microbiome by providing a snapshot of host-microbial co-metabolism. However, the progress of integrating microbial taxonomic and metabolomics profiles is delayed due to difficulties in annotating mass spectrometry data into meaningful compounds. A method that bypasses this bottleneck, Mummichog, can be used to directly infer biological insights from untargeted metabolomics data, but is hindered by its inability to consider all products of host-microbe and microbe-microbe interactions within the gut. Previous Mummichog implementations use a single organism's metabolic network to define the metabolic space from which peak annotations are obtained. Instead, we create community-scale metabolic networks (CMNs), combining both human and microbial metabolic networks, tailored specifically to a provided microbial taxonomic profile. CMNs are then leveraged to contextualize microbiome-specific metabolomics data and model community-wide metabolism. Using this method, we show how CMNs improve biological insights obtained from a multi-omic investigation of Inflammatory Bowel Disease as compared to previous Mummichog implementations. Overall, our improved algorithm is a powerful tool that can be added to the microbiome investigators toolbox to integrate multi-omics microbiome data.

Background

The human gut microbiome plays a vital role in shaping human health, performing several roles otherwise inaccessible to its host including digestion, synthesis of vitamins, stimulation of proper immune system development, and protection against opportunistic pathogens (Belkaid & Hand, 2014; Clarke et al., 2014; Heintz-Buschart & Wilmes, 2018). Humans and their resident gut microbiota have evolved together over time and maintain a symbiotic relationship (Backhed et al., 2012). Disruptions to this relationship, known as dysbiosis (Hooks & O'Malley, 2017), has been attributed to several diseases including immune-mediated inflammatory diseases such as Crohn's Disease (CD), Multiple Sclerosis and Rheumatoid arthritis (Forbes, Van Domselaar, & Bernstein, 2016), as well as others including allergies, asthma, and cardiovascular disease (CVD) (Carding et al., 2015).

Advances in high-throughput 16S rRNA sequencing technologies have revolutionized the ease and speed to which taxonomic profiles can be obtained from microbial communities (Bharti & Grimm, 2021; Jovel et al., 2016; M. G. I. Langille, 2018). These have led to our current understanding of the vast taxonomic array of microbes across different biological systems and their influences on host phenotypes. However, while taxonomic profiles do provide knowledge about which microbes are present, why certain microbes associated with disease or healthy states is not well understood (Doolittle & Booth, 2017; M. G. I. Langille, 2018). Intuitively, the rationale for a microorganism's presence in a biological system lies not with *who* they are but what they *do*, with such functions having downstream effects on their host (Inkpen et al., 2017; Klassen, 2018; Krause et al., 2014). Gaining a better understanding of microbial function is

essential for research to move beyond characterizing lists of microbes towards actionable insights.

Metabolic function is believed to underlie taxonomic variations in the microbiome (Louca et al., 2016). Gut microbes contribute to the production of biologically active small molecules, termed microbial-derived metabolites, that could enhance host fitness (Lavelle & Sokol, 2020; W.-J. Lee & Hase, 2014; Nicholson et al., 2012). These metabolites serve as signaling molecules, mediating vital host-microbial interactions through a dynamic crosstalk involving numerous molecular pathways. These small molecule products greatly impact biological processes with important consequences to human health including digestion, immune system development and regulation, inflammation, and neurodevelopment (Hsiao et al., 2013; Levy et al., 2016; Sharon et al., 2014). A classic example of a gut-microbial derived metabolite with negative consequences for its host is trimethylamine N-oxide, a derivative of dietary L-carnitine and choline, which are found in red meat, that has been systematically linked to CVD (Koeth et al., 2013; Z. Wang et al., 2011), colorectal cancer (CRC) (Xu, Wang, & Li, 2015), and non-alcoholic fatty liver disease (NAFLD) (Tremaroli & Bäckhed, 2012). Another notable group of metabolites are the short chain fatty acids (SCFAs) acetate, propionate, and butyrate, which are metabolized by gut microbes from dietary fibers and starches. These metabolites confer several beneficial effects, such as maintaining intestinal homeostasis, providing protection against inflammation, and maintaining the blood brain barrier integrity (Braniste et al., 2014; Dalile, Van Oudenhove, Vervliet, & Verbeke, 2019; Parada Venegas et al., 2019; Silva, Bernardi, & Frozza, 2020; van der Hee & Wells, 2021). SCFAs are known to be decreased in patients with Inflammatory Bowel Disease (IBD) (Parada Venegas et al., 2019), provide

protection against diabetes (Mariño et al., 2017), and mediate microbial-gut–brain axis crosstalk (Dalile et al., 2019; Silva et al., 2020). Without a doubt, microbial-derived metabolites can have either beneficial or devastating consequences for the host.

Perturbations in cellular processes are rapid and leave metabolic fingerprints, representing the physiological state of the host and its resident microbiota (Gilbert et al., 2016; Zierer et al., 2018). Metabolomics, which leverages mass-spectrometry technologies, can be used to obtain a comprehensive profile of metabolites within a biological system. Metabolomics can either be targeted, which accurately quantifies a pre-determined set of metabolites, or untargeted, which aims to profile all metabolites present within a sample without prior knowledge. Because human fecal samples are incredibly complex, containing the products of metabolism of food, medication, and even one's environment, untargeted metabolomics is becoming routinely used to explore all molecules within the human gut (Melnik et al., 2017). An important challenge to interpreting untargeted metabolomics data is the sheer amount of data as well as the difficulty to properly annotate each peak (Bauermeister, Mannochio-Russo, Costa-Lotufo, Jarmusch, & Dorrestein, 2021). One solution to peak annotation is the Mummichog algorithm (S. Li et al., 2013a), which bypasses the step of metabolite identification to obtain biological insights from high-resolution untargeted metabolomics data. In brief, the method leverages an organism's genome-scale metabolic network (GEM) to define the metabolic space from which to pull metabolite annotations. It uses a peak's mass-to-charge ratio and matches it to a compound based on its similarity to the compound's molecular weight. From here, the putatively annotated peaks are used as input for pathway enrichment, which summarizes the important peaks into

interpretable biological pathways. However, this method is limited to a single GEM, which restricts the metabolic universe from which metabolite annotations are obtained.

GEMs are comprehensive metabolic reconstructions of an organism, containing the entire set of metabolic reactions that can occur within said organism. Traditionally, GEMs have been used for constraint-based modeling of the microbiome, with their applications ranging from predicting gene expression, metabolic engineering to enhance the production of target chemicals, drug design, and modeling community dynamics (Chowdhury & Fong, 2020; Gu, Kim, Kim, Kim, & Lee, 2019; Magnúsdóttir & Thiele, 2018; Oberhardt, Palsson, & Papin, 2009). High-quality GEMs exist for >6000 organisms, the majority of which are bacteria species (Gu et al., 2019). As GEMs are designed to be accurate representations of an organism's metabolism, a natural application would be to directly combine these models into a community metabolic network (CMN). Moreover, current work in prediction of microbiome function focusses solely on the contributions of the gut microbiota, losing important interactions between the host and its native gut microbiome (Douglas et al., 2020; Noecker et al., 2016; Wemheuer et al., 2020). Metabolites produced by host cells can be used or transformed by gut microbes, and vice-versa. We therefore posit that including host functional information will improve the predictive potential of an integrated metabolic network.

Here, we introduce microMum, a method for knowledge-based integration of microbial taxonomic and metabolomics profiles. This method enables àla carte creation of CMNs from a provided taxonomic microbial profile which is later used as scaffolding for the metabolomics data. This microbiome-focused analytical pipeline for untargeted metabolomics data enables fast and more accurate biological insights into the human gut microbiome than its predecessors. In

addition, this method associates metabolic changes with specific microbiota and/or the human host, providing jumping off points for further investigations of important host-microbiota interactions.

Methods

The microMum workflow is illustrated in **Figure 6** and requires a multi-omic dataset consisting of a taxonomic microbial profile and an untargeted metabolomics dataset from the same set of samples. As a first step, the provided taxonomic microbial profile is matched to the internal microMum library of GEMs. Matched GEMs, consisting of both the microbial and host GEMs, are directly merged to create a community metabolic network. This method assumes that all metabolites are shared (i.e. excreted across cell membranes). Next, the untargeted metabolomics data is processed to identify differentially abundant peaks between the user's groups of interest (e.g. cases versus controls). Differentially abundant peaks are then used as input for putative compound annotation, using all metabolites from the community metabolic network as potential matches. Putatively annotated compounds are then overlayed onto metabolic pathways for functional enrichment. These steps are further detailed in the methods below. All results and visualizations can be directly used for publication purposes. All core functions are written in R (V4.0.2) and visualizations were created using the ggplot (Wickham, Chang, & Wickham, 2016) and plotly R packages (https://plot.ly).

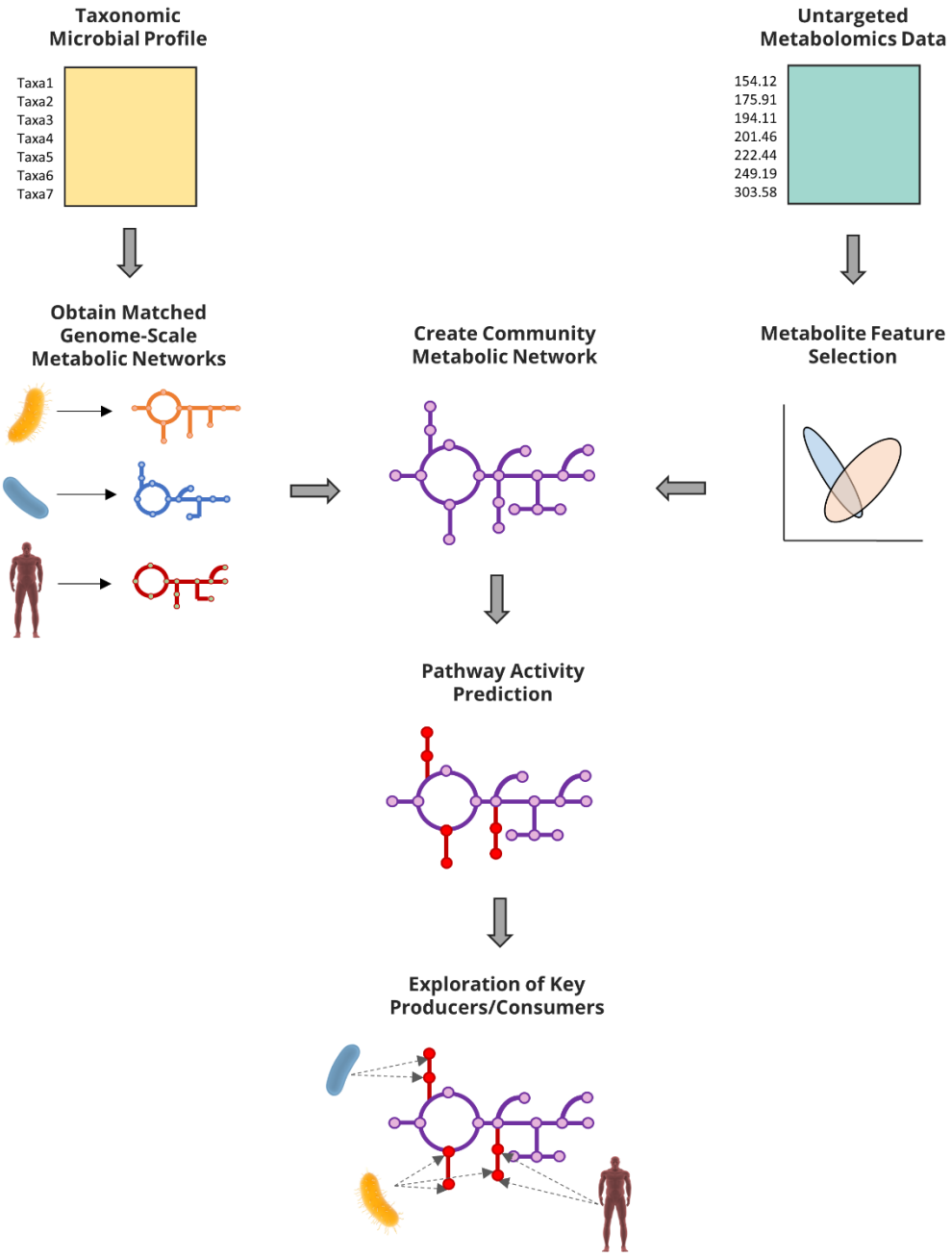


Figure 6. microMum workflow. The first step begins with matching the user's taxonomic microbial signature to the internal database of genome-scale metabolic models (GEMs). The matched GEMs are then combined to create a unique community metabolic model. Meanwhile, feature selection is performed on the untargeted metabolomics data. Differentially abundant peaks are then used as input for putative compound annotation. Putatively annotated compounds are overlayed onto the community metabolic network and pathway enrichment is performed.

Further exploration of pathway activity results can be used to identify key organisms who contribute to enriched metabolic functions.

Genome-scale metabolic models

Two high-quality and independent resources of microbial metabolic reconstructions (Gu et al., 2019; Mendoza, Olivier, Molenaar, & Teusink, 2019), AGORA (Magnusdottir et al., 2017) and CarveMe (Machado, Andrejev, Tramontano, & Patil, 2018), are incorporated into microMum. Briefly, AGORA models are built using a bottom-up approach, where they are first assembled using genome annotations of the organism of interest and then further manually refined using experimental data and gap-filling algorithms. In comparison, CarveMe models are built using a top-down approach, whereby a universal metabolic model is constructed and from which organism-specific models are "carved". To investigate the metabolome coverage by each GEM resource, enrichment analysis using chemical class ontologies was performed (Fahy & Subramaniam, 2020; Pang et al., 2021) (**Supplementary Materials – Metabolome Coverage**). Overall, each resource covered different chemical classes so both databases were kept ensuring that comprehensive metabolomic insights could be inferred. As for host GEMs, the most up to date human metabolic model, Human1, is also integrated into microMum (Robinson et al., 2020). Metabolic annotations for all GEM models were manually enhanced using HMDB (Wishart et al., 2017a), PubChem (S. Kim et al., 2016), and BIGG (King et al., 2016). Finally, GEMs were converted into matrices and stored in an SQLite database.

Matching to genome-scale metabolic models

Dependent on the type of taxonomic microbial features, the internal matching function, *match2gems*, will be employed. For signatures that are species/strain names, NCBI Taxonomy IDs, Greengenes taxonomy/OTU IDs, an exact match search against the microMum GEM database will be performed. For SILVA taxonomy, a helper function was implemented to improve matching results by synchronizing taxonomy names in the user's data such as by replacing "p__Bacteroidetes" to "p__Bacteroidota" or cleaning semi-colons from the user's inputs. Finally for ASVs, sequence variants are converted to individual fasta files that are then used as input to vsearch (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) to identify matches to GEMs. The 16S rRNA sequence database used as reference for AGORA and CarveMe models were obtained from the Borenstein lab (https://borenstein-lab.github.io/MIMOSA2shiny/downloads.html).

Creation of a community metabolic network

Following GEM matching, a community metabolic network will be created using the *CreateCMN* function. The simplest method to create such a network is to treat the microbiome as an "enzymatic soup" and ignore species-specific boundaries. In this sense, CMNs are created by directly merging host and gut microbial GEMs. Here, all metabolites present within each GEM (0 = not present, 1 = present) is pulled from the SQLite database and merged to create a single table of metabolite names as columns and their presence in each matched GEM as rows. This

merged table, serving as the community metabolic network, will be used as the compound universe from which metabolite annotations will be pulled from.

Available pathway libraries

A recently published paper has demonstrated that the choice of pathway database used in enrichment analysis can have a greater effect on enrichment results than statistical corrections used in these analyses (Peter D Karp, Midford, Caspi, & Khodursky, 2021). We therefore have decided to support two popular metabolic pathway databases, KEGG (Minoru Kanehisa, Furumichi, Sato, Ishiguro-Watanabe, & Tanabe, 2021; Minoru Kanehisa & Goto, 2000; M. Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012) and MetaCyc (Ron Caspi et al., 2020). The KEGG resource was initiated in 1995 and now consists of 544 reference pathways with over 18000 compounds. In comparison, MetaCyc consists of 2749 base pathways with over 15000 compounds. Using the KEGGREST R package (Tenenbaum, 2021), we obtained 164 generic metabolic pathways containing 5917 compounds. As the number of MetaCyc pathways were much greater than in KEGG, we focused on pathways found in only bacteria and humans. Using the MetaCyc SmartTables tool, we thus obtained 1510 metabolic pathways belonging to bacteria and 350 belonging to humans. Compound annotations from both databases were enhanced using the internal compound database of MetaboAnalyst (Pang et al., 2021), and valid PubChem identifiers (CIDs) were first enhanced using WebChem R package and manual searching of ChEBI, HMDB, KEGG and PubChem.

Predicted pathway activity

Following community metabolic network creation, the final step is to perform functional enrichment of the network using the *PerformMicroPSEA* function. The functional enrichment answers the question of which metabolic pathways are differentially active between cases and controls. Specifically, this function dynamically builds an internal library of compounds based on the provided CMN, including all potential adducts and a mass to compound dictionary for fast look-up. The function can also perform the original mummichog, GSEA or both methods for functional enrichment (Chapter 3).

Various scenarios were investigated to evaluate whether the community metabolic network improves biological insights. First, to evaluate the composition of the CMN, we investigated whether using all microbes from a taxonomic profile versus only significantly enriched microbes between cases and controls would have any impact on the results. Second, to evaluate the value of the integrated CMN, we compared results obtained using only the human GEM, a CMN consisting only of all microbes, and the combined host and microbial CMN. Third, a robust permutation method was implemented to investigate whether the list of microbes used to create the CMN truly influences the results. This was achieved by randomly permuting the selection of GEMs and subsequently the predicted metabolite contents of the CMN, thereby removing the association between the taxonomic profile and disease phenotype. To accomplish this, functional enrichment was first performed using the provided list of microbes. Then, using a randomly permuted list of microbes the same length as the total number of matched GEMs from the user's data, enrichment analysis results are obtained using the new randomly permuted CMNs. By default, the permutation was set to 1000 rounds. The empirical p-value is calculated

as the number of times the functional enrichment results using the permutations were more significant than using the original data.

Data visualization

We have added two functions to visualize the enrichment results. The first is the *microMumLollipopChart* function, which creates a lollipop chart that summarizes the results of the pathway enrichment. It includes a color-coding system to easily interpret KEGG of MetaCyc pathways into their higher categories. It can be used to plot the results of a single enrichment result, but also to compare enrichment results of the same CMN using either AGORA or CarveMe models. By plotting the comparison of the GEM model databases, one can easily see which pathways are consistently enriched across both databases and the rank of all enriched pathways according to either database. The second function is *microMumBubbleChart*, which outputs a bubble chart that is dynamically created based on a selected pathway. This plot shows all matched GEMs and the presence of the pathway metabolites across the organisms. The intention behind this plot is to permit visual exploration of the enrichment results and identify important organisms that contribute significantly to a certain function. For instance, one could see that organism X uniquely contributes 5 metabolites to an enriched pathway, suggesting it could play a key role in disease pathogenesis.

Case Study

To demonstrate the utility of microMum, we used a large-scale investigation of Crohn's Disease in the United States of America (Franzosa et al., 2019). Untargeted metabolomics data

(C8 pos and C18 neg) was obtained from the Integrative Human Microbiome Project (https://www.hmpdacc.org/ihmp/) and processed using the MetaboAnalystR workflow as described in Chapter 3. Species-level metagenomics data was directly obtained from the supplementary material included with the manuscript (e.g. Supplementary Dataset 4: Per-subject microbial species relative abundance profiles). From the original manuscript, the taxonomic classifications were obtained using MetaPhlAn2 version 2.2.0 (Truong et al., 2015). The data was further analyzed on MicrobiomeAnalyst (Chong et al., 2020) using the LEfSe module to identify discriminatory microbes of CD from healthy controls. Following this, the list of species names (all or discriminatory only) were used as input to microMum and matched to GEMs using the species/strain names option of the internal matching algorithm.

Results

As a very first step to evaluate the benefit of creating a community metabolic network, we checked the number of metabolites that are shared between the human GEM and all microbial GEMs. In total, there are only 584 metabolites shared between human and microbial GEMs. There are 860 metabolites uniquely of microbial origin and 913 metabolites found only within the human. Of note, these numbers are current estimates from our library of GEMS as the origin of metabolites from microbes may not have yet been discovered. Nonetheless, these numbers confirmed our motivation to create an integrated community metabolic network that encompasses all potential metabolites for functional enrichment. Next, to illustrate the

microMum workflow, we showcase an investigation of Crohn's Disease that is available as an example from the R package.

Crohn's Disease Case Study

The taxonomic signature of Crohn's Disease in adult Americans was obtained from a recently published large-scale multi-omics study (Franzosa et al., 2019). The aim of this study was to understand the gut-metabolome mediated interactions between a host and its gastrointestinal microbiota. Initially, the entire taxonomic signature was used as input to the microMum workflow, consisting of 119 microbes. Of these, 86 had corresponding GEM matches to the AGORA database. This set of GEMs were used for the section below.

Comparison of human-only, microbial-only and community metabolic networks

The validity of findings obtained using the Mummichog algorithm predicates largely on the selected metabolic model used to putatively annotate MS peaks. Therefore, we first evaluated whether there were differences in functional enrichment results when using either (i) only the human GEM, (ii) a community metabolic network consisting of only matched microbial GEMs, or (iii) the combined human and microbial community metabolic network as the underlying GEM model for peak annotation. For the human-only analysis, 13 pathways had a Gamma-adjusted p-value < 0.05 (**Supplementary Table 8**). Using the microbe-only network, 21 pathways had a Gamma-adjusted p-value < 0.05 (**Supplementary Table 9**). For the combined CMN, 15 pathways had a Gamma-adjusted p-value < 0.05 (**Supplementary Table 10**). Steroid biosynthesis, Steroid degradation, Linoleic acid metabolism, alpha-Linolenic acid metabolism,

Ether lipid metabolism, and Sphingolipid metabolism were uniquely enriched in the human-only and CMN results. Meanwhile, Folate biosynthesis and Ubiquinone and other terpenoid-quinone biosynthesis were unique to the microbe-only and CMN results. This motivates the use of a combined community metabolic network that captures different functional potential from the untargeted metabolomics data versus human-only and microbe-only networks.

Differences between using all or significantly enriched microbes for community metabolic network creation

Once the motivation to use a combined community metabolic network was clear, we next evaluated whether all microbes (allCMN) or only differentially abundant microbes (deCMN) should be used for GEM matching. Identifying differentially abundant microbes is the first step towards understanding how certain microbial taxa are associated with various phenotypes (Lin & Peddada, 2020). If the metabolism of a microbe plays a large role as to why it is linked to a disease, we posit that the metabolic changes between phenotypes can be captured solely using differentially abundant microbes. Therefore, of the 38 differentially abundant taxa between Crohn's disease and healthy controls as identified using LEfSe (Segata et al., 2011), 29 had matches to the AGORA GEM database and were used to build the CMN. The results of the functional enrichment analysis using the KEGG pathway library can be found in **Supplementary Table 11** and visualized as a lollipop chart in **Supplementary Figure 4**. Overall, the number and rank of enriched metabolic pathways using the allCMN and deCMN were identical. Moreover, the differences in Gamma-adjusted p-values were inconsequential (e.g. for Steroid biosynthesis, a p-value of 0.0219 in allCMN and 0.0212 in deCMN). Together, this shows that using only

differentially abundant microbes to create the CMN can capture the same metabolic variation as using all available microbes.

Contrasting results obtained using AGORA and CarveMe metabolic models

As shown in the Supplementary Materials, the metabolome coverage between AGORA and CarveMe were different enough to warrant keeping both sets of GEM models as options. We thus compared the pathway activity prediction using either the AGORA or CarveMe models. Of the 38 differentially abundant microbes, 27 had matches to the CarveMe GEM models. The same 15 significantly enriched metabolic pathways were identified when using the CMN made from CarveMe models as compared to the AGORA models (**Supplementary Table 12**). There were minor changes to the rank of the pathways, owing to different numbers in compound hits across the pathways. The comparison of the pathway activity results is also visualized as a lollipop chart in **Figure 7**. From this plot, predicted pathway activity is consistent using either AGORA or CarveMe for this Crohn's Disease use-case.

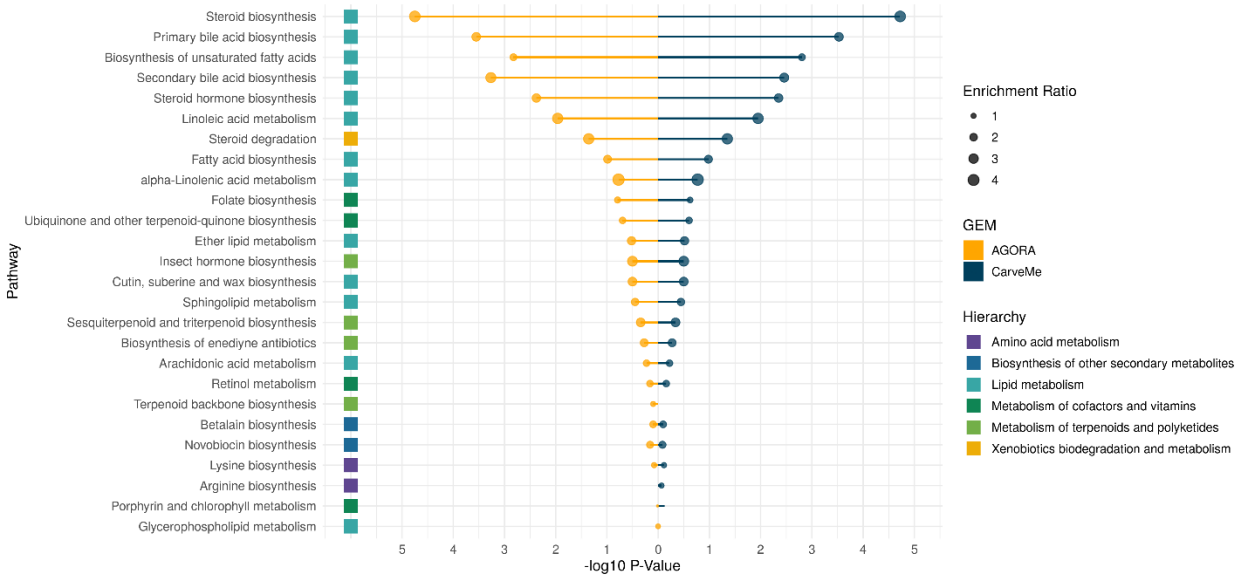


Figure 7. Lollipop chart of altered metabolic pathways using microMum. The Lollipop chart shows the enrichment of KEGG metabolic pathways in AGORA (yellow lines) and CarveMe (blue lines) matched microbes implicated in Crohn's Disease. The x-axis is the -log 10 p-value (scaled from 0-5), and the y-axis are the pathways. The square colored boxes on the left of the plot shows the metabolic hierarchy of that pathway to aid interpretation. The length of the lines corresponds to the -log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits). Lines with an asterisk at the end are pathways that were also observed to be significantly altered using real metabolomics data.

Comparison with previous Mummichog analyses

Next, we compared functional enrichment results obtained using microMum and its preceding Mummichog implementation (ogMum) to determine whether microMum provided greater insights. Notably, ogMum permits the selection of a single GEM for peak annotation, with only 2 bacterial species known to be found in the human gut (*Escherichia coli K-12 MG1655* and *Bacillus subtilis*) as valid microbial options. For ogMum, we therefore performed the analysis three times using one of three GEMs; *hsa_mfn* (a manually curated human GEM from the original implementation), *hsa_kegg* (the KEGG human metabolic model), and *eco_kegg*

(the KEGG *Escherichia coli K-12 MG1655* metabolic model). Comparing between the microMum (15 pathways < Gamma-adjusted p-value 0.05) and ogMum *hsa_mfn* analysis (13 pathways < Gamma-adjusted p-value 0.05) (**Supplementary Table 13**), both identified alterations in bile acid and fatty acid metabolisms, as well as Vitamin E (microMum: Ubiquinone and other terpenoid-quinone biosynthesis) and Vitamin D3 (microMum: Steroid biosynthesis). ogMum *hsa_mfn* uniquely identified changes in Carnitine shuttle, Vitamin B6 and Glycerophospholipid metabolism. Meanwhile microMum uniquely pinpointed changes to Folate and Sphingolipid metabolism that were not captured using the previous implementation. Meanwhile, the ogMum *hsa_kegg* analysis had zero pathways with a Gamma-adjusted p-value < 0.05 (**Supplementary Table 14**). Finally, while the *eco_kegg* analysis also did not have any pathways with a Gamma-adjusted p-value < 0.05, there were 10 pathways when the p-value threshold was increased to 0.1, including those related to amino acids, sugar, biotin, and porphyrin metabolisms (**Supplementary Table 15**). Importantly, alterations in bile and fatty acids were not captured using the ogMum KEGG analyses.

Permutation of community metabolic networks

Finally, we evaluated whether the list of microbes used to create the CMN has a significant impact on the predicted pathway activity results. To achieve this, GEM models were randomly selected to build a CMN, the same number as the length of matched microbes from the input. The microMum algorithm was then performed using the permuted CMN and repeated 1000 times. An empirical p-value was calculated as the number of times a pathway was more significant using the permuted CMNs versus the original CMN. The top 11 results of the permutation can be found in **Table 4**. Interestingly, Tyrosine metabolism was the only metabolic

pathway with a p-value < 0.05 (Empirical p-value 0.022). All other metabolic pathways had Empirical p-values greater than 0.05, with 7 pathways approaching statistical significance (Empirical P-Value < 0.15). This suggests that for some pathways, the taxonomic identity of microbial GEMs does not influence the interpretation of the predicted pathway activity results. The full table of results can also be found in **Supplementary Table 16**.

Table 4. Top 11 altered metabolic pathways using microMum. The top 11 predicted pathway activity of untargeted metabolomics data showing the results of the community metabolic model (CMN) permutation.

KEGG Pathway	Hits	FET	Gamma	Empirical Hits	Empirical P-Value
Tyrosine metabolism	47	0.79	0.235	22	0.022
Secondary bile acid biosynthesis	17	8.4E-07	0.0212	119	0.119
Steroid biosynthesis	25	1.7E-09	0.0212	120	0.12
Primary bile acid biosynthesis	25	2.6E-07	0.0212	120	0.12
Biosynthesis of unsaturated fatty acids	22	5.3E-06	0.0212	120	0.12
Steroid hormone biosynthesis	53	3.4E-05	0.0212	120	0.12
Linoleic acid metabolism	12	1.9E-04	0.0213	121	0.121
Steroid degradation	5	2.4E-03	0.0227	131	0.131
Fatty acid biosynthesis	8	1.1E-02	0.0241	168	0.168
Folate biosynthesis	16	2.5E-02	0.0249	249	0.249
Ubiquinone and other terpenoid-quinone biosynthesis	22	3.7E-02	0.0258	479	0.479

Note: The Gamma P-Value is from the original CMN model, while the Empirical Hits represents the number of times the pathway activity results using the randomly permuted CMNs was better than the original. The Empirical P-Value represents the Number of Empirical Hits divided by the number of permutations (1000 permutations).

Visual exploration of microMum results

Another important plot to help visualize the results is the bubble chart. The intent of this plot was to explore the metabolic contributions of the microbiome and human host to each enriched pathway. **Figure 8** depicts the presence of each metabolite belonging to the secondary

bile acid biosynthesis pathway, which was significantly enriched between Crohn's disease patients and healthy controls. From this plot, *Ruminococcus gnavus* is an important contributor to this pathway, providing Isochendeoxycholic, Ursocholic and Ursodeoxycholic acids that were not present from the human host. This plot, which can be dynamically created for each enriched pathway, aides the comprehension of the enrichment results by allowing one to hone down on pathway-specific important organisms and serves as a jumping off point for hypothesis generation.

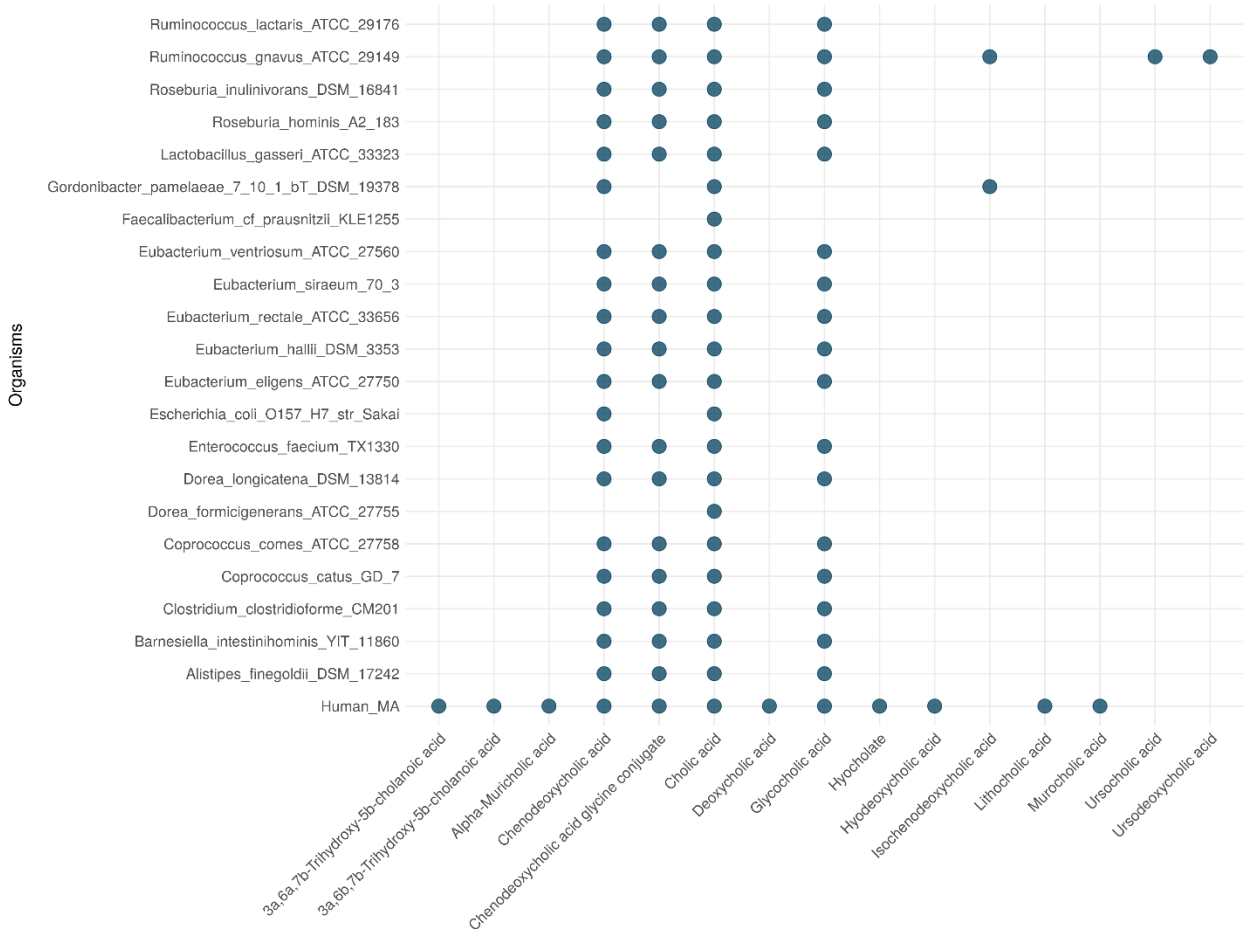


Figure 8. microMum Bubble chart of Secondary Bile Acid Biosynthesis. The bubble chart shows the predicted presence of metabolites belonging to the secondary bile acid biosynthesis

pathway across all AGORA matched microbes implicated in Crohn's Disease and the human genome-scale metabolic network.

Discussion

This work is among the first efforts to integrate microbial sequencing and untargeted metabolomics data using a knowledge-based framework. We began by systematically breaking down the different ways to build the community metabolic networks, ultimately landing on using differentially abundant microbes and human host. The merit of the microMum algorithm was then demonstrated using a large-scale multi-omics investigation of Crohn's Disease. The original investigation identified 8 metabolite classes that were significantly over-abundant in CD, including sphingolipids, bile acids, long chain fatty acids and cholesterols (Franzosa et al., 2019). This was highly consistent with our findings, and alterations in sphingolipids was uniquely identified using microMum as compared to previous Mummichog implementations. Importantly, sphingolipids are mediators of inflammation, thought to be required for intestinal homeostasis, and have become a novel therapeutic target for IBD treatment (Brown et al., 2019; Sukocheva, Lukina, McGowan, & Bishayee, 2020).

Within the case study, we performed a robust permutation to create CMNs, randomly selecting GEMs from our internal database (the same number of microbes used to create the original CMN). The intent was to evaluate whether the taxonomic identity of microbes would impact the interpretation of microMum results. A p-value of 0 would indicate that the taxonomic identity significantly impacts the results, and a p-value of 1 would indicate that the taxonomic identity has no impact on the results. Ranking by the Empirical p-value, Tyrosine metabolism

was the only pathway with a p-value < 0.05, indicating that the set of original microbes did greatly impact pathway prediction for this pathway. Meanwhile, the permutation results for the remaining microbes were not significant, suggesting that the taxonomic identity of microbes had minimal/moderate influence on the interpretation of functional enrichment results.

Non-significant p-values could be attributed to poor power or small effects. As we used 1000 permutations, the sample-size should have been large enough to capture significant results. One possible reason for non-significant results could stem from the GEMs themselves, whereby the metabolic diversity of the microbes was not large enough. From the original publication, the average metabolic distance was 0.48 (Jaccard distance between the list of reactions between microbe pairs) and the largest metabolic distance was 0.78. This is not unexpected, as microbes living within the same environment should be metabolically similar enough to survive within their shared habitat (Mazumdar, Amar, & Segrè, 2013). Additionally, from the Human Microbiome Project, it has been previously reported that there is high metabolic similarity amongst gut microbial taxa, owing to the functional redundancy of core functions such as carbohydrate and amino-acid metabolism (Lozupone et al., 2012). Therefore, metabolic pathways that are common across the gut microbiome should have poor permutation p-values, while pathways that are unique to a few microbes would be greatly impacted by the taxonomic identity of microbes used to create the CMN. In our case-study, alterations in bile acid metabolism were amongst the top-identified metabolic changes in CD as compared to healthy controls but had non-significant permutation p-values (Empirical p-values were ~0.12). However, a publication from the authors of the AGORA resource identified 232 GEMs (from a total of 773 reconstructions) with bile acid reactions (Heinken et al., 2019). Due to the relative

commonality of bile acid metabolites amongst AGORA GEMs, it is not unanticipated that the bile acid pathways had Empirical p-values ~ 0.12.

The benefit of integrating microbial data within the microMum framework is that results can now be attributed to specific microbes, which helps researchers plan future experiments. Previous Mummichog implementations required the selection of a single organism's metabolic database, which limits the universe of compounds for peak annotation and thus the scope of results obtained from this method. With microMum, researchers can now visually explore whether enriched pathways were largely driven by the human host or by microbes, as well as identify which microbes likely contributed to the changes. Future enhancements to microMum include adding a methodological manner to attribute functional changes to specific organisms, as well as to support other hosts such as mice. Furthermore, the current algorithm only identifies whether a certain metabolite is present within an organism, disregarding whether it can produce or consume the metabolite. Another improvement would be to integrate this knowledge into the microMum visualizations to get a fuller and more dynamic picture of the results.

The importance of multi-omics investigations towards better understanding of the gut microbiome has been continuously stressed over the years (Chong & Xia, 2017; Jansson & Baker, 2016; Son, Shoaie, & Lee, 2020; Q. Wang et al., 2019), with few methods tailored specifically towards integrating metabolomics and microbiome sequencing data. One important and well-used method is MIMOSA (Noecker et al., 2016) and its more recent upgrade MIMOSA2(Noecker, Eng, & Borenstein, 2021). MIMOSA/2 is intended for paired microbiome and metabolomic data, with the goal of linking changes in taxonomy to changes in metabolite measurements. Like microMum, it builds a taxon specific CMN, which underpins the rest of the

analysis. From the CMN, a metabolic profile is predicted, based on whether the taxa are capable of synthesizing or utilizing the metabolites. A linear regression is then performed between the predicted metabolic profile and actual metabolomic measurements. Finally, the model fit is decomposed into the predicted contribution from each taxon. Unlike microMum however, only microbial metabolic models are used to build the CMN. Moreover, while able to identify microbially-important metabolites and pinpoint important taxa to said metabolites, it does not provide further functional insights (e.g. changes in specific pathways) and requires already annotated metabolomics data (e.g. targeted).

Another method for integrating multi-omics microbiome data is mmvec, which is a machine learning algorithm that builds a neural network to predict metabolite abundance from a single microbial sequence (Morton et al., 2019). Briefly, the neural network is trained on the microbiome sequencing data to best predict the actual metabolite abundances. In the end, the method was shown to outperform existing statistical methods such as Pearson's correlation and SparCC (Friedman & Alm, 2012) to identify co-occurring microbe-metabolites. Unlike microMum and MIMOSA/2, this method is purely statistical and does not integrate any previous knowledge to make predictions. Moreover, the use of a black-box method makes interpretation of microbe-metabolite interactions difficult.

Overall, microMum was able to capture true changes in the metabolism of CD (from the original publication), as well as provide hints as to how microbes may contribute to alterations in metabolism and ultimately disease pathogenesis. While a valuable tool, it is not without its limitations. First, as the underlying knowledgebase are GEMs, any insights obtained using our method are greatly impacted by the quality of the GEMs. Further, not all microbes had

corresponding genome-scale metabolic model matches. Whilst the matching algorithm has been refined to maximize matches between a user's list of taxa to our GEM database, there may not yet exist GEMs for those microbes, or inconsistency between taxonomic annotations could lead to no matches. As high-quality GEMs become publicly available, we will continuously incorporate these into our tool to improve our GEM-matching percentage. Moreover, taxonomic classifications must at least be to species or better, strain level resolution. Microbial GEMs do not exist in our tool at higher taxonomic levels such as genus or class as there is too much functional variation between different species. Finally, while a community metabolic network is created, methods that leverage network topology other than pathway enrichment is not considered. Use of topological-based methods could improve insights such as identifying hub metabolites/species (Layeghifard, Hwang, & Guttman, 2017) or understanding the modularity of the network (Greenblum, Turnbaugh, & Borenstein, 2012).

Conclusion

microMum is a freely available R package and is the first, to our knowledge, that provides a means to integrate a microbial taxonomic profile and untargeted metabolomics data using a knowledge-based approach. We demonstrate its utility on a well-known IBD cohort study and successfully replicate their findings. microMum also produces several publication-ready key figures that enable users to investigate microbial contributions to specific changes in metabolic function. Finally, its simplicity allows for interpretable functional insights and empowers hypothesis generation for future experimental validation.

Supplementary Materials: Metabolome Coverage of AGORA Versus CarveMe Metabolic Models

The goal of the supplementary materials shown here is to compare the metabolome coverage of AGORA and CarveMe metabolic models. As both databases were created using different resources, an objective method to compare their metabolite contents was needed. Therefore, enrichment analysis of the metabolic contents of each database underwent enrichment analysis using the RefMet chemical class ontology (Fahy & Subramaniam, 2020). PubChem Identifiers (CIDs) were used as input (Supplementary Table 1). The results are briefly summarized below.

Supplementary Table 1. Number of metabolites in AGORA and CarveMe databases with valid PubChem Chemical Identifiers.

Database	**# PubChem CIDs**	**Database**	**# PubChem CIDs**
AGORA	1137	**CarveMe**	1414

Comparison of Super Chemical Class

The top three super chemical classes enriched (*FDR-adjusted p-value* < 0.05) were consistent between AGORA and CarveMe. Organic nitrogen compounds were only enriched in AGORA, and not CarveMe model metabolites (**Supplementary Tables 2, 3, Supplementary Figure 1**).

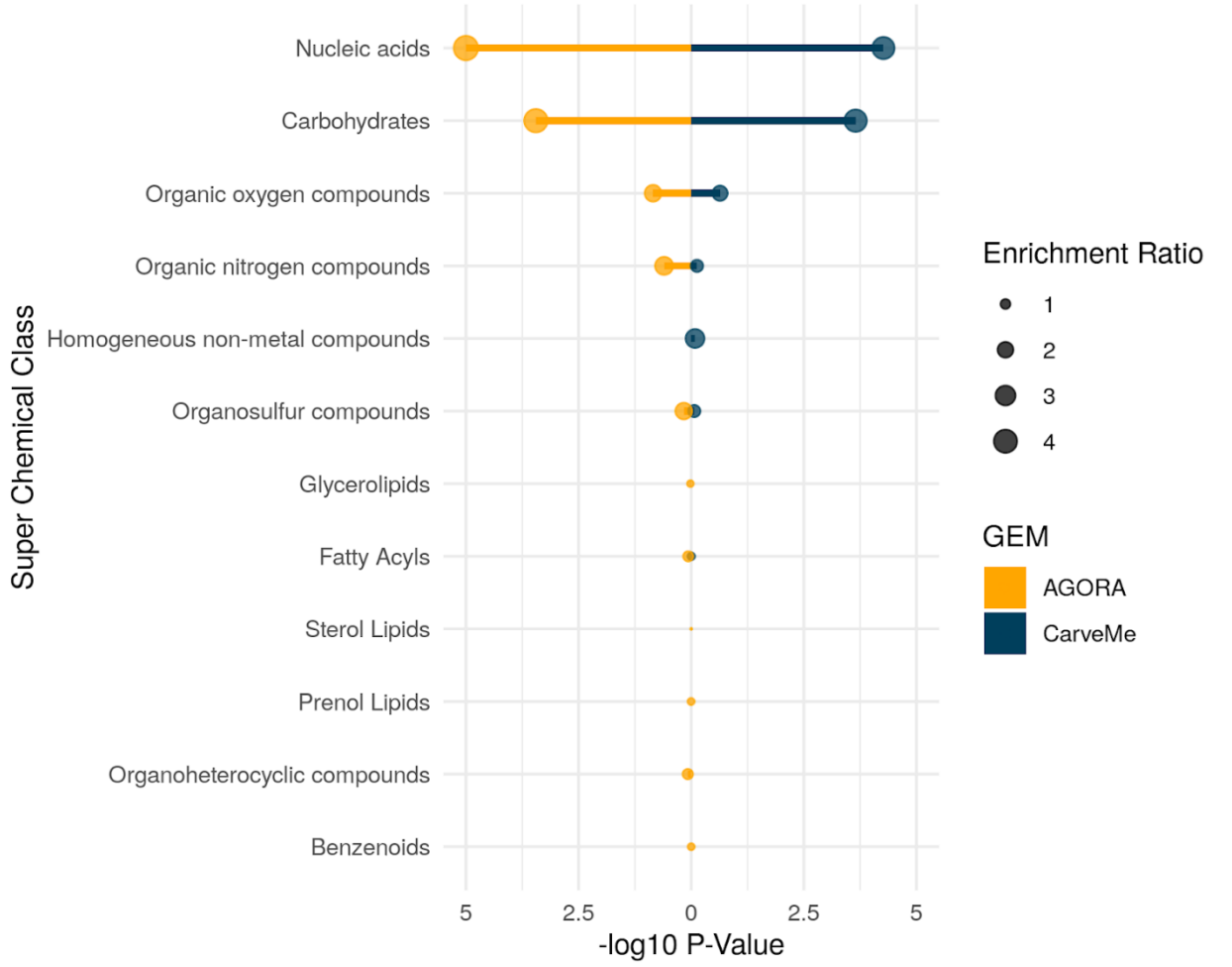
Supplementary Table 2. Top 5 enriched chemical super classes in AGORA models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Nucleic acids	374	18.4	81	2.73E-30	5.73E-29	5.73E-29
Carbohydrates	306	15	61	4.08E-21	8.16E-20	4.28E-20
Organic oxygen compounds	322	15.8	35	9.85E-06	0.000187	6.9E-05
Organic nitrogen compounds	145	7.12	18	0.000275	0.00495	0.00144
Organosulfur compounds	37	1.82	4	0.106	1	0.447

Supplementary Table 3. Top 5 enriched chemical super classes in CarveMe models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Nucleic acids	374	22.8	84	5.96E-26	1.25E-24	1.25E-24
Carbohydrates	306	18.7	70	2.65E-22	5.31E-21	2.79E-21
Organic oxygen compounds	322	19.7	37	0.000168	0.0032	0.00118

Organic nitrogen compounds	145	8.86	12	0.176	1	0.925
Homogeneous non-metal compounds	6	0.366	1	0.315	1	1



Supplementary Figure 1. Lollipop chart showing the enrichment of Super Chemical Classes in AGORA as compared to CareveMe metabolic models. The length of the lines correspond to the -log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits).

Comparison of Main Chemical Class

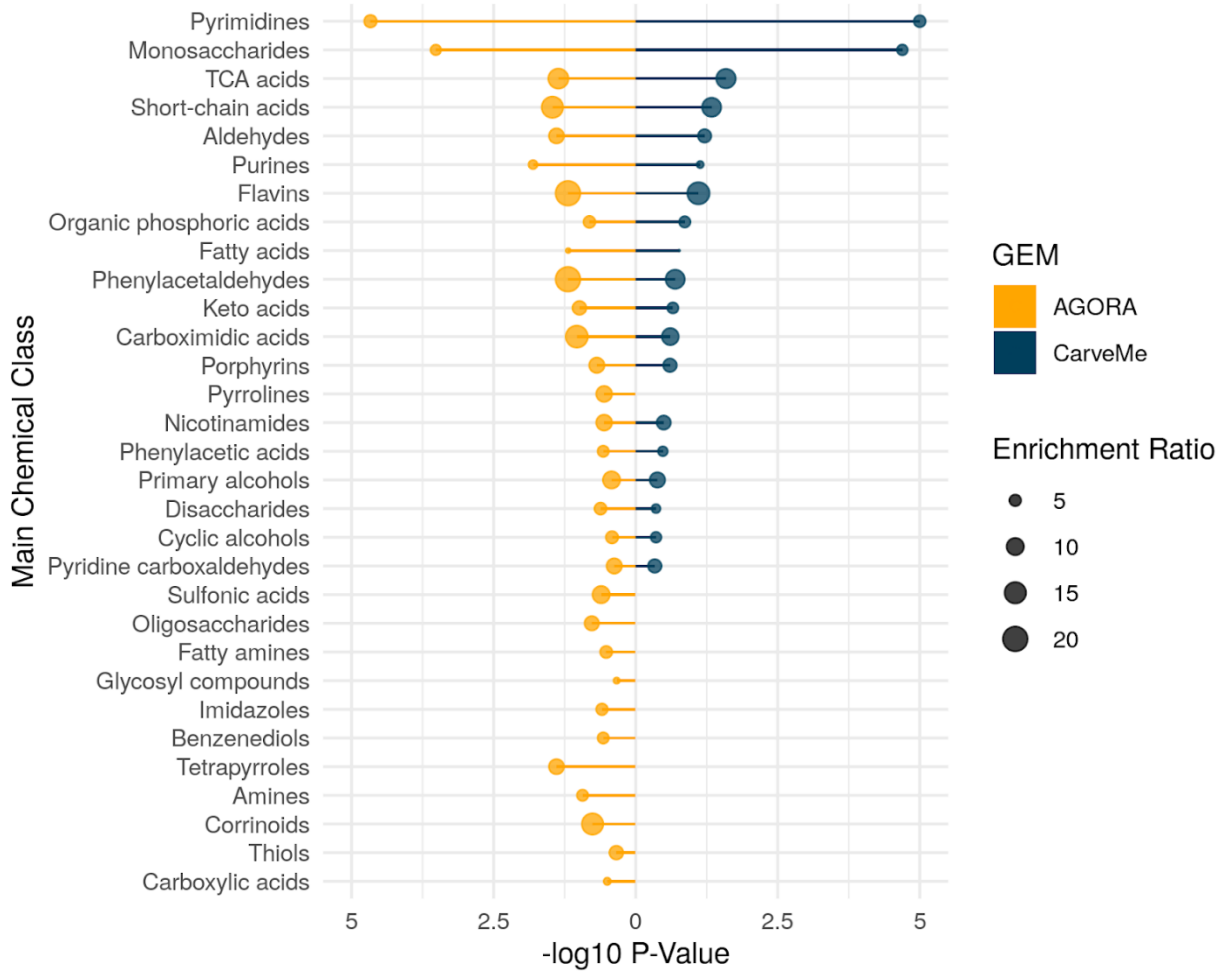
For AGORA models, 26 main chemical classes were enriched (*FDR-adjusted p-value* < 0.1) whereas CarveMe model metabolites were enriched in 13 main chemical classes (**Supplementary Tables 4, 5**). All enriched chemical classes from CarveMe models were also enriched in AGORA models. Main chemical classes enriched in AGORA and not CarveMe models include Tetrapyrroles, Amines and Oligosaccharides (**Supplementary Figure 2**).

Supplementary Table 4. Top 5 enriched chemical main classes in AGORA models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Pyrimidines	152	7.59	43	3.96E-21	8.63E-19	8.63E-19
Monosaccharides	202	10.1	43	4.24E-16	9.19E-14	4.62E-14
Purines	165	8.24	28	1.27E-08	2.75E-06	9.26E-07
Short-chain acids	8	0.4	6	3.94E-07	8.46E-05	2.14E-05
Aldehydes	23	1.15	9	8.11E-07	0.000174	2.95E-05

Supplementary Table 5. Top 5 enriched chemical main classes in CarveMe models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Pyrimidines	152	9.44	49	1.39E-22	3.04E-20	3.04E-20
Monosaccharides	202	12.5	55	3.04E-21	6.61E-19	3.32E-19
TCA acids	9	0.559	7	1.13E-07	2.45E-05	8.24E-06
Short-chain acids	8	0.497	6	1.43E-06	0.000307	7.78E-05
Aldehydes	23	1.43	9	4.95E-06	0.00106	0.000216



Supplementary Figure 2. Lollipop chart showing the enrichment of Main Chemical Classes in AGORA as compared to CareveMe metabolic models. The length of the lines correspond to the -log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits).

Comparison of Sub Chemical Class

37 sub chemical classes were enriched (*FDR-adjusted p-value* < 0.1) in AGORA models as compared to 22 sub chemical classes in CarveMe models (**Supplementary Tables 6, 7**). 17 sub chemical classes were enriched in AGORA but not in CarveMe models, including Metallotetrapyrroles, Acyl CoAs and Oligosaccharides (**Supplementary Figure 3**). Meanwhile Monosaccharides and Pyrimidine dNDP were enriched in CarveMe models but not AGORA models.

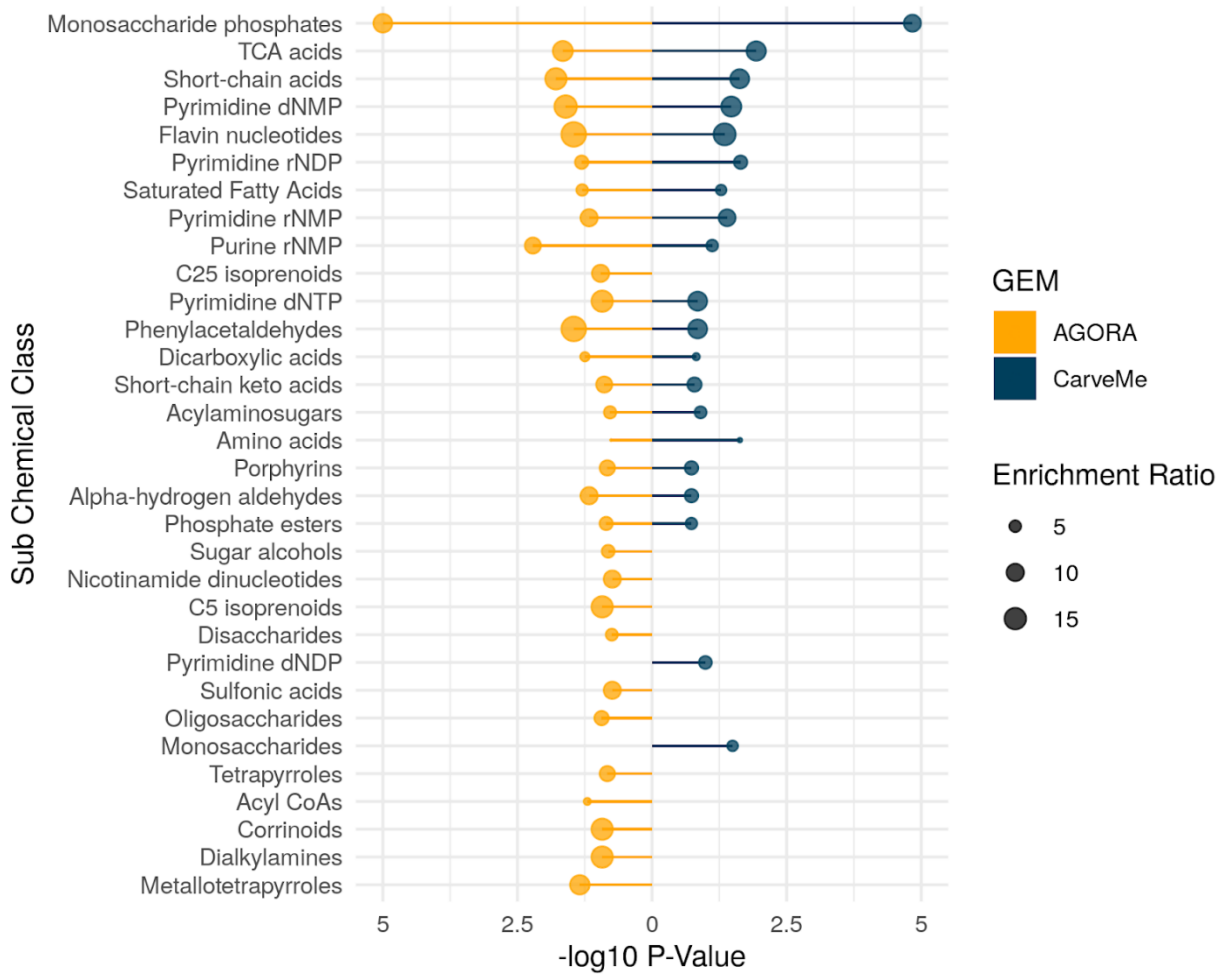
Supplementary Table 6. Top 5 enriched chemical sub classes in AGORA models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Monosaccharide phosphates	36	1.84	21	1.73E-18	8.15E-16	8.15E-16
Purine rNMP	25	1.28	11	1.38E-08	6.47E-06	3.24E-06
Short-chain acids	8	0.409	6	4.54E-07	0.000213	7.12E-05
TCA acids	9	0.461	6	1.3E-06	0.000609	0.000153
Pyrimidine dNMP	6	0.307	5	2E-06	0.000932	0.000188

Supplementary Table Table 7. Top 5 enriched chemical sub classes in CarveMe models.

Pathway	**Total**	**Expected**	**Hits**	**Raw *p***	**Holm *p***	**FDR *p***
Monosaccharide phosphates	36	2.29	22	6.59E-18	3.1E-15	3.1E-15
TCA acids	9	0.573	7	1.34E-07	6.28E-05	3.15E-05

Pyrimidine rNDP	25	1.59	10	1.43E-06	0.000668	0.000155
Amino acids	545	34.7	64	1.61E-06	0.000753	0.000155
Short-chain acids	8	0.509	6	1.65E-06	0.000768	0.000155



Supplementary Figure 3. Lollipop chart showing the enrichment of Sub Chemical Classes in AGORA as compared to CareveMe metabolic models. The length of the lines correspond to the -

log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits).

Summary

From this analysis, we can see the similarities and differences between the metabolome coverage of AGORA and CarveMe models. Overall, it seems that AGORA models had a greater coverage of chemical classes as compared to CarveMe, despite more CarveMe metabolites having a PubChem CID. Particularly at the main and sub chemical classes, these differences were more pronounced. Due to such observations, it is recommended that both model databases should be used to obtain better biological insights.

Supplementary Materials: Crohn's Disease Case Study

Below are supplementary tables for the comparison of human-only, microbe-only and community metabolic network microMum results.

Supplementary Table 8. Top 10 predicted pathway activity of untargeted metabolomics data using the human GEM.

KEGG Pathway	Hits	FET	Gamma
Steroid biosynthesis	25	1.8E-08	0.0152
Primary bile acid biosynthesis	25	1.9E-06	0.0152
Biosynthesis of unsaturated fatty acids	22	2.8E-05	0.0152
Secondary bile acid biosynthesis	14	7.9E-05	0.0152
Steroid hormone biosynthesis	53	3.3E-04	0.0152
Linoleic acid metabolism	12	5.4E-04	0.0153
Steroid degradation	5	4.2E-03	0.0170
Fatty acid biosynthesis	8	2.0E-02	0.0188
alpha-Linolenic acid metabolism	3	3.8E-02	0.0289
Ether lipid metabolism	11	1.2E-01	0.0313

Supplementary Table 9. Top 10 predicted pathway activity of untargeted metabolomics data using the microbial GEMs.

KEGG Pathway	Hits	FET	Gamma
Secondary bile acid biosynthesis	11	9.8E-06	0.0083
Primary bile acid biosynthesis	10	4.5E-03	0.0088
Fatty acid biosynthesis	7	3.5E-03	0.0089
Ubiquinone and other terpenoid-quinone biosynthesis	14	8.0E-03	0.0090
Folate biosynthesis	12	1.4E-02	0.0096
Steroid hormone biosynthesis	5	5.0E-02	0.0161
Porphyrin and chlorophyll metabolism	19	1.3E-01	0.0181
Cutin, suberine and wax biosynthesis	2	3.6E-02	0.0218
Retinol metabolism	3	9.4E-02	0.0320
Sesquiterpenoid and triterpenoid biosynthesis	3	9.4E-02	0.0320

Supplementary Table 10. Top 10 predicted pathway activity of untargeted metabolomics data using the community metabolic network.

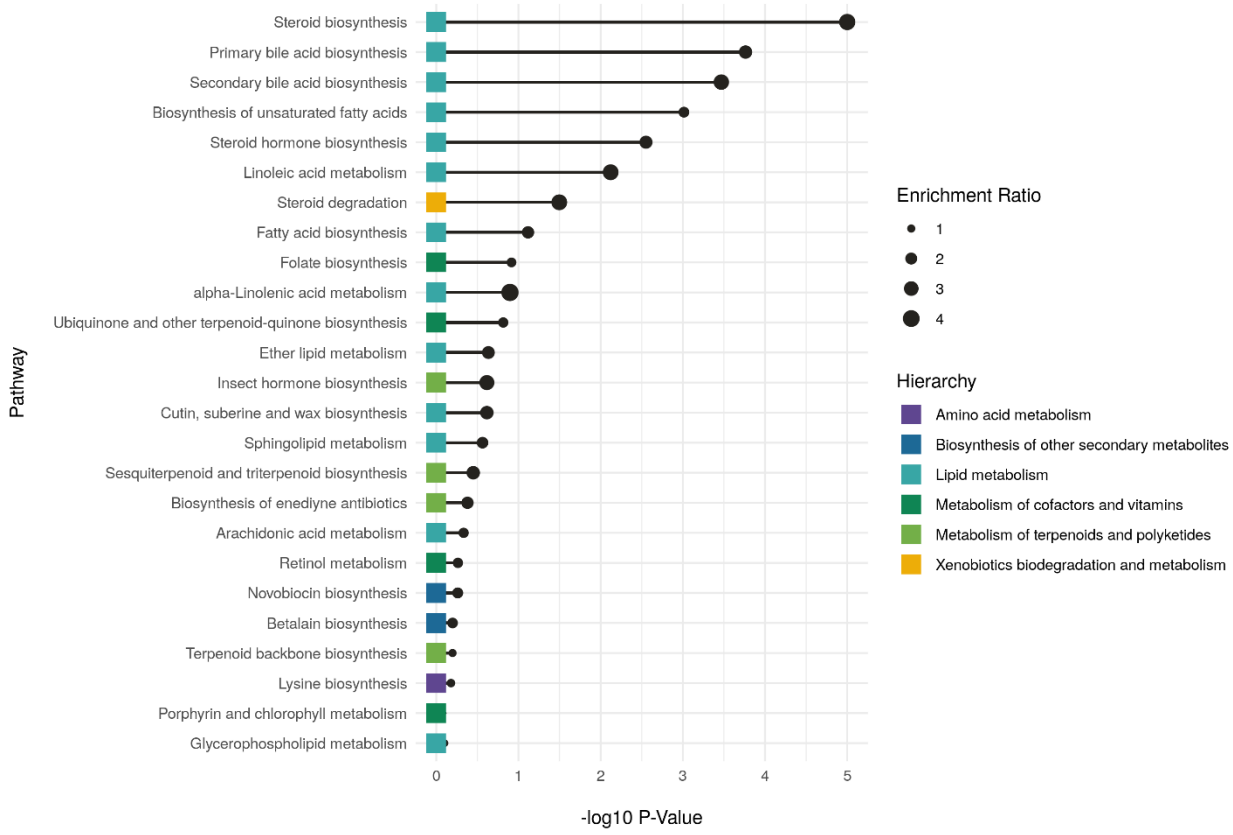
KEGG Pathway	Hits	FET	Gamma
Steroid biosynthesis	25	1.3E-09	0.0219

Primary bile acid biosynthesis	25	2.0E-07	0.0219
Secondary bile acid biosynthesis	17	7.0E-07	0.0219
Biosynthesis of unsaturated fatty acids	22	4.4E-06	0.0219
Steroid hormone biosynthesis	53	2.6E-05	0.0219
Linoleic acid metabolism	12	1.7E-04	0.0220
Steroid degradation	5	2.2E-03	0.0234
Fatty acid biosynthesis	8	1.0E-02	0.0247
Folate biosynthesis	16	2.3E-02	0.0254
Ubiquinone and other terpenoid-quinone biosynthesis	22	3.4E-02	0.0263

Below is the supplementary table of microMum results using only significantly different microbes when creating the community metabolic network using AGORA GEM models.

Supplementary Table 11. Top 10 predicted pathway activity of untargeted metabolomics data using the community metabolic network created using differentially abundant microbes from the AGORA GEM database.

KEGG Pathway	Hits	FET	Gamma
Steroid biosynthesis	25	1.7E-09	0.0212
Primary bile acid biosynthesis	25	2.6E-07	0.0212
Secondary bile acid biosynthesis	17	8.4E-07	0.0212
Biosynthesis of unsaturated fatty acids	22	5.3E-06	0.0212
Steroid hormone biosynthesis	53	3.4E-05	0.0212
Linoleic acid metabolism	12	1.9E-04	0.0213
Steroid degradation	5	2.4E-03	0.0227
Fatty acid biosynthesis	8	1.1E-02	0.0241
Folate biosynthesis	16	2.5E-02	0.0249
Ubiquinone and other terpenoid-quinone biosynthesis	22	3.7E-02	0.0259



Supplementary Figure 4. Lollipop chart of the functional enrichment results of KEGG metabolic pathways using the community metabolic network of differentially abundant microbes between Crohn's Disease and healthy controls. The x-axis is the -log 10 p-value (scaled from 0-5), and the y-axis are the pathways. The square colored boxes on the left of the plot shows the metabolic hierarchy of that pathway to aid interpretation. The length of the lines corresponds to the -log10 raw p-value and the dot at the tip represents the enrichment ratio (number of observed hits / number of expected hits).

Below is the supplementary table of microMum results using only significantly different microbes when creating the community metabolic network using CarveMe GEM models.

Supplementary Table 12. Top 10 predicted pathway activity of untargeted metabolomics data using the community metabolic network created using differentially abundant microbes from the CarveMe GEM database.

KEGG Pathway	Hits	FET	Gamma
Steroid biosynthesis	25	1.9E-09	0.0206
Primary bile acid biosynthesis	25	2.8E-07	0.0206

Biosynthesis of unsaturated fatty acids	22	5.7E-06	0.0206
Steroid hormone biosynthesis	53	3.8E-05	0.0206
Secondary bile acid biosynthesis	14	2.4E-05	0.0206
Linoleic acid metabolism	12	2.0E-04	0.0207
Steroid degradation	5	2.5E-03	0.0221
Fatty acid biosynthesis	8	1.1E-02	0.0235
Ubiquinone and other terpenoid-quinone biosynthesis	23	5.5E-02	0.0270
Folate biosynthesis	15	5.0E-02	0.0277

Below are the supplementary tables of predicted pathway activity results using the previous Mummichog implementation from MetaboAnalyst.

Supplementary Table 13. Top 10 predicted pathway activity of untargeted metabolomics data using the *hsa_mfn* manually curated metabolic model.

Pathway	**Hits**	**FET**	**Gamma**
Bile acid biosynthesis	55	1.06E-10	0.0129
Squalene and cholesterol biosynthesis	42	2.03E-07	0.0129
Vitamin D3 (cholecalciferol) metabolism	14	5.93E-05	0.0130
Vitamin E metabolism	41	0.001413	0.0131
Carnitine shuttle	38	0.002926	0.0132
De novo fatty acid biosynthesis	20	0.002805	0.0133
C21-steroid hormone biosynthesis and metabolism	88	0.012006	0.0136
Omega-6 fatty acid metabolism	7	0.040724	0.0193
Linoleate metabolism	36	0.15926	0.0246
Biopterin metabolism	12	0.16209	0.0311

Supplementary Table 14. Top 10 predicted pathway activity of untargeted metabolomics data using the human KEGG metabolic model.

KEGG Pathway	**Hits**	**FET**	**Gamma**
Steroid hormone biosynthesis	84	8.32E-14	0.127
Steroid biosynthesis	41	1.37E-05	0.127
Primary bile acid biosynthesis	34	0.000217	0.127
Glycosaminoglycan degradation	17	0.023798	0.141
Aminoacyl-tRNA biosynthesis	21	0.069181	0.157
Biosynthesis of unsaturated fatty acids	11	0.04866	0.159
Lysine degradation	14	0.066413	0.163

Arginine biosynthesis	10	0.070926	0.173
Arginine and proline metabolism	34	0.14518	0.177
Retinol metabolism	16	0.10724	0.177

Supplementary Table 15. Top 10 predicted pathway activity of untargeted metabolomics data using the *Escherichia coli* KEGG metabolic model.

KEGG Pathway	Hits	FET	Gamma
Arginine and proline metabolism	28	0.000786	0.0569
Aminoacyl-tRNA biosynthesis	21	0.001869	0.0573
Arginine biosynthesis	13	0.007416	0.0599
Porphyrin and chlorophyll metabolism	14	0.019747	0.0635
Ubiquinone and other terpenoid-quinone biosynthesis	11	0.023862	0.0660
Lysine biosynthesis	11	0.023862	0.0660
Arachidonic acid metabolism	5	0.025532	0.0756
Galactose metabolism	30	0.12706	0.0834
Amino sugar and nucleotide sugar metabolism	29	0.16819	0.0929
Biotin metabolism	4	0.053361	0.0944

Supplementary Table 16. Predicted pathway activity of untargeted metabolomics data using microMum showing the results of the community metabolic model permutation.

	Hits	FET	Gamma	Empirical Hits	Empirical P-Value
Steroid biosynthesis	25	2E-09	0.021	120	0.12
Primary bile acid biosynthesis	25	3E-07	0.021	120	0.12
Secondary bile acid biosynthesis	17	8E-07	0.021	119	0.119
Biosynthesis of unsaturated fatty acids	22	5E-06	0.021	120	0.12
Steroid hormone biosynthesis	53	3E-05	0.021	120	0.12
Linoleic acid metabolism	12	2E-04	0.021	121	0.121
Steroid degradation	5	2E-03	0.023	131	0.131
Fatty acid biosynthesis	8	1E-02	0.024	168	0.168
Folate biosynthesis	16	2E-02	0.025	249	0.249
Ubiquinone and other terpenoid-quinone biosynthesis	22	4E-02	0.026	479	0.479
Ether lipid metabolism	11	8E-02	0.034	434	0.434
alpha-Linolenic acid metabolism	3	3E-02	0.035	264	0.264
Sphingolipid metabolism	17	1E-01	0.035	575	0.575
Cutin, suberine and wax biosynthesis	4	8E-02	0.047	408	0.408
Insect hormone biosynthesis	4	8E-02	0.047	408	0.408
Arachidonic acid metabolism	36	3E-01	0.052	900	0.9
Sesquiterpenoid and triterpenoid biosynthesis	5	2E-01	0.065	572	0.572
Retinol metabolism	13	3E-01	0.086	467	0.467
Terpenoid backbone biosynthesis	21	4E-01	0.102	846	0.846
Biosynthesis of enediyne antibiotics	3	2E-01	0.108	561	0.561
Lysine biosynthesis	15	5E-01	0.124	172	0.172
Betalain biosynthesis	8	4E-01	0.142	819	0.819
Novobiocin biosynthesis	4	3E-01	0.149	678	0.678
Porphyrin and chlorophyll metabolism	21	6E-01	0.173	691	0.691
Glycerophospholipid metabolism	18	7E-01	0.197	922	0.922
Arginine biosynthesis	14	6E-01	0.199	902	0.902
Monobactam biosynthesis	10	6E-01	0.209	862	0.862
Glucosinolate biosynthesis	10	6E-01	0.209	862	0.862
Tyrosine metabolism	47	8E-01	0.235	22	0.022
Vitamin B6 metabolism	16	8E-01	0.262	632	0.632
Nicotinate and nicotinamide metabolism	21	8E-01	0.281	933	0.933
Phenylalanine metabolism	22	8E-01	0.310	785	0.785
Biosynthesis of various secondary metabolites - part 2	13	8E-01	0.318	482	0.482
Tropane, piperidine and pyridine alkaloid biosynthesis	18	8E-01	0.326	662	0.662

Fatty acid degradation	8	7E-01	0.330	815	0.815
D-Arginine and D-ornithine metabolism	8	7E-01	0.330	815	0.815
Limonene and pinene degradation	8	7E-01	0.330	815	0.815
Isoquinoline alkaloid biosynthesis	8	7E-01	0.330	815	0.815
Tryptophan metabolism	40	9E-01	0.339	949	0.949
Phenylpropanoid biosynthesis	10	9E-01	0.414	445	0.445
Riboflavin metabolism	10	9E-01	0.414	828	0.828
Phenylalanine, tyrosine and tryptophan biosynthesis	21	9E-01	0.420	270	0.27
Biosynthesis of various secondary metabolites - part 3	17	9E-01	0.457	897	0.897
Cyanoamino acid metabolism	12	9E-01	0.490	837	0.837
Styrene degradation	12	9E-01	0.490	850	0.85
Alanine, aspartate and glutamate metabolism	19	1E+00	0.519	896	0.896
Valine, leucine and isoleucine degradation	13	9E-01	0.524	945	0.945
Valine, leucine and isoleucine biosynthesis	20	1E+00	0.547	954	0.954
Histidine metabolism	21	1E+00	0.575	895	0.895
Arginine and proline metabolism	47	1E+00	0.589	969	0.969
Glutathione metabolism	16	1E+00	0.615	867	0.867
Lysine degradation	24	1E+00	0.648	777	0.777
Cysteine and methionine metabolism	34	1E+00	0.726	497	0.497
Metabolism of xenobiotics by cytochrome P450	80	1E+00	0.742	952	0.952
Methane metabolism	26	1E+00	0.812	299	0.299
Amino sugar and nucleotide sugar metabolism	28	1E+00	0.838	927	0.927
Purine metabolism	45	1E+00	0.863	924	0.924
Pyrimidine metabolism	31	1E+00	0.869	852	0.852
Glycolysis / Gluconeogenesis	14	1E+00	1.000	1000	1
Fatty acid elongation	5	8E-01	1.000	1000	1
Caffeine metabolism	2	5E-01	1.000	1000	1
Glycine, serine and threonine metabolism	27	1E+00	1.000	1000	1
Penicillin and cephalosporin biosynthesis	3	7E-01	1.000	1000	1
Prodigiosin biosynthesis	2	5E-01	1.000	1000	1
Phenazine biosynthesis	2	5E-01	1.000	1000	1
beta-Alanine metabolism	17	1E+00	1.000	1000	1
D-Glutamine and D-glutamate metabolism	5	8E-01	1.000	1000	1
D-Alanine metabolism	4	8E-01	1.000	1000	1
O-Antigen nucleotide sugar biosynthesis	7	9E-01	1.000	1000	1
Peptidoglycan biosynthesis	4	8E-01	1.000	1000	1
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	1	3E-01	1.000	1000	1
Glyoxylate and dicarboxylate metabolism	14	1E+00	1.000	1000	1

Thiamine metabolism	16	1E+00	1.000	1000	1
Carotenoid biosynthesis	3	7E-01	1.000	1000	1
Nitrogen metabolism	3	7E-01	1.000	1000	1
Sulfur metabolism	12	1E+00	1.000	1000	1
Biosynthesis of siderophore group nonribosomal peptides	5	8E-01	1.000	1000	1
Biosynthesis of vancomycin group antibiotics	2	5E-01	1.000	1000	1

Chapter 5. Discussion

Throughout all my work in this thesis and beyond was the theme to create user-friendly and freely accessible tools for omics data analysis, interpretation, and visualization to bridge the gap between bench and data scientists. With rapid advances in high-throughput omics technologies and decreasing costs, there is an ever-growing number of omics-centric studies. Multiple methods exist for data pre-/processing, which can be overwhelming for novice researchers. Thoughtful analytical pipelines that are intuitive, innovative, and easy-to-use are urgently needed.

MetaboAnalyst, as discussed in Chapter 2, is a comprehensive, web-based tool suite for metabolomics data analysis, visualization, and functional interpretation. It was first released in 2009 with a single module for metabolomic data processing and statistical analysis (Jianguo Xia, Nick Psychogios, Nelson Young, & David S Wishart, 2009), with substantial updates in Version 2.0 for functional analysis and data interpretation (Jianguo Xia, Rupasri Mandal, Igor V Sinelnikov, David Broadhurst, & David S Wishart, 2012), and Version 3.0 for biomarker analysis, power analysis, and joint pathway analysis (Jianguo Xia, Igor V Sinelnikov, Beomsoo Han, & David S Wishart, 2015). As the field of metabolomics continues to evolve, so should MetaboAnalyst. We therefore developed Version 4.0, in tandem with its companion R package (MetaboAnalystR), towards a more transparent and integrative metabolomic data analysis. This upgrade consisted of: (i) a major overhaul to its user interface towards a more modern design, (ii) improved reproducibility/transparency with the inclusion of a R Command History throughout a user's session that can be used to recreate all analyses locally using the MetaboAnalystR

package, (iii) support for meta-analysis and multi-omics data analysis, (iv) an update of underlying knowledgebases in collaboration with HMDB 4.0 (Wishart et al., 2017a), and (v) a new module based on the mummichog algorithm for pathway activity prediction from untargeted metabolomics data (S. Li et al., 2013a). Together, these updates ensured that MetaboAnalyst remains on the forefront of computational metabolomics and enables researchers to make novel and insightful discoveries.

Untargeted metabolomics, also known as global metabolomics, aims to measure all possible metabolites within samples without *a priori* knowledge of the metabolome. A typical LC-MS based metabolomics experiment can generate 10,000s peaks (features) characterized by their mass and retention times. However, as a single peak can potentially match multiple compounds within the given mass range, peak annotation requires significant efforts to search through compound databases and perform tandem MS experiments. Due to this challenge, functional interpretation of global metabolomics data is not straightforward, as classical metabolic pathway enrichment analysis requires named metabolites as input, not MS peaks. To address this bottleneck, Li et al proposed a novel approach, named mummichog, to directly infer pathway activities from peak lists by leveraging the collective power of metabolic pathways, without requiring accurate metabolite identification(S. Li et al., 2013a). This algorithm assumes that a certain degree of random errors during individual peak assignment will not change the collective behavior jointly determined by all metabolites involved in the pathways. This concept has been recently adapted to the popular Gene Set Enrichment Analysis algorithm (Subramanian et al., 2005) in MetaboAnalystR 2.0 (Chapter 3). The original mummichog algorithm is based on over representation analysis to test if certain pathways are enriched in the significant peaks as compared

to null models based on peak lists of the same size randomly drawn from the inputted peak list. In comparison, GSEA is a cut-off free method that evaluates the overall differences of two distributions based on Kolmogorov-Smirnov tests. The manuscript in Chapter 3 showcased the differences and similarities in functional interpretations of MS peaks using both versions of the algorithm.

Since the publications of MetaboAnalyst 4.0 and MetaboAnalystR 2.0, substantial upgrades have been made to the mummichog algorithm, including support for multi-modal pathway prediction (i.e. positive and negative ion modes). This was an important enhancement as a mass spectrometer is run in either positive or negative mode, and different molecules will be ionized dependent on the mode. For example, nitrogen and oxygen bases are detected in positive mode, whereas molecules with acidic functional groups such as carboxylic acids are detected in negative mode (J. Liigand et al., 2020; P. Liigand et al., 2017). In addition to this, adduct and currency metabolite customization was enabled, as different adducts may be formed dependent on the LC-MS system or solutions used. Other modifications to the mummichog code were the inclusion of retention times to strengthen peak annotations, and support for meta-analysis of multiple untargeted metabolomics datasets. These updates have been detailed in subsequent publications, MetaboAnalyst 5.0 (Pang et al., 2021) and MetaboAnalystR 3.0 (Pang et al., 2020). Ultimately, the ability to obtain biological insights from MS peaks has been refined and maximized.

Despite the plethora of microbiome research, gaining biological insights beyond a list of significantly different taxa is not yet the norm. microMum, described in Chapter 4, is an R package for predicting changes in microbial metabolism by integrating paired untargeted

metabolomics and taxonomic microbial sequencing data. Bacteria in the gut communicate amongst themselves as well as with their host via the production of metabolites. Microbial-derived metabolites, including SCFAs and bile acids, are known to have systemic effects on their host's health. Therefore, if one has a list of taxa differentially abundant between two groups as well as corresponding untargeted metabolomics data, he or she can use microMum to predict alterations in microbial metabolism. This helps to answer why a set of microbes is linked to a disease and steers researchers towards understanding how that set of microbes affects host phenotype. Moreover, insights obtained are easily interpretable – there is no black box of complicated algorithms that can overwhelm users when trying to understand the results. Finally, the IBD use-case in Chapter 4 showcased how inferred metabolic function through microMum captured metabolomic differences in the original publication and provided further hints not previously reported on how certain microbes could have contributed to disease pathogenesis.

The validity of results obtained from microMum predicate upon several different aspects. One, that the majority of a user's list of microbes is captured within the microMum GEM database. If only a small percentage of a user's list has GEM matches, the insights obtained using microMum would not be close to the truth due to the lack of information. Second, it requires that the untargeted metabolomics data analysis to identify important peaks between cases and controls was properly executed. If done improperly, it could lead to false positives (signal there that does not truly exist) or inversely, false negatives (signal truly exists but was not picked up). Along the same thread, the differential abundance analysis to identify important microbes between disease phenotypes should also be done correctly to properly annotate changes in metabolism to the right microbes. If these prerequisites are met, microMum would be suitable for

a user's multi-omic microbiome data analysis. Overall, the novel framework developed in Chapter 4 will enable fast and interpretable integration of untargeted metabolomics and taxonomic microbial signatures.

Chapter 6. Conclusion

Bioinformatic pipelines created in Chapters 2 and 3 bring forward cutting-edge and user-friendly platforms for metabolomics data processing and analysis. Moreover, the knowledge-based platform created in Chapter 4 will help users functionally interpret and integrate their paired metabolomics and taxonomic sequencing data for further biological insights. For instance, say a user has collected untargeted metabolomics and identified a list of microbes enriched in their study of rheumatoid arthritis (RA). After using microMum, they identify nitrogen and sulfur metabolism to be different between RA and healthy controls, which could have negative consequences to human health (X. Zhang et al., 2015). Using microMum's visualizations, they could then examine this pathway and find that it is largely driven by microbe X, making it a potential target for therapeutic interventions. The platform therefore fills an urgently needed gap by providing users a powerful tool to integrate paired untargeted metabolomics and taxonomic microbial signatures to perform function-driven hypothesis generation.

Moving forward, I envision several steps that can be undertaken to enhance the utility of all tools developed throughout my PhD. For MetaboAnalyst/R, putative compound identification is a key step when performing functional interpretation of untargeted metabolomics data. This is a challenge as if solely the molecular weight of a molecule is used to predict peak identity, there will be numerous matches. For instance, searching for a compound in the Human Metabolome Database (Wishart et al., 2017a) with a range of 180.0 to 180.5 results in 81 potential metabolites. This can lead to an inflated false-positive identification rate. To address this, retention times, which is the time taken for a molecule to pass through a chromatography

column, can be used in tandem with chemical similarity metrics to improve this step. Mummichog (S. Li et al., 2013a), the original underlying algorithm for this module, leverages organism-specific metabolic pathways for functional interpretation. Because a pathway is essentially the chemical transformation of a metabolite, metabolites within the same pathway should be similarly structured, and therefore have similar retention times. My idea is therefore to incorporate retention time and chemical similarity cut-offs when performing the putative compound identification step. Altogether, I envision this would improve biological insights obtained using this module.

In-depth understanding of community-wide metabolism is essential to gain mechanistic insights to community-level microbial ecology and further inform therapeutic manipulations of the gut microbiome. With microMum, the next step would be to leverage the custom community metabolic networks to further integrate other microbiome-specific multi-omics data, such as transcriptomics or proteomics. The goal here would be to refine the community metabolic network to add/remove genes that are not expressed and apply network-based topological algorithms to pinpoint key features of dysbiosis and investigate host-microbe interactions. Moreover, differentially enriched genes and metabolites can then be mapped (incorporated as weighted edges/nodes) to create representative disease-networks. Standard graph theory (e.g. hubs and modularity), and the prize-collecting Steiner tree (PCST) algorithm can then be used to identify important links, key players, and sub-networks within the microbiome. Ultimately, this integrative approach of CMNs, multi-omics data, and topological network-based algorithms will provide system-level resolution and deep mechanistic insights to microbiome function. Such updates would allow for the integration of multiple omics technologies to provide a holistic

overview of the gut microbiome, thereby taming the complexity of the system to identify core ecological principles and generate testable hypotheses for microbiome manipulation of metabolic function to improve health.

References

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., . . . Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol, 8*(6), e1002358. doi:10.1371/journal.pcbi.1002358

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., . . . Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res, 46*(W1), W537-W544. doi:10.1093/nar/gky379

Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC bioinformatics, 14*(1), 112.

Arts, R. J., Novakovic, B., ter Horst, R., Carvalho, A., Bekkering, S., Lachmandas, E., . . . Wang, S.-Y. (2016). Glutaminolysis and fumarate accumulation integrate immunometabolic and epigenetic programs in trained immunity. *Cell metabolism, 24*(6), 807-819.

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., . . . Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc Natl Acad Sci U S A, 101*(44), 15718-15723.

Backhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., . . . Finlay, B. B. (2012). Defining a Healthy Human Gut Microbiome: Current Concepts, Future Directions, and Clinical Applications. *Cell host & microbe, 12*(5), 611-622. doi:10.1016/j.chom.2012.10.012

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., . . . Zhong, H. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell host & microbe, 17*(5), 690-703.

Bajaj, J. S., Thacker, L. R., Fagan, A., White, M. B., Gavis, E. A., Hylemon, P. B., . . . Gillevet, P. M. (2018). Gut microbial RNA and DNA analysis predicts hospitalizations in cirrhosis. *JCI insight, 3*(5). doi:10.1172/jci.insight.98019

Baquero, F., & Nombela, C. (2012). The microbiome as a human organ. *Clinical Microbiology and Infection, 18*, 2-4.

Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P., & Thiele, I. (2015). Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome, 3*(1), 55.

Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K., & Dorrestein, P. C. (2021). Mass spectrometry-based metabolomics in microbiome investigations. *Nature Reviews Microbiology*, 1-18.

Beger, R. D., Dunn, W., Schmidt, M. A., Gross, S. S., Kirwan, J. A., Cascante, M., . . . Hankemeier, T. (2016). Metabolomics enables precision medicine:"a white paper, community perspective". *Metabolomics, 12*(9), 149.

Belkaid, Y., & Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell, 157*(1), 121-141.

Benton, H. P., Want, E. J., & Ebbels, T. M. (2010). Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics, 26*(19), 2488-2489. doi:10.1093/bioinformatics/btq441

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics, 22*(1), 178-193.

Blazewicz, S. J., Barnard, R. L., Daly, R. A., & Firestone, M. K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *Isme Journal, 7*(11), 2061-2068. doi:10.1038/ismej.2013.102

Blum, T., & Kohlbacher, O. (2008). MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics, 24*(18), 2108-2109.

Braniste, V., Al-Asmakh, M., Kowal, C., Anuar, F., Abbaspour, A., Tóth, M., . . . Kundu, P. (2014). The gut microbiota influences blood-brain barrier permeability in mice. *Science translational medicine, 6*(263), 263ra158-263ra158.

Brown, E. M., Ke, X., Hitchcock, D., Jeanfavre, S., Avila-Pacheco, J., Nakata, T., . . . Franzosa, E. A. (2019). Bacteroides-derived sphingolipids are critical for maintaining intestinal homeostasis and symbiosis. *Cell host & microbe, 25*(5), 668-680. e667.

Cambiaghi, A., Ferrario, M., & Masseroli, M. (2016). Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Briefings in bioinformatics, 18*(3), 498-510.

Candela, M., Biagi, E., Maccaferri, S., Turroni, S., & Brigidi, P. (2012). Intestinal microbiota is a plastic factor responding to environmental changes. *Trends Microbiol, 20*(8), 385-391. doi:10.1016/j.tim.2012.05.003

Cangelosi, G. A., & Meschke, J. S. (2014). Dead or alive: molecular assessment of microbial viability. *Appl Environ Microbiol, 80*(19), 5884-5891. doi:10.1128/AEM.01763-14

Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., & Owen, L. J. (2015). Dysbiosis of the gut microbiota in disease. *Microb Ecol Health Dis, 26*(1), 26191. doi:10.3402/mehd.v26.26191

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., . . . Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res, 42*(Database issue), D459-471. doi:10.1093/nar/gkt1103

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., . . . Karp, P. D. (2020). The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research, 48*(D1), D445-D453.

Charitou, T., Bryan, K., & Lynn, D. J. (2016). Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution, 48*(1), 27.

Chong, J., Liu, P., Zhou, G., & Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature Protocols, 15*(3), 799-821.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., . . . Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res, 46*(W1), W486-W494. doi:10.1093/nar/gky310

Chong, J., & Xia, J. (2017). Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. *Metabolites, 7*(4), 62. doi:10.3390/metabo7040062

Chong, J., & Xia, J. (2018). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics, 34*(24), 4313-4314.
Chong, J., Yamamoto, M., & Xia, J. (2019). MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. *Metabolites, 9*(3), 57.
Chowdhury, S., & Fong, S. S. (2020). Leveraging genome-scale metabolic models for human health applications. *Current opinion in biotechnology, 66*, 267-276.
Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., & Dinan, T. G. (2014). Minireview: gut microbiota: the neglected endocrine organ. *Molecular endocrinology, 28*(8), 1221-1238.
Consortium, I. H. i. R. N. (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe, 16*(3), 276.
Cox, A. G., Hwang, K. L., Brown, K. K., Evason, K. J., Beltz, S., Tsomides, A., . . . Chhangawala, S. (2016). Yap reprograms glutamine metabolism to increase nucleotide biosynthesis and enable liver growth. *Nature cell biology, 18*(8), 886.
Cryan, J. F., & Dinan, T. G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat Rev Neurosci, 13*(10), 701-712. doi:10.1038/nrn3346
Cui, X., Ye, L., Li, J., Jin, L., Wang, W., Li, S., . . . Cai, J. (2018). Metagenomic and metabolomic analyses unveil dysbiosis of gut microbiota in chronic heart failure patients. *Sci Rep, 8*(1), 635. doi:10.1038/s41598-017-18756-2
d'Hennezel, E., Abubucker, S., Murphy, L. O., & Cullen, T. W. (2017). Total Lipopolysaccharide from the Human Gut Microbiome Silences Toll-Like Receptor Signaling. *mSystems, 2*(6), e00046-00017.
Dalile, B., Van Oudenhove, L., Vervliet, B., & Verbeke, K. (2019). The role of short-chain fatty acids in microbiota–gut–brain communication. *Nature reviews Gastroenterology & hepatology, 16*(8), 461-478.
David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., . . . Fischbach, M. A. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature, 505*(7484), 559.
Davidson, R. L., Weber, R. J., Liu, H., Sharma-Oates, A., & Viant, M. R. (2016). Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *Gigascience, 5*(1), 10.
Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., . . . Kolaczkowski, B. (2014). *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front Microbiol, 5*, 678.
De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., . . . Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A, 107*(33), 14691-14696.
Di Guida, R., Engel, J., Allwood, J. W., Weber, R. J., Jones, M. R., Sommer, U., . . . Dunn, W. B. (2016). Non-targeted UHPLC-MS metabolomic data processing methods: a

comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics, 12*(5), 93.
Doolittle, W. F., & Booth, A. (2017). It's the song, not the singer: an exploration of holobiosis and evolutionary theory. *Biology & Philosophy, 32*(1), 5-24.
Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., . . . Langille, M. G. (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology, 38*(6), 685-688.
Drewes, J. L., White, J. R., Dejea, C. M., Fathi, P., Iyadorai, T., Vadivelu, J., . . . Sears, C. L. (2017). High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes, 3*(1), 34. doi:10.1038/s41522-017-0040-3
Duboc, H., Rajca, S., Rainteau, D., Benarous, D., Maubert, M.-A., Quervain, E., . . . Despras, G. (2013). Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. *Gut, 62*(4), 531-539.
Fahy, E., & Subramaniam, S. (2020). RefMet: a reference nomenclature for metabolomics. *Nature methods, 17*(12), 1173-1174.
Fernández-Albert, F., Llorach, R., Andrés-Lacueva, C., & Perera, A. (2014). An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics, 30*(13), 1937-1939.
Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. In *Functional Genomics* (pp. 155-171): Springer.
Fierer, N., Hamady, M., Lauber, C. L., & Knight, R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A, 105*(46), 17994-17999.
Fitz-Gibbon, S., Tomida, S., Chiu, B. H., Nguyen, L., Du, C., Liu, M., . . . Li, H. (2013). Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *J Invest Dermatol, 133*(9), 2152-2160. doi:10.1038/jid.2013.21
Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., . . . Fierer, N. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol, 15*(12), 531. doi:10.1186/s13059-014-0531-y
Forbes, J. D., Van Domselaar, G., & Bernstein, C. N. (2016). The gut microbiota in immune-mediated inflammatory diseases. *Frontiers in microbiology, 7*, 1081.
Forsberg, E. M., Huan, T., Rinehart, D., Benton, H. P., Warth, B., Hilmers, B., & Siuzdak, G. (2018). Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nat Protoc, 13*(4), 633.
Foster, K. R., Schluter, J., Coyte, K. Z., & Rakoff-Nahoum, S. (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature, 548*(7665), 43-51.
Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., . . . Huttenhower, C. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A, 111*(22), E2329-2338. doi:10.1073/pnas.1319284111
Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., . . . Xavier, R. J. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol, 4*(2), 293-305. doi:10.1038/s41564-018-0306-4

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput Biol, 8*(9), e1002687. doi:10.1371/journal.pcbi.1002687
Gardinassi, L. G., Xia, J., Safo, S. E., & Li, S. (2017). Bioinformatics Tools for the Interpretation of Metabolomics Data. *Current Pharmacology Reports, 3*(6), 374-383. doi:10.1007/s40495-017-0107-0
Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., . . . Jacob, D. (2014). Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31*(9), 1493-1495.
Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., . . . Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature, 535*(7610), 94-103. doi:10.1038/nature18850
Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., . . . Ley, R. E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. *Cell host & microbe, 19*(5), 731-743. doi:10.1016/j.chom.2016.04.017
Goveia, J., Pircher, A., Conradi, L. C., Kalucka, J., Lagani, V., Dewerchin, M., . . . Carmeliet, P. (2016). Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO molecular medicine, 8*(10), 1134-1142.
Greenblum, S., Turnbaugh, P. J., & Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A, 109*(2), 594-599.
Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome biology, 20*(1), 1-18.
Haidich, A.-B. (2010). Meta-analysis in medical research. *Hippokratia, 14*(Suppl 1), 29.
Hanash, S. M., Pitteri, S. J., & Faca, V. M. (2008). Mining the plasma proteome for cancer biomarkers. *Nature, 452*(7187), 571-579.
Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., . . . Williams, M. (2012). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research, 41*(D1), D456-D463.
Heinken, A., Ravcheev, D. A., Baldini, F., Heirendt, L., Fleming, R. M., & Thiele, I. (2019). Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome, 7*(1), 1-18.
Heintz-Buschart, A., & Wilmes, P. (2018). Human Gut Microbiome: Function Matters. *Trends Microbiol, 26*(7), 563-574. doi:10.1016/j.tim.2017.11.002
Hofmann, A., & Hagey, L. (2008). Bile acids: chemistry, pathochemistry, biology, pathobiology, and therapeutics. *Cellular molecular life sciences, 65*(16), 2461-2483.
Holman, J. D., Tabb, D. L., & Mallick, P. (2014). Employing ProteoWizard to convert raw mass spectrometry data. *Current protocols in bioinformatics, 46*(1), 13.24. 11-13.24. 19.
Hooks, K. B., & O'Malley, M. A. (2017). Dysbiosis and its discontents. *MBio, 8*(5).
Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., . . . Petrosino, J. F. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell, 155*(7), 1451-1463.
Huan, T., Forsberg, E. M., Rinehart, D., Johnson, C. H., Ivanisevic, J., Benton, H. P., . . . Poole, F. L. (2017). Systems biology guided by XCMS Online metabolomics. *Nature methods, 14*(5), 461.

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207-214. doi:10.1038/nature11234
Inkpen, S. A., Douglas, G. M., Brunet, T., Leuschen, K., Doolittle, W. F., & Langille, M. G. (2017). The coupling of taxonomy and function in microbiomes. *Biology & Philosophy, 32*(6), 1225-1243.
Integrative, H. M. P. R. N. C. (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe, 16*(3), 276-289. doi:10.1016/j.chom.2014.08.014
Jackson, M. A., Jeffery, I. B., Beaumont, M., Bell, J. T., Clark, A. G., Ley, R. E., . . . Steves, C. J. (2016). Signatures of early frailty in the gut microbiota. *Genome medicine, 8*(1), 8.
Jansson, J. K., & Baker, E. S. (2016). A multi-omic future for microbiome studies. *Nature microbiology, 1*(5), 1-3.
Jewison, T., Su, Y., Disfany, F. M., Liang, Y., Knox, C., Maciejewski, A., . . . Arndt, D. (2013). SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Research, 42*(D1), D478-D484.
Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology, 17*(7), 451-459.
Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol, 17*(7), 451-459. doi:10.1038/nrm.2016.25
Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular cell biology, 17*(7), 451.
Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., . . . Madsen, K. L. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in microbiology, 7*, 459.
Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic acids research, 49*(D1), D545-D551.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research, 45*(D1), D353-D361.
Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research, 28*(1), 27-30.
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research, 40*(D1), D109-D114.
Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res, 40*(Database issue), D109-114. doi:10.1093/nar/gkr988
Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., . . . Subhraveti, P. (2017). The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform*. doi:10.1093/bib/bbx085
Karp, P. D., Midford, P. E., Caspi, R., & Khodursky, A. (2021). Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC genomics, 22*(1), 1-11.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature, 474*(7351), 327.
Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Shoemaker, B. A. (2015). PubChem substance and compound databases. *Nucleic Acids Research, 44*(D1), D1202-D1213.
Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Res, 44*(D1), D1202-1213. doi:10.1093/nar/gkv951
Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*, 234. doi:10.1186/1471-2105-7-234
Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics, 7*(1), 234.
Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics, 8*, 105. doi:10.1186/1471-2105-8-105
King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., . . . Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research, 44*(D1), D515-D522.
Klassen, J. L. (2018). Defining microbiome function. *Nature microbiology, 3*(8), 864-869.
Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., . . . Li, L. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature medicine, 19*(5), 576-585.
Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyotylainen, T., Hamalainen, A. M., . . . Xavier, R. J. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe, 17*(2), 260-273. doi:10.1016/j.chom.2015.01.001
Krause, S., Le Roux, X., Niklaus, P. A., Van Bodegom, P. M., Lennon, J. T., Bertilsson, S., . . . Bodelier, P. L. (2014). Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in microbiology, 5*, 251.
Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., & Neumann, S. (2011). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry, 84*(1), 283-289.
Kuntz, T. M., & Gilbert, J. A. (2017). Introducing the Microbiome into Precision Medicine. *Trends Pharmacol Sci, 38*(1), 81-91. doi:10.1016/j.tips.2016.10.001
Kurilshikov, A., Wijmenga, C., Fu, J., & Zhernakova, A. (2017). Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends Immunol, 38*(9), 633-647. doi:10.1016/j.it.2017.06.003
Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., . . . Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol, 31*(9), 814-821. doi:10.1038/nbt.2676
Langille, M. G. I. (2018). Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *mSystems, 3*(2), e00163-00117. doi:10.1128/mSystems.00163-17

Lavelle, A., & Sokol, H. (2020). Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature reviews Gastroenterology & hepatology, 17*(4), 223-237.

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends Microbiol, 25*(3), 217-228. doi:10.1016/j.tim.2016.11.008

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics, 12*(1), 253.

Lee, W.-J., & Hase, K. (2014). Gut microbiota–generated metabolites in animal health and disease. *Nature chemical biology, 10*(6), 416-424.

Lee, Y. K., & Mazmanian, S. K. (2010). Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science, 330*(6012), 1768-1773.

Levy, M., Thaiss, C. A., & Elinav, E. (2016). Metabolites: messengers between the microbiota and the immune system. *Genes & development, 30*(14), 1589-1597. doi:10.1101/gad.284091.116

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., . . . Hoffmann, C. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell host & microbe, 18*(4), 489-500.

Li, D., Kirsop, J., & Tang, W. H. W. (2015). Listening to Our Gut: Contribution of Gut Microbiota and Cardiovascular Risk in Diabetes Pathogenesis. *Current diabetes reports, 15*(9), 63. doi:10.1007/s11892-015-0634-1

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., . . . Pulendran, B. (2013a). Predicting network activity from high throughput metabolomics. *PLoS Comput Biol, 9*(7), e1003123.

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., . . . Pulendran, B. (2013b). Predicting network activity from high throughput metabolomics. *PLoS computational biology, 9*(7), e1003123.

Li, S. S., Zhu, A., Benes, V., Costea, P. I., Hercog, R., Hildebrand, F., . . . Voigt, A. Y. (2016). Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science, 352*(6285), 586-589.

Li, X., Watanabe, K., & Kimura, I. (2017). Gut Microbiota Dysbiosis Drives and Implies Novel Therapeutic Strategies for Diabetes Mellitus and Related Metabolic Diseases. *Front Immunol, 8*, 1882. doi:10.3389/fimmu.2017.01882

Li, Z., Lu, Y., Guo, Y., Cao, H., Wang, Q., & Shui, W. (2018). Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Analytica chimica acta, 1029*, 50-57.

Liigand, J., Wang, T., Kellogg, J., Smedsgaard, J., Cech, N., & Kruve, A. (2020). Quantification for non-targeted LC/MS screening without standard substances. *Scientific reports, 10*(1), 1-10.

Liigand, P., Kaupmees, K., Haav, K., Liigand, J., Leito, I., Girod, M., . . . Kruve, A. (2017). Think negative: finding the best electrospray ionization/MS mode for your analyte. *Analytical chemistry, 89*(11), 5665-5668.

Limketkai, B. N., Mullin, G. E., Limsui, D., & Parian, A. M. (2017). Role of vitamin D in inflammatory bowel disease. *Nutrition in Clinical Practice, 32*(3), 337-345.
Lin, H., & Peddada, S. D. (2020). Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms and Microbiomes, 6*(1), 1-13.
Lommen, A., & Kools, H. J. (2012). MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics, 8*(4), 719-726.
Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science, 353*(6305), 1272-1277.
Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature, 489*(7415), 220-230.
Luna, R. A., & Foster, J. A. (2015). Gut brain axis: diet microbiota interactions and implications for modulation of anxiety and depression. *Current opinion in biotechnology, 32*, 35-41.
Machado, D., Andrejev, S., Tramontano, M., & Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research, 46*(15), 7542-7553.
Magnusdottir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., . . . Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol, 35*(1), 81-89. doi:10.1038/nbt.3703
Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current opinion in biotechnology, 51*, 90-96.
Manor, O., & Borenstein, E. (2017). Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. *Cell host & microbe, 21*(2), 254-267.
Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., . . . Magis, A. T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications, 11*(1), 1-12.
Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., & Andres-Lacueva, C. (2018). Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC bioinformatics, 19*(1), 1.
Marcobal, A., Kashyap, P. C., Nelson, T. A., Aronov, P. A., Donia, M. S., Spormann, A., . . . Sonnenburg, J. L. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *Isme Journal, 7*(10), 1933-1943. doi:10.1038/ismej.2013.89
Mariño, E., Richards, J. L., McLeod, K. H., Stanley, D., Yap, Y. A., Knight, J., . . . Rossello, F. J. (2017). Gut microbial metabolites limit the frequency of autoimmune T cells and protect against type 1 diabetes. *Nature immunology, 18*(5), 552-562.
Maurice, C. F., Haiser, H. J., & Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell, 152*(1), 39-50.
Mazmanian, S. K., Round, J. L., & Kasper, D. L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature, 453*(7195), 620.
Mazumdar, V., Amar, S., & Segrè, D. (2013). Metabolic proximity in the order of colonization of a microbial community. *PLoS One, 8*(10), e77617.
Melnik, A. V., da Silva, R. R., Hyde, E. R., Aksenov, A. A., Vargas, F., Bouslimani, A., . . . Alexandrov, T. (2017). Coupling targeted and untargeted mass spectrometry for

metabolome-microbiome-wide association studies of human fecal samples. *Analytical chemistry, 89*(14), 7549-7559.
Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome biology, 20*(1), 1-20.
Mirzaei, M. K., & Maurice, C. F. (2017). Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nat Rev Microbiol, 15*(7), 397.
Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., & Kanehisa, M. (2010). PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res, 38*(Web Server issue), W138-143. doi:10.1093/nar/gkq318
Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., . . . Vazquez-Baeza, Y. (2019). Learning representations of microbe–metabolite interactions. *Nature methods, 16*(12), 1306-1314.
Nash, W. J., & Dunn, W. B. (2018). From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends in Analytical Chemistry*.
Natividad, J. M., & Verdu, E. F. (2013). Modulation of intestinal barrier by intestinal microbiota: pathological and therapeutic implications. *Pharmacol Res, 69*(1), 42-51. doi:10.1016/j.phrs.2012.10.007
Nayfach, S., & Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell, 166*(5), 1103-1116.
Newsom, S. N., & McCall, L.-I. (2018). Metabolomics: Eavesdropping on silent conversations between hosts and their unwelcome guests. *PLoS Pathogens, 14*(4), e1006926.
Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012). Host-gut microbiota metabolic interactions. *Science*, 1223813.
Nielsen, J. (2017). Systems biology of metabolism. *Annual Review of Biochemistry, 86*, 245-275.
Noecker, C., Eng, A., & Borenstein, E. (2021). MIMOSA2: A metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. *Bioinformatics*.
Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., . . . Borenstein, E. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems, 1*(1), e00013-00015.
Noecker, C., McNally, C. P., Eng, A., & Borenstein, E. (2017). High-resolution characterization of the human microbiome. *Translational Research, 179*, 7-23.
O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO reports, 7*(7), 688-693. doi:10.1038/sj.embor.7400731
Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular systems biology, 5*(1), 320.
Paglia, G., Miedico, O., Cristofano, A., Vitale, M., Angiolillo, A., Chiaravalle, A. E., . . . Di Costanzo, A. (2016). Distinctive pattern of serum elements during the progression of Alzheimer's disease. *Scientific Reports, 6*, 22769.
Pang, Z., Chong, J., Li, S., & Xia, J. (2020). MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. *Metabolites, 10*(5), 186.

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., . . . Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic acids research*.

Parada Venegas, D., De la Fuente, M. K., Landskron, G., González, M. J., Quera, R., Dijkstra, G., . . . Hermoso, M. A. (2019). Short chain fatty acids (SCFAs)-mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. *Frontiers in immunology, 10*, 277.

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology, 13*(4), 263-269.

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A., . . . Falony, G. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature, 535*(7612), 376.

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics, 11*(1), 395. doi:10.1186/1471-2105-11-395

Pluskal, T., Castillo, S., Villar-Briones, A., & Orešič, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics, 11*(1), 395.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*(7285), 59-65. doi:10.1038/nature08821

Quercia, S., Candela, M., Giuliani, C., Turroni, S., Luiselli, D., Rampelli, S., . . . Pirazzini, C. (2014). From lifetime to evolution: timescales of human gut microbiota adaptation. *Front Microbiol, 5*, 587. doi:10.3389/fmicb.2014.00587

Quigley, E. M., & Gajula, P. (2020). Recent advances in modulating the microbiome. *F1000Research, 9*.

Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., . . . Luchinat, C. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE [Electronic Resource], 10*(5), e0124219.

Ravikrishnan, A., Nasre, M., & Raman, K. (2018). Enumerating all possible biosynthetic pathways in metabolic networks. *Scientific reports, 8*(1), 9932.

Ren, Z., Li, A., Jiang, J., Zhou, L., Yu, Z., Lu, H., . . . Zhang, R. (2018). Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut*, gutjnl-2017-315084.

Reynolds, L. A., Redpath, S. A., Yurist-Doutsch, S., Gill, N., Brown, E. M., van der Heijden, J., . . . Woodward, S. E. (2017). Enteric helminths promote Salmonella coinfection by altering the intestinal metabolome. *The Journal of infectious diseases, 215*(8), 1245-1254.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet, 16*(2), 85-97. doi:10.1038/nrg3868

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., . . . Billa, V. (2020). An atlas of human metabolism. *Science signaling, 13*(624).

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ, 4*, e2584.

Rojo, D., Mendez-Garcia, C., Raczkowska, B. A., Bargiela, R., Moya, A., Ferrer, M., & Barbas, C. (2017). Exploring the human microbiome from multiple perspectives: factors altering its composition and function. *FEMS Microbiol Rev, 41*(4), 453-478. doi:10.1093/femsre/fuw046

Rooks, M. G., Veiga, P., Wardwell-Scott, L. H., Tickle, T., Segata, N., Michaud, M., . . . Garrett, W. S. (2014). Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. *Isme Journal, 8*(7), 1403-1417. doi:10.1038/ismej.2014.3

Rosen, C. E., & Palm, N. W. (2017). Functional classification of the gut microbiota: the key to cracking the microbiota composition code: functional classifications of the gut microbiota reveal previously hidden contributions of indigenous gut bacteria to human health and disease. *BioEssays, 39*(12), 1700032.

Rost, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., . . . Kohlbacher, O. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods, 13*(9), 741-748. doi:10.1038/Nmeth.3959

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., . . . Bar, N. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature, 555*(7695), 210.

Ryan, P. M., London, L. E., Bjorndahl, T. C., Mandal, R., Murphy, K., Fitzgerald, G. F., . . . Caplice, N. M. (2017). Microbiome and metabolome modifying effects of several cardiovascular disease interventions in apo-E−/− mice. *Microbiome, 5*(1), 30.

Salosensaari, A., Laitinen, V., Havulinna, A. S., Meric, G., Cheng, S., Perola, M., . . . Watrous, J. D. (2021). Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nature Communications, 12*(1), 1-8.

Sanguinetti, E., Collado, M. C., Marrachelli, V. G., Monleon, D., Selma-Royo, M., Pardo-Tendero, M. M., . . . Iozzo, P. (2018). Microbiome-metabolome signatures in mice genetically prone to develop dementia, fed a normal or fatty diet. *Scientific reports, 8*(1), 4907.

Schmidt, T. S., Raes, J., & Bork, P. (2018). The human gut microbiome: from association to modulation. *Cell, 172*(6), 1198-1215.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol, 12*(6), R60. doi:10.1186/gb-2011-12-6-r60

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology, 14*(8), e1002533.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv*, 060012.

Shafquat, A., Joice, R., Simmons, S. L., & Huttenhower, C. (2014). Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends Microbiol, 22*(5), 261-266. doi:10.1016/j.tim.2014.01.011

Sharon, G., Garg, N., Debelius, J., Knight, R., Dorrestein, P. C., & Mazmanian, S. K. (2014). Specialized metabolites from the microbiome in health and disease. *Cell metabolism, 20*(5), 719-730. doi:10.1016/j.cmet.2014.10.016

Silva, Y. P., Bernardi, A., & Frozza, R. L. (2020). The role of short-chain fatty acids from gut microbiota in gut-brain communication. *Frontiers in endocrinology, 11*, 25.

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., . . . Zhu, T. H. (2017). Influence of diet on the gut microbiome and implications for human health. *Journal of translational medicine, 15*(1), 1-17.

Sjögren, K., Engdahl, C., Henning, P., Lerner, U. H., Tremaroli, V., Lagerquist, M. K., . . . Ohlsson, C. (2012). The gut microbiota regulates bone mass in mice. *Journal of bone and mineral research, 27*(6), 1357-1367.

Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., . . . Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring, 27*(6), 747-751.

Son, J. W., Shoaie, S., & Lee, S. (2020). Systems Biology: A Multi-Omics Integration Approach to Metabolism and the Microbiome. *Endocrinology and Metabolism, 35*(3), 507.

Strati, F., Cavalieri, D., Albanese, D., De Felice, C., Donati, C., Hayek, J., . . . Calabrò, A. (2017). New evidences on the altered gut microbiota in autism spectrum disorders. *Microbiome, 5*(1), 24.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

Sukocheva, O. A., Lukina, E., McGowan, E., & Bishayee, A. (2020). Sphingolipids as mediators of inflammation and novel therapeutic target in inflammatory bowel disease. *Advances in protein chemistry and structural biology, 120*, 123-158.

Surana, N. K., & Kasper, D. L. (2017). Moving beyond microbiome-wide associations to causal microbe identification. *Nature, 552*(7684), 244.

Tang, W. W., Wang, Z., Kennedy, D. J., Wu, Y., Buffa, J. A., Agatisa-Boyle, B., . . . Hazen, S. L. (2015). Gut microbiota-dependent trimethylamine N-oxide (TMAO) pathway contributes to both development of renal insufficiency and mortality risk in chronic kidney disease. *Circulation research, 116*(3), 448-455.

Tautenhahn, R., Boettcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics, 9*(1), 504.

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry, 84*(11), 5035-5039.

Tayyari, F., Gowda, G. N., Olopade, O. F., Berg, R., Yang, H. H., Lee, M. P., . . . Mohammed, S. I. (2018). Metabolic profiles of triple-negative and luminal A breast cancer subtypes in African-American identify key metabolic differences. *Oncotarget, 9*(14), 11677.

Tenenbaum, D. (2021). KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). *R package version 1.32.0*.

Thevenot, E. A., Roux, A., Xu, Y., Ezan, E., & Junot, C. (2015). Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. *J Proteome Res, 14*(8), 3322-3335. doi:10.1021/acs.jproteome.5b00354

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal, 474*(11), 1823-1836.
Tian, L., Wang, X.-W., Wu, A.-K., Fan, Y., Friedman, J., Dahlin, A., . . . Liu, Y.-Y. (2020). Deciphering functional redundancy in the human microbiome. *Nature Communications, 11*(1), 1-11.
Tremaroli, V., & Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature, 489*(7415), 242-249.
Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., . . . Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods, 12*(10), 902-903.
Tseng, G. C., Ghosh, D., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research, 40*(9), 3785-3799.
Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., . . . Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods, 12*(6), 523-526. doi:10.1038/nmeth.3393
Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., . . . Affourtit, J. P. (2009). A core gut microbiome in obese and lean twins. *Nature, 457*(7228), 480.
Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., & Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci Transl Med, 1*(6), 6ra14. doi:10.1126/scitranslmed.3000322
Tzoulaki, I., Ebbels, T. M., Valdes, A., Elliott, P., & Ioannidis, J. P. (2014). Design and analysis of metabolomics studies in epidemiologic research: a primer on-omic technologies. *American journal of epidemiology, 180*(2), 129-139.
Uppal, K., Walker, D. I., Liu, K., Li, S., Go, Y.-M., & Jones, D. P. (2016). Computational metabolomics: a framework for the million metabolome. *Chemical research in toxicology, 29*(12), 1956-1975.
Ursell, L. K., Haiser, H. J., Van Treuren, W., Garg, N., Reddivari, L., Vanamala, J., . . . Knight, R. (2014). The Intestinal Metabolome: An Intersection Between Microbiota and Host. *Gastroenterology, 146*(6), 1470-1476. doi:10.1053/j.gastro.2014.03.001
van der Hee, B., & Wells, J. M. (2021). Microbial regulation of host physiology by short-chain fatty acids. *Trends in microbiology*.
Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., . . . Jansson, J. K. (2009). Shotgun metaproteomics of the human distal gut microbiota. *Isme Journal, 3*(2), 179-189. doi:10.1038/ismej.2008.108
Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., & Yanes, O. (2016). Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *Trac-Trends in Analytical Chemistry, 78*, 23-35. doi:10.1016/j.trac.2015.09.005
Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., . . . Blennow, K. (2017). Gut microbiome alterations in Alzheimer's disease. *Scientific reports, 7*(1), 13537.
Wadi, L., Meyer, M., Weiser, J., Stein, L. D., & Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature methods, 13*(9), 705-706.

Walsh, C. J., Hu, P., Batt, J., & Santos, C. C. D. (2015). Microarray meta-analysis and cross-platform normalization: integrative genomics for robust biomarker discovery. *Microarrays, 4*(3), 389-406.
Wang, F., Meng, J., Zhang, L., Johnson, T., Chen, C., & Roy, S. (2018). Morphine induces changes in the gut microbiome and metabolome in a morphine dependence model. *Scientific reports, 8*(1), 3596.
Wang, Q., Wang, K., Wu, W., Giannoulatou, E., Ho, J. W., & Li, L. (2019). Host and microbiome multi-omics integration: applications and methodologies. *Biophysical reviews, 11*(1), 55-65.
Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., . . . Chung, Y.-M. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature, 472*(7341), 57-63.
Wanichthanarak, K., Fan, S., Grapov, D., Barupal, D. K., & Fiehn, O. (2017). Metabox: A toolbox for metabolomic data analysis, interpretation and integrative exploration. *PLoS ONE [Electronic Resource], 12*(1), e0171046.
Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., & Wemheuer, B. (2020). Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome, 15*, 1-12.
Wen, B., Mei, Z., Zeng, C., & Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. *BMC bioinformatics, 18*(1), 183.
Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. *Create Elegant Data Visualisations Using the Grammar of Graphics. Version, 2*(1), 1-189.
Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery, 15*(7), 473.
Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov, 15*(7), 473-484. doi:10.1038/nrd.2016.32
Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., . . . Karu, N. (2017a). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res, 46*(D1), D608-D617.
Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., . . . Karu, N. (2017b). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research, 46*(D1), D608-D617.
Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Dong, E. (2012). HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Research, 41*(D1), D801-D807.
Xia, J. (2017). Computational Strategies for Biological Interpretation of Metabolomics Data. *Adv Exp Med Biol, 965*, 191-206. doi:10.1007/978-3-319-47656-8_8
Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. *Nucleic Acids Res, 40*(Web Server issue), W127-133. doi:10.1093/nar/gks374
Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research, 40*(W1), W127-W133.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res, 40*(W1), W127-W133.
Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res, 37*(Web Server issue), W652-660. doi:10.1093/nar/gkp356
Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research, 37*(suppl_2), W652-W660.
Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res, 37*(suppl_2), W652-W660.
Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic Acids Res, 43*(W1), W251-257. doi:10.1093/nar/gkv380
Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Research, 43*(W1), W251-W257.
Xia, J., Sinelnikov, I. V., Han, B., & Wishart, D. S. (2015). MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res, 43*(W1), W251-W257.
Xia, J., Sinelnikov, I. V., & Wishart, D. S. (2011). MetATT: a web-based metabolomics tool for analyzing time-series and two-factor datasets. *Bioinformatics, 27*(17), 2455-2456.
Xia, J., & Wishart, D. S. (2010a). MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics, 26*(18), 2342-2344.
Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res, 38*(Web Server issue), W71-77. doi:10.1093/nar/gkq329
Xia, J., & Wishart, D. S. (2010b). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research, 38*(suppl_2), W71-W77.
Xu, R., Wang, Q., & Li, L. (2015). A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat. *BMC genomics, 16*(7), 1-9.
Yao, Q., Xu, Y., Yang, H., Shang, D., Zhang, C., Zhang, Y., . . . Han, J. (2015). Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Scientific Reports, 5*.
Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., . . . Anokhin, A. P. (2012). Human gut microbiome viewed across age and geography. *Nature, 486*(7402), 222.
Zhang, G., Dervishi, E., Dunn, S. M., Mandal, R., Liu, P., Han, B., . . . Ametaj, B. N. (2017). Metabotyping reveals distinct metabolic alterations in ketotic cows and identifies early predictive serum biomarkers for the risk of disease. *Metabolomics, 13*(4), 43.
Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., . . . Wang, J. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med, 21*(8), 895-905. doi:10.1038/nm.3914

Zhao, Y., Cong, L., Jaber, V., & Lukiw, W. J. (2017). Microbiome-derived lipopolysaccharide enriched in the perinuclear region of Alzheimer's disease brain. *Front Immunol, 8*, 1064.

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., . . . Vieira-Silva, S. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science, 352*(6285), 565-569.

Zierer, J., Jackson, M. A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., . . . Steves, C. J. (2018). The fecal metabolome as a functional readout of the gut microbiome. *Nat Genet*, 1.

Zmora, N., Suez, J., & Elinav, E. (2019). You are what you eat: diet, health and the gut microbiota. *Nature reviews Gastroenterology & hepatology, 16*(1), 35-56.