Metabolic modeling-based tools for integrative microbiome data analysis

Cecilia Anne Buuck Noecker

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Elhanan Borenstein, Chair

David Fredricks

William Noble

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Metabolic modeling-based tools for
integrative microbiome data analysis

Cecilia Anne Buuck Noecker

Chair of the Supervisory Committee:
Elhanan Borenstein, Associate Professor
Department of Genome Sciences

Complex communities of microbes reside in and on humans, where they closely interact with their hosts by performing a massively diverse array of metabolic reactions. Genomic and metabolomic technologies can now describe both the taxonomic profile of these communities and their metabolic products in unprecedented detail. By measuring both microbial composition and metabolite phenotypes from the same samples, and using the resulting datasets to make and evaluate predictions on the links between microbes and metabolites, it may be possible to infer and characterize metabolic mechanisms occurring in complex natural communities. However, relatively few computational analysis tools exist to integrate and make sense of such "microbiome-metabolome" datasets. In this dissertation, I describe the development and application of methods that use these datasets and reference databases to identify and evaluate relationships between microbes and metabolites. After introducing the current state of knowledge

and available tools in the study of how microbial metabolites impact human health and disease, I present an initial framework for integrating microbiome and metabolomics datasets using metabolic modeling. I demonstrate its ability to predict and explain metabolic shifts in bacterial vaginosis, and further illustrate its application in two case studies, deciphering diet-microbiome interactions in mice and characterizing metabolic mechanisms in the microbiota of children with autism spectrum disorder. In order to compare this approach with alternatives and gain a better understanding of the limiting factors in microbiome-metabolome data analysis, I next describe a comprehensive framework for defining gold-standard mechanistic links between microbes and metabolites and using simulations to evaluate and compare our ability to recover them across different datasets and analysis methods. Finally, informed by the previous applications and evaluations, I introduce MIMOSA2, an updated software tool for inferring mechanistic links from microbiome-metabolome datasets. Together, this work reinforces and advances the utility of metabolic modeling for the analysis and interpretation of large-scale microbiome-metabolome studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without support and contributions from many people:

- Elhanan, for his consistent support, confidence, patience, discerning questions, and unfailing helpful feedback.
- All current and past Borenstein lab members, for teaching me countless new things, asking good questions, and setting excellent examples as computational microbiome scientists.
- Numerous collaborators who collected, shared, explained, and discussed much of the data described below: Sujatha Srinivasan, David Fredricks, Casey Theriot, Colin Brislawn, Janet Jansson, Young-Mo Kim, Tom Metz, Gil Sharon, Sarkis Mazmanian, and others.
- My committee members, for their helpful efforts and insights: David Fredricks, Johanna Lampe, Bill Noble, and Cole Trapnell.
- Everyone in the Genome Sciences department, for contributing to a warm, positive, and cutting-edge scientific environment, and a generally wonderful place to work. Special thanks to my GS grad cohort for being a steady source of encouragement and friendship, and to Brian Giebel and the GS-IT staff for their seamless ability to resolve my minor crises.
- The eScience Institute and the Data Science IGERT program, for funding and training support.
- Members of campus organizations that have enriched and supported my graduate school experience: Women in Genome Sciences, Husky Triathlon, UAW4121, the Genomics Salon.
- My family, in-laws, and friends near and far, for their perpetual understanding and enthusiastic support.
- Finally, I'm very grateful to be on this journey with Micah, without whom it would not be nearly as meaningful or fun.

Chapter 1. INTRODUCTION

Microbes are the most ancient, pervasive and successful forms of life on this planet. Two billion years ago, the production of oxygen gas by photosynthetic cyanobacteria facilitated the establishment of the planet's atmosphere as we know it and the rise of all forms of multicellular life (Holland, 2006). Communities of microbes can be found in virtually every environment, including in and on macro-scale hosts such as humans. Any such community and its habitat are referred to as a *microbiome* (Burge, 1988).

The human microbiome, the set of microbes residing in and on humans, consists of approximately as many microbial cells as there are human cells in the body (Sender et al., 2016). The composition of microbial taxa in the human microbiome is extremely variable across individuals and populations, and has been linked to a vast array of diseases and conditions (Falony et al., 2016; Gilbert et al., 2016). The highest density of microbes is found in the human gut, where a few hundred microbial species interact to play crucial roles in metabolism and immunity. Just as ancient cyanobacteria shaped the atmosphere by producing oxygen, members of host-associated microbiomes modify their environments by performing diverse metabolic reactions and producing varied small molecules known as *metabolites* (Wikoff et al., 2009). These metabolites can act as vital signaling molecules and provide fuel for other microbes and/or their host (Donia and Fischbach, 2015). Recent technological developments have enabled scientists to measure both microbes and their metabolic environment on an unprecedented scale (Franzosa et al., 2015). However, many aspects of the mechanisms by which microbial metabolism can promote human health or disease, as well as the ability to control and modify those mechanisms, remain poorly understood. In this dissertation, I describe the development and application of computational tools

to analyze the relationships between variation in the human microbiome, its metabolic activity, and its impacts on health and disease.

1.1 THE ROLE OF MICROBIAL METABOLISM IN HUMAN HEALTH AND DISEASE

While its renown has greatly increased in the last 15 or so years, human microbiome research has a long history. Antony van Leeuwenhoek examined bacteria scraped from his mouth under an early microscope in 1683 (Lane, 2015). In the early 1900s, Arthur Kendall cultured microbes from fecal samples and published a detailed overview of the field of "intestinal bacteriology", describing the major metabolic reactions performed by gut microbes and their dependence on host diet, and inferring, "the multiplicity of types and the variety of physiological requirements of this intestinal flora are…a strong reminder of the influence which the unrestrained activity of these organisms might conceivably exercise upon the general condition of the host" (Kendall, 1909).

Kendall was correct that the human gut microbiome performs a vast array of metabolic processes that strongly impact its host. These metabolic activities form the basis for a mutualistic symbiosis between humans and their microbes, in which the host provides a regular nutrient supply which the microbes utilize to carry out various beneficial metabolic processes (Walter and Ley, 2011). A recent database of metabolic capabilities of the human host and 818 resident gut microbial species (Noronha et al., 2019) includes over 4,000 microbial reactions that are not performed by human cells. I describe below some non-comprehensive examples of these processes. It is likely that many such processes are yet to be discovered, because many associations with no known mechanism have been observed between microbiome features and host disease or physiology (Gilbert et al., 2016), and moreover many compounds measured in metabolomics assays of the gut microbiome remain unidentified and uncharacterized (Wissenbach et al., 2016).

First, one of the most abundant and significant metabolic contributions of the microbiota is the fermentation of indigestible polysaccharides into short chain fatty acids (SCFAs) that serve as an energy source for gut cells (Donohoe et al., 2011). These SCFAs also play extremely important regulatory roles in the immune system and in energy balance and metabolism, having been identified as regulators of T cell differentiation (Smith et al., 2013b), barrier function (Kelly et al., 2015), and host serotonin production (Reigstad et al., 2014), among other processes. Both the total production of SCFAs and the relative amounts of specific SCFA compounds have been recognized as factors to which host cells are responsive (Perry et al., 2016). Dysregulation of SCFA metabolism has been particularly well-characterized in obesity, diabetes, and metabolic syndrome (Kasubuchi et al., 2015).

Members of the human microbiome synthesize amino acids that can be utilized by the host (Metges, 2000; Smith et al., 2013a) and can also metabolize amino acids via many different pathways, some well-understood and others only recently characterized (Dodd et al., 2017). In the gut, microbial amino acid metabolism can produce compounds with known beneficial impacts: in particular, tryptophan metabolites such as indole perform important signaling functions, in a phenomenon first identified in the 1970s (Agus et al., 2018; Whitt and Demoss, 1975; Wlodarska et al., 2017; Zelante et al., 2013). Notably, some of these compounds, such as serotonin, can be produced by either the host or the microbiome, and the relative contributions from each source in the gut are not fully known (Yano et al., 2015). Some microbial amino acid metabolites have negative effects. For example, in the vaginal microbiome, metabolism of amino acids generally occurs as a component of bacterial vaginosis (Wolrath et al., 2001), producing foul-smelling polyamines including putrescine and cadaverine.

In a clear demonstration of the personalized impact of microbiome composition on human nutrition and disease risk, members of the microbiome produce trimethylamine from

phosphatidylcholine and L-carnitine in red meat (Koeth et al., 2013; Wang et al., 2011). Trimethylamine (TMA) is oxidized by the liver enzyme *FMO* to produce trimethylamine oxide (TMAO), in a process that has been shown to contribute to atherosclerosis in a dose-dependent manner (Qi et al., 2018). Importantly, these pathways are distributed across many microbial taxa but not present universally in human microbiota (Campo et al., 2015; Rath et al., 2017), meaning that characterizing their level of activity and the microbial producers in any given individual may be important for assessing and addressing their risk of cardiovascular disease.

These examples are just a few of the best-understood mechanisms by which the composition and metabolism of the mammalian gut microbiota can modulate health and disease. Beyond these, the composition of the microbiome can also impact the metabolism and mechanism of xenobiotic drugs (Spanogiannopoulos et al., 2016), and produce essential vitamins for the host including folate and riboflavin (LeBlanc et al., 2013). From a broader perspective, microbial ecology researchers have described metabolic ecosystem engineering functions performed by microbial communities across many different hosts and environments (Burgin et al., 2011). Across these examples, it is notable that host responses can be extremely sensitive to the levels of particular microbial metabolites (as seen for SCFAs in particular). Moreover, these metabolites can be established or modified by different microbial taxa in different individuals, meaning that identifying universal microbial drivers may be challenging or futile.

1.2 TOOLS FOR STUDYING MICROBIOMES AND THEIR METABOLISM

The study of the metabolic activities of the human microbiome has rapidly expanded, in large part due to recent technology developments. In particular, the accessibility of large-scale methods for comprehensively profiling microbial taxa, genes, expression, proteins, and metabolites has facilitated documentation of these features across human and non-human populations, disease states, lifestyles, environments, and perturbations. The use of these "omics"

technologies has vastly contributed to our understanding of human microbiome variation, although limitations remain in their interpretability and reproducibility. In this section, I give an overview of the use of several of these technologies in the context of the human microbiome.

The ability to survey and characterize microbial populations in their natural environment is particularly important because so many microbes cannot be isolated and grown in a lab. In particular, the natural habitat of microbial residents of the gut is anaerobic, on a precisely defined pH gradient, extremely dense with nutrients and other microbes (the concentration of bacterial cells in the colon is approximately 10^{11} cells/mL (Berg, 1996)), and shaped by host immune cells (Donaldson et al., 2018), conditions that are clearly challenging to recreate in laboratory culture. One recent study estimated that 81% of microbial cells on the earth are from uncultured genera, although this number is lower for human-associated microbes (Lloyd et al., 2018), and technology development continues to facilitate culturing of new taxa (Lau et al., 2016; Rettedal et al., 2014).

1.2.1 *Profiling microbiome taxonomic composition with amplicon sequencing*

DNA sequence data provides a concise tool for identifying the organisms in an environment. The most widely used approach for characterizing the microbial taxa present in an environment is amplicon sequencing of the 16S rRNA gene. Ribosomal RNA genes were first proposed as markers of phylogenetic divergence between life forms in the 1970s, since they are universally conserved across all forms of life but contain highly variable and continually evolving regions. Carl Woese used an early analysis of sequence divergence between ribosomal sequences to propose the existence of the archaea (Woese and Fox, 1977; Woese et al., 1977). Norman Pace and colleagues then combined this insight with newer sequence technologies to perform the first 16S rRNA microbial surveys (Olsen et al., 1986).

A modern 16S rRNA sequencing assay proceeds by first amplifying a region of the gene using a PCR reaction with primers designed to target highly conserved sequence while spanning

one or more hypervariable regions, and then sequencing the resulting products, most commonly using Illumina technologies (Goodrich et al., 2014). A common approach to analyzing these data is to group sequences into clusters of 97% similarity, known as operational taxonomic units (OTUs). A more recent alternative approach is to simply discard or correct for reads likely to be the product of sequencing errors (Amir et al., 2017; Callahan et al., 2016), and then quantify the abundance of the individual remaining amplicon sequence variants (ASVs). Taxonomic and phylogenetic information about a 16S rRNA dataset can be inferred by comparing sequences against existing databases of ribosomal sequence variation, including Greengenes (McDonald et al., 2012), SILVA (Quast et al., 2013), and the Ribosomal Database Project (Cole et al., 2009). Notably, these assays only capture bacterial and archaeal diversity – other genes must be sequenced to assay eukaryotic microbes or viruses. Additionally, 16S rRNA sequence divergence is not a perfect representation of taxonomic or genomic divergence: some bacterial genomes encode multiple distinct 16S rRNA gene sequences (Tikhonov et al., 2015), and horizontal gene transfer can result in functional differences between bacterial strains with similar or identical 16S rRNA sequences (Brito et al., 2016), although many genomic features can be predicted with high accuracy from 16S rRNA-based taxonomy (Iwai et al., 2016; Langille et al., 2013). Overall, 16S rRNA surveys are an accessible and reliable tool for documenting microbial diversity.

1.2.2 *Metagenomic and metatranscriptomic sequencing of microbiomes*

Instead of amplifying and sequencing a targeted region, some microbiome studies simply sequence all of the DNA present in a microbiome sample, in what is known as shotgun metagenomic sequencing. This method, pioneered for the study of metabolic potential in soil bacterial communities (Handelsman et al., 1998), has been applied to a wide variety of samples and environments. Unlike 16S rRNA sequencing, metagenomics surveys the DNA of all organisms in an environment, and can provide information on the functional capacities of a

microbiome as well as its taxonomic composition. Tools to annotate shotgun metagenomic reads with both of these categories of information have advanced recently (Franzosa et al., 2018a; Greenblum et al., 2015), and their use extracts great detail from these datasets on the relationship between the ecology of a community and its metabolic functions (Manor and Borenstein, 2017). Full microbial genomes can also be assembled from metagenomic reads (Sharon and Banfield, 2013).

Analogously, in a metatranscriptomic study, community RNA is isolated and reverse transcribed, and can then be sequenced and annotated using many of the same methods as for metagenomics. However, metatranscriptomic assays are typically resource-intensive and computationally demanding: microbial mRNA is technically challenging to enrich, and the resulting dataset, consisting of measurements of thousands of genes across many taxa, is highly complex (Klingenberg and Meinicke, 2017).

1.2.3 *Metabolomics for the microbiome*

Beyond genomics, multiple metabolomics technologies can be applied to characterize small molecules from microbiome samples. The field of microbiome science has increasingly recognized the importance of measuring not just the metabolic potential of a community, but also its ultimate metabolite environment and phenotype (Turnbaugh and Gordon, 2008).

These measurements are most commonly made using either gas or liquid chromatography paired with mass spectrometry (GC-MS or LC-MS), although nuclear magnetic resonance (NMR) spectroscopy is also used and can be more precise and reproducible (Cai et al., 2017; Markley et al., 2017). In a typical mass spectrometry metabolomics experiment, chromatography is used to separate components of a complex sample, and the resulting analytes are then ionized and passed through a mass spectrometer, which detects mass-to-charge ratios of the analytes, usually with high resolution (Smith et al., 2014). The mass spectrometer can either be targeted towards

measuring specific compounds of interest, or non-targeted (global) in order to detect as many compounds as possible. The choices of chromatography and ionization methods also affect the ability to measure different classes of compounds. The resulting data are generally processed into a listing of spectral peaks along with their retention times, mass-to-charge (*m/z*) ratios, and peak areas, which indicate concentration or abundance. The next challenge is to chemically identify these peaks, which can be done with varying levels of rigor, from a mass-based search of the *m/z* values to comparisons of spectra against library standards (Sumner et al., 2007). Several competing databases and tools can be used for metabolomics processing and identification (Chong et al., 2018; Kind et al., 2009; Smith et al., 2005; Wang et al., 2016; Wishart et al., 2009), presenting challenges for standardization and comparison across studies and platforms. Importantly, in most complex samples, including human feces and serum, the majority of peaks detected by untargeted mass spectrometry cannot be confidently identified (Pedersen et al., 2016; Yen et al., 2015), although with the continual growth of reference databases, steady progress has been made towards solving this problem (Wishart et al., 2018).

1.3 Documenting, modeling, and manipulating variation in microbiome metabolism

The omics technologies described above can measure taxonomic and metabolic features of any microbiome. However, measuring the component parts of a community does not necessarily produce an explanation of how they are put together, or how they might be modified in order to treat a disease. Addressing those questions requires developing and evaluating predictive models, which is challenging in the context of the microbiome's extensive diversity and variability. Therefore, two significant branches currently exist in the study of human microbiome metabolism: firstly, applying the omics technologies described above to document microbiome features and associations in a growing set of populations and conditions ("top-down"); and secondly, detailed

predictive modeling and experimentation to describe the metabolic activities of simple communities and well-studied taxa ("bottom-up"). Strategies to integrate insights from both of these branches may facilitate the eventual development of microbiome-based optimizations and therapies.

1.3.1 *Multi-omic studies of population variation in microbiome metabolism (top-down)*

First, broad observational studies using omics tools have begun to elucidate enormous diversity and variation in the human microbiome, and its role in health. Metagenomic sequencing studies have uncovered thousands of human-associated microbial strains and hinted that more diversity remains uncharacterized (Pasolli et al., 2019). Large-scale population studies have surveyed the microbiome across broad swaths of individuals, mainly in Western countries (Falony et al., 2016; Huttenhower et al., 2012; Lloyd-Price et al., 2017; McDonald et al., 2018a; Qin et al., 2010; Zhernakova et al., 2016). These survey studies have found that microbiome composition covaries with an assortment of traits and lifestyle factors, including human genetic variation (Goodrich et al., 2016; Rothschild et al., 2018), diet (De Filippis et al., 2014; Wu et al., 2016), medication (Forslund et al., 2015), age (Gibson et al., 2016; Shen et al., 2018), socioeconomic status (Miller et al., 2016; Stagaman et al., 2018), and geography (Vangay et al., 2018; Yatsunenko et al., 2012), among others. Natural fluctuations in gut microbiome features over time have also been tracked at multiple levels of resolution (David et al., 2014; DiGiulio et al., 2015; Flores et al., 2014; Gibbons et al., 2017). Finally and most notably, associations between the microbiome and a wide range of diseases have been observed, suggesting that microbial features may be informative biomarkers of disease onset or progression (Duvallet et al., 2017; Kang et al., 2013; Morgan et al., 2012; O'Keefe, 2016; Wen et al., 2017; Yassour et al., 2016).

Increasingly, these studies include "multi-omic" components, applying more than one of the technologies described above, and in particular combining microbiome taxonomic or

functional profiling with metabolomics assays (De Filippis et al., 2015; Franzosa et al., 2018b; Org et al., 2017; Pedersen et al., 2016; Sinha et al., 2016; Stewart et al., 2017; Zierer et al., 2018). This strategy has also been facilitated by improvements in sample preparation methods for multiple platforms (Melnik et al., 2017). It has been repeatedly recognized that this "microbiome-metabolome" study design may be useful for not just identifying biomarkers, but also understanding the mechanistic implications of microbiome variation on the host (Franzosa et al., 2015; iHMP Research Network Consortium, 2014; Knight et al., 2018; Ursell et al., 2014).

Indeed, multi-omic studies have found strong statistical associations between microbiome features and metabolite levels in feces, serum, and other tissues (Califf et al., 2017; Org et al., 2017; Srinivasan et al., 2015; Zierer et al., 2018). In some cases, these associations may be suggestive of particular mechanisms by which microbes interact with their metabolic environment and with the host (Franzosa et al., 2018b; Lin et al., 2018; McHardy et al., 2013). However, bioinformatic tools to assess such mechanisms in an unbiased manner, and generally for multi-omic data integration in the context of microbial communities (as opposed to single organisms (Rezola et al., 2014)) are relatively scarce, meaning that the mechanisms that give rise to the observed associations are unclear in many cases. Moreover, it is also unclear what study designs and sample sizes may best promote the discovery of functional relationships between microbes, their metabolic activity, and human physiology and disease using these technologies.

1.3.2 *Modeling metabolic mechanisms in the microbiome (bottom-up)*

An alternative approach to studying microbial community metabolism is through detailed modeling and experimentation to understand individual organisms and simple model communities, using computational simulations and model systems.

In particular, constraint-based metabolic modeling approaches (van der Ark et al., 2017; Schellenberger et al., 2011) provide systematic computational frameworks for organizing

knowledge of a microbial species' metabolism and using it to predict its growth, metabolic phenotypes, and/or interactions with other taxa. Several platforms can automatically reconstruct the metabolic capacities of a species from its genome (Arkin et al., 2018; Henry et al., 2010), although using the resulting models to make reliable predictions typically still requires some expert manual curation. Collections of high-quality metabolic models are also becoming available for an increasing number of species (Machado et al., 2018; Magnúsdóttir et al., 2016). The resulting metabolic network models can be analyzed to make predictions using Flux Balance Analysis (Varma and Palsson, 1994) and related approaches, by calculating the maximal growth rate of a species within the constraints of its particular set of metabolic reactions and a specific nutrient environment.

This approach has been used to describe and predict gut microbial metabolic responses to dietary variation (Shoaie et al., 2015) as well as interactions with host metabolism (Heinken and Thiele, 2015a; Mardinoglu et al., 2015). Some progress has been made towards using these models to predict the dynamics and interactions of communities of microbial taxa with differing metabolic capacities (Chan et al., 2017; Chiu et al., 2014; Erbilgin et al., 2017). However, most of these efforts have focused on small-scale model communities, rather than the full diversity of the human microbiome. Many methods have also been developed to integrate individual genome-scale metabolic models with transcriptomic and/or metabolomic data from single organisms grown in culture (Machado and Herrgård, 2014; Sajitz-Hermstein et al., 2016; Vijayakumar et al., 2017), mainly for the purpose of producing more constrained and accurate models. No such integration methods currently exist for community models.

These computational modeling approaches are complemented by experimental model systems, which can validate predictions from both omics studies and computational simulations. These systems include culture-based study of the metabolic and growth capacities of individual

strains (Tramontano et al., 2018) and of small model communities of representative taxa (Biggs et al., 2016). Anaerobic bioreactors are capable of supporting increasingly complex gut communities and can be finely tuned and manipulated (Williams et al., 2015). Finally, experiments with gnotobiotic animals (which are exposed to a controlled set of microbes in a sterile lab environment) are a powerful system for evaluating causality in host-microbiome interactions, at the level of either individual taxa or entire communities (Goodman et al., 2011), although mice are clearly not identical to humans, and the structural differences in their digestive tract can be a potential confounder for such studies (Nguyen et al., 2015).

1.3.3 *Merging top-down and bottom-up approaches to advance rational manipulation of microbiome metabolism*

A stated goal of much of human microbiome research is to develop strategies to modify and control microbial composition in order to optimize health or treat disease (Stegen et al., 2018; Zeevi et al., 2015). This concept has a long history: Eli Metchnikoff proposed in the early 1900s that consuming live bacteria from yogurt could promote health and extend lifespan (Schlesinger and Metchnikoff, 1908). It has gained traction with increased awareness of the microbiome's connections to health and disease, and is supported by the idea that microbiome composition is more transient and responsive than the human genome (Nayak and Turnbaugh, 2016).

General probiotics and fecal microbiome transplants have been moderately successful at modifying the human microbiome to treat infectious and inflammatory diseases (Bisanz et al., 2014; Fuentes et al., 2017; McDonald et al., 2018b; Suez et al., 2018). However, a rationally designed and personalized approach may ultimately prove more fruitful. In a promising example, early stage experiments have found that small molecule inhibitors of microbial TMA-producing enzymes can reduce circulating TMAO levels and correspondingly thrombosis potential in animal models (Roberts et al., 2018). Notably, this therapy would only be relevant to those who possess

this pathway in their microbiome (Rath et al., 2017), and its effectiveness might further depend on the specific taxa involved and their ecological interactions within the community.

Therefore, the broader use of systematic microbiome-targeted therapies requires ongoing improvements in both measurement and simulation technologies, as well as the ability to integrate measurements and models to accurately assess mechanisms. Any rationally designed microbiome manipulation will typically first require establishing a) what microbial process to target, b) which microbes perform the process of interest, and c) whether the answers to a) and b) are universal or (more likely) variable across the human population. Some initial steps have been taken to be able to answer these questions systematically using omics data: for example, a recent study described how the interspecies metabolic interactions, as estimated by community metabolic modeling, differ between colorectal cancer subtypes (Hale et al., 2018). Other studies have used *ad hoc* analysis approaches to microbiome-metabolome data to assess differentially abundant metabolic processes in a set of disease-associated microbiome samples and identify taxa that might be performing them (Franzosa et al., 2018b; McHardy et al., 2013; Pedersen et al., 2016) However, the validity and accuracy of these approaches in the context of complex and variable communities remains undetermined, an important obstacle to any future translational application.

1.4 AIMS OF THIS WORK

Advances in both omics technologies and metabolic modeling present an opportunity to start developing strategies to synthesize insights from both of these research directions, and to infer the key workings of a microbiome from a list of its constituent parts. Validating that such strategies can feasibly and accurately answer mechanistic questions will also require new approaches. The objective of this work is therefore to assess whether metabolic modeling and metabolic reference databases can inform integrative analysis of microbiome-metabolome data, and to introduce broadly applicable conceptual frameworks and software for doing so.

Specifically, I first present an initial modeling approach for linking paired microbiome-metabolome datasets with each other and with metabolic reference data. The method assesses whether measured variation in each metabolite is consistent with expectations based on taxonomic abundances and estimated community metabolic potential, and if so, identifies potential species and reaction contributors to the observed variation. I found that this approach can inform data interpretation for a high number of metabolites, more than expected from randomized reference data, in the vaginal and mouse microbiome, and that this variation displays consistent properties across datasets and environments.

Secondly, I demonstrate the utility of this method, implemented as an R package named MIMOSA, in two additional case studies to inform two different sets of research questions: the first on the linked impacts of diet and microbiome on the metabolome, and the second on metabolic contributions of the microbiota in autism spectrum disorder.

Thirdly, I introduce frameworks for evaluating microbiome-metabolome studies. I describe a metric for quantifying the flux-based contribution of each member of a microbial community to variation in the environmental concentration of any metabolite, and report the calculation of these contributions from dynamic simulation data to identify a set of gold-standard microbe-metabolite links. Using these contributions, I evaluated how well a commonly used analysis approach, species-metabolite correlation analysis, can recover these gold-standard links and how this is influenced by properties of the species, metabolites, and communities in question.

Finally, I describe an improved version of the MIMOSA software for inferring mechanistic links between taxa and metabolites from microbiome-metabolome data. This new tool is informed by the previous case studies and evaluation framework, has been tested for its ability to identify true key contributors from simulated data, and includes additional improvements in usability and compatibility with reference databases.

Chapter 2. METABOLIC MODEL-BASED INTEGRATION OF MICROBIOME TAXONOMIC AND METABOLOMIC PROFILES ELUCIDATES MECHANISTIC LINKS BETWEEN ECOLOGICAL AND METABOLIC VARIATION

This chapter is based on the following manuscript published in *mSystems*:

Noecker, C., A. Eng, S. Srinivasan, C. Theriot, V. Young, J. Jansson, D. Fredricks and E. Borenstein. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems*. *1(1)*:e00013-15, doi:10.1128/mSystems.00013-15, 2016.

Multiple molecular assays now enable high-throughput profiling of the ecology, metabolic capacity, and activity of the human microbiome. However, analyses of such multi-meta-omic data typically focus on statistical associations, often ignoring extensive prior knowledge of the mechanisms linking these various facets of the microbiome. In this chapter I introduce a comprehensive framework to systematically link variation in metabolomic data with community composition by utilizing taxonomic, genomic, and metabolic information. Specifically, I integrate available and inferred genomic data, metabolic network modeling, and a method for predicting community-wide metabolite turnover to estimate the potential of a given community to synthesize and/or degrade individual metabolites. The framework then compares variation in predicted metabolic potential with variation in measured metabolites' abundances to evaluate whether community composition can explain observed shifts in the community metabolome, and to identify

key taxa and gene contributors. Focusing on two independent vaginal microbiome datasets, each pairing 16S rRNA community profiling with large-scale metabolomics, we demonstrate that our framework successfully recapitulates observed variation in a substantial share of metabolites (37%). Well-predicted metabolite variation tends to result from disease-associated metabolism. We further identify several disease-enriched species that significantly contribute to these predictions. Interestingly, our analysis also detects metabolites for which predicted variation negatively correlates with measured variation, suggesting environmental control points of community metabolism. Applying this framework to gut microbiome datasets reveals similar trends, including prediction of bile acid metabolite shifts. This framework is an important first step towards a system-level multi-omic integration and an improved mechanistic understanding of the microbiome activity and dynamics in health and disease.

2.1 BACKGROUND

The human microbiome carries out a plethora of metabolic processes which are often vital to the health of the host. Microbiome metabolic activity can, for example, impact energy harvest, inflammation, and infection susceptibility (Ferreyra et al., 2014; Smith et al., 2013b; Turnbaugh et al., 2006), suggesting that alterations in community metabolism may be an important mechanism underlying an array of poorly understood associations between the composition of the microbiome and disease (Cox et al., 2014; Koeth et al., 2013; Stefka et al., 2014). Indeed, the metabolic capacity of the gut microbiome appears to be relatively constant across healthy individuals (Huttenhower et al., 2012), yet can vary dramatically in response to perturbations such as antibiotic treatment or diet changes (David et al., 2013; Pérez-Cobas et al., 2013) or in a variety of disease states (Morgan et al., 2012; Qin et al., 2012).

Understanding the relationship between the composition of the microbiome and its metabolic activity (and ultimately, the development of microbiome-associated diseases) is

therefore an important task. To this end, numerous recent studies have paired comprehensive taxonomic characterization (based on, for example, 16S rRNA gene assays) with metabolomic profiling, aiming to reveal and evaluate the mechanisms underlying taxonomic and metabolic shifts in the microbiome across diverse environments and disease states (Cowan et al., 2014; Daniel et al., 2014; Edlund et al., 2015; Erickson et al., 2012; He et al., 2015; Kim et al., 2015; Lu et al., 2014; Mao et al., 2014; Marcobal et al., 2013; Shankar et al., 2015; Srinivasan et al., 2015; Theriot et al., 2014; Tong et al., 2014; Walker et al., 2014a; Weir et al., 2013; Zhang et al., 2015). To date, however, methods for integrating taxonomic and metabolomic data are lacking and consequently the vast majority of these studies have analyzed community composition and metabolite profiles independently or focused on identifying statistical associations between these two data types.

While the discovery of such associations is clearly an important first step in describing the function and dynamics of the microbiome in health and disease, association analyses ignore extensive prior knowledge of genomic capacities and metabolic mechanisms that link community ecology and metabolism and may accordingly fall short of gaining a systems-level mechanistic understanding of such complex ecosystems. For example, a strong correlation between a species and a metabolite may have very different interpretations depending on whether the species in question is known to degrade that metabolite or to synthesize it. Integrating the taxonomic and metabolomic profiles of the system under study therefore requires not only linking these two datasets but also the incorporation of prior reference information about the metabolic capacities of various community members and the way such capacities interact. Specifically, an integrated analysis could shed light on the extent to which variation in a metabolite of interest can be explained by observed shifts in community ecology and metabolic capacity as opposed to alternative environmental factors. This is crucial for gaining a comprehensive understanding of the

microbiome and for future efforts to modulate metabolic phenotypes via microbiome-based interventions.

Several recent studies have taken initial steps to address this challenge. One avenue of research aims to reconstruct predictive metabolic models of community metabolism in various settings (using, for example, constraint-based modeling), which can then potentially be validated by metabolomic profiling (Chiu et al., 2014; Heinken and Thiele, 2015a; Shoaie et al., 2015). This approach, however, depends on relatively complete and high-quality metabolic models of the species involved, and may therefore not scale well to complex communities with partially characterized taxa. Other studies have used information about enzymatic reactions to infer metabolic turnover potential from taxonomic composition and metagenome content (Larsen and Dai, 2015; Larsen et al., 2011). Yet, these studies focused on comparing predicted metabolic potential to environmental parameters or community dysbiosis, rather than to detailed, large-scale metabolomic phenotypes. McHardy et al. (2013) used correlation network analysis to cluster metabolites and evaluated the correspondence between the resulting clusters and metabolically related pathway abundances, an approach that successfully quantified relationships between functional pathways and metabolites but that was still primarily association-based and difficult to interpret. Sridharan et al. (2014) similarly focused on a small subset of metabolism, constructing a reference genome-based supraorganism metabolic network model and applying a pathway construction algorithm to predict bioactive aromatic microbial metabolites likely to be found in the human gut. These studies all show the tremendous promise of linking microbial composition to metabolomic variation based on prior knowledge of the various metabolic processes, yet are still limited in scale and call for the development of a systematic, mechanistic approach for evaluating the relationships between the community ecology and metabolite shifts.

Here, we therefore present a comprehensive analytical multi-omic framework for integrating community structure and metabolic profile, aiming to elucidate mechanisms underlying metabolic variation in the human microbiome. This framework first infers community gene content based on available and inferred genomic information, and adapts a method originally developed to interpret environmental metagenomes (Larsen et al., 2011) to approximate the potential effect of the microbiome on each metabolite. We systematically compare these estimates to measured metabolome variation and interpret the results in terms of metabolic mechanisms based on taxonomic shifts. We apply this framework to two datasets pairing community taxonomic composition and global metabolite profiles from the vaginal microbiome, as well as to datasets from the gut microbiome of humans and mice. Using this framework, we identify a large number of metabolites whose variation across samples can be explained (or "predicted") by shifts in microbial community composition and the metabolic capacity of the various member species. We further use this approach to identify species and reactions that are key contributors to the calculated community-wide metabolic potential, and highlight putative alternative mechanisms for poorly predicted metabolites. Importantly, our analysis detects broad trends in metabolite predictability across datasets and serves as a proof-of-concept of the use of systematic mechanism-based integration of multi-omic data to gain new insight into microbial community metabolism.

2.2 METHODS

2.2.1 *Assembling and processing datasets*

We obtained several previously published datasets (Erickson et al., 2012; Jansson et al., 2009; Jozefczuk et al., 2010; Srinivasan et al., 2015; Theriot et al., 2014) from publicly available databases or through a collaboration, each pairing 16S rRNA gene-based taxonomic data with

metabolomic profiles. For vaginal samples, DNA was extracted for 16S rRNA gene analysis from vaginal swabs, and cervicovaginal lavage fluid was collected for metabolomic analysis. Samples from the first dataset (Dataset 1) were analyzed for taxonomic composition using quantitative PCR (qPCR) with primers and probes specific for 14 vaginal bacterial species, and for metabolites using global liquid (LC-MS) and gas chromatography (GC-MS) coupled with mass spectrometry for metabolomics. Samples from the second dataset (Dataset 2) were analyzed by using broad-range 16S rRNA gene PCR coupled with high-throughput 454 sequencing of the 16S rRNA gene (Roche 454), and targeted metabolomics using LC-MS with multiple reaction monitoring for 180 compounds, chosen partially based on findings from Dataset 1. In Dataset 3, taxonomic composition was assayed using 454 FLX Titanium sequencing of V3-V5 regions of the 16S rRNA gene, and metabolites were measured using global LC-MS and GC-MS metabolomics. Dataset 4 paired Roche 454 shotgun sequencing of sample DNA with FT-ICR-MS metabolomics. See Table A1 for details about each dataset. Metabolite and transcript (microarray) data for the E. coli dataset were downloaded, profiling E. coli grown in culture, treated with five different stress-based perturbation, and assayed before and after perturbation. We included only one time point before perturbation and one immediately after for each biological replicate in our analysis. We mapped identified metabolite names to KEGG IDs following the same approach as for Dataset 2.

We re-processed 16S rRNA taxonomic sequencing datasets via a standard closed-reference OTU picking pipeline using QIIME v1.8.0 (Caporaso et al., 2010a, 2010b; DeSantis et al., 2006; Edgar, 2010; McDonald et al., 2012) and rarefied the resulting OTU tables to the number of reads in the lowest coverage sample. For Dataset 2, we confirmed that this pipeline produced similar taxonomic profiles as the pplacer method used in the original publication (Pearson correlation across samples of all genera quantified by both methods was 0.97). We did not normalize the 16S rRNA gene qPCR data of Dataset 1. A subset of samples in Dataset 1 also had associated 16S

rRNA gene sequencing data, on the basis of which we removed one outlier sample whose sequencing results were dominated by a species not profiled by qPCR. For the E. coli gene expression data, we used the KEGG API to map gene IDs to KEGG orthology groups (KOs) for consistency with the other datasets, and used the provided normalized microarray intensities.

Processing of metabolomic data varied depending on the technology used. For Datasets 1 and 3, in which metabolomic profiles were produced by Metabolon, Inc. and included detailed metabolite identifications, we filtered out compounds without KEGG identifications and used the raw peak area values. For Dataset 2 (generated at the Northwest Metabolomics Research Center) and the E. coli dataset, we used the KEGG API (Kanehisa and Goto, 2000) to associate named and measured compounds with KEGG metabolite IDs. For Dataset 4, which lacked confident library-based metabolite identifications, we used MetaboSearch (Zhou et al., 2012) to perform a mass-based search against the Metlin, MMCD, LipidMaps, and HMDB databases (Cui et al., 2008; Smith et al., 2005; Sud et al., 2007; Wishart et al., 2009), with a matching threshold of 1 ppm. The KEGG identifications with the smallest mass difference were assigned as the putative identification, following (McHardy et al., 2013; Tong et al., 2014). When multiple putative identifications had the same difference in mass from a peak, preference was given to metabolites in the metabolic network generated based on species abundances, using genomic information as additional support for the presence of that metabolite. When multiple putative KEGG identifications remained, one was randomly assigned to that peak. If multiple peaks mapped to the same KEGG metabolite ID, their abundances were summed. Metabolites with non-zero abundance in <5 samples were discarded from downstream analysis.

2.2.2 *Predicting metagenome content from taxonomic composition*

For Datasets 2 and 3 we used PICRUSt (Langille et al., 2013) to predict metagenome content across samples based on taxonomic composition, and normalized the resulting predictions

using MUSiCC (Manor and Borenstein, 2015) (Release 1.0). To predict genome content from the qPCR species abundances in Dataset 1, we searched IMG for available reference genomes. For 11 out of the 14 species profiled, at least one reference genome was available. When multiple reference genomes were available for a given species, we selected either the highest quality genome or a genome from a vaginal isolate (in consultation with the researchers that generated this dataset). See Table S2 for details about the genomes used. We downloaded KEGG orthology (KO) annotations for these genomes from IMG (January 2014) and predicted the metagenome as a product of the reference genome KO annotations and species abundances. For Dataset 4, orthology group abundances were estimated directly from shotgun sequencing reads using a BLAST-based annotation pipeline (Carr and Borenstein, 2014).

2.2.3 *Metabolic network reconstruction and CMP score calculation*

We adapted the PRMT (Predicted Relative Metabolomic Turnover) method developed by (Larsen et al., 2011) to estimate the metabolic potential of a microbial community based on measurements of gene content. This method does not predict metabolite fluxes or concentration directly; instead, it synthesizes and integrates information about gene abundances in terms of KEGG orthology groups and a stoichiometric matrix describing quantitative relationship between genes and metabolites to provide an estimate of the way community composition may impact each metabolite's abundance. To this end, we first created a modified stoichiometric matrix M in which each row represents a particular metabolite and each column represents a particular gene (KO), such that each cell M_{ij} represents the combined relative capacity for enzymatic gene j to modify metabolite i (see (Larsen et al., 2011)). To create this matrix, we utilized pathway reaction information and stoichiometric coefficients from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Specifically, for each reaction catalyzed by an enzyme coded by gene x that transforms metabolite A into metabolite B with stoichiometric coefficients c and d

respectively, we subtract c from M_{Ax} and add d to M_{Bx}. To focus our analysis on the primary transformations catalyzed by each enzyme, we only linked genes to reactions and metabolites that are annotated in KEGG metabolic pathways, using the *reaction_mapformula.lst* file from the KEGG database (2013 version). We then filtered this matrix to only include reactions annotated as occurring in a single direction, ignoring all reversible reactions that cannot contribute to metabolite predictions and all metabolites involved in only reversible reactions. Lastly, we performed two additional modifications: First, following previous studies (Greenblum et al., 2012; Taxis et al., 2015), we excluded "currency" metabolites that are involved in reactions associated with 30 or more genes from the final matrix. Second, following (Larsen et al., 2011), we normalized each row of M such that all negative elements sum to -1 and all positive elements sum to 1. The resulting matrix accordingly estimates the relative contribution of each gene to the accumulation or depletion of each metabolite. We then multiply this M matrix with a matrix G that represents the abundance of each gene in each sample to obtain community-wide metabolic potential (CMP) scores, capturing the relative capacity of the metagenome content of each sample to create or deplete each metabolite.

2.2.4 *Comparing CMP scores with metabolomic data*

Notably, since CMP scores represent relative predictions, they can only be interpreted in the context of comparisons between samples (assuming some baseline metabolite profile across samples). Accordingly, to assess how the obtained CMP scores compare to measured metabolomic variation, we performed a Mantel test for each metabolite, assessing the correlation between pairwise differences (across all pairs of samples) in CMP scores and the corresponding pairwise differences in measured metabolite abundances. We further corrected for multiple hypothesis testing using a local FDR approach implemented in the R package *qvalue* (Dabney and Storey, 2015), and classified metabolites with both a Mantel *p-value* and FDR *q-value* less than 0.01 as

well-predicted. We evaluated the significance of negative pairwise correlations in a similar manner, classifying metabolites as anti-predicted based on the same significance cutoffs.

2.2.5 *Testing significance with randomly shuffled networks and metabolite labels*

Given the covariance structure of the dataset, we also wished to quantify whether our framework identified more well-predicted metabolites than expected by chance. To this end, we repeatedly generated randomized metabolic networks, ran our framework as detailed above using the randomized network to link genes to metabolites, and compared the number of well-predicted metabolites obtained with these randomized networks to the number of well-predicted metabolites obtained with the original network. To preserve the core structural characteristics of the original network, random networks were generated following the edge shuffling approach outlined in (Milo, 2002) (exchanging edges 5000 times to produce each network).

We also used a permutation-based approach to evaluate whether anti-predicted metabolites are linked by a metabolic reaction to well-predicted metabolites more frequently than expected by chance. To this end, we repeatedly permuted the labels of every metabolite in the network while maintaining a fixed network topology. We counted the number of times an anti-predicted metabolite was connected to a well-predicted metabolite by a synthesizing reaction, a depleting reaction, or a reversible reaction in these permuted networks and compared the resulting distribution with the numbers obtained using the original data.

2.2.6 *Identifying key species and gene contributors*

To quantify the contribution of each species to the calculated CMP scores of each metabolite and to identify key species contributors, we examined the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the entire community and the CMP scores calculated based on each species by itself (i.e., recalculating the metagenome

content and CMP scores based solely on the abundance of each species separately). Species for which this correlation coefficient for a given metabolite was >0.5 were considered key species contributors for that metabolite.

To compare key species contributors between Datasets 1 and 2, we identified corresponding species across the two datasets by searching the Greengenes 97% OTU representative sequence set for exact matches with the PCR primers used by (Srinivasan et al., 2015) to generate Dataset 1. Notably, this mapping identified OTU 4377809 as *Mageeibacillus indolicus* (previously known as BVAB3), OTU 227000 (mistakenly characterized as *Shuttleworthia*) as BVAB1, and OTU 133178 as *Eggerthella sp 1*.

To identify key gene contributors to the calculated CMP scores of for each metabolite, we examined the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the original stoichiometric matrix and the CMP scores calculated when using a matrix in which the link between the gene in question and the metabolite was deleted (i.e., zeroing the corresponding entry in the matrix). Genes for which this correlation was <0.5 were considered key contributors for that metabolite. We further defined any reaction catalyzed by the enzyme coded by that gene as a key reaction contributor. In addition, if all of the key reaction contributors of a given metabolite produce that metabolite, we classified that metabolite's CMP scores as driven primarily by synthesis. We similarly classified a metabolite whose key reaction contributors all deplete that metabolite as driven primarily by degradation.

2.3 RESULTS

2.3.1 *A metabolic model-based framework for integrating taxonomic and metabolomic data*

We developed a computational framework to systematically link variation in community ecology with observed variation in its metabolic phenotype (Figure 2.1). Our framework

specifically assesses whether the measured between-sample variation in metabolite abundances can be explained by observed shifts in species composition and information about the metabolic capacity of each species.

Briefly, our framework first infers the metagenome content for each sample based on taxonomic composition and available or inferred reference genome information (Langille et al., 2013). Inferred metagenomes are then normalized using a previously introduced method (MUSiCC; (Manor and Borenstein, 2015)), resulting in an estimate of the average copy number of each gene across microbiome genomes. Next, our framework applies a method for predicting relative metabolic turnover (Larsen et al., 2011), using a metabolic network model to translate the resulting enzymatic gene abundance estimates into community-based metabolite potential (CMP) scores. These scores represent the relative capacity of the community in a given sample to generate or deplete each metabolite, based on metabolic reference information that links enzymes to their substrates and products (Kanehisa and Goto, 2000). To evaluate these scores, our framework then compares for each metabolite the differences in CMP scores between all pairs of samples with the differences in the corresponding measured metabolite abundance. Using these pairwise comparisons and a statistical test for correlation between two distance matrices, our framework evaluates whether there is an agreement between variation in predicted CMP scores and variation in measured metabolite abundances. We term those metabolites for which this agreement is statistically significant “well-predicted”. Finally, our framework uses a perturbation-based approach to identify the bacterial species, genes, and reactions that are the key mechanistic contributors to calculated CMP scores. A more detailed description of this framework can be found in the Materials and Methods section.

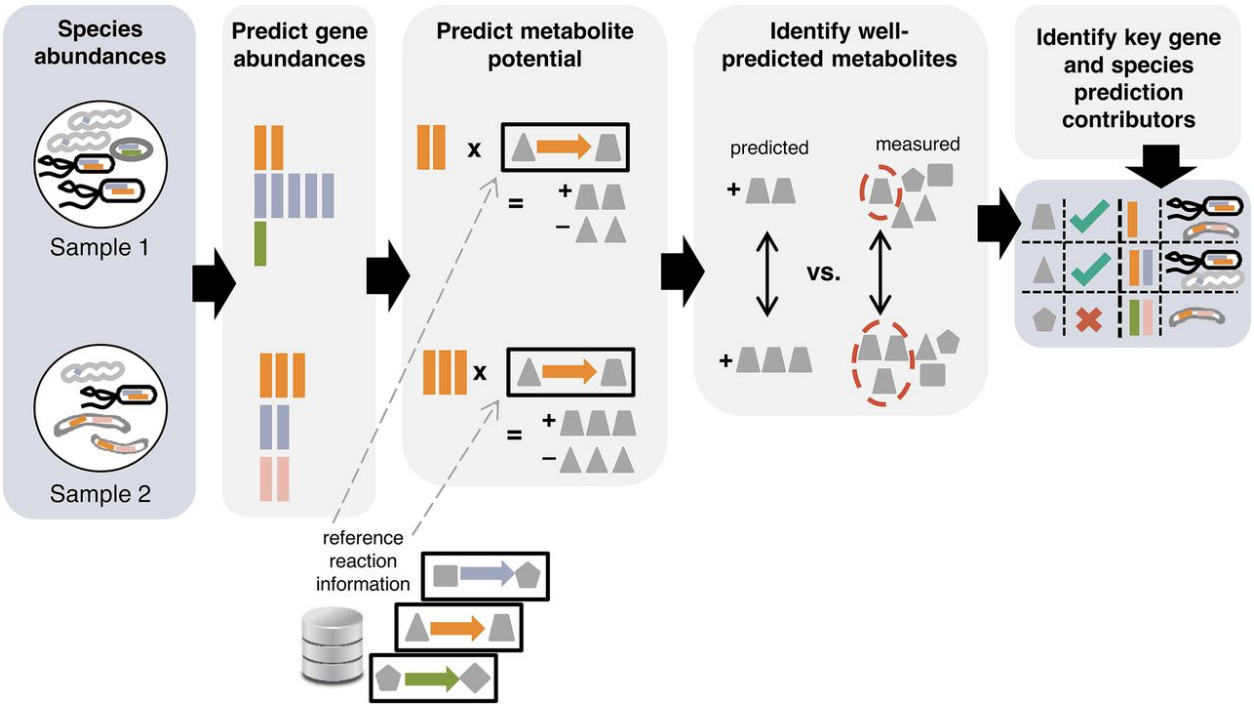

Figure 2.1. **Framework for integrating taxonomic and metabolomic data.**

Species composition is first used to predict the metagenome's gene content, which is then paired with reaction information to estimate the community metabolic potential (CMP) for each sample and metabolite. Variation in predicted CMP scores is compared to variation in measured metabolite abundances (using pairwise differences) to identify well-predicted metabolites. A perturbation-based approach is used to additionally identify key species, gene, and reaction contributors to CMP scores.

2.3.2 *Metabolic model-based prediction explains metabolite variation in the vaginal microbiome based on taxonomic shifts*

We first applied our framework to datasets pairing bacterial community and metabolomic profiles from the vaginal microbiome, a relatively simple community typically dominated by a limited number of species. We specifically analyzed two independently obtained datasets (each consisting of ~70 samples; Table A1), characterizing the vaginal microbiome and metabolome of healthy women and women with bacterial vaginosis (BV) (Srinivasan et al., 2015). Samples from the first dataset (Dataset 1) were analyzed for taxonomic composition using quantitative PCR (qPCR) for 14 vaginal bacterial species and for metabolites using global LC-MS and GC-MS,

whereas samples from the second dataset (Dataset 2) were analyzed using broad-range 16S rRNA gene sequencing and targeted LC-MS (see Methods).

In each of these datasets, we used our framework to calculate the CMP score for each metabolite and in each sample. Of the metabolites assayed in each dataset, roughly 50% could not be associated with a CMP score due to missing or non-informative linked reference gene and reaction data, and were accordingly discarded from downstream analysis. CMP scores of the remaining metabolites were compared to measured metabolite abundances as described above to examine whether the observed variation in the metabolite abundances across samples can be explained mechanistically by variation in the set of species comprising the community. Surprisingly, we found that 40.2% of the analyzed metabolites in Dataset 1 and 34.5% of metabolites analyzed in Dataset 2 were well-predicted, suggesting that for a substantial fraction of metabolites information about the metabolic capabilities for the member species is sufficient to explain observed differences in metabolite abundance. We further confirmed that the identification of well-predicted metabolites and the correlations observed between calculated CMP scores and measured abundances are not artifacts of the data covariance structure, using randomized metabolic networks to generate a predictability null model. We found that randomized networks produced a consistently lower proportion of well-predicted metabolites compared to the real network ($p<0.01$ for both datasets). Metabolites analyzed in both datasets were generally predictable at similar levels ($\rho=0.63$, Spearman correlation test; Figure 7.1). Finally, we also observed a significant overlap between metabolites for which variation in CMP scores was significantly correlated with variation in measured metabolite abundance in both Datasets 1 and 2 and in a simple monoculture-based *E. coli* dataset ($p=0.04$; Fisher exact text; see Appendix A and Figure 7.2). This finding suggests that our framework may identify consistent control points in microbial metabolism.

We next examined whether well-predicted metabolites tend to be associated with specific metabolic categories or host state. We found that well-predicted metabolites spanned a range of metabolic categories (Figure 2.2A). Specifically, well-predicted metabolites represent all major metabolic categories, with many well-predicted metabolites associated with amino-acid metabolism, an important category of microbe-mediated processes in this environment. Additionally, 60% and 40% of the strongly BV-enriched metabolites, including known metabolic markers of BV (Wolrath et al., 2001) such as the amino acid catabolites N-acetylputrescine, spermidine, and citrulline, were predicted well in each dataset respectively (Figure 2.2B).

Interestingly, we also observed a substantial portion of metabolites in which variation in CMP score was strongly negatively correlated with variation in measured abundances (25.6% in Dataset 1 and 29.3% in Dataset 2). These anti-predicted metabolites were often linked to a well-predicted metabolite either by a reversible reaction (which is not factored into CMP score calculation; 7 and 4 metabolite pairs in Dataset 1 and 2 respectively) or by a reaction synthesizing the anti-predicted metabolite from a well-predicted metabolite (6 and 2 metabolite pairs). For example, in Dataset 1, glutamate is well-predicted while glutamine, a metabolite that can be synthesized from glutamate, is anti-predicted, suggesting that other unaccounted-for factors influence its abundance in this environment. Overall, anti-predicted metabolites were adjacent to well-predicted metabolites more frequently than expected by chance (15 and 8 metabolite pairs, $p < 0.005$ and $p < 0.03$ in Dataset 1 and 2 respectively, permutation-based test). Such anti-predicted metabolites may be the result of missing information about community composition or genomic capacities. However, it may also point to environmentally-regulated points in metabolism (as opposed to microbiome-controlled metabolites), where an environmental change in metabolite abundance and nutrient availability give rise to taxonomic shifts in the microbiome. Put differently, whereas for well-predicted metabolites an increase in their abundance correlates with an increase

in the abundance of species with the capacity to synthesize them, increase in the abundance of anti-predicted metabolites can be potentially introduced by the environment, leading to selection for species with the capacity to degrade them (and see also Discussion).

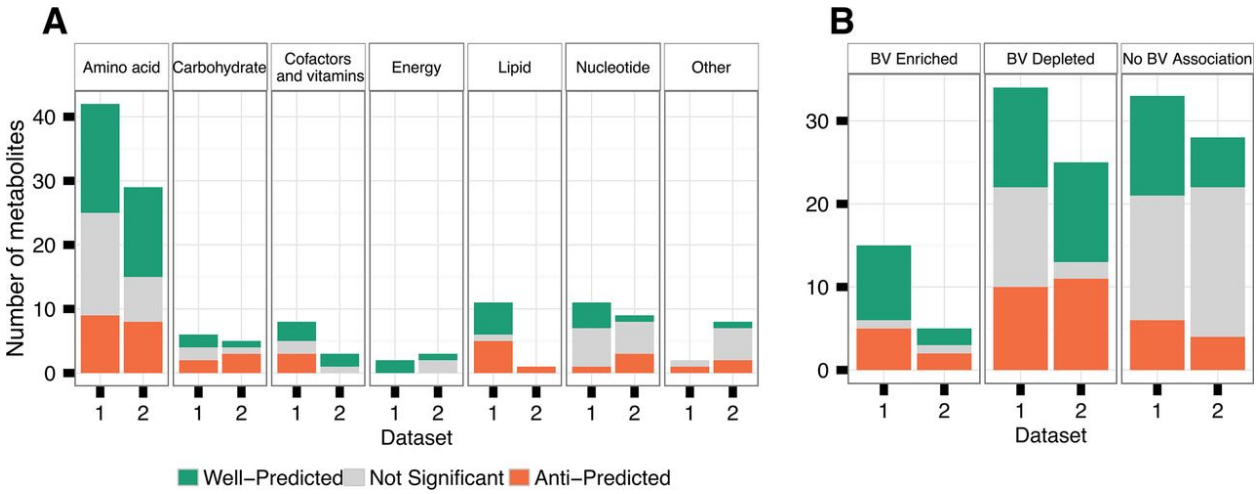

Figure 2.2. **Metabolite predictability across metabolic categories (A) and disease states (B) in the vaginal microbiome.**

Well-predicted metabolites are defined as those for which variation in CMP scores is significantly correlated (using a Mantel test) with variation in measured metabolite abundance at a false discovery rate (FDR) of 0.01. Anti-predicted metabolites are similarly defined as those for which variation in CMP scores is significantly negatively correlated with variation in measured metabolite abundances (FDR 0.01). Metabolic categorization is based on KEGG data, and disease enrichment is based on a Wilcoxon rank sum test for association with bacterial vaginosis (BV) with a Bonferroni-corrected P value of <0.1.

2.3.3 *A small set of BV-enriched bacterial species explains a large portion of metabolome variation*

We next examined the contribution of individual species to the calculated CMP scores of each metabolite. We quantified each species' contribution as the correlation between a CMP score that is calculated based on that species alone (e.g., ignoring all other species in the community) with the community-wide CMP score described above. We defined species for which this correlation was above 0.5 as key contributors. We use an effect size cutoff for this designation in

order to maintain a consistent definition across datasets of varying sample sizes, and because the significance of the overall correspondence between CMP scores and metabolite abundances has already been assessed. We first focused on Dataset 1, in which only a small number of species was assayed but the availability of absolute concentration data (owing to the use of qPCR) may better distinguish key species in the community. In total, we found that 10 out of the 11 species profiled in Dataset 1 were key contributors to at least one metabolite. Importantly, the vast majority of metabolites (93.9%) had 4 or fewer key contributors, yet the particular combination of species varied widely across metabolites. This suggests that shifts in the abundance of each metabolite (and in particular shifts associated with the BV state) may be attributed to a small number of species rather than to community-wide dysbiosis. For instance, although both N-acetylputrescine and citrulline are BV-enriched polyamine metabolites, the increased abundance of N-acetylputrescine in BV is driven mostly in both datasets by the genomic capacities and variation in the abundance of *Prevotella* species, while citrulline's enrichment is driven primarily by *Atopobium vaginae* and *Eggerthella*. Species contributing to the CMP scores of anti-predicted metabolites also recover known processes: for example, *Lactobacillus iners* is the only key species contributor driving the anti-prediction of glycerol in Dataset 1 (due to *L. iners* genome encoding glycerol utilization genes). A recent metatranscriptomic study of vaginal *L. iners* found evidence that this species is the only member of this community known to use glycerol as a carbon source (Macklaim et al., 2013), which combined with our results suggests that a vaginal environment with glycerol availability may promote *L. iners* growth.

We further examined the number of metabolites (and specifically well-predicted metabolites) for which each species was a key contributor to CMP score calculation. We found that in Dataset 1, *Eggerthella sp. 1* and *Megasphaera* type 1 were key contributors to a particularly high number of metabolites, relative to other species (Figure 2.3). BV-enriched metabolites

predicted well primarily by these two species alone include N-acetylneuraminate, ethanolamine, and the lipid metabolites 4-trimethylaminobutanoate/gamma-butyrobetaine and 3-methyl-2-oxobutanoate (see also Figure 7.3). Notably, these are neither the most abundant nor the most variable species in this dataset, although *Eggerthella sp. 1* is the most differentially abundant species between healthy and BV samples based on Wilcoxon rank-sum tests ($p < 10^{-8}$), whereas *Megasphaera* is fifth-most differentially abundant ($p < 10^{-7}$). *Eggerthella* also has the largest genome of any of the profiled species in terms of the number of protein coding genes (2936 genes). Combined, these findings illustrate that the species contributing most significantly to potential shifts in disease-associated metabolic phenotypes may not necessarily be the most abundant or most variable species, and that observed metabolic shifts are the product of complex dependencies between ecological dynamics and metabolic capacity.

These trends are partially recapitulated in Dataset 2 (Figure 7.4). Specifically, 31 out of the 171 OTUs in this dataset were key contributors to at least one metabolite. Again, most metabolites (62%) had 4 or fewer key contributors, but the combination of OTUs varied across metabolites. Out of the 42 metabolites analyzed in both datasets, 24 share at least one key contributing genus, including 9 out of the 11 metabolites predicted well in both datasets (Figure 7.4B). Interestingly, however, the OTUs that contributed to the CMP scores of the most well-predicted metabolites in Dataset 2 included an OTU (4377809) corresponding to the bacterium *Mageeibacillus indolicus* (previously known as BVAB3), a *Prevotella* OTU (403822), and an OTU (227000) identified as BVAB1 (of which, only *M. indolicus* was analyzed in Dataset 1). An OTU corresponding to the *Eggerthella* species noted in Dataset 1 was also a key contributor to many well-predicted metabolites. Relatively low contributions to CMP scores by *Lactobacillus crispatus* (typically associated with health) and *Atopobium vaginae* were consistent between the two datasets. Given the differences in taxonomic profiling methods between the two datasets (qPCR versus 16S rRNA

gene sequencing), difference in the way genomic content was inferred (reference genomes versus PICRUSt-based predictions), missing reference genomic information for three species assayed in Dataset 1, and the focus on selected metabolites of interest in Dataset 2, variation in the obtained key contributors is perhaps not surprising. For example, the increased importance of *M. indolicus* in Dataset 2 could be a function of differences in the features of metabolites assayed and analyzed between the two datasets, or from differences in reference information; a total of 88 out of 732 KOs differed in copy number between the reference genome used for prediction in Dataset 1 and the predicted genome content for the corresponding OTU in Dataset 2.

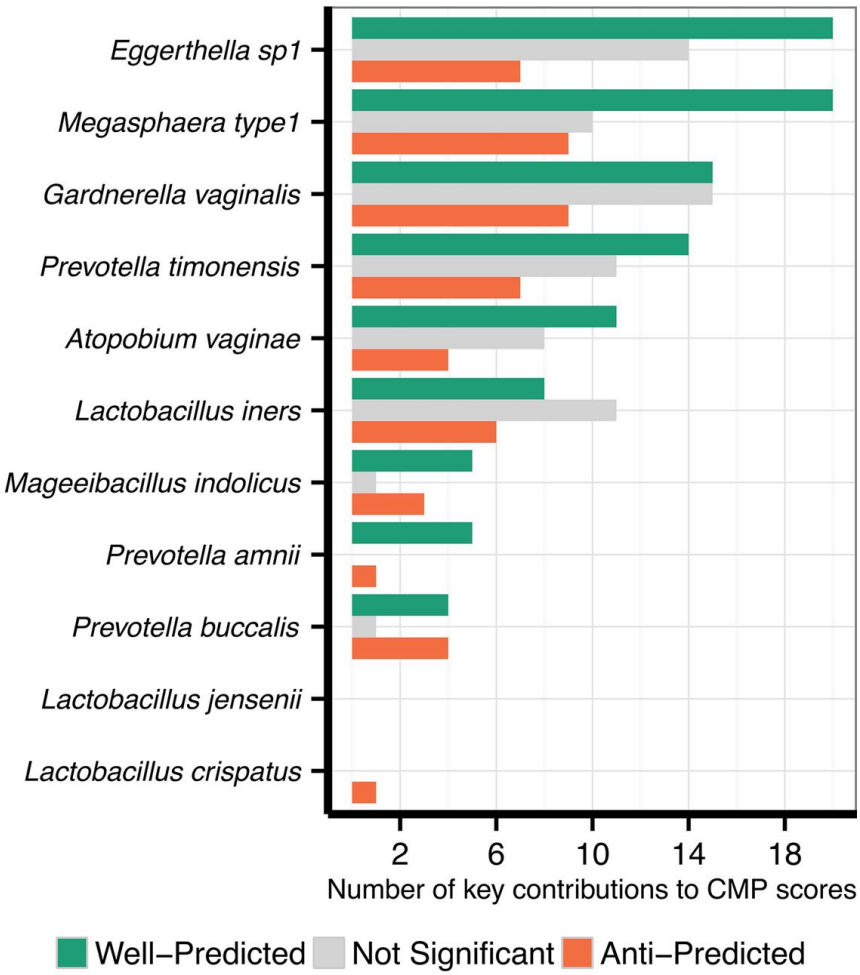

Figure 2.3. **Key species contributors to metabolites in the vaginal microbiome.**

Each species that participated in the calculation of CMP scores in data set 1 is shown along the y axis. The x axis indicates the numbers of well-predicted and anti-predicted metabolites (as

well as those with nonsignificant predictions) for which that species was a key contributor (see Methods).

2.3.4 *Well-predicted metabolites tend to be involved in condition-specific metabolism.*

We next set out to identify key gene contributors to each metabolite's CMP score, by calculating the correlation between the original CMP scores and a CMP score calculated when the link between the gene in question and the metabolite is deleted from the metabolic model (Methods). Genes for which this correlation was less than 0.5 were considered key contributors for that metabolite and any reaction catalyzed by the enzyme encoded by that gene was considered a key reaction contributor. This analysis relates specific combinations of reaction information and genomic shifts to the predicted potential for metabolite variation, allowing us to examine whether our approach recovers known metabolic mechanisms (Figure 7.3). For example, CMP scores for well-predicted amino acid derivatives including N-acetylputrescine and citrulline were driven by synthesis enzymes forming part of amino acid catabolism pathways and encoded by BV-associated bacteria (Figure 7.3A,B). A subset of amino acids, including glutamate and phenylalanine, were predicted well on the basis of a combination of available biosynthesis pathways and predicted abundance of tRNA synthetase genes and degradation pathways. Pyruvate levels were slightly lower in BV samples and predicted well primarily by acetolactate synthase, which catalyzes the first step diverting pyruvate to branched chain amino acid synthesis. This mechanism is consistent with the overall shift from carbohydrate-based to amino acid-based metabolism typical of the BV state. In another example, (Srinivasan et al., 2015) have noted that the depletion of reduced glutathione in BV samples is surprising as the BV vagina is a relatively reduced environment (Holmes et al., 1985). Our framework predicts this shift in glutathione well in both datasets (prediction levels of 0.49 and 0.30 in Datasets 1 and 2, respectively), and attributes it to a lack of

glutathione peroxidase genes in *Lactobacillus* species that predominate in healthy vaginal samples (Figure 7.3C). Genes in cofactor synthesis pathways also tended to contribute to predictive CMP scores for metabolites in these pathways including nicotinate, NAD+ and FAD+. In yet another example, the depletion of NAD+ in BV samples is explained primarily by genes in a NAD+ salvage pathway found selectively in *Prevotella* species.

We further characterized the set of key gene contributors of each metabolite and explored their relationship with metabolite predictability (Figure 7.5). Most metabolites had only a small number of genes with the potential to enzymatically impact them, and of these most were identified as key contributors. Interestingly, well-predicted metabolites tended to have a higher proportion of the set of relevant genes as key contributors in both datasets (p=0.002 and p=0.09 in Datasets 1 and 2, respectively; Wilcoxon rank-sum test). Surprisingly, the key genes for well-predicted metabolites are less variable across samples than those for other metabolites (p=0.08 and p=0.002 in Datasets 1 and 2, respectively; Wilcoxon rank-sum test). We also examined whether the key gene contributors for each metabolite encoded enzymes solely catalyzing reactions synthesizing the metabolite in question, degrading it, or both (see Methods). We found that BV-enriched metabolites with key contributor genes that are associated only with synthesis enzymes were almost always well-predicted (11 out of 13 metabolites across both datasets; Figure 2.4 and Figure 7.6). Conversely, metabolites depleted in the BV state and with key contributor genes encoding only degradation enzymes also tended to be well-predicted (18 out of 31 metabolites across both datasets; Figure 2.4 and Figure 7.6). These trends suggest that the most predictable variation resulted from the transition between these two conditions, and in particular the impact of the presence or absence of novel metabolic synthesis and degradation capacities in BV, rather than shifts in the abundance of more widely found metabolic pathways.

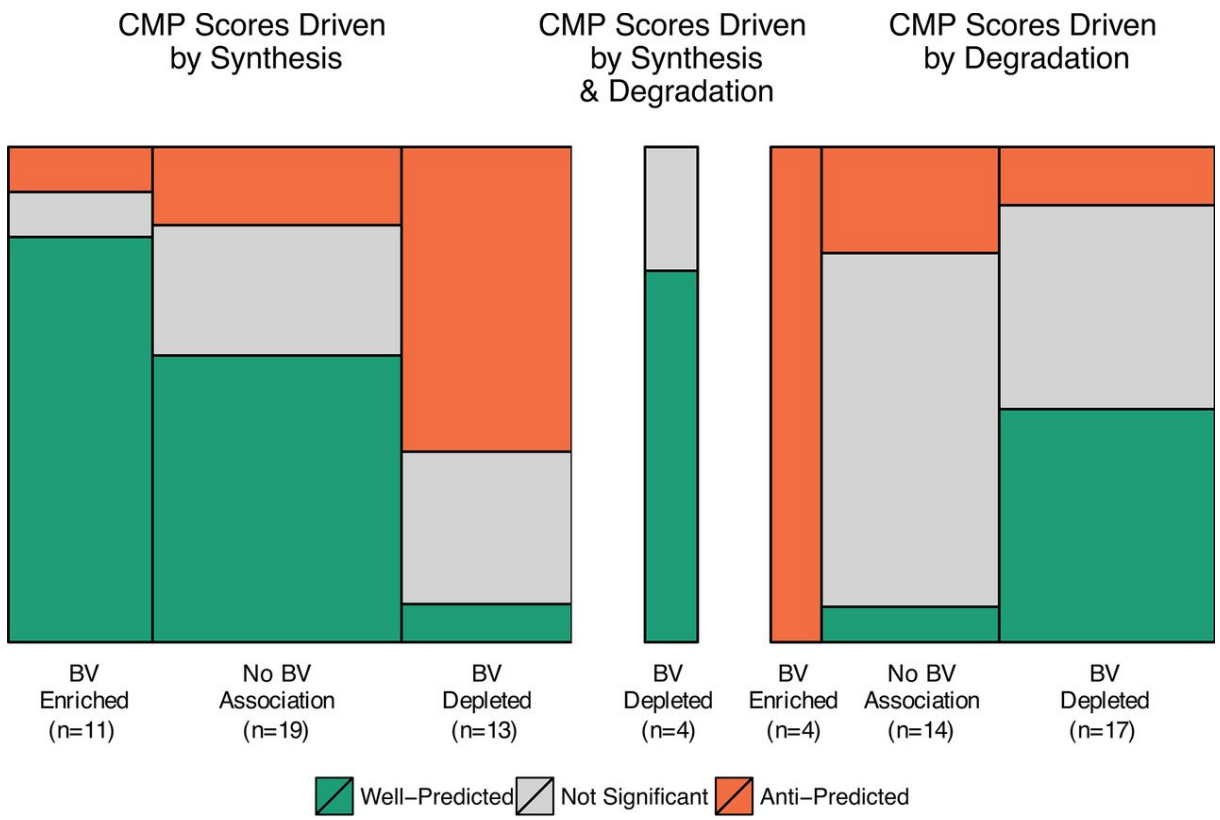

Figure 2.4. **Trends in metabolite predictability in terms of key gene contributors.** Area plots depict the numbers of metabolites in Dataset 1 whose CMP scores are driven by synthesis, by degradation, or by both in relation to their association with the host state and their predictability. The width of each box corresponds to the number of metabolites associated with each host disease state (enriched in BV samples, depleted in BV samples, or neither), and the height corresponds to the number of metabolites that are well-predicted, anti-predicted, or not significantly predicted (also indicated by color). See Figure 7.6 for a similar plot describing metabolite prediction in Dataset 2.

2.3.5 *Application to gut microbiome communities reiterates metabolic trends and highlights community complexity*

Finally, we explored the application of this framework to samples from gut microbial communities, bearing in mind the caveats of increased environmental influences resulting from diet as well as increased community complexity. Specifically, we applied this framework to two additional datasets: one evaluating the impact of antibiotic treatment with cefoperazone on the cecal contents of SPF mice (Theriot et al., 2014) (Dataset 3), and another profiling the microbiome

and metabolome of humans with inflammatory bowel disease and healthy controls (Dataset 4) (Erickson et al., 2012; Jansson et al., 2009). Because the second study used shotgun metagenomic sequencing, in its analysis we did not need to predict metagenome content from community composition and instead estimated gene abundances directly (Carr and Borenstein, 2014) (see Methods). As expected given the more complex and unstable community environment, we observed a lower proportion of well-predicted metabolites in these datasets (Figure 7.7), but also identified interesting patterns in the relationships between variation in community composition and metabolism in these settings.

Dataset 3 recapitulated many metabolite predictability trends observed in our analysis of the vaginal microbiota, including successful prediction of metabolite variation and the effect of perturbation on community ecology and metabolism. In total, 39 of the 116 metabolites (33.6%) assayed and analyzed in this dataset were well-predicted. As in Datasets 1 and 2, this fraction was higher than expected under a null model using randomized metabolic networks ($p < 0.001$). Interestingly, we observed substantial overlap in the identity of the metabolites that were well-predicted and anti-predicted in this dataset with those predicted similarly in Datasets 1 and 2, as well as a general positive correlation between prediction levels across datasets (Figure 2.5). One well-predicted metabolite of interest is gamma-aminobutyrate (GABA), which was enriched in the subset of samples from mice six weeks after antibiotic treatment. Key contributor analysis indicated that increased synthesis from 4-aminobutanal by an OTU in the genus *Oscillospira* drove the CMP score variation for this metabolite. Several products of carbohydrate metabolism were also well-predicted, including the sugars stachyose and mannose. Analysis of key contributors revealed that the oligosaccharide stachyose is predicted on the basis of its depletion by glycosidases from diverse Firmicutes taxa including *Ruminococcus* and *Turicibacter*, while mannose is predicted on the basis of increased production via glycan degradation from

mannoglycans by several OTUs in the Clostridiales order in healthy samples. These shifts reflect the impacts of increased glycan degradation potential in the microbiome of mice from the control cohort compared to those treated with antibiotics. As in the BV datasets, synthesis products found to be more abundant in the more diverse microbiome of the control cohort were most likely to be predictable (48% of 29 such metabolites were well-predicted).

In Dataset 4, only a very low proportion of the metabolites analyzed were predicted well (6 out of 31), which is likely due to a markedly smaller sample size and potentially noisier metabolomic data and identifications. Interestingly, however, four out of these six metabolites (chenodeoxycholate, glycochenodeoxycholate, glycocholate, and taurocholate) are primary conjugated or unconjugated bile acids, which form part of a tightly regulated pathway of host-microbial co-metabolism with hormonal signaling functions. This enrichment of bile acid-associated products among the well-predicted metabolites (p=0.03; Fisher exact test) highlights the important role of microbiome ecology in microbial metabolism of bile acids in the gut. Specifically, higher levels of bile acid metabolites in IBD cases have been noted previously in this dataset (Jansson et al., 2009). Our results show that this shift in bile acids is concordant with variation in the abundances of microbial bile salt hydrolase genes.

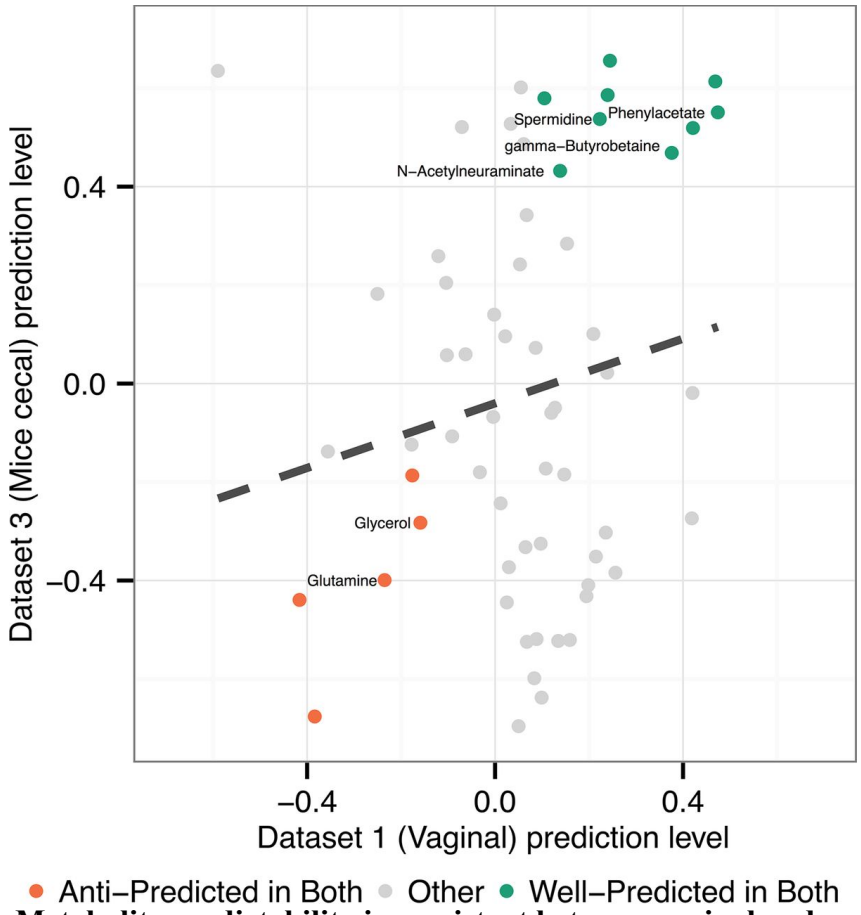

Figure 2.5. **Metabolite predictability is consistent between vaginal and mouse cecal data sets.**

The plot shows the relationship between the level of predictability for each metabolite (measured as the Spearman correlation between pairwise differences in calculated CMP scores and pairwise differences in measured metabolite abundances) in data set 1 (human vaginal microbiome samples) and data set 3 (mouse gut samples). Colors indicate metabolites that are well-predicted in both data sets or anti-predicted in both data sets. Metabolites that are well-predicted in both data sets are enriched for amino acid catabolites, including phenylacetate, spermidine, and beta-alanine.

2.4 Discussion

Above, we have introduced a novel analytical framework that represents an important step towards a principled, systematic and mechanistic integrative analysis of microbial community composition and metabolomic data. Our framework goes beyond *ad hoc* correlation-based analysis

and aims to assess the correspondence between ecology and metabolic phenotype based on the existing body of knowledge about microbial genomes and metabolic capacities. By evaluating metabolite variation in terms of the functional implications of ecological shifts, we identified a large share of the vaginal metabolome whose variation can be explained by shifts in ecology-based and community-wide enzymatic potential. This high predictability is somewhat surprising as our framework ignores many factors that could potentially impact this link, including strain variation, gene and protein expression, and metabolic fluxes (Edlund et al., 2015; Shi et al., 2014; Sonnenburg et al., 2005). This finding suggests that ecological dynamics and their impact on community metabolic capacities likely play a major role in mediating broad metabolic differences between microbiomes.

Further, our characterization of key species and gene contributors to calculated CMP scores and consequently to the predictability of each metabolite provides evidence that particular BV-associated species have substantial effects on the metabolome. By comprehensively identifying species whose enzymatic capacity and variation across samples are consistent with the observed shift in the abundance of a particular metabolite, we were able to gain deeper insight into the drivers of species-metabolite dynamics in the vaginal microbiota and bacterial vaginosis. Specifically, our analysis of key contributor species identified a subset of BV-associated species (*Eggerthella*, *Megasphaera*, and *Mageeibacillus indolicus*) as particularly likely to be important drivers of metabolic variation in this environment. The low contributions of lactobacilli suggest that their importance in the vaginal microbial ecosystem is not described well by current reference knowledge of their role in canonical pathways. Alternatively, this can be attributed to having twice as many women with BV compared to women without BV in our datasets, and moreover only some of the women without BV had abundant *L. crispatus*. More generally, we observed that the abundance of metabolic capacities (based on taxonomic composition) is often sufficient to explain

measured variability in the abundance of many BV-associated metabolites. This intriguing result suggests that while information about ecological shifts may not necessarily provide a comprehensive understanding of changes in flux in core metabolic reactions, it is often sufficient to account for the accumulation or depletion of many more peripheral metabolites that vary most dramatically between health and dysbiosis.

We also extended this method to analyze datasets from the gut microbiota of mice and humans and identified preliminary mechanistic links in these complex environments. The lower predictability in this context likely reflects the greater complexity of this community and the plethora of factors, both external and internal to the community that can potentially affect metabolite abundances. Studying the impact of such factors on various metabolic processes is an important direction for future research. Nevertheless, the overlap observed here in the set of metabolites that are well-predicted across a single organism in culture (*E. coli*), a simple community (the vaginal microbiome), and a complex host-associated community (gut microbiome) may represent shared control points in microbial metabolic networks. This consistency indicates that across multiple environments, the limiting factor for accumulation or depletion of these metabolites is the presence or abundance of microbial enzymatic potential that can directly act on them. This finding further reinforces the credibility of our framework and the shared features of microbial metabolic regulation across all of these settings. In addition, the predictability of the metabolic shifts associated with major ecological perturbations across datasets is consistent with previous metabolic regulation findings that core reactions tend to be regulated by a precise balance between precursor metabolite concentrations and enzyme concentrations, and that intracellular concentrations of core metabolites are generally robust in response to perturbations (Ishii et al., 2007; Jozefczuk et al., 2010).

One obvious caveat of our framework and of the resulting findings is the inability to distinguish between failing to predict due to missing reference information (e.g., incomplete genome annotation) versus failing due to a range of alternative mechanisms regulating metabolite shifts and environmental inputs and outputs. For example, our framework currently does not capture host metabolism, and future work may extend our model to include human gut metabolic processes. Similarly, our model does not consider signaling processes, transcriptional regulation, or bounds on metabolic fluxes. This limitation is further compounded by the use of a broad reaction database such as KEGG. For example, our model only assigns effects for enzymes catalyzing nonreversible reactions. This approach presumably captures major metabolic fluxes for well-characterized microbes, but the information lost from reversible reactions may hinder our ability to predict metabolites in other pathways. An extended framework could, for example, infer reaction directionality from pathway context, constraint-based modeling, or directly from metabolomic data using a machine learning approach.

Such improvements could also help clarify the interpretation of anti-predicted metabolites, which spanned roughly a third of all predictions across datasets and can be explained by several potential mechanisms. Anti-predicted metabolites whose CMP scores are driven by degradation reactions, especially with downstream well-predicted metabolites, are suggestive of environmentally-regulated metabolite changes that cause taxonomic shifts based on nutrient availability, such as the example of glycerol anti-predicted by *L. iners*. Other anti-predicted metabolites may be explained by missing reaction information. For example, putrescine and cadaverine are both anti-predicted based on a high correlation with the abundance of genes coding for enzymes that utilize these metabolites to synthesize further polyamine derivatives (N-acetylputrescine and aminopropylcadaverine, respectively). This finding suggests that other enzymes that are currently not incorporated into our predictions (including synthesis reactions that

are present but lacking information on reaction directionality), or enzymes from other unmeasured microbes in these samples, are likely important for driving the accumulation of these metabolites. In other cases, anti-prediction may suggest alternative metabolic mechanisms controlling metabolite variation, beyond direct enzymatic regulation.

Finally, the trends revealed by our analysis highlight the tight coordination of various metabolic processes even in the context of complex communities. Our framework evaluated each metabolite independently, but the resulting predictability trends, as well as evidence from other studies (Edlund et al., 2015; Srinivasan et al., 2015), show that dramatic shifts in metabolite abundances occur in a strongly coordinated fashion, through a combination of changes in substrate and enzyme concentrations mediated by a variable range of taxa. The analysis framework presented here is an important first step towards deconstructing and interpreting these relationships in mechanistic detail from comprehensive multi-omic data. In turn, this mechanistic understanding will be vital to ultimately enable the rational design of strategies to modify the microbiome and its metabolic phenotype (Greenblum et al., 2013; Waldor et al., 2015).

2.5 ACKNOWLEDGMENTS

We thank Keith Bayer and Colin Brislawn for technical support in obtaining various data sets. We are grateful to all members of the Borenstein lab for helpful discussions.

Chapter 3. CASE STUDIES IN MICROBIOME-METABOLOME INTEGRATION: CHARACTERIZING DIET-MICROBIOME INTERACTIONS AND ASD-LINKED MICROBIAL METABOLITES

This chapter is based on excerpts of the following two publications:

Snijders, A.M., Langley, S.A., Kim, Y.-M., Brislawn, C.J., **Noecker, C.**, Zink, E.M., Fansler, S.J., Casey, C.P., Miller, D.R., Huang, Y., Karpen, G.H., Celniker, S.E., Brown, J.B., Borenstein, E., Jansson, J.K., Metz, T.O. and Mao, J.-H. (2016). Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. Nature Microbiology *2*, 16221.

Sharon, G., Cruz, N.J., Kang, D.W., Gandal, M.J., Wang, B., Kim, Y.-M., Zink, E.M., Casey, C.P., Taylor, B.C., Lane., C.J., Bramer, L.M., Isern, N.G., Hoyt, D.W., **Noecker, C.**, Sweredoski, M.J., Moradian, A., Borenstein, E., Jansson, J.K., Knight, R., Metz, T.O., Lois, C., Geschwind, D.H., Krajmalnik-Brown, R., and Mazmanian, S.K. (2019). Human microbiomes from autism spectrum disorder induce behavioral symptoms in mice.

In Chapter 2, I described a new computational method for comparing the reference-based estimated metabolic potential of a set of microbial communities with their metabolic output. I subsequently introduced an R package implementation of this framework, named MIMOSA, freely available on GitHub (www.github.com/borenstein-lab/MIMOSA). MIMOSA performs two key analyses: identifying metabolites whose variation is statistically consistent with expectations based on microbial metabolic potential (well-predicted), and identifying specific microbial taxa and reaction contributors to metabolic potential that might be responsible for the observed variation. Here, I describe two case studies applying MIMOSA to microbiome-metabolome datasets, one focused on each of these two key analysis tasks. In the first case study, fecal microbiomes and

metabolomes were compared between genetically divergent mice reared in two different lab environments and fed two different standard chows. By using MIMOSA to identify putative microbial metabolites and comparing the results with metabolites known to be present in each chow, we distinguished metabolites varying depending on the microbiome, the diet, or both. In the second case study, we analyzed metagenomic and metabolomic data from gnotobiotic mice colonized with human gut microbiota from either children with autism spectrum disorder (ASD) or neurotypical controls. MIMOSA identified specific taxa and reactions as potential contributors to metabolite differences between the two groups, and these identifications were used to inform conclusions and follow-up validation studies. In both of these examples, analysis with MIMOSA provided a specific mechanistic summary of complex ecological and metabolic trends, in terms of the current reference knowledge of microbial metabolism.

3.1 BACKGROUND

Although the gut microbiome plays important roles in host physiology, health and disease (Clemente et al., 2012), we lack understanding of the complex interplay between host genetics, early life environment, and diet on the microbial and metabolic composition of the gut. Previous studies have identified associations and mechanisms linking the mammalian microbiome and its activity with host gene variants (Beaumont et al., 2016; Benson, 2016; Blekhman et al., 2015; Goodrich et al., 2016), with early life exposures (Dominguez-Bello et al., 2010, 2016), and with diet content (Gentile and Weir, 2018), but the relative importance and interactions between these remains unclear. We studied the impacts of early life history, dietary changes, and genetic variation on the fecal microbiome and metabolome of mice from the genetically diverse Collaborative Cross system (Collaborative Cross Consortium, 2012).

Autism spectrum disorders (ASD) are a group of neurodevelopmental conditions with a broad range of manifestations involving altered social communication and interaction, as well as

repetitive, stereotyped behaviors. The gut microbiome has been suggested to play a role in ASD. Gut bacterial communities differ between individuals diagnosed with ASD and those that are typically-developing (TD) (Coretti et al., 2018; De Angelis et al., 2013; Gondalia et al., 2012; Kang et al., 2013, 2018; Kushak et al., 2017; Son et al., 2015; Strati et al., 2017; Williams et al., 2011), as well as in mouse models of ASD (Buffington et al., 2016; Coretti et al., 2017; Hsiao et al., 2013; Kim et al., 2017; Liu et al., 2018; Sgritta et al., 2019; de Theije et al., 2014). Fecal microbiome profiles are most divergent in ASD subjects presenting with GI dysfunction (Gondalia et al., 2012; Son et al., 2015), a common comorbidity of autism (Chaidez et al., 2014). In addition, microbial-based interventions, including fecal transplantation, antibiotics, and probiotics, have shown promise in a limited number of open-label human trials (Grimaldi et al., 2018; Kang et al., 2017; Sandler et al., 2000). Some gut microbes have demonstrated therapeutic potential in animal models of ASD (Buffington et al., 2016; Hsiao et al., 2013; Sgritta et al., 2019; Tabouy et al., 2018). Furthermore, changes in the microbiome often result in altered metabolic profiles, impacting the availability and diversity of nutrients and microbial secondary metabolites (Dodd et al., 2017; Maier et al., 2017; Melnik et al., 2017; Sharon et al., 2014; Ursell et al., 2014; Wikoff et al., 2009). Indeed, metabolomic analyses of serum, feces, and urine from ASD subjects have uncovered differences in various molecules compared to TD controls, with many dysregulated compounds being of microbial origin (De Angelis et al., 2013; Ming et al., 2012; Mussap et al., 2016). Notably, amino-acid transport and degradation capacities differ between TD and ASD individuals (Aldred et al., 2003; Evans et al., 2008), intriguing because amino acids serve as precursors for many potent neuroactive molecules, as well as classic neurotransmitters. In order to extend these associations toward understanding disease etiology, we sought to define functional contributions by the gut microbiome to behaviors associated with ASD.

In the first case study, I used MIMOSA to putatively differentiate between metabolic variation arising from microbial variation, diet differences, or both. In the second, we hypothesized that the gut microbiomes of individuals with ASD contribute to behavioral and metabolic differences from that of the TD population, finding that colonization of germ-free (GF) wild-type mice with fecal microbiomes from ASD subjects is sufficient to induce core features of ASD in their offspring. MIMOSA analysis characterized specific microbiome features putatively involved in the production of microbial metabolites that ultimately may affect brain function and regulate behavior.

3.2 METHODS

3.2.1 *Case Study 1: Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome*

Methods for mouse husbandry, sample collection, microbiome profiling, and metabolome profiling and identification are described in Appendix B.

We produced a closed-reference OTU table using *vsearch* to align reads from all 77 samples with both sequencing and metabolomics data to the pre-clustered Greengenes database. We rarefied the OTU table to 4,000 reads and used it as input to MIMOSA, a framework for integrating taxonomic and metabolomic microbiome data (Noecker et al., 2016). MIMOSA uses genomic data, metabolic information and taxonomic composition to predict the community-wide biosynthetic and degradation potential for each metabolite in each sample and identifies metabolites whose variation across samples is consistent with (and can be explained by) variation in this predicted metabolic potential. Metagenome content was inferred for each sample using PICRUSt (Langille et al., 2013) and normalized using MUSiCC (Manor and Borenstein, 2015). From these data, a community-wide metabolic model was constructed for each sample and community metabolic potential (CMP) scores were calculated, representing the relative capacity

of the predicted community enzyme content in that sample to synthesize or degrade each metabolite. We then compared variation in these scores across samples to variation in measured metabolite concentrations using a rank-based Mantel test, to identify metabolites for which variation in concentration across samples is positively correlated (consistent) with variation in community metabolism (as predicted by the CMP scores), using a local FDR *q*-value less than 0.01 as the significance threshold. We similarly identified metabolites for which variation in concentration across samples is negatively correlated (contrasting) with CMP scores, with the same significance threshold. To identify potential contributing OTUs for each metabolite, we calculated the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the entire community and the CMP scores generated based on each species by itself (that is, recalculating the metagenome content and CMP scores based solely on the abundance of this species). OTUs for which this correlation coefficient for a given metabolite was greater than 0.5 were classified as potential contributing OTUs for that metabolite.

3.2.2 *Case Study 2: Links between human ASD microbiota, metabolites, and behavior in gnotobiotic mice*

Methods for collection of human fecal samples, mouse husbandry, mouse colonization, sample collection, microbiome profiling and analysis, and metabolome profiling are described in Appendix B.

We used a metabolic modeling-based framework, MIMOSA, to identify metabolites whose variation across samples can be explained by variation in the metabolic potential of the microbiome (Noecker et al., 2016). Specifically, we used MetaPhlAn2 (Truong et al., 2015) and HUMAnN2 (Franzosa et al., 2018a) to generate taxonomic and functional annotations of metagenomic sequencing data from the colon contents of oASD and oTD mice (n=35). The HUMAnN2 default UniRef90 gene family database (Suzek et al., 2015) was used for functional

annotation, and the results were then mapped to the KEGG database using the *humann2_regroup_table* utility. We mapped metabolite names to KEGG compound IDs using the KEGG API via the KEGGREST R package, resulting in 114 metabolites with KEGG IDs from the fecal GC-MS analysis, 66 metabolites with KEGG IDs from the fecal NMR analysis, and 125 metabolites with KEGG IDs from the serum GC-MS analysis. We then analyzed each metabolomics dataset separately using the R package MIMOSA version 1.0.1 (http://elbo.gs.washington.edu/software_MIMOSA.html), identifying consistent metabolites as those for which metabolite concentrations and community metabolic potential (CMP) scores displayed a significant rank-based correlation (local FDR q-value < 0.01). Calculation of CMP scores, comparisons between CMP scores and metabolite concentrations, and identification of potential taxa contributors were performed as described previously (Section 3.2.1). Gene families were identified as potential contributors to a consistent metabolite by calculating the Pearson correlation between the vectors of CMP scores obtained with and without reactions linked to that gene. Metabolites for which all contributor genes were exclusively involved in synthesis reactions were classified as "Primarily predicted by synthesis", and those whose potential contributor genes were exclusively linked to degradation reactions were classified as "Primarily predicted by degradation".

3.3 Results and Discussion

3.3.1 *Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome*

To decipher the respective contributions of host genetics, early life history and diet on the gut microbiome we leveraged 30 independent, genetically distinct Collaborative Cross (CC) mouse strains, a large multi-parental panel of recombinant inbred strains with defined single nucleotide polymorphisms (SNPs) that captures ~90% of the known variation in laboratory mice.

Sixteen strains were maintained in a specific pathogen-free (SPF) facility (Built Environment 1, BE1), and 14 additional strains were maintained in a barrier facility that screens for additional infectious agents, including *Pasteurella pneumotropica* and *Helicobacter* (Built Environment 2, BE2). Mice were fed the same water and food sources at both locations. Faecal samples were collected at 12 weeks of age (Figure 3.1A) and the gut microbiome composition was characterized by sequencing 16S rRNA genes (V4 hypervariable region).

Investigation of the faecal metabolite composition allowed us to determine the influence of early life environment and diet on the gut metabolome. For these analyses we focused on 24 CC strains that were housed in BE1 and BE2 (fed Diet 2, Labdiet Prolab 3500) and in BE3 (fed Diet 1, Labdiet Picolab 5053). Although the two diets have similar macronutrient compositions, the metabolite profiles are quite distinct. Extracts from the stool samples were analyzed by gas chromatography–mass spectrometry (GC–MS) and metabolites were identified by comparison to a reference library containing mass spectral and retention index information for over 850 metabolites (Kind et al., 2009). A total of 122 unique metabolites were identified, including amino acids, sterols, mono- and disaccharides, glycolytic and tricarboxylic acid cycle intermediates, short- and long-chain fatty acids, and products of microbial metabolism. An additional 110 peaks were detected but not identified. The metabolites significantly clustered by diet, with differences in relative abundances of proteinogenic amino acids, mono- and disaccharides, sterols and fatty acids driving the separation (Figure 3.1B).

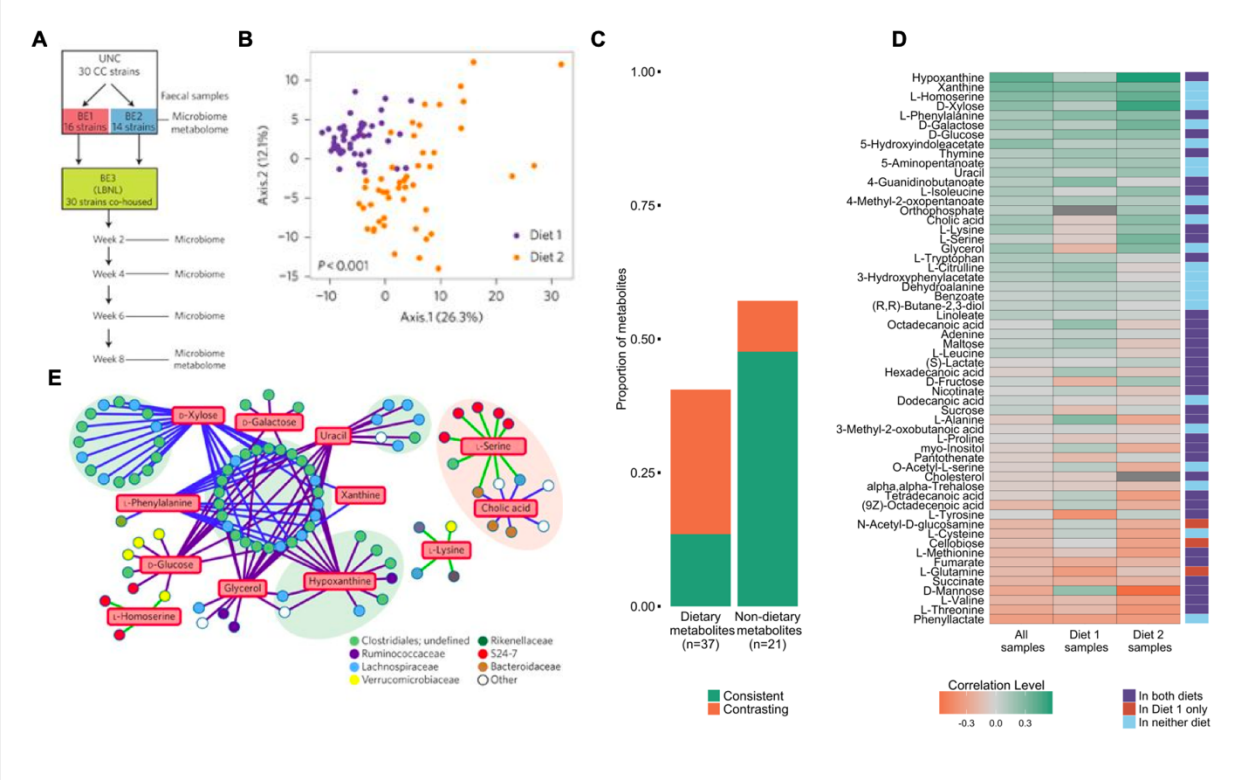

Figure 3.1. **Dietary and microbial influences on fecal metabolite profiles.** **A)** Schematic of the study design. **B)** PCoA of metabolite profiles were measured in faecal samples of 24 CC strains maintained on two different diets ($P < 0.001$, $R^2 = 0.16597$, ADONIS). **C)** Proportions of dietary and non-dietary metabolites whose measured variation is consistent (positively correlated) or contrasting (negatively correlated) with community metabolic potential (as predicted by metabolic modeling with MIMOSA). **D)** Correspondence between variation in community metabolic potential and variation in measured metabolite concentration across all samples and across sample subsets from each diet and facility. **E)** Metabolic modelling-based taxonomic contributors to metabolite variation for mice on the autoclaved Labdiet Prolab 3500 chow (BE1 and BE2). Individual OTUs shown (circles; coloured at the family level) are those whose metabolic capacity and variation across samples are consistent with the metabolic potential of the entire community and with measured variation in the linked metabolites (squares). Green and orange clouds behind OTU sub-networks indicate Clostridiales and Bacteroidales enrichment. Edge colour indicates whether a given OTU potentially impacts a certain metabolite variation via synthesis (blue edges), degradation (green edges) or both (purple edges).

We used a metabolic modelling-based framework, MIMOSA (Noecker et al., 2016), to identify metabolites whose variation across samples is explained by variation in the metabolic potential of the microbiome, based on differences in species composition and estimated gene composition. By applying MIMOSA to the pooled set of metabolome samples from both diets, we found that variation in dietary metabolites (compounds detected in chow pellets by metabolomics) was poorly explained by microbial community composition (Figure 3.1C). However, the variation in a high proportion of non-dietary metabolites (47.6%; 10 out of 21 metabolites not detected in chow) was consistent with predicted community metabolic potential (CMP), suggesting a substantial role for microbial metabolism in metabolite synthesis and/or degradation. Specifically, the observed variation in many gut metabolites was consistent with the predicted CMP, including hypoxanthine, L-homoserine, 5-hydroxyindoleacetate and cholate (Figure 3.1D). More metabolites varied consistently with predicted CMP in samples from the nutritionally simpler Diet 2, suggesting that the microbiome may have a larger and more direct impact on the faecal metabolome in this context. The predicted CMP was driven by the metabolic potential of a diverse set of taxa, including OTUs from the phyla Firmicutes, Bacteroidetes and Actinobacteria (Figure 3.1E). Interestingly, the measured concentrations of several metabolites present in one or both diets were negatively correlated with predicted CMP (mostly on the basis of microbial degradation enzymes), indicating that food containing these metabolites could drive the expansion of microbes that use them efficiently. These findings highlight the combined impacts of diet and microbiome composition on the gut metabolome and the complex interactions between them.

Our studies using the CC mouse cohort and an integrated, systematic analysis paradigm revealed how gut microbiome composition and metabolic function are shaped by interactions between host genotype, early life environment and diet. We found that early life history impacts the microbiome composition, whereas dietary changes have only a moderate effect. By contrast,

the gut metabolome was shaped mostly by diet, with specific non-dietary metabolites explained by microbial metabolism. This study provides a foundation for future investigations of how reciprocal interactions between environmental factors and gut microbiome and metabolome compositions contribute to a wide spectrum of mammalian traits and disease susceptibility.

3.3.2 *Links between human ASD microbiota, metabolites, and behavior in gnotobiotic mice*

Microbiome profiles of individuals with ASD, especially that of ASD children with intestinal symptoms, are different from TD controls (De Angelis et al., 2013; Gondalia et al., 2012; Kang et al., 2018; Son et al., 2015; Strati et al., 2017). While suggestive, these association studies are unable to resolve cause-and-effect relationships to the disorder. Accordingly, we tested whether altered human microbiomes, and their metabolic processes, functionally contribute to ASD behaviors in mice. Fecal samples from male TD and ASD donors selected based on Autism Diagnostic Observation Schedule (ADOS) (Gotham et al., 2007) and GI severity index scores (GSI) (Schneider et al., 2006). As ASD is a developmental disorder with evidence for prenatal effects (Hallmayer, 2011; Lyall et al., 2014), we colonized GF male and female C57BL/6J mice with TD or ASD donor samples (recipient TD and ASD, annotated rTD and rASD, respectively), and subsequently mated these animals (Figure 3.2A). Adult offspring mice (offspring TD and ASD, annotated oTD and oASD) that inherited human donor microbiota were either sampled (feces, serum, brains) or behavior tested (Figure 3.2A).

To explore mechanisms for gut-brain connections, we performed untargeted metabolomics analyses of colon contents from oTD and oASD mice using proton nuclear magnetic resonance (^{1}H NMR) spectroscopy and gas chromatography-mass spectrometry (GC-MS), and their corresponding serum by GC-MS. Twenty-seven out of 313 detected metabolites are significantly different in the colon contents of oASD mice, compared to oTD mice, with 24 detected by GC-MS (11 identified) and four by NMR. Differentiating metabolites were diverse:

amino acids and their derivatives, isoflavonoids, carbohydrates and their derivatives, and fatty acids. We observed higher concentrations of amino acids in colon contents of oASD mice, similar to reports studying ASD in children (De Angelis et al., 2013). Notably, we find differences in several agonists and antagonists of the inhibitory gamma-aminobutyric acid (GABA) and glycine receptors (Figure 3.2B). Specifically, 5-aminovaleric acid (or 5-aminopentanoate; 5AV) is significantly lower in oASD mice. 5AV is a weak gamma-aminobutyric acid (GABA) receptor agonist, and is significantly lower in children with ASD, compared to TD controls (Ming et al., 2012). Lower levels of another weak GABA agonist (and a potent glycine receptor agonist), taurine , are also found in a subset of ASD subjects (Adams et al., 2011; Park et al., 2017; Tu et al., 2012). Intriguingly, oASD mice have 50% less taurine compared to oTD mice, as measured by NMR. Together, lower levels of 5AV and taurine suggest that gut microbes may impact inhibitory signaling through GABA and glycine receptors.

To assess the contribution of the gut microbiome to the metabolome we analyzed metagenomic data from colon contents by HUMAnN2 (Franzosa et al., 2018a) and coupled these data with MIMOSA (Noecker et al., 2016) – a metabolic model-based framework for inferring the contribution of bacterial species and genes to the production and degradation of metabolites measured by NMR and GC-MS analysis (Figure 3.2D). Interestingly, MIMOSA analysis highlighted that amino acids were putatively predominantly degraded or utilized by the microbiome (by specific *Akkermansia, Alistipes* and *Bacteroides* species), in contrast with other metabolites, whose variation was consistent with both synthesis and utilization potential from various other bacteria (Figure 3.2C). The metagenomic analysis identified specific KEGG pathways differentially present in oTD and oASD gut microbiomes, with various chemical structure transformation maps highly represented in oASD mice. The KEGG amino acid metabolism pathway was slightly increased in oASD mice, but with a small effect. We further

explored KEGG pathways related to amino-acid metabolism and found that the metabolism of various amino-acids, and specifically that of proline, taurine, and glutamate and glutamine, was differentially represented in the metagenomes of oASD mice, compared to oTD mice. These data indicate that the microbiome differentially metabolizes dietary amino acids in oASD mice.

We also used MIMOSA to explore potential pathways by which taurine and 5AV are produced or utilized. Taurine is predicted to be produced by either deconjugation of tauro-conjugated bile acids, such as taurocholic acid, or by decarboxylation of L-cysteate (Figure 3.2E). HUMAnN2 analysis identified that both bile-salt hydrolase (K01442) and glutamate decarboxylase (K01580) are present at lower abundance in oASD metagenomes. Interestingly, MIMOSA analysis predicted that taurine concentrations are consistent with differential synthesis potential from 3 specific species: *Alistipes* sp. HGB5, *Alistipes finegoldii,* and *Bacteroides xylanisolvens* (Figure 3.2D). 5AV is predicted to be the product of Stickland fermentation of proline (Huang et al., 2018). However, we found no difference in the abundance of the proline reductase. Δ1-pyrolline-5-carboxylate (P5C) is an intermediate of proline production from *trans*-4-hydroxyproline (Hyp) by *HypD* (Huang et al., 2018; Levin et al., 2017), which is subsequently reduced to L-proline by P5C reductase (K00286), as well as the oxidation of L-proline by the proline dehydrogenase *putA* (K13821) with L-glutamate as the end product (Liu et al., 2017). While the abundance of *hypD* is not significantly decreased in oASD metagenomes, we found a significant increase (driven mainly by two oASD donors) in *putA* and a significant decrease in P5C reductase abundance (Figure 3.2F) in oASD mice. Additionally, MIMOSA identified *trans*-4-hydroxproline as a consistent microbial metabolite on the basis of the abundance of K00286. These data suggest that, in oTD microbiomes, L-proline may be produced from *trans*-4-hydroxproline via P5C and subsequently donates an electron to produce 5AV. Conversely, in

oASD microbiomes, the balance shifts towards production of P5C and glutamate, and away from 5AV.

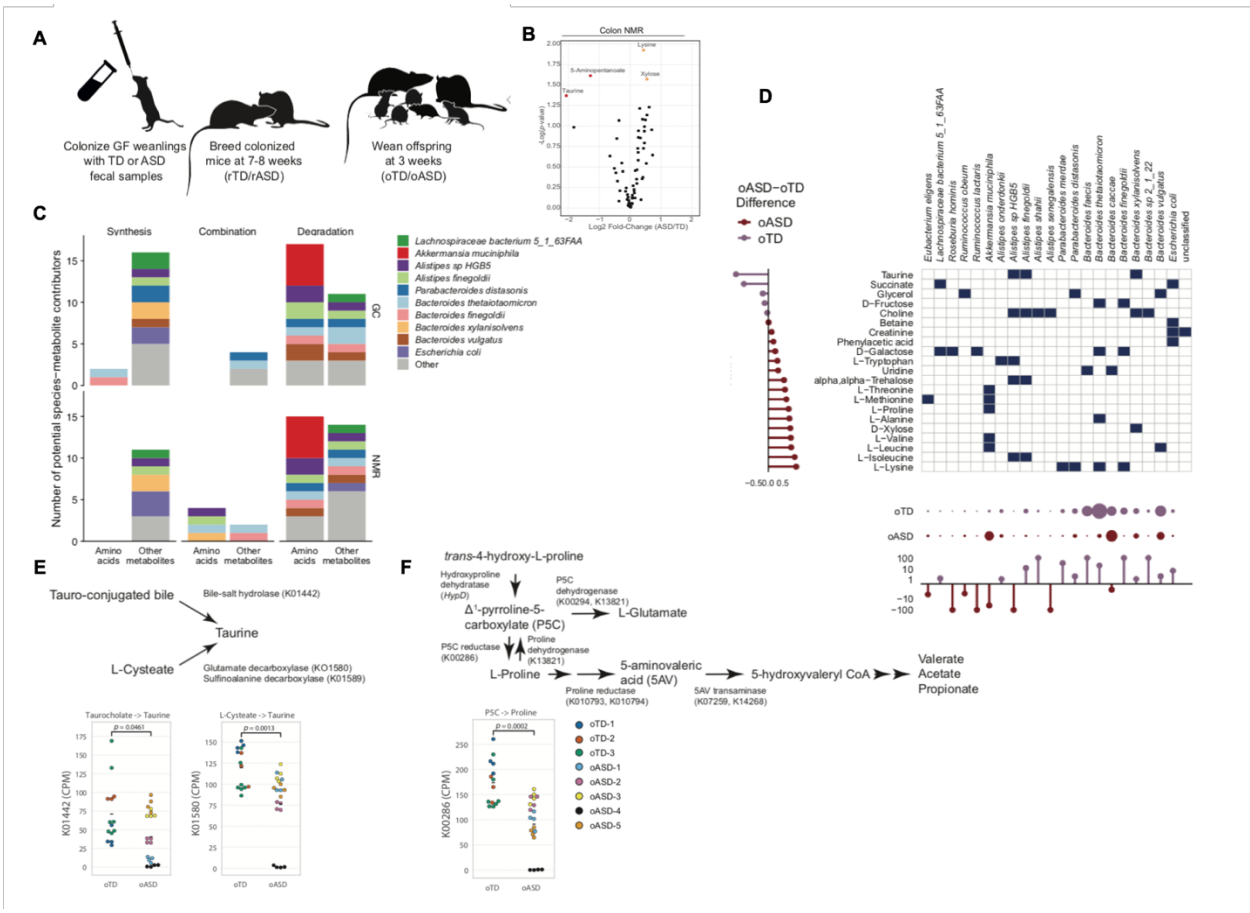

Figure 3.2. **Metabolic variation across ASD and neurotypical microbiomes.**

A) Schematic of gnotobiotic model establishment, **B)** Volcano plots of differentially abundant metabolites identified by an untargeted metabolomics of colon contents by ^{1}H NMR. Significantly different metabolites with more than 50% difference are marked in red, and those with modest effects (<50%) are marked in yellow. *N*oASD = 20, *N*oTD = 15 (4-7 mice per donor). p-values were calculated using the maximum likelihood test of a mixed effect linear model. **C)** Putative bacterial contributors to variation in amino acids and other metabolites identified by a MIMOSA analysis, separated by their mechanism of action (synthesis/production, degradation/utilization, or both), and based on GC-MS (top) and NMR (bottom) metabolomic analysis and metagenomic analysis of colon contents in oASD mice, **D)** MIMOSA-model prediction of strains involved in production or degradation of specific metabolites. Columns correspond to MetaPhlAn2-identified species in oTD and/or oASD metagenomes. Rows correspond to the metabolites across detected by NMR that were significantly consistent with

metabolic potential at q-value ≤ 0.1. Blue squares indicate that the estimated metabolic potential of the strain in question is consistent with contributing to variation in that metabolite (production or utilization/degradation). The area of the colored points along the bottom indicates the relative abundances of each taxon in oTD and oASD samples. The segments along the bottom indicate the relative ratio of each taxon in oTD versus oASD samples. The segments along the left side show the average difference in metabolite concentration Z-scores between the control and ASD donor samples. The column of colored tiles indicates the MIMOSA correlation between metabolic potential scores and metabolite concentrations for each metabolite. **E)** Taurine production in oASD mice is deficient. The pathway diagram shows reactions identified by MIMOSA as possible sources of taurine and KEGG orthologs involved. Plots show K01442 ($H = 3.9784$, $p = 0.0461$) and K01580 ($H = 10.3641$, $p = 0.0013$) copy abundances in oTD and oASD mice quantified by HUMAnN2. Differences in means were analyzed by a Kruskal-Wallis non-parametric test. **E)** Pathways providing substrates for Stickland fermentation to produce 5AV. The diagram shows pathways upstream to 5AV production and KEGG orthologs involved. The plot shows K00286 ($H = 13.7584$, $p = 0.0002$) copies in oTD and oASD mice quantified by HUMAnN2. Differences in means were analyzed by a Kruskal-Wallis non-parametric test.

3.4 CONCLUSIONS

This chapter describes two examples of using MIMOSA to infer potential microbial metabolic impacts from a set of microbiome and metabolome profiles. In both examples, MIMOSA identified metabolite variation that was consistent with microbial metabolic potential, as well as putative species and reaction contributors to the observed variation, which informed the conclusions of the study. In Case Study 1, analysis with dimensionality reduction approaches indicated that most variation in fecal metabolomes was driven by the assigned chow (Figure 3.1.B). Using MIMOSA, we confirmed that variation in chow-linked compounds was poorly explained by microbial metabolism, but we additionally identified several metabolites that appeared to vary depending on the metabolism of specific microbial taxa in both chow groups, or only in one chow group, indicating an interaction between chow content and microbial activity (Figure 3.1D-E).

This analysis additionally highlights the importance of a controlled nutrient environment for detecting the effects of microbial metabolism. In Case Study 2, we applied MIMOSA to hypothesize microbial features that might explain metabolic differences between gnotobiotic mice colonized with the microbiota of either children with ASD or neurotypical controls, with a focus on metabolites with putative differing physiological impacts between the two groups. We characterized an overall shift in amino acid metabolism between the groups, inferred that differences in taurine concentrations are more consistent with shifts in microbial synthesis, rather than utilization, on the part of specific taxa, and found evidence of a pathway producing 5AV from *trans*-4-hydroxyproline. These findings have informed follow-up experiments towards modifying the microbiota to control the concentration of these metabolites.

These case studies also illustrate the inherent limitations of the MIMOSA approach. Specifically, while MIMOSA identifies putative mechanisms and quantifies the strength of an association between predicted metabolic potential and metabolite measurements, it provides no measure of confidence in the true occurrence of an inferred mechanism. This lack of performance benchmarks was recently highlighted as an area in need of improvement for genomics analysis methods more generally (Lotterhos et al., 2018). Also, the evidence for any MIMOSA inference is contingent on the particularities of the KEGG reaction annotations for a given metabolite. For example, in Case Study 2, MIMOSA linked variation in *trans*-4-hydroxyproline concentrations to the abundance of the P5C reductase gene family on the basis of a pathway characterized mainly in eukaryotes. However, a different pathway, involving the same gene family but with an alternate role for this metabolite, was recently found to be active in gut bacteria (Huang et al., 2018; Levin et al., 2017), and is therefore a more plausible hypothetical mechanism. Thus, although MIMOSA can identify mechanisms that might be overlooked by manual curation, assessing its results still

requires careful follow-up interpretation in light of the incomplete reference knowledge on which this method is based.

3.5 ACKNOWLEDGMENTS

3.5.1 *Case Study 1*

The authors thank S.E. Cates, N.N. Robinson and G.D. Shaw in the Systems Genetics Core at UNC for technical assistance and M.H. Stoiber for helpful discussions, especially regarding statistical analysis. This work was primarily supported by funding from the Office of Naval Research under ONR contract N0001415IP00021 (J.J., J.H.M. and A.M.S.). Additional support was provided by the Low Dose Scientific Focus Area, Office of Biological and Environmental Research, US Department of Energy (G.K., J.H.M. and A.M.S.) and the Lawrence Berkeley National Laboratory Directed Research and Development (LDRD) program funding under the Microbes to Biomes (M2B) initiative (S.C., B.B., G.K., J.H.M. and A.M.S.). C.N. was supported by an NSF IGERT DGE-1258485 fellowship and in part by New Innovator Award DP2 AT007802-01 to E.B. Partial support was also provided under the Microbiomes in Transition (MinT) Initiative as part of the Laboratory Directed Research and Development Program at PNNL. Metabolomic measurements were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the US DOE OBER and located at PNNL in Richland, Washington. PNNL and LBNL are multi-program national laboratories operated by Battelle for the DOE under contract DE-AC05-76RLO 1830 and the University of California for the DOE under contract DE AC02-05CH11231, respectively.

3.5.2 *Case Study 2*

The authors would like to thank Drs. Hiutung Chu, Gal Lenz, Catherine Schretter, Daniel Dar, and members of the Mazmanian laboratory for critical discussions. We also thank Yolanda

Huang for *hypD* reference sequences. We would like to thank Dr. James Adams for patient recruitment efforts and critical review on the manuscript. We also thank Dr. Juan Maldonado and Morgan Bennett at the Microbiome analysis laboratory in Arizona State University for their support on DNA extraction, library preparation, and bacterial 16S rRNA gene sequencing. We acknowledge the help of animal and veterinary technicians at the Caltech Office of Laboratory Animal Resources. Metabolomics analyses were supported by the Microbiomes in Transition (MinT) Initiative as part of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). Metabolomics measurements were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy Office of Biological and Environmental Research and located at PNNL in Richland Washington. PNNL is a multi-program national laboratory operated by Battelle for the DOE under contract DE-AC05-76RLO 1830. Autism Speaks Postdoctoral Fellowship in Translational Research #9718 and Human Frontiers Science Program Long-Term Fellowship 2012/65 (to G.S). Funding includes grants from NIH 1R01GM124312-01 (to E.B.), NIMH ACE CENTER and NETWORK (to D.H.G.), Autism Research Institute and the BHARE (Brenen Hornstein Autism Research & Education) Foundation (to D.W.K and R.K.B), the Simons Foundation (to S.K.M.), Heritage Medical Research Institute (to S.K.M.), and NIH MH100556 (to S.K.M.).

Chapter 4. DEFINING AND EVALUATING MICROBIAL CONTRIBUTIONS TO METABOLITE VARIATION IN MICROBIOME-METABOLOME ASSOCIATION STUDIES

This chapter is based on the following manuscript:

Noecker, C. H.-C. Chiu, C.P. McNally, and E. Borenstein. Defining and evaluating microbial contributions to metabolite variation in microbiome-metabolome association studies. (2019). *bioRxiv*, doi:10.1101/402040.

4.1 SUMMARY

Correlation-based analysis of paired microbiome-metabolome datasets is becoming a widespread research approach, aiming to comprehensively identify microbial drivers of metabolic variation. To date, however, the limitations of this approach have not been comprehensively evaluated. To address this challenge, we introduce a mathematical framework to quantify the contribution of each taxon to metabolite variation based on uptake and secretion fluxes. We additionally use a multi-species metabolic model to simulate simplified gut communities, generating idealized microbiome-metabolome datasets. We then compare observed taxon-metabolite correlations in these datasets to calculated ground-truth taxonomic contribution values. We find that in simulations of both a model 10-species community and of complex human gut microbiota, correlation-based analysis poorly identifies key contributors, with extremely low predictive value despite the idealized setting. We further demonstrate that the predictive value of correlation analysis is strongly influenced by both metabolite and taxon properties, as well as exogenous environmental variation. We finally discuss the practical implications of our findings for interpreting microbiome-metabolome studies.

4.2 INTRODUCTION

Microbial communities have a tremendous impact on their surroundings, ranging from the degradation of environmental toxins (Hazen et al., 2010) to the production of climate change-relevant metabolites (Shi et al., 2014). Host-associated communities, in particular, have a substantial impact on their hosts, and often produce a diverse set of metabolites that interact with numerous host pathways. In humans, such microbiome-derived metabolites have been identified as contributing factors to a wide array of diseases including heart disease (Koeth et al., 2013), autism (Hsiao et al., 2013), non-alcoholic fatty liver disease (Dumas et al., 2006), colon cancer (Louis et al., 2014), inflammatory bowel disease (Wlodarska et al., 2017), and susceptibility to infection (Ferreyra et al., 2014). Characterizing the ways microbial communities modulate their environments and the relationship between community structure and metabolic impact is therefore a major, timely, and complex challenge with promising implications for human health, as well as to environmental stewardship, agriculture, and industry.

When facing this challenge, perhaps the most important task is identifying specific community members that drive variation in metabolites of interest. Taxa responsible for observed metabolic differences across communities may be ideal targets for interventions aiming to modify metabolic phenotypes. Their identification, however, can be a daunting task. Complex microbial communities are often composed of hundreds or thousands of poorly characterized species, each with a unique and frequently unknown complement of metabolic capacities. Even when multiple species are known to possess the potential to synthesize or degrade a metabolite of interest, the metabolic activity of each species (and consequently, its contribution to metabolic variation) may be different (Rath et al., 2017). Moreover, community ecology, interspecies interactions, and nutrient availability (e.g., via diet) can all regulate and influence the metabolic

activity of each species, rendering the link between community members and metabolic products extremely complex and challenging to infer (David et al., 2013; De Filippis et al., 2015; Snijders et al., 2016).

To address this challenge and to identify community members that play an important role in metabolic variation, a growing number of studies are now comprehensively assaying multiple facets of community structure across samples, including, most notably, taxonomic and metabolite compositions (Shaffer et al., 2017). For example, to investigate the links between taxonomic shifts and metabolic phenotypes in the healthy vaginal microbiome and in bacterial vaginosis, a recent study used a combination of 16S rRNA qPCR, sequencing, and both global and targeted metabolomics (Srinivasan et al., 2015). Another study, aiming to identify taxonomic and metabolic features of resistance and susceptibility to *C. dificile* infection in the mouse gut similarly applied 16S rRNA sequencing and global metabolomics (Theriot et al., 2014). In another example, researchers characterized metabolic and microbial features of periodontitis in the oral microbiome before and after treatment, combining 16S rRNA sequencing, shotgun metagenomic sequencing, and metabolomics (Califf et al., 2017). These are just a few examples of a plethora of recent microbiome-metabolome studies, investigating the metabolic effects of microbiome variation in the contexts of chronic and infectious disease, agriculture, precision medicine, nutrition, fermented food science, and more (Antharam et al., 2016; Garg et al., 2017; Heintz-Buschart et al., 2016; Hua et al., 2017; Price et al., 2017; Vandeputte et al., 2017; Walsh et al., 2016; Weir et al., 2013). Such multi-omic studies are also a major focus of several large-scale initiatives to study both host-associated and environmental microbiomes (Alivisatos et al., 2015; iHMP Research Network Consortium, 2014).

Given the taxonomic and metabolomic profiles obtained via such microbiome-metabolome assays, the vast majority of studies rely on simple univariate correlation-based

analyses to link variation in community ecology to variation in metabolic activity (Choo et al., 2017; De Filippis et al., 2015; Kang et al., 2018; Lin et al., 2018; Melnik et al., 2017; Srinivasan et al., 2015; Theriot et al., 2014). Such analyses specifically aim to identify species whose abundance across samples is correlated with the concentration of metabolites, often assuming that highly significant correlations reflect a direct mechanistic link between the taxon and metabolite in question. These studies further regularly assume that positive correlations imply synthesis and negative correlations imply degradation, or that targeting the microbe in question could be used to modulate the concentrations of the metabolites with which it is correlated. For example, a recent study characterizing the microbiome and metabolome in Spleen-yang-deficiency syndrome (Lin et al., 2018) concluded that a positive correlation between Bacteroides and mannose likely resulted from extracellular degradation of mannan into mannose by that taxon. Similarly, a study of antibiotic perturbations to the microbiome and metabolome stated that the presence of several weak positive and negative correlations between genera and arginine supported the conclusion that arginine levels may be affected by many community members with high functional redundancy (Choo et al., 2017).

Yet, to date, the extent to which a correlation-based analysis effectively detects direct metabolic relationships between taxa and metabolites is unclear. Obviously, a strong correlation between the abundance of a certain species and the concentration of a metabolite across samples *could* reflect direct synthesis or degradation of the metabolite by that species, but could also arise due to environmental effects, precursor availability, selection, random chance, or co-occurrence between species. Similarly, cross-feeding, external host processes, and varying enzymatic regulation can mask a correlation even when this species does in fact contribute to observed metabolite variation. Indeed, previous studies have suggested that microbe-metabolite correlations must have a high rate of false positives (Chong and Xia, 2017), and a recent

experimental study pairing microbiome-metabolome correlation analysis with *in vitro* monoculture validations found anecdotally that several observed correlations were in fact false positives (Hoyles et al., 2018). The limitations of correlation analysis have also been discussed and well-characterized in other data types (for example (Weiss et al., 2016; Werhli et al., 2006)). Importantly, however, the extent of such limitations in the context of microbiome-metabolome studies, the way they are shaped by microbial community metabolism, and their impact on data interpretation in this context have not been systematically evaluated.

Importantly, two crucial challenges hinder a comprehensive and systematic evaluation of correlation-based analysis. The first is the lack of a rigorous general definition of a microbe's contribution to metabolite variability. While establishing the main taxonomic contributors to metabolite variation may be straightforward for specialized, well-characterized metabolites that are synthesized by just a single taxon, it can be much less clear for metabolites that can be synthesized (and/or degraded or modified) by many different taxa in the community. The second challenge is the absence of ground truth data on the nature of microbe-metabolite relationships. While limited data on the taxa driving metabolite shifts can be obtained from comparative mono- and co-culture studies (Biggs et al., 2016; Hoyles et al., 2018; Kešnerová et al., 2017), large-scale and comprehensive datasets that link species and metabolite abundances in the context of a complex community, for which the precise impact of each species on observed metabolite variation is known, are currently not available.

In this study, we address these two challenges, combining a novel framework for quantifying microbial contributions with a model-based simulated dataset. Specifically, we first introduce a generalizable and rigorous mathematical framework for decomposing observed metabolite variation and quantifying the contribution of each community member to this variation based on uptake and secretion fluxes. Second, we use a dynamic multi-species genome-

scale metabolic model to simulate the metabolism of microbial communities of varying complexity and to generate idealized datasets of paired taxonomic and metabolomic abundances, with complete information on metabolite fluxes, microbial growth, interspecies interactions, and environmental influences. Applying our mathematical framework to these simulated datasets, we could then compare calculated contribution values to observed taxon-metabolite correlations and evaluate the ability of correlation-based analyses to identify key microbial contributors. We were additionally able to investigate factors that shape the relationship between community composition and metabolism in depth and to identify specific properties and mechanisms that impact the performance of microbiome-metabolome correlation studies.

Notably, given the objectives of this study, we intentionally focus on characterizing microbiome-metabolome relationships in a model-based, tractable, and well-defined setting. Indeed, our metabolic model may not perfectly capture all the complex and diverse mechanisms that are at play in host-associated communities; however, considering the scope of this study, accurately modeling the metabolism of a specific community may not be crucial. Rather, for our analysis, we want our simulated data to recapitulate broad trends observed in naturally occurring microbial ecosystems, as indeed has been observed in similar models (Bauer et al., 2017; Garza et al., 2018; Heinken and Thiele, 2015b; Magnúsdóttir et al., 2016; Shoaie et al., 2015). Moreover, utilizing this model-based approach allows us to dissect the relationship between community composition and metabolic phenotypes without the complexities inherent to *in vivo* communities (including spatial heterogeneity, measurement error, inter-microbial signaling, or strain-level variation), and with variation in the concentrations of environmental metabolites resulting exclusively from microbial metabolic activity. Analyzing the ability of a correlation-based analysis to detect true microbial drivers of metabolite variation in these simplified, best-

case settings provides a baseline for the expected performances of such analyses in real microbiome-metabolome studies.

4.3 Results

4.3.1 *Quantifying the impact of individual microbial species on variation in metabolite concentrations*

In this study, we consider a microbial community as an idealized system, consisting of a population of multiple microbial species in a shared, well-mixed, biochemical environment. Each species uptakes necessary metabolites from the shared environment, performs a variety of metabolic processes to promote its growth, and secretes certain metabolites back into the shared environment. We additionally assume that certain nutrients flow into the environment and that microbial cells and metabolites are diluted over time. These processes can represent, for example, the inflow of dietary nutrients and the transit through the gut in the context of the gut microbiome. For simplicity, we primarily consider a constant inflow and dilution rate, as in a chemostat setting. Accordingly, a microbiome-metabolome study can be conceived as analyzing a set of several such communities (at a certain point in time), each with a different composition of microbial species and correspondingly variable environmental metabolite concentrations. We focus initially on a controlled setting with identical nutrient inflow across all microbiomes, but later examine the impacts of differences in nutrient inflow between communities.

Given this setting, we first sought to establish a rigorous and quantitative framework for defining the impact of each microbial species (or any taxonomic grouping) in the community on the variation observed in the concentration of a given metabolite across community samples. We focused on species that *directly* modulate the environmental concentration of a given metabolite via synthesis or degradation, ignoring indirect effects via, for example, the synthesis of a precursor substrate that could impact the metabolic activity of other species. We noted that the

total concentration of a metabolite in the environment can be represented as the sum of cumulative synthesis or degradation fluxes of this metabolite by each of the *n* species in the community, as well as cumulative environmental fluxes (e.g., total nutrient inflow and dilution). Formally, the metabolite concentration, *M,* can therefore be expressed as a sum of *n* dependent random variables m_i, where each m_i denotes the overall synthesis or degradation of the metabolite by each species, along with an additional random variable m_{env}, denoting the overall impact of environmental processes.

$$M = \sum_{i=1}^{n} m_i + m_{env} \qquad (4.1)$$

As discussed above, when analyzing microbiome-metabolome datasets, the goal is often to identify taxa responsible for *changes* in the concentration of a metabolite of interest across a set of samples. Accordingly, here we wish to quantify the *contribution* of each species to the *variance* in the concentration of that metabolite across samples. Specifically, in the formulation above, *var(M)* depends on the variance in the constituent microbial and environmental factors, as well as the covariance between these components. This variance can then be linearly separated into *n+1* terms, representing the contribution of each species (denoted c_i), and of any environmental nutrient fluxes (denoted c_{env}) to the total variation in the metabolite:

$$var(M) = \sum_{i=1}^{n} c_i + c_{env}\ ;\ c_i = var(m_i) + \sum_{j \neq i} cov(m_i, m_j) + cov(m_i, m_{env}) \qquad (4.2)$$

If the nutrient inflow is constant across samples, its effect can be ignored and its contribution to the variance c_{env} is 0. Additionally, in a chemostat setting, the dilution of each metabolite can be accounted for in the calculation of each contribution, as it depends strictly on the dilution rate and on previous metabolite concentrations (Methods). Finally, in order to compare species contributions across metabolites and to represents the relative share of the total

variance of a given metabolite that is attributable to species *I*, we defined the *relative* contribution to variance $\hat{c}_i$ of each species *i* to metabolite *M* by normalizing contribution values by the metabolite's total variance:

$$\hat{c}_i = \frac{c_i}{var(M)} \qquad (4.3)$$

This framework for calculating microbial contribution values provides a systematic measure of the causal impact of each taxon on observed variation in the environmental concentration of each metabolite, distilling the effect of complex ecological and metabolic interactions to a concise and interpretable set of quantities. Moreover, the obtained contribution profile is a linear decomposition of observed metabolic variation, wherein the sum of contributions of all species equals the observed variation in the metabolite. Notably, when a species' activity has large negative covariances with the activities of other community members, contribution values can be negative. Such negative contribution values indicate that a species' secretion or uptake of that metabolite varies in a way that mitigates the activity of others. Correspondingly, contribution values can be greater than 1, reflecting scenarios in which a species in fact generates more variation of this metabolite than is ultimately observed, but that its impact is mitigated by other species.

It is also worth noting that our analytical decomposition of contributions to variance is mathematically equivalent to calculating the Shapley values for the variance in metabolite concentrations (see Methods and Figure S1). Shapley value analysis is a game theory technique that defines an individual's contribution to a collective outcome, and has been shown to be the only general definition that is efficient, linear, symmetric, and assigns zero values to null contributors (Shapley, 1953). A similar, Shapley value-based approach was recently applied to address the related problem of identifying the primary taxonomic contributors to differential functional abundances in metagenomic data (Manor and Borenstein, 2017).

4.3.2 *A multi-species metabolic model for generating complex microbiome-metabolome data*

We next set out to generate a large-scale dataset of microbiome-metabolome profiles with complete information about metabolite uptake and secretion fluxes. To this end, we used a multi-species metabolic model to simulate the growth, dynamics, metabolism, and environment of a simple microbial community. This model is based on a previously introduced genome-scale framework for modeling the metabolism of multi-species communities and for tracking the metabolic activity of each community member over time (Chiu et al., 2014; Manor et al., 2014). Briefly, this framework assumes that each species optimizes its growth selfishly given available nutrients in the shared environment and predicts the metabolic activity for each species in short time increments using Flux Balance Analysis (Varma and Palsson, 1994). After each increment, the model uses the predicted metabolic activities of the various species to update the biomass of each species and the concentration of metabolites in the shared environment (hence, potentially impacting the growth and metabolism of other species in subsequent time steps). Importantly, this model allows for the natural emergence of metabolic competition and exchange between species, as well as selection for taxa with the most efficient growth rate in a given nutrient environment. Full details of this model and simulation parameters can be found in the Methods.

We specifically modeled a simplified gut community that was previously explored experimentally (Faith et al., 2011). This community includes 10 representative gut species, spanning the major clades found in the human gut and collectively encoding the key metabolic processes taking place in this environment, including breakdown of complex dietary polysaccharides, amino acid fermentation, and removal of fermentation end products via sulfate reduction and acetogenesis. Genome-scale metabolic models of these 10 species were obtained from the AGORA collection (Magnúsdóttir et al., 2016) – a recently introduced set of high-quality gut-specific metabolic models. To mimic the experimental gnotobiotic mouse setting

(Faith et al., 2011), we simulate growth in a chemostat, with a nutrient inflow mimicking the content of a standard corn-based mouse chow, and a dilution rate consistent with mouse transit time and gut volume. While maintaining this nutritional environment, we systematically explored the landscape of possible community compositions, varying the initial relative abundance of each species from 10% to 60% (with a consistent total abundance equal to the community carrying capacity), resulting in a total of 61 different community compositions. For the analysis below, we simulated growth for 144 hours (as 576 15-minute time steps). For most community compositions considered, this simulation time consisted of an initial stabilization period followed by a transition to a near-steady-state equilibrium with little change in community composition (Figure 4.1A). Notably, across the various simulations, some species maintained high abundances throughout the course of the simulation, while others reverted to lower levels.

Throughout the course of each simulation, we recorded the abundances of each species, the secretion and uptake rate of each metabolite by each species (as well as internal reaction fluxes), and the concentration of each metabolite in the environment (Figure 4.1A-B), thereby obtaining a comprehensive dataset describing species composition, metabolic activities, and metabolite concentrations across 61 different communities. To mirror the typical structure of a microbiome-metabolome cross-sectional dataset, we specifically considered the abundances of species and the concentrations of metabolites in the environment at the end of each simulation (i.e., after the final time point; see Figure 4.1). 60 of the 68 metabolites present in the nutrient inflow exhibited at least some variation across communities, as did 18 additional microbially-produced metabolites. Metabolite variation was generally low (median coefficient of variation 0.021), reflecting a relatively stable nutrient environment, yet 25 metabolites (32%) did have a coefficient of variation greater than 0.1. For downstream analysis, we excluded metabolites

without substantial measurable variance across samples, filtering those with variance at or below the 25th percentile. This resulted in a dataset of 52 variable metabolites, of which 14 are purely microbially-produced metabolites, 9 are microbially-produced but also present in the nutrient inflow, and 29 are introduced only through the nutrient inflow. Of these 52 variable metabolites, 47 are utilized by any member of the community (including 18 that are cross-fed in at least one simulation). The final species compositions and the final concentrations of several key metabolites across all simulations are shown in Figure 4.2A-F, and ordination plots of species and metabolite data are shown in Appendix C, Figure 9.2.

Exploring this dataset, we found that species composition and metabolite concentrations exhibited complex patterns and biologically reasonable distributions (Figure 9.3) (Unterseher et al., 2011). Several metabolic processes known to occur in the mammalian gut were replicated by our simulations, including, for example, conversion of acetate to butyrate by *E. rectale* (Rivière et al., 2015), and production of key microbial metabolites such as 4-aminobutyric acid (GABA), indole, and succinate. Cross-feeding relationships were observed frequently (18 metabolites), including cross-feeding of 6 amino acids, whose exchange is widespread in host-associated microbiota (Mee et al., 2014). Additionally, we ran several sets of simulations with introduced fluctuations in the nutrient inflow concentrations, and found that the resulting species compositions partially recapitulated the diet responses observed by (Faith et al., 2011) (Appendix C, Supplementary Results).

Clearly, the model and simulations described above represent a gross simplification of the microbiome's structure, dynamics, and function. Importantly, however, this simplification is also an important strength. Specifically, the data obtained from these simulations provide a unique opportunity to examine the relationship between community dynamics and metabolic activity in a realistic, yet tractable model of community metabolism where complete information

about the activity and fluxes of each microbial species is available (Figure 9.4). Indeed, our multi-species model captures many of the intricacies of bacterial genome-scale metabolism and the interconnectedness (both within and between species) of multiple metabolic processes, yet without additional complexities inherent to *in vivo* communities. Furthermore, in our simulations, variation in the concentrations of environmental metabolites results exclusively from microbial metabolic activity, with no variation in nutrient inflow or other non-microbial sources, providing a controlled setting for evaluating the relationship between community members and metabolite concentrations.

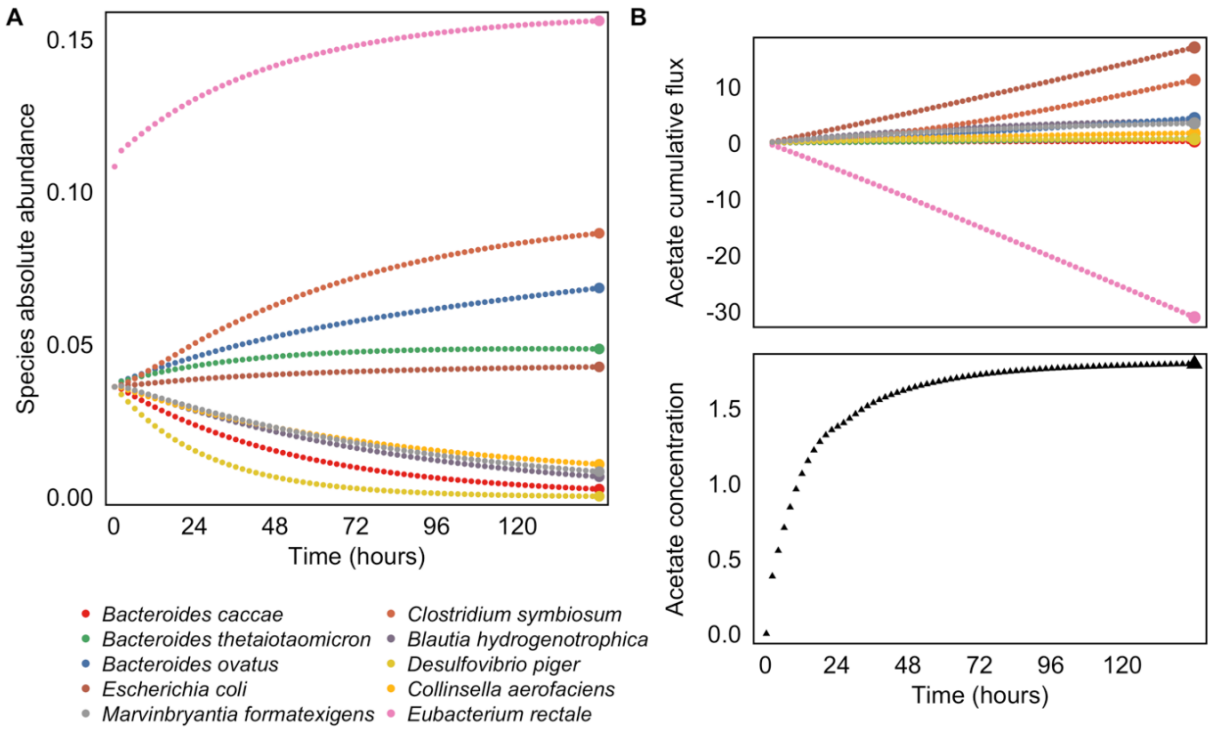

Figure 4.1. **Simulating multi-omic data with a dynamic multi-species genome-scale framework.**

(A) Community species abundances throughout a single simulation run. Abundances were quantified in units of microbial biomass. In this simulation, community composition was initialized with a high relative abundance of *Eubacterium rectale*. For visual clarity, only every eighth time step is illustrated. Species abundances at the final time point (highlighted with larger colored circles) were used for calculating species-metabolite correlations. **(B)** Cumulative

secretion and uptake of acetate by each community member, throughout the same simulation run illustrated in panel A. Acetate was synthesized by several species and consumed by *E. rectale* over the course of the simulation. Total cumulative fluxes (highlighted with larger colored circles) were used for calculating species contributions to metabolite variation. The bottom plot illustrates the resulting environmental concentration of acetate at each time point. The metabolite concentration at the final time point (highlighted with a larger black triangle) was used for calculating species-metabolite correlations.

4.3.3 *Metabolite variation is driven by diverse microbial mechanisms*

Given the simulated dataset described above (for which uptake and secretion fluxes are known), we applied our contribution framework to calculate the contribution of each species to the variation observed in each of the 52 variable metabolites (Figure 9.5). The resulting contribution values can be used as ground-truth information about the link between microbial activity and environmental metabolites.

To highlight the nature and utility of such contribution values, and to demonstrate how metabolic fluxes translate into contribution profiles, we first describe our results for several example metabolites (Figure 4.2). Putrescine, an amino acid fermentation product, is an example of the simplest case, in which one microbial species – *E. coli* – synthesizes a metabolite that is not utilized or modified by other community members. Variation in the environmental concentration of putrescine was hence fully determined by the level of secretion from *E. coli*, which is therefore assigned a relative contribution of 1 (Figure 4.2B). Tetradecanoic acid, in contrast, was introduced (at a constant rate) via the nutrient inflow and utilized by the three *Bacteroides* species in the community to varying degree (primarily by *B. ovatus* and to a slightly lesser extent by *B. thetaiotaomicron*). The calculated contribution values successfully attributed variation in the environmental concentration of this metabolite to these three species, and correctly captured the difference in the magnitude between their effects (Figure 4.2C). Variation

in uracil, another metabolite introduced via the nutrient inflow, was mainly driven by large shifts in its uptake by *B. ovatus,* but this effect is partially masked by *E. rectale,* which reduced its uptake when *B. ovatus*' flux was high and vice versa. Other species also utilized uracil, but at relatively similar levels across samples, and accordingly with relatively little impact on its variation. These complex patterns were all captured by the contribution profile obtained by our framework, with *B. ovatus* assigned a high positive contribution, *E. rectale* assigned an intermediate *negative* contribution, and other species assigned relatively negligible contribution values (Figure 4.2D). More complex species-metabolite relationships were also accurately and effectively summarized. Contribution values for acetate, for example, reflected the cross-feeding interactions that underlie variation in its concentration (Figure 4.2E). It was introduced to the shared environment by several species (primarily *C. symbiosum*), but most of its variation ultimately depended on the level of uptake by *E. rectale.* Finally, the contribution profile of succinate demonstrates how extremely strong interspecies interactions can produce contribution values much greater than the observed variance (Figure 4.2F). In the simulated data, this metabolite was synthesized by *B. hydrogenotrophica,* but was almost always fully utilized by other community members. The calculated contributions suggest that if the synthesis of succinate by *B. hydrogenotrophica* would not have been offset by uptake from other species, the variance in succinate concentration across samples would have been 71.7 times higher than is actually observed. (Note that the difference between positive and negative is always 1.)

Examining the complete set of variable metabolites and calculated contribution values revealed similar patterns of interactions (Figure 9.5). Specifically, as for the metabolites discussed above, negative contributions and/or contribution values greater than 1 were widespread. Nearly all metabolites (50 out of 52) had at least one species with a negative contribution value, and 36 had at least one species with a contribution value greater than 1. Of

the 32 other metabolites with negative contributions, 29 were present in the nutrient inflow and their negative contributions result from competition between species for their uptake. This prevalence of negative and extreme values suggests that strong negative interspecies interactions have substantial impacts on metabolite concentrations, and that often, observed variation in a given metabolite's concentration is the complex outcome of multiple species generating and offsetting much higher variation.

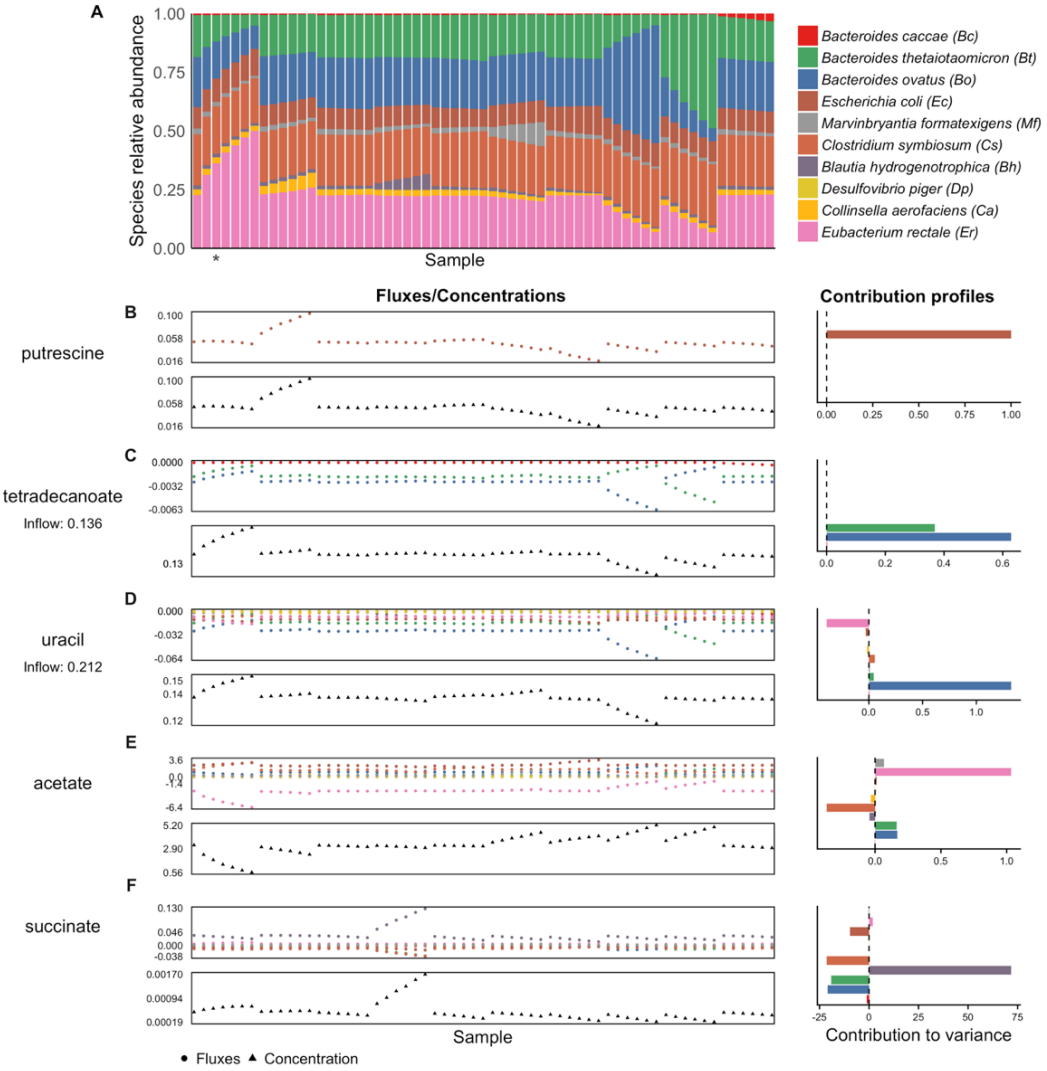

Figure 4.2. **Species abundances, cumulative fluxes, and contributions to variance in metabolite concentrations in our simulated dataset.**

(A) The dataset of species abundances at the final time point of 61 simulation runs. Each bar represents a simulation run, with the colors indicating relative abundance of each species. The

abundance profile from the simulation runs highlighted in Figure 1 is indicated with an asterisk. **(B-F)** For five example metabolites, the upper plot shows the total cumulative secretion or uptake of that metabolite by each species across all 61 simulation runs (or samples). The lower plot shows the corresponding environmental concentration at the final time point. The bar plot on the right shows the contribution values for each species and metabolite, calculated from the flux values and describing each species' linear contribution to the overall metabolite variance.

It is also important to note that while the average metabolic uptake/secretion flux of each species and the magnitude of its contribution to a given metabolite were generally significantly correlated (Spearman, $p < 0.01$ for 49 of the 52 metabolites), the species with the highest flux was often *not* the largest contributor to variation (26 of the 52 metabolites). Similarly, the variance in a species' flux was significantly correlated with its contribution for 48 of the metabolites, but for 9 metabolites the species with the most variable flux was still not the largest contributor (due to differences in whether variable flux generated by one species is compensated by variation in the flux of another). These findings suggest that even if the magnitude and variation of species uptake and secretion fluxes across a set of microbiome samples are known (rather than just the abundances of species, which is the only measure usually assayed), metabolic interdependence between species would still make true contributor species challenging to identify.

Combined, the observations above highlight the complex relationship between species activity and measured metabolite concentrations, demonstrating the important role of both direct and indirect species interactions. This complex relationship, observed even in the idealized settings of our simulation model, is potentially markedly more complex than what is assumed by many microbiome-metabolite association-based analyses.

4.3.4 *Correlation analysis fails to detect true microbial contributors to metabolite variation*

Given our observations above, we next set out to comprehensively assess how well pairwise correlation analysis (commonly used for analyzing microbiome-metabolome data) can detect true taxonomic contributors to metabolite variance. Put differently, we evaluated the extent to which a correlation between species abundance and metabolite concentration across samples captures the true causative contribution of a species' metabolic activity to observed metabolite variation.

Following numerous microbiome-metabolome studies (Kang et al., 2018; McHardy et al., 2013; Srinivasan et al., 2015; Walsh et al., 2016), we considered identifying species-metabolite relationships as a classification task, aiming to identify for each metabolite the set of species that are primarily responsible for the variation observed in its concentration across samples. To this end, we defined *key contributor* species for each metabolite as those with a contribution value greater than 10% of the total positive contribution values. This resulted in a set of 83 species-metabolite key contributor pairs, representing true links between species activity and metabolite variation. On average, each metabolite had only 1.6 contributors (Figure 9.6), although 7.5 species on average had utilized or synthesized each metabolite at any point. 31.3% of these contributions occurred via synthesis reactions, 66.3% via utilization, and 2.4% (2 instances) via both processes. We then calculated the Spearman rank correlations between species abundances and metabolite concentrations across samples, and used a p-value threshold of 0.01 to define significant correlation between species and metabolites. This produced a set of 191 significant species-metabolite correlations, representing putative species-metabolite links. Scatter plots of these species-metabolite abundance relationships are shown for several example pairs in Figure 9.7.

Comparing this set of significant species-metabolite correlations to the set of species-metabolite key contributors clearly illustrated the difficulty of using univariate associations to infer mechanistic contributions (Figure 4.3). Indeed, of the 191 significant species-metabolite correlations, the vast majority (141) were false positives (corresponding to a positive predictive value of only 26.2%), and did not represent true contributor relationships (Figure 4.3A). Moreover, more than a third of these false positive species-metabolite pairs (51 out of 141) had *no* mechanistic connection; i.e., the species did not ever use or produce the metabolite in question. Furthermore, for 12 variable metabolites (out of 52), none of the key contributors were successfully detected by a correlation analysis. The overall accuracy was somewhat higher (66.5%), reflecting the high number of non-contributors that are also not correlated. Using a stricter cutoff ($p < 0.0001$, equivalent to a Bonferroni-corrected value of 0.05) only improved the positive predictive value to 33% and the accuracy to 77.1%. Indeed, a ROC curve analysis (Figure 4.3B) produced an area under the curve of 0.72, and overall correlations and scaled contribution values were only weakly associated (Figure 4.3C), suggesting that these findings can only be partially mitigated by changing classification thresholds. Metabolites of different classes had generally similar correspondence between correlations and contributions (Figure 4.3D).

Notably, key contributors for purely microbially-produced metabolites were not identified more accurately than those for metabolites in the nutrient inflow (66% versus 67%), which is perhaps not surprising since we used a constant inflow across samples (but see also our analysis below with variable inflow). Moreover, the total variance in a metabolite was not associated with the accuracy or predictive value with which key contributors for that metabolite were identified (Spearman rho, $p > 0.1$). Across species, contributions were identified most

accurately for *D. piger*, which had a relatively low number of contributions (Figures 4.3E and C.5C), but the positive predictive value was nonetheless <50% for all species.

We obtained similar results across several variants of this analysis (Appendix C, Supplementary Results, Figures C.6, C.8, and C.9). To assess the impact of dynamic shifts over the duration of each simulation, we calculated an alternative set of contribution values based on the net steady-state metabolite flux rates at the final time point of each simulation, finding extremely similar results as for contributions to cumulative variation in concentration. We also evaluated the use of an alternative classification task, aiming to detect all microbes that affect variation in a given metabolite across samples regardless of whether their effects are ultimately reflected in the observed concentrations (i.e. those with large positive or negative contributions), again resulting in similar findings (Supplementary Results, Figure 9.6). Finally, we profiled the effects of model simulation parameters on correlation results, including the simulation length and the maximum enzymatic rate *V_max*, again finding minimal effects on contribution and correlation results (Supplementary Results, Figures C.8-9).

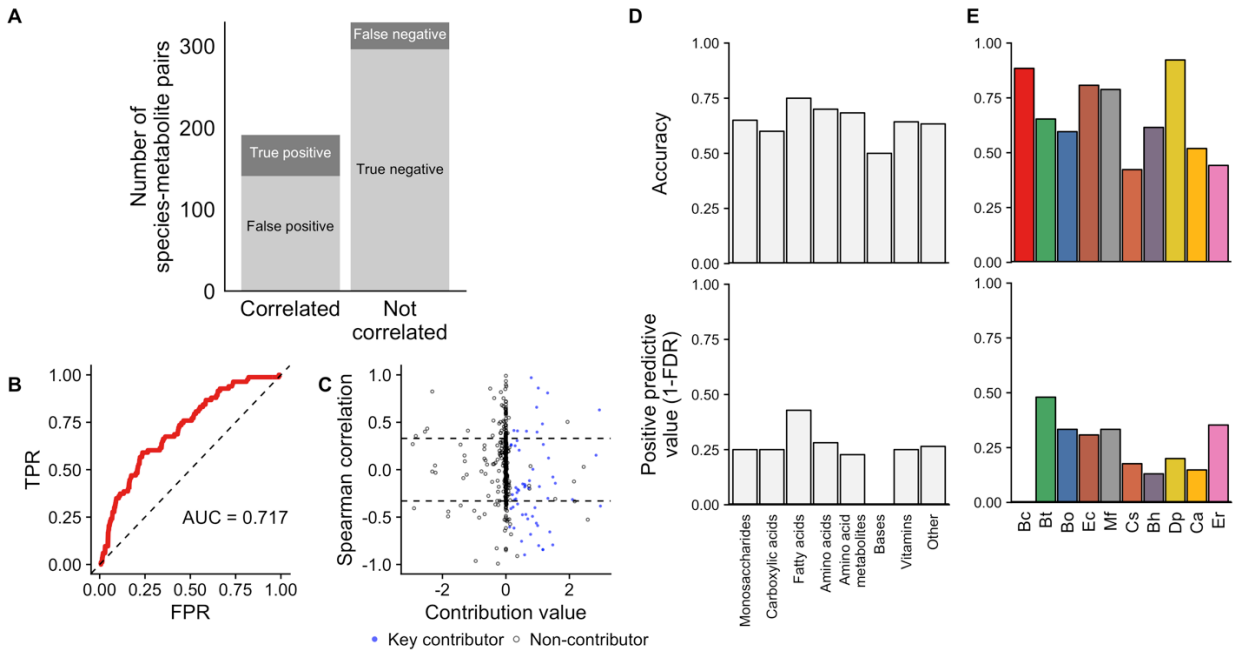

Figure 4.3. **Species-metabolite correlations poorly predict species contributions to metabolite variation.**

(A) The number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (indicated by shade of gray). **(B)** Receiver operating characteristic (ROC) plot, showing the ability of absolute Spearman correlation values to classify key contributors among all species-metabolite pairs. **(C)** Scatter plot of species-metabolite pairs, showing the poor correspondence between true contribution values (x-axis) and Spearman correlation (y-axis). Key contributors are plotted as blue points, others as hollow circles. Dashed lines show significant correlations (p<0.01). There are 65 species-metabolite pairs with a contribution value greater than 3 in magnitude whose values are not shown. **(D-E)** Accuracy and positive predictive value of Spearman correlation analysis for detecting true key contributors across metabolite classes (Panel D) and for each of the 10 species (Panel E).

4.3.5 *Species and metabolite properties explain discrepancies between correlations and contributions*

Our analysis above demonstrated that correlations between species abundances and metabolite concentrations can often be only poorly associated with true contribution of species to

metabolite variation. We therefore next investigated the origins of such discrepancies. We examined whether individual metabolites or species are predisposed to produce a significant species-metabolite correlation when the species in fact does *not* contribute to that metabolite variation (i.e., false positives), or to mask such correlation when the species *does* in fact contribute to this metabolite variation (i.e., false negatives), and if so, what species and metabolite properties are linked to those outcomes.

To determine whether the identity of the species or metabolite in question can explain inaccurate identifications of key contributors, we used a regression-based analysis. Specifically, we considered all species-metabolite non-contributor pairs, and fitted a logistic regression model to predict whether a species-metabolite pair exhibited significant correlation (false positive), based on either species identities, metabolite identities, or both (Methods). We then compared these three models using a likelihood ratio test to assess whether species and/or metabolite identities are informative. We similarly considered all species-metabolite key contributor pairs separately, again fitting a logistic regression model based on species identities, metabolite identities, or both to predict whether a pair failed to exhibit significant correlation (false negative).

For non-contributors, we found that false positives can be explained largely by species identity (likelihood ratio test (LRT) for inclusion of species terms $p < 10^{-13}$). Incorporating both species and metabolite identities did not significantly improve the model (LRT for metabolite terms p=0.72). This finding suggests that false positives – correlations observed between species and metabolites to which they in fact did not contribute – are the outcome of interactions at the species level, regardless of the metabolite in question. This impact of strong interactions between dataset features on association test results has been described extensively in other data types (33, 34). Indeed, examining the 141 false positives identified above, we found that many can be

explained by the relationships between the three dominant species in this community: *E. rectale*, *B. thetaiotaomicron*, and *B. ovatus*. These species competed strongly for carbon sources (and utilized their maximum allocation of sucrose, glucose, and fructose at nearly every step of the simulation), and their abundances were therefore negatively correlated. As a result, metabolites that varied due to the activity of one of these species were also frequently correlated with the other two. In total, 32 false positive correlations paired one of these species with a metabolite for which another species in this trio was a key contributor. More generally, we found that the probability of a false positive correlation for a particular species and metabolite depended on the species' correlation with the true key contributors for that metabolite (p=0.006, Spearman rho between share of false positives and interspecies correlation; Figure 4.4A). Moreover, the maximum correlation each species had with any other species is a strong predictor of its overall specificity, which varies widely from 33.3% for *E. rectale* to 92% for *D. piger* (Spearman rho=-0.84, p=0.002). We also found that species identity was similarly predictive of whether a significantly correlated metabolite-species pair represented a true contributor versus a false positive (Appendix C, Supplementary Results).

In the case of key contributors, we found that false negative correlations can be explained largely by metabolite identity (LRT for metabolite terms p=0.002; although the species involved was also somewhat informative with LRT p=0.08). Put differently, a lack of correlation between the abundance of a key contributor species and the concentration of the metabolite to which it contributed was determined mainly by the nature of the metabolite in question. This lack of correlation between a given metabolite and its contributors could have resulted from competition or exchange of a metabolite between multiple species, such that none of the involved species end up strongly associated with the final outcome on their own. Indeed, across all metabolites, the average correlation between a metabolite and its key contributors is negatively associated with

its number of key contributors (Spearman rho=-0.45, p=0.0008). The number of key contributors for any metabolite was also thus negatively associated with the sensitivity of contributor detection for that metabolite (Spearman rho=-0.48, p=0.0004; Figure 4.4B). We further hypothesized that false negative outcomes might be more common for metabolites with more or larger negative species contributions, since these, by definition, mask or compensate for the activity of key contributor species. While all metabolites with a false negative outcome did have at least one species with a negative contribution value, as mentioned above, this was true for nearly all analyzed metabolites (50/52), and the number of negative contributing species was not associated with the occurrence of a false negative correlation (p=0.86, Wilcoxon rank sum test). Moreover, we also did not observe any effect of the average concentration of a metabolite on the sensitivity and accuracy of its detection via correlation analysis, nor of whether it is secreted, utilized, or cross-fed (Figure 4.4C). In summary, our analysis suggests that the largest factor explaining whether a metabolite's key contributor can be detected by a correlation analysis is simply whether there are other community members (key contributors) that also impact the observed concentration of that metabolite.

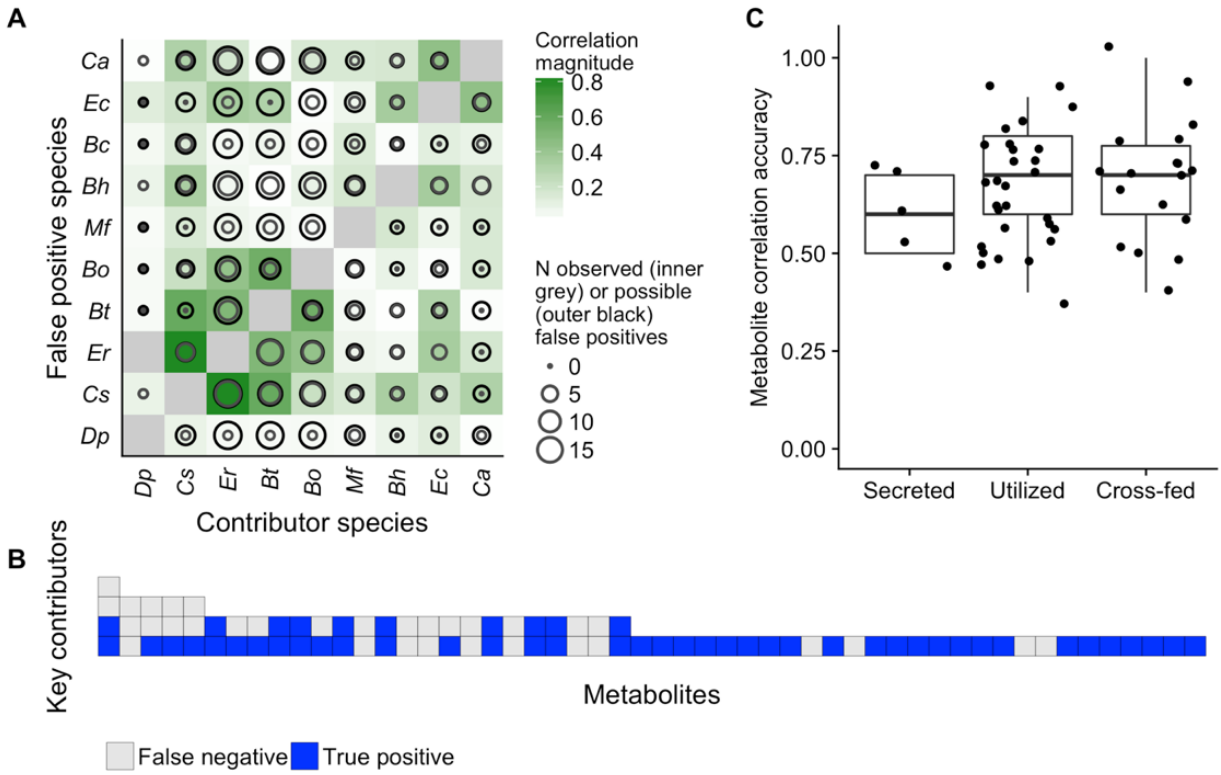

Figure 4.4. **Metabolite and species properties explain correlation-contribution discrepancies.**

(A) Strongly correlated species pairs produced more false positive metabolite correlations. In this plot, the color of each tile indicates the strength of correlation in the abundances of each pair of species. The size of the outer black circle in each cell represents the number of metabolites for which the species on the x-axis is a key contributor and the species on the y-axis is not. The size of the inner circle represents the share of those metabolites for which a false positive is observed for the species on the y-axis. It can be seen that many false positive correlations involve the taxa with the strongest interspecies associations: *E. rectale, B. ovatus,* and *B. thetaiotaomicron.* **(B)** Metabolites with more microbial key contributors were more prone to false negative correlations. Each column represents an analyzed metabolite, ordered by its number of key microbial contributors, which are represented by each tile. The tiles are coded by the correlation outcome for each contributor. **(C)** Correlations detected key contributors equally accurately regardless of whether a metabolite is secreted, utilized, or cross-fed by the species. Each point represents the accuracy of correlations for a single metabolite across its comparisons with all 10 species.

4.3.6 *Environmental fluctuations in metabolite concentrations impact detection of key contributors*

Our analyses above all focused on a single simulated dataset in which the nutrient inflow was constant across all samples, meaning that metabolite variation was fully governed by microbial activity. However, in reality, metabolite variation can and does arise also from non-microbial sources, potentially affecting both the landscape of key microbial contributors and our ability to detect them via correlation-based analyses. To explore the impact of environmental fluctuations, we therefore ran several sets of additional simulations with varying degrees of nutrient fluctuation, designed to emulate a range of levels of experimental diet control and variation in host absorption across the simulated mouse gut communities. In these simulations, we maintained the same set of 61 initial species compositions but added small amounts of stochastic noise to the nutrient inflow, sampling inflow concentrations for each compound in each simulation from a normal distribution with a mean equal to the compound's original inflow rate and a standard deviation ranging from 0.5% to 10% of the mean in 8 increments (Methods). For each of the resulting 8 datasets, we again calculated contribution values (with the added element of the nutrient inflow as a potential contributor to variance), identified key contributors, and compared them with the results of a correlation analysis.

Examining the obtained contribution values, we found, as expected, that variation in inflow quantities can outweigh the variation in microbial fluxes, and that as the variation in inflow increases, its contribution to metabolite variation increased at the expense of the contributions of community members (Figure 4.5A). As a result, the number of key contributions attributed to each species decreased for metabolites in the nutrient inflow (Figure 4.5B). Interestingly, however, some species lost their contributions more gradually than others, and in some cases even became key contributors for additional metabolites (Figure 4.5B). For most

metabolites, the relative ranking of species with the highest contribution values was unchanged with increasing fluctuations (Appendix C, Supplementary Results).

We next examined how correlation-based detection of key microbial contributors was affected by these inflow fluctuations. We assigned each of the 52 metabolites in each of the 9 datasets (the original dataset with no inflow fluctuations and the 8 datasets with varying degree of fluctuations) to bins according to the level of contribution attributed to the inflow for this metabolite at that degree of fluctuation (see Methods). We then evaluated the performance of correlation analysis for each bin separately. The share of true key contributors naturally decreased rapidly with increasing environmental contribution, as did the number of significantly correlated species-metabolite pairs (Figure 4.5C). Importantly, however, the sensitivity of correlations decreased substantially with the level of contribution attributed to the inflow, but the specificity in fact increased from 67.7% to 92.3% (Figure 4.5D). This suggests that while environmental fluctuations disrupted the signal linking microbial species with the metabolites they impact, they also disrupted indirect associations between species and metabolites (false positives). Overall, however, the AUC did not change significantly with increasing environmental contribution (Figure 9.10A), and the positive predictive value is similarly relatively stable (and never rose higher than 37%). Interestingly, the detection of some metabolites not present in the inflow was also affected by inflow fluctuations in a similar manner (Appendix C, Supplementary Results, Figure 9.10B).

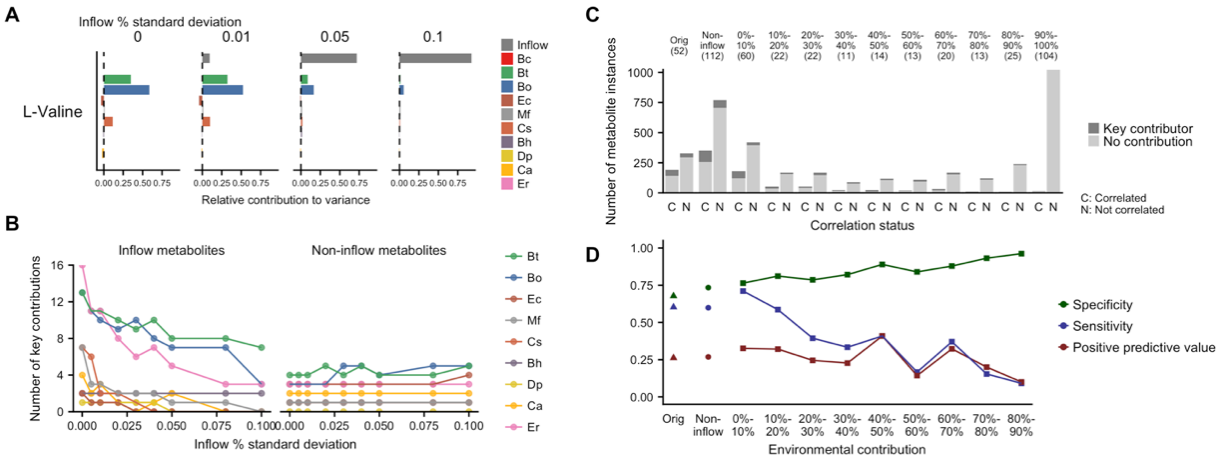

Figure 4.5. **Environmental fluctuations impact correlation-contributor sensitivity and specificity.**

(A) Example set of contribution profiles for a single inflow metabolite, L-valine, with increasing fluctuations in its inflow. The relative contribution values for each species and for the inflow are shown for 4 sets simulation runs, each with a different degree of fluctuation. The label on each plot describes the relative standard deviation (coefficient of variation) of inflow concentrations for that set of simulations. The microbial contributions to variance in L-valine concentrations became smaller with increasing variation from the external environment. **(B)** Shifts in key microbial contributors with increasing environmental inflow fluctuations. The number of key contributions of each species to the 52 analyzed metabolites is shown, separately for metabolites present in and absent from the nutrient inflow. Microbial contributors to inflow metabolites decreased as environmental contributions increased, but this effect varied between taxa. **(C)** Correlation analysis failed to detect key microbial contributors regardless of the size of contribution from external inflow. Across all sets of simulations, metabolites were binned based on the percent of total positive contribution from the external inflow. The bar plots shown have the same format as Figure 4.3A, showing the number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (shade of gray). The first two bars, labeled "Orig" describe the original set of simulations (replicating Figure 3A). The next two show the results for non-inflow metabolites across all levels of inflow fluctuations. The remaining bars show the results for metabolites with increasing levels of environmental contribution. **(D)** Correlation analysis detected key microbial contributors with increased specificity, decreased sensitivity, and

generally consistent positive predictive value with increasing contribution from the external inflow. Statistics are shown for same environmental contribution bins as in Panel C.

4.3.7 *Correlation analysis is similarly limited in simulations of more complex and diverse human gut microbiota*

Our results have illustrated consistent discrepancies between microbe-metabolite correlations and microbial contributions to metabolite variation in a model ten-species community. We lastly addressed the question of whether these findings generalize to more complex mammalian gut microbiota, communities with many times more taxa and a more uneven distribution across individuals. To do so, we ran an additional set of simulations emulating human gut microbiota transplanted into gnotobiotic mice. We first mapped 16S rRNA sequence variants from the Human Microbiome Project (Huttenhower et al., 2012) to the genomes of the AGORA model collection at 97% sequence identity (Magnúsdóttir et al., 2016), and selected 57 samples with a successful mapping rate greater than 25% relative abundance. The total share of mapped reads averaged 36.7% across these samples, with a maximum of 73.5%. Despite this variation, mapped reads displayed features typical of Western gut microbiomes, including a predominance of Bacteroidetes and Firmicutes phyla along with varying lower abundances of Actinobacteria and Proteobacteria (Figure 4.6A). The number of species identified in each sample ranged from 23 to 62, with a median of 42. We ran a simulation based on each sample by setting the initial species relative abundances according to the relative abundances of mapped reads, while maintaining the same physical parameters as previous simulations (see Methods for additional details). We used nutrient inflow quantities with 1% standard deviation between samples. Initial species compositions displayed characteristic shifts in abundance over the simulation time course (Figure 9.11A). Metabolites were also highly variable, with a median coefficient of variation of 71% across 222 metabolites (Figure 9.11B).

We calculated contribution values for this dataset, finding a smaller share of key contributions (only 392 out of 29,082 possible species-metabolite pairs). Only 35.1% of species (46 out of 131) were identified as key contributors to any metabolite. The genera with the most contributions were *Bacteroides*, *Ruminococcus*, and *Enterobacter*, which were also three of the four most abundant genera in the final dataset (Figure 4.6B).

In this noisier and more layered dataset, only a small share of species-metabolite pairs was significantly correlated. In order to fairly compare with the previous dataset while accounting for the larger number of hypothesis tests, we defined significance based on an equivalent Benjamini-Hochberg estimated false discovery rate (0.027) as the $p < 0.01$ cutoff used for the previous dataset. 2.2% of species-metabolite pairs displayed significant correlations at this cutoff ($p < 0.00058$). This level of correlation is comparable to a recent microbiome-metabolome study of the colon of healthy humans (McHardy et al., 2013), in which 1.4% of OTU-metabolite pairs displayed Spearman correlation coefficients of the same effect size. In our dataset, correlation analysis detected contributors with high specificity (98.4%), and an area under the ROC curve of 0.89. However, the positive predictive value was still only 29.0%, rising as high as 57% with a significance cutoff of $p < 10^{-10}$. We compared these classification results with the original dataset, finding that despite the difference in overall AUC, sensitivity, sensitivity, and predictive value are similar or worse for the two datasets at commonly used FDR thresholds between 0.1 and 0.01 (Figure 4.6C), and sensitivity and predictive value are both highly dependent on the choice of significance threshold. As in the ten-species dataset, a large share of false positive species-metabolite pairs (65.4%, 291 out of 445) also involved species with no capacity to impact the metabolite in question.

The outcomes of correlation analysis were influenced by the same factors as observed in the model community dataset, but also by several additional characteristics. False positive

classifications were, again, driven by interspecies covariance: Species significantly correlated (at 10% FDR) with a true key contributor for a metabolite were 13.6 times more likely to have a false positive correlation with that metabolite than species with no such link ($p < 10^{-16}$). Notably, the false positive rate of a given species was also substantially affected by its prevalence: the number of samples in which a species was present was negatively associated with its specificity (Spearman rho = -0.57, p=0.002, Figure 9.11C), among species with at least 3 key contributions. In other words, widely prevalent species were more prone to false positive correlations than rarer species.

Properties of both metabolites and species were again linked to false negative contributions. As in the ten-species dataset, species contributions to metabolites with more than one key contributor were 5.2 times more likely to not be correlated than those that were the sole key contribution for a metabolite ($p < 10^{-10}$, Fisher exact test). In this dataset, an elevated share of these metabolites with multiple key contributors were cross-fed between different species (p=0.00007, Fisher exact test), and correspondingly, key contributors for cross-fed metabolites were also 1.6 times less likely to be significantly correlated (p=0.02). Both cross-feeding and false negative outcomes occur variably across metabolite classes, with nucleotide metabolites having the highest rates of both phenomena (Figure 9.11D). Taken overall, our simulations and analysis of this realistic microbiota simulation demonstrates that correlation analysis can have greater utility in a microbial community dataset with greater complexity and variability, but the results are again strongly influenced by properties of individual metabolites and species.

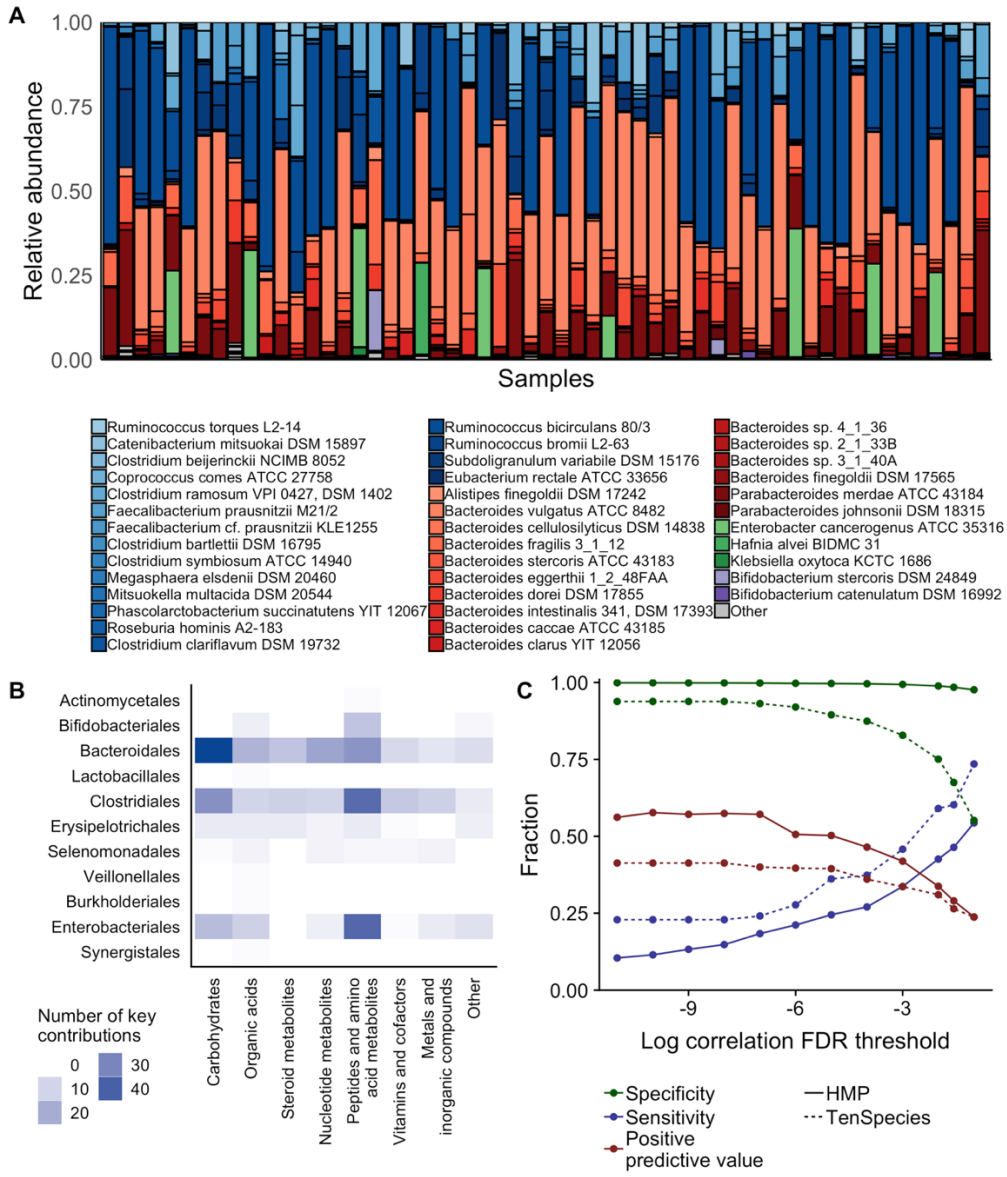

Figure 4.6. **Correlation-contribution discrepancies persist in simulations of complex human gut-based microbiota.**

(A) Species abundances of the 57 Human Microbiome Project (HMP) based-simulations at the 144 hour time point. Shades of blue indicate species in the phylum Firmicutes; red, Bacteroidetes; green, Proteobacteria; and purple, Actinobacteria. **(B)** Key contributions to metabolite variation across the HMP-based dataset, summarized at the level of taxonomic orders and metabolite categories. **(C)** Performance of correlation analysis for identifying key species-metabolite contributors in the HMP-based dataset (solid lines) compared with the original 10-

species dataset (dashed lines) across varying significance levels, using Benjamini-Hochberg false discovery rate (FDR) corrected *p*-values.

4.4 Discussion: Insights and implications for microbiome-metabolome analyses

Above, we have investigated the ability of correlation-based analyses to detect key microbial contributors responsible for variation in metabolite concentrations across samples. Our findings suggest that microbe-metabolite correlation analysis may be a useful approach for exploratory analyses, but they highlight some of the limitations and caveats of such microbiome-metabolome studies and identify several factors that impact the relationship between community composition and metabolite concentrations. Below, we elaborate on a set of practical conclusions and their implications for the analysis and interpretation of microbiome-metabolome studies.

Association-based analyses of microbiome-metabolome assays have low predictive value for detecting direct species-metabolite relationships and require conservative interpretation. Microbiome-metabolome association studies have been previously proposed as a powerful tool for the identification of causal mechanisms of microbiome metabolism (Gilbert et al., 2016), and indeed, such studies often present detected associations as evidence for mechanistic relationships (Choo et al., 2017; De Filippis et al., 2015; Lin et al., 2018). However, our analysis suggested that the positive predictive value of significant species-metabolite correlations for identifying true microbial contributors can be extremely low: less than 50% across all settings, as low as 10% in the context of large environmental fluctuations, and 29% in simulations based directly on human gut composition. Recent experimental studies pairing microbiome-metabolite correlation analysis with *in vitro* monoculture validations have similarly anecdotally observed many false positive correlations (Hoyles et al., 2018). Additionally, given the somewhat low sensitivity observed in our analysis, a lack of association is not necessarily

sufficient to reject a hypothesis that a particular microbial taxon impacts a particular metabolite. The choice of correlation threshold should therefore be chosen carefully, taking into account the complexity of the community and the environmental context. In general, identified correlations between microbial taxa and metabolites should be interpreted very conservatively and used mostly to prioritize microbe-metabolite relationships for follow-up validation studies (e.g., via culture-based studies or germ-free model organism colonization). One potential approach for improving the predictive value of such correlation-based analyses is to examine whether they replicate across multiple conditions. Indeed, we found that a correlation does provide stronger evidence for a contributor relationship if it persists across different contexts. Across our 9 simulated datasets with varied environmental fluctuations, the 43 species-metabolite pairs that were significantly correlated in every dataset were 2.1 times more likely to denote true key contributor relationships than other significant correlations (Fisher exact test, p=0.05), although their positive predictive value was still relatively low (39.5%). Of the limited number of significant correlations shared between our original and HMP-based datasets (n=5), all were false positives in both datasets, reiterating the need for caution.

The predictive power of correlation-based analysis is species-, metabolite-, and context- dependent. In our datasets, metabolites varied widely in both contribution profiles and in their detectability via correlation analysis. In particular, the key contributors for metabolites acted upon by fewer species, and potentially those that are not exchanged between different species, were identified more readily. Moreover, in our simulations of human gut communities, contributions by less prevalent species were identified much more accurately than those by widely-found species, indicating that hypotheses based on associations of rarer species should potentially be prioritized. Correlation analysis may thus identify microbes involved in specialized secondary metabolic processes (e.g. products of complex biosynthetic pathways)

more readily than those involved in more widespread processes. Therefore, correlation-based approaches may be more informative for analyzing compounds that are specific to a small number of rare taxa, but accurate dissection of the taxa controlling variation in widely-trafficked metabolites may require more detailed analysis and experimentation. Similarly, we found that species-metabolite correlations for species that are strongly associated with other taxa (e.g., those with tight interactions with other community members) are often spurious, suggesting that such correlations should be regarded less confidently.

External metabolic fluctuations can strongly impact the detection of microbial contributions. Our analysis of the impact of environmental fluctuations suggested that the presence of environmental variability from a diverse set of samples could in fact increase correlation specificity. We also found that the sensitivity of correlation analysis rapidly decreased with increasing environmental fluctuations (from 60% to 9%). These observations suggest that while a tightly controlled environment (e.g., identical diets) is intuitively expected to increase the strength of microbiome-metabolome studies, its value depends on the study priorities. Specifically, if the goal is to identify clear-cut microbial drivers of healthy- and disease-associated metabolite shifts, stochastic variation in nutrient availability could be beneficial as it may reduce the rate of false positive associations. In contrast, for studies searching for a particular microbial taxon's involvement in a particular process (e.g. aiming to determine whether an ingested probiotic impacts aspects of gut metabolism), a more controlled environment may be favorable. It should, however, be noted that our findings were based on environmental fluctuations that were uniform and independent, which may not hold for real-life environmental fluctuations such as diet variation. It is also worth noting that in our simulations, microbial fluxes for some environmental metabolites could be drowned out by as little as 0.5% variation in nutrient inflow quantities, while others still had substantial microbial contributions

even with 10% variation in inflow. When interpreting an observed association, the scale of possible microbial variation relative to external variation should therefore be taken into account.

Mechanistic reference information can improve the predictive power of microbiome-metabolome studies. In our simulated dataset, 36% of the false positive correlations occurred between a metabolite and a species that was in fact not capable of uptaking or secreting that metabolite. Ruling out such falsely detected links would substantially improve the positive predictive value of a correlation-based analysis. One approach for doing so is by utilizing genomic information, which can be obtained or predicted for many microbial taxa (Langille et al., 2013). By coupling such genomic information with metabolic databases such as KEGG or MetaCyc (Caspi et al., 2014; Kanehisa and Goto, 2000), researchers can filter out correlation-based links that are likely not feasible causative relationships. Further improvement can be obtained by integrating such reference information directly into the analysis. Indeed, we previously introduced a computational framework, termed MIMOSA (Noecker et al., 2016), that utilizes a simple community-wide metabolic model to assess whether measured metabolite variation is consistent with shifts in community metabolic potential, and to identify potential contributing taxa. MIMOSA has been applied to varied host-associated microbiomes from varied body sites and from human and mouse hosts (Casero et al., 2017; Snijders et al., 2016; Stewart et al., 2017). Applying MIMOSA to the simulated ten-species dataset analyzed above (Methods), we found that it indeed identified key contributors significantly more accurately than a correlation-based analysis, with an AUC of 0.89 (Figure 4.7). Notably, in this analysis, we assumed MIMOSA has access to the correct set of metabolic reactions possessed by each species. Using standard less-complete information obtained directly from the KEGG database (as done regularly when using this tool) reduced the number of metabolites that could be analyzed from 52 to 39, with improved specificity (96%) and positive predictive value (61%) and an

ultimately comparable AUC (0.74). Combined, these findings suggest that reference model-based approaches can provide stronger evidence for mechanistic relationships than strictly correlation-based methods, but their use depends on complete and high-quality metabolic reference databases.

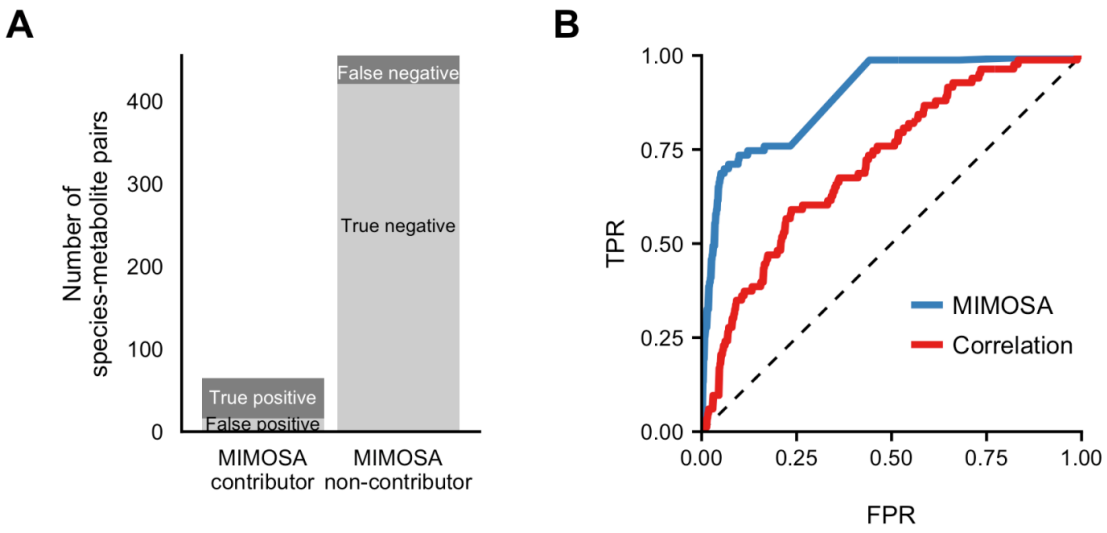

Figure 4.7. **MIMOSA identified key microbial contributors more accurately than correlation analysis.**

(A) The number of species-metabolite pairs that were identified as potential contributors (left bar) or not (right bar) by MIMOSA, and its correspondence with true key contributors. **(B)** Receiver operating characteristic (ROC) plot, showing the ability of both MIMOSA and absolute Spearman correlation values to classify key contributors among all species-metabolite pairs.

4.5 Future Opportunities and Challenges

Microbiome-metabolome studies have an important role in microbial ecology research. They specifically have great potential to dissect the metabolic interactions of complex microbial communities, and to unify "top down" and "bottom up" microbiome research approaches by providing mechanistic information at a systems level. Moreover, from a translational perspective, microbiome-metabolome studies can inform efforts to design targeted therapies to alter specific microbial or metabolic features of a community (Shaffer et al., 2017). Such interventions require

first identifying putative targets, which in many cases may entail identifying the key contributor species that drive observed shifts in a particular beneficial or detrimental metabolic phenotype.

Importantly, while we show here that a correlation-based analysis may be limited in its ability to identify these key microbe-metabolite links, this does not necessarily imply an inherent limitation of microbiome-metabolome data. For example, analyzing our data, we found that species abundance is in fact a very good proxy for metabolic activity (median correlation of 0.996 between abundance and flux for all species-metabolite pairs), meaning that the variance in total species abundance drastically outweighs the individual-level variance in flux rates. When we further examined whether false negative associations in our original dataset stem from a disconnect between the abundance of a species and its metabolite uptake or secretion rates, we identified only 2 undetected key contributor pairs that could be explained by such a discrepancy. This analysis suggests that taxonomic abundance data is sufficient to explain and model community metabolic variation to great extent, despite common concerns about potential discrepancies between community composition and function. It also suggests that metatranscriptomic expression data may not provide much additional value for this purpose, as other studies have indicated (Franzosa et al., 2014; Iwai et al., 2016; Langille et al., 2013).

Given the increasing prevalence of microbiome-metabolome studies, their promise, and the caveats of association-based research discussed above, further development of computational and statistical methods for analyzing such datasets is clearly needed. Possible directions include the use of multi-species dynamic metabolic models that can replicate experimental observations (Magnúsdóttir and Thiele, 2018), multivariate approaches for deconvolving interactions between species and the environment (Doledec and Chessel, 1994; Randolph et al., 2015), and probabilistic methods that can integrate prior information while allowing for other unknown mechanisms (Chong and Xia, 2017; Zhu et al., 2012). The conceptual framework of taxon-

metabolite contributions, and the use of dynamic simulations demonstrated here, can both inform the future development and evaluation of such methods.

There is also a continued need for gold standards to evaluate new methods. This study is only a first step in that direction and has analyzed one specific type of research question: identifying microbial taxa directly responsible for variation in metabolite concentrations between samples in a cross-sectional study design. Although this focus describes many recent microbiome-metabolome studies, other studies may address a wide range of complementary research questions, and correspondingly, the desired "ground truth" can take different forms. Moreover, depending on the objective, an alternative definition of a taxon-metabolite relationship may be required. For example, it may be valuable to identify key contributors that act via alternative mechanisms, such as by modifying substrate availability or environmental conditions (e.g. (Keren et al., 2015)), or to distinguish metabolite variation arising in response to a perturbation from variation due to differences in steady-state metabolism between communities. Additionally, our findings rely on an *in silico* system that may not capture many aspects of community ecology and metabolism, and it is possible that the predictive value of correlation analysis, as well as of other analytical methods, differs fundamentally in this system as compared to true biological systems. Further studies should also consider additional variables such as community diversity, sample size, measurement error, and other types of environmental variation. Ongoing technology developments in mass spectrometry and stable isotope probing will ideally enable future evaluation analyses using experimental, quantitative, species-specific community flux data to define key microbial contributors (Berry et al., 2013; Kurczy et al., 2016). Such evaluations can also take advantage of datasets comparing community microbiome-metabolome data with *in vitro* monoculture or mono-colonization data (Biggs et al., 2016; Hoyles et al., 2018; Kešnerová et al., 2017).

Ultimately, much remains to be learned about the many processes through which complex microbial communities shape their environment. The first major call for the application of metabolomics to microbiome research, published 10 years ago (Turnbaugh and Gordon, 2008), noted that new methods will be necessary to integrate genomic and metabolic data and inform the prediction of community metabolic properties from metagenomes. Now that microbiome-metabolome datasets are widely available, ongoing development of analysis methods for these studies has great potential to generate new knowledge. Moreover, future work in this area stands to benefit from the utility of dynamic, multiscale metabolic modeling. Detailed mechanistic simulations are used widely in astronomy, climate science, and other fields to make methodological choices and assess possible experimental outcomes when ground truth measurements are unavailable or difficult to obtain (Collins et al., 2006; Connolly et al., 2014). An analogous strategy in microbiome research may be similarly fruitful.

4.6 METHODS

4.6.1 *Derivation of species contributors to variation*

We derived an expression representing the contribution of each species to the variance in the concentration of each metabolite. While we describe this calculation in terms of species, a similar calculation could be done at the level of phyla, strains, or any grouping of the community for which metabolite secretion and uptake fluxes are available.

The concentration of a given metabolite M at the end of a single simulation run is a function of the uptake and secretion fluxes (responding to the species' degradation and synthesis activities) of the n species, the environmental inflow over all time steps m_{in}, and the dilution m_{out} out of the chemostat over all time steps:

$$M = \sum_{i=1}^{n} m_i + m_{in} - m_{out} \qquad (4.4)$$

The value of m_{out} at a given time step t is the product of the dilution rate D and the metabolite concentration at the previous time point (see above). This fact can be used to express m_{out} in terms of all the previously recorded environmental inflow and microbial activities. The metabolite concentration at any time point t, $M(t)$, is then equal to:

$$M(t) = \sum_{k=1}^{t-1}[(1-D)^{t-k-1}\sum_{i=1}^{n} m_{ik}] + m_{in}\sum_{k=1}^{t-1}(1-D)^{k} \quad (4.5)$$

where m_{ik} represents the activity of species i at a single time point k. We can then ignore dilution outflow by replacing each activity value m_i in the final concentration calculation above with a value corrected for the mitigating effect of chemostat dilution over the course of the simulation up to time t, defined here as m_i*. m_i* represents the total amount of a compound secreted or uptaken by species i, minus the share of that quantity that is eventually diluted out over the course of the simulation.

$$m_i^* = \sum_{k=1}^{t-1}(1-D)^{t-k-1} m_{ik} \quad (4.6)$$

and thus,

$$M = m_{in} + \sum_{i=1}^{n} m_i^* \quad (4.7)$$

In this work, we refer to "environmental fluctuations" as the effect of the independently parameterized nutrient inflow, m_{in}, and where not otherwise specified we use m_i to imply m_i*, a species activity quantity that accounts for the corresponding subsequent dilution out of the system.

Using the expression above, *var(M)* can then be clearly expressed as a sum of correlated environmental and microbial random variables:

$$var(M) = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(m_i, m_j) + \sum_{i=1}^{n} cov(m_i, m_{env})$$
$$= \sum_{j=1}^{n} var(m_j) + var(m_{env}) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} cov(m_i, m_j) + 2\sum_{i=1}^{n} cov(m_i, m_{env})$$
(4.8)

This expression can then be partitioned additively into *n+1* terms representing the contribution of each microbial species and of fluctuations in the environmental nutrient inflow.

$$c_i = \sum_{j=1}^{n} cov(m_i, m_j) + cov(m_i, m_{env}) = var(m_i) + \sum_{j\neq i} cov(m_i, m_j) + cov(m_i, m_{env})$$
(4.9)

4.6.2 *Multi-species Dynamic Flux Balance Analysis modeling*

In this study, we simulated the growth and metabolism of a community of 10 representative gut species that was previously explored experimentally (Faith et al., 2011). We specifically utilized a previously introduced multi-scale framework for modeling the dynamics and metabolism of multiple microbial species in a well-mixed shared nutrient environment (Chiu et al., 2014; McNally and Borenstein, 2018). This framework assumes that each species in the community aims to maximize its own growth on a short time scale given available nutrients, and uses Flux Balance Analysis to predict the growth and metabolic activity of each species at this short time scale (Varma and Palsson, 1994). The shared environment is then iteratively updated based on the species' predicted growth, uptake, and secretion rates, such that metabolic interactions are mediated via the environment as a natural byproduct of species activities, rather than being explicitly modeled (Manor et al., 2014).

We used genome-scale metabolic model reconstructions of the 10 community members from the AGORA collection version 1.01 (Magnúsdóttir et al., 2016), which have been consistently curated to remove or modify thermodynamically unfavorable reactions, remove futile

cycles, and confirm growth in anaerobic environments on expected carbon sources, with additional curation for several biosynthesis pathways. The COBRA toolbox was used to convert each AGORA model to MATLAB format (Schellenberger et al., 2011). The growth and metabolism of the 10-species community were simulated in a chemostat setting in 15-minute time intervals. We set the chemostat volume to be approximately equal to a mouse gut (0.00134 liter (Casteleyn et al., 2010)). We similarly set metabolite inflows to emulate the macronutrient and micronutrient quantities in a corn-based mouse chow (Faith et al., 2011) (provided in Supplementary Data 1).

The simulations were performed following a previously introduced procedure (Chiu et al., 2014), repeated for each time step t_n: First, the maximum uptake rate for all metabolites by all species, denoted as v_{jk} for metabolite *j* and species *k,* were calculated based on Michaelis-Menten single-substrate kinetics, with assumed universal values for maximum rate V_{max} and transporter affinity K_m for all metabolites (provided in Supplementary Data 1). v_{jk} was further constrained based on an allocation of the metabolite's environmental concentration to each species in proportion with its biomass. Then, the steady state reaction fluxes for each species *k* at time point t_n were determined by maximizing the growth rate μ_k, within the obtained constraints on environmental metabolite uptake. To obtain a single and consistent flux solution for each species, the total flux activity for each species (i.e., the sum of absolute fluxes given the predicted optimal growth rate) was minimized, under the assumption that organisms prefer to operate their metabolism with minimal enzymatic cost (Holzhütter, 2004). The optimal flux solutions were solved using linear programming with GLPK (www.gnu.org/software/glpk). With the resulting flux and growth rate information, the total biomass of each species *k*, $bio_k(t_n)$, was updated for the next time point t_{n+1}, using a standard exponential growth function incorporating dilution:

$$bio_k(t_{n+1}) = bio_k(t_n)e^{\mu_k \Delta t} - bio_k(t_n)D\Delta t \qquad (4.10)$$

where D is the dilution rate. We set D to 0.0472 per hour, in order to obtain community growth rates consistent with the observed average growth rate of the three most abundant species growing under 47 different carbon conditions (McNulty et al., 2013). The total amount of uptake or secretion for each species k and metabolite j over a single time step was then calculated as previously derived (Chiu et al., 2014):

$$m_{FBA}^{jk}(t_n) = \frac{v_{jk}}{\mu_k} * \ bio_k(t_n)(e^{\mu_k \Delta t} - 1) \quad (4.11)$$

where v_{jk} is the rate of uptake or secretion specified by the FBA solution for that species and metabolite at that time point, μ_k is the species growth rate, $bio_k(t_n)$ is the species abundance, and Δt is the size of the time step. Finally, combining the flux solutions of all species, nutrient inflow, and dilution, along with the steady state assumption of no intracellular metabolite accumulation, the concentration of a given metabolite in the shared nutrient environment at the next time point, $M_j(t_{n+1})$ can be updated as:

$$M_j(t_{n+1}) = M_j(t_n) + m_{FBA}^{j}(t_n) + m_{in}^{j}\Delta t - M_j(t_n)D\Delta t \qquad (4.12)$$

where $m_{FBA}^{j}(t_n)$ is the metabolic impact from all species considering their abundance and their uptake and secretion rates of metabolite j, and m_{in}^{j} is the inflow rate of metabolite j. This process of calculating uptake rates, Flux Balance Analysis solutions, and updated metabolite concentrations was then repeated iteratively for the duration of the simulation.

Each simulation was run for a period of 144 hours or 576 time steps. This time period was long enough for most simulation runs to approach a steady state composition: specifically, in >65%

of the simulations analyzed in our study, the change in abundance in any species over the final 3 hours was less than 0.01% of the carrying capacity (see below), and all had no changes greater than 0.3% of the capacity over that period. The concentrations of species and metabolites, the species growth rates, and the solved rates of all reactions for each species (including uptake and secretion) were recorded in each step of each simulation and used for subsequent analyses (Supplementary Data 1 and 2).

4.6.3 *Simulation initialization parameters*

We fixed the initial total abundances of microbes to the carrying capacity for this system and media, which was estimated to be 0.433 units of biomass. This capacity was calculated as the average final total abundance from a set of simulations with varying compositions and low initial abundances. We then varied the relative abundances, increasing the abundance of one species at a time at the expense of all other species equally. Specifically, for each species, we ran simulations in which the ratio of that species' initial abundance relative to all other species was 2, 3, 4.5, 6, 9, and 13 times (equating to a range in relative abundance of 10% to 60% for each species). This resulted in a total of 61 simulation runs (one with all species starting at equal abundance and 6 with increased abundance of each species). We chose this sample size to approximately represent the sample sizes of published cross-sectional microbiome-metabolome association studies (Califf et al., 2017; Srinivasan et al., 2015). We set the initial inflow concentrations to the amount that would dilute in over one hour under the calculated inflow rates.

4.6.4 *Calculation of contribution values for variable metabolites*

We calculated contribution values for all metabolites with variance in concentration above the 25th percentile. We chose this threshold in order to include as many metabolites as possible

while excluding those that only varied at all in fewer than half of the simulation runs, or whose variation would be subject to potential numerical errors.

4.6.5 *Comparison with Shapley values*

We implemented an approximate Shapley value algorithm (Manor and Borenstein, 2017) as an alternative strategy to calculate contributions for the simulated dataset. Briefly, 15,000 random orderings of the 10 species were randomly generated. For each ordering, the variance in metabolite activity is calculated for subsets of size 1 to 10, adding in species according to the specified ordering. The difference in variance as a given species is added to the subset, denoting the *marginal* contribution of that species to variation, is recorded. The average marginal contribution across all orderings for each species is then defined as its contribution to variance.

4.6.6 *Species-metabolite correlation analysis*

We calculated Spearman correlations between absolute species abundances (quantified as total biomass) and concentrations of variable metabolites. We used absolute abundances in order to evaluate the relationships between species and metabolites under the hypothetically best possible measurements of both data types. We also compared correlation results using relative abundances and found very minimal differences in the main simulation dataset: only 7 species-metabolite pairs (1.3%) are significantly correlated using absolute abundances but not relative, and only 4 pairs (0.8%) are correlated using relative abundances but not absolute.

We used a *p*-value threshold of 0.01 to classify "significant" associations for binary comparisons. For interpretability, we refer to *p*-values not corrected for multiple hypothesis testing, since the number of tests remained constant across nearly all of our analyses (520 possible species-metabolite pairs). The 0.01 threshold we use to define significantly correlated pairs is

equivalent to a Benjamini-Hochberg corrected false discovery threshold of 0.027, calculated using the R function *p.adjust* (Benjamini and Hochberg, 1995).

4.6.7 *Logistic regression modeling of correlation outcomes*

We used logistic regression models to identify factors that can be used to predict whether a non-contributing species-metabolite pair displays a significant correlation (false positive), and whether a key contributor species-metabolite pair fails to be correlated (false negative). We used the *glm* function in R to fit models of the log odds of whether a non-contributing species is correlated with its corresponding metabolite (false positive or true negative), using as predictors grouped indicator values for species and metabolite identities. We separately fit another set of logistic regression models to predict whether a key contributor species is correlated (true positive or false negative), with the same predictors. Models were compared using likelihood ratio tests using the *anova* function in R.

4.6.8 *Simulations with varied inflow quantities*

We ran 8 additional sets of simulations with the same set of 61 different initial species compositions but with varying degrees of inflow fluctuations. Specifically, the nutrient inflow quantities were sampled independently from a normal distribution, with a mean of the original inflow concentration and the standard deviation equal to a set percent of the mean. The 8 levels of deviation were 0.5%, 1%, 2%, 3%, 4%, 5%, 8%, or 10%. In the comparison of correlation results across samples, we evaluated the same set of 52 variable metabolites as for the original dataset for consistency, although given the added noise, additional metabolites met the same variance cutoff we used to define variable metabolites.

To evaluate correlation performance as a function of increasing environmental contribution, we binned the 38 analyzed inflow metabolites across the 8 datasets based on the size

of the environmental contribution to variance for the metabolite in that dataset. In other words, metabolites in any dataset with an environmental contribution greater than 0 but less than 10% of the total positive variance contributions were binned into a single category, those with an environmental contribution between 10% and 20% were binned into the next category, and so on. We analyzed the 52 metabolites in the original constant-environment dataset as a separate category, and did the same for the 14 non-inflow metabolites in each of the 8 environmentally-varying datasets.

Confidence intervals for AUC values were calculated using the *pROC* package in R (Robin et al., 2011), using a bootstrap method with 500 resamplings.

4.6.9 *Simulations of Human Microbiome Project-based microbiota*

To simulate more complex gut microbiota, we downloaded the 16S rRNA sequence variant abundance tables from the Human Microbiome Project (Huttenhower et al., 2012), processed with deblur (Amir et al., 2017), from Qiita (Gonzalez et al., 2018). We also downloaded ribosomal RNA sequences for all of the 818 genomes corresponding with AGORA v1.0.2 models from NCBI RefSeq and GenBank using the *biomartr* R package (Drost and Paszkowski, 2017). We used *vsearch* version 2.8.1 (Rognes et al., 2016) to map the HMP sequences to the AGORA ribosomal sequences with 97% identity, with the max_rejects parameter set to 0 in order to obtain the highest identity match for each sequence variant. We chose to model a subset of 57 samples for which at least 25% of their total read counts successfully mapped to an AGORA genome. We normalized species abundances based on the 16S rRNA copy number of the corresponding genome, and initialized 57 simulations with the starting relative abundances determined based on the AGORA-mapped relative abundances of these samples. We updated the nutrient inflow to enable growth by most models. We assessed whether the additional of each individual metabolite to the original nutrient inflow had a growth-promoting effect on any species, specifying proportions similar to

the average European diet in the Virtual Metabolic Human database where possible (Noronha et al., 2018). Metabolites that promoted growth in at least one species were retained in the revised nutrient inflow, and the process of testing for increased growth with the addition of any single metabolite was repeated. After two rounds of adding metabolites to the inflow, 15 models, representing 3.4% of the total normalized abundance across all samples, still displayed zero growth. We removed these from the simulations and used the final updated nutrient inflow with the 131 remaining models. All other simulation parameters were the same as for the original 10-species community simulations. When analyzing the role of interspecies correlation in this dataset, we excluded species that appear in fewer than 4 samples.

4.6.10 *Application of MIMOSA to simulated data and comparison with correlation analysis*

We applied MIMOSA v1.0.2 (github.com/borenstein-lab/MIMOSA) (Noecker et al., 2016) to the obtained set of metabolite and species abundances. To construct the community metabolic network model required by MIMOSA, we merged the 10 species-level models used in the simulations into a single stoichiometric matrix. If a reversible reaction only ever proceeded in a single direction in any simulation, we encoded it as non-reversible. To apply the KEGG-based version of MIMOSA, we converted the model metabolite IDs to KEGG IDs (Kanehisa and Goto, 2000), downloaded KEGG Orthology gene annotations for the 10 modeled species from the IMG/M database (Markowitz et al., 2012), and ran a MIMOSA analysis using the KEGG metabolic network model encoded in *reaction_mapformula.lst* (KEGG version downloaded 2-2018).

4.6.11 *Code and data availability*

Code for all the analyses presented in this study is available online in the form of R notebooks at https://github.com/borenstein-lab/microbiome-metabolome-evaluation. The code

and media files for performing dynamic FBA co-culture simulations is available from https://borensteinlab.com/download.html. All data generated and analyzed in this study and displayed in the figures are included in Supplementary Data 1 through 4 (available at the same URL).

4.7 ACKNOWLEDGMENTS

C.N. was supported in part by a National Science Foundation (NSF) IGERT DGE-1258485 fellowship. C.P.M. was funded by NHGRI grant T32 HG000035. This work was supported in part by NIH New Innovator Award DP2 AT007802–01 and NIH grant 1R01GM124312–01 to E.B.

Chapter 5. MIMOSA2: A METABOLIC NETWORK-BASED TOOL FOR INFERRING MECHANISTIC LINKS FROM MICROBIOME-METABOLOME DATA

5.1 SUMMARY

Recent technology developments have facilitated a recent expansion of microbiome-metabolome studies, in which researchers measure the composition of a set of microbial communities together with metabolite measurements from the same samples. A common goal of many of these studies is to identify microbial features (species or genes) that may contribute to differences in metabolite phenotypes between the communities of interest. Previous work showed that integrating these datasets with reference knowledge on microbial metabolic capacities may enable more precise and confident inference of such microbe-metabolite links. Here, we present MIMOSA2, an R package and web server for model-based integrative analysis of microbiome-metabolome datasets. MIMOSA2 uses reference data from multiple sources to construct a community metabolic model from microbiome data, which is then compared with the metabolomics data to identify putative microbiome-controlled metabolites and specific taxonomic contributors to metabolite variation. MIMOSA2 is flexible to a variety of input data types and can be customized to incorporate user-defined metabolic pathways. We demonstrate MIMOSA2's ability to identify true microbial mechanisms from simulation data in comparison with alternative approaches, and we describe an example application to a microbiome-metabolome study of genetically divergent mice.

5.2 BACKGROUND

Microbial community metabolism is dynamic, wide-ranging and impactful. Microbial processes drive global nutrient cycling (McGuire and Treseder, 2010), metabolic dysregulation in

human disease (Kasubuchi et al., 2015), and detoxification of pollutants (Hazen et al., 2010), among other roles. Two common questions across the study of host-associated and environmental microbial ecology, as well as metabolic design and engineering, are whether metabolic differences between environments can be explained by differences in microbial composition and ecology, and if so, which specific microbial community members might be the key actors generating differences in metabolism between healthy and disease samples, or across varying environments.

Sequencing and metabolomics technologies have enabled population surveys of microbial taxa and metabolites across various environments (Shaffer et al., 2017). These multi-omic microbiome-metabolome studies have great potential utility to link microbial taxa to their metabolic impacts, by uncovering associations between the abundances of each across a set of similar environments. However, we previously found that analyzing such datasets using simple correlation analysis can produce high rates of false positive and/or false negative links between taxa and metabolites, and that these rates are strongly influenced by individual features of the microbial taxon and metabolite in question (Noecker et al., 2019). Comparisons of microbe-metabolite associations with metabolite production in monoculture have also suggested a high false positive rate for this approach (Hoyles et al., 2018). Metabolic reference databases such as KEGG (Kanehisa and Goto, 2000), as well as collections of genome-scale metabolic reconstructions such as AGORA (Magnúsdóttir et al., 2016), can inform this analysis to generate more precise hypotheses on the relationships between microbes and metabolites, and/or to test whether a specific taxon-metabolite mechanistic hypothesis (for example, observed *in vitro* or in another dataset) is detectably supported by the dataset in question. Methods to integrate microbiome and metabolome data with metabolic model reference databases have been recently described (Garza et al., 2018; McHardy et al., 2013; Noecker et al., 2016; Pedersen et al., 2018), but their utility and interpretation have not been consistently evaluated, and may not be available

in an easy-to-use format or compatible with the current variety of microbiome and metabolome profiling data formats and reference data resources.

Here, we introduce MIMOSA2, an R package and web server (http://elbo-spice.gs.washington.edu/shiny/MIMOSA2shiny/) for model-based integration of microbiome and metabolome data. MIMOSA2 uses metabolic reference information from multiple sources with a simple model, to estimate community metabolic potential to produce and utilize any metabolite, to assess whether metabolite variation across a dataset is consistent with estimated community metabolic potential, and if so which taxa and reactions are the primary contributors. We demonstrate MIMOSA2's performance using simulation data and describe an example application to a microbiome-metabolome dataset of genetically divergent mice.

5.3 METHODS AND IMPLEMENTATION

MIMOSA2 is an R package and Shiny-based web server for analyzing and visualizing paired microbiome-metabolome datasets. It uses genome-scale metabolic reference data to interpret measurements of microbial taxa and metabolites. The input file paths and parameters for any MIMOSA2 analysis run are encoded in a configuration table, which can be utilized to reproduce the analysis via either the server interface or the *run_mimosa2* function in the R package. An overview of the input options and steps in the full analysis is shown in Figure 5.1.

5.3.1 *Data input options*

The basic input requirement for MIMOSA2 is a pair of datasets from the same set of samples: one of microbiome measurements and the other of metabolites. Each of these datasets may take a variety of forms, depending on the study's design, specific experimental assays, and choice of processing techniques.

Users can provide microbiome data generated from 16S rRNA or shotgun metagenomic sequencing studies. 16S rRNA data can be provided as a feature table of amplicon sequence variants, or of closed-reference operational taxonomic units (OTUs) using either the Greengenes or SILVA databases. Data from a shotgun metagenomic study can be provided in the form of a table of KEGG Ortholog functional abundances, which can also be stratified by microbial taxa, as produced by a HUMAnN2 analysis (Franzosa et al., 2018a). If datasets from both platforms are provided (16S rRNA and metagenomics), MIMOSA2 will run a full analysis with both and provide results comparing the two.

Metabolite data can be produced from any metabolomics platform, and must contain putative metabolite identifications in the form of either metabolite names or KEGG compound IDs. If metabolite names are provided, they are mapped to KEGG IDs for the main analysis using the mapping utility from the *MetabolAnalystR* package (Chong et al., 2018).

5.3.2 *Reference data and metabolic model generation*

MIMOSA2 uses reference data to estimate how metabolic potential differs across a set of microbiome samples. It takes advantage of genome-scale metabolic model data from multiple sources to allow for use of the most appropriate reference data for a given dataset. Specifically, MIMOSA2 can currently generate a metabolic network model from one of two sources. For gut microbiome datasets, it can utilize the AGORA 1.0.2 collection of genome-scale metabolic models of gut microbial species (Magnúsdóttir et al., 2016). For other datasets or to capture a wider selection of taxa, it generates a metabolic network model using the *reaction_mapformula.lst* curated set of reactions from the KEGG database. Upcoming updates will also provide options for the use of EMBL GEMs, a library of genome-scale metabolic models for all 5,587 reference and

representative bacterial genomes in RefSeq (Machado et al., 2018), as well as the MetaCyc database (Caspi et al., 2014).

Abundances of microbiome taxa are mapped to metabolic reactions in various ways, depends on the input data type (Figure 5.1). 16S rRNA sequence variants are mapped using *vsearch* to either RNA genes for the AGORA genome collection (see Appendix D), or to Greengenes 99% OTU representative sequences. Greengenes OTUs are mapped to KEGG using the precomputed genome inferences from PICRUSt 1.1.3 (Langille et al., 2013), and are linked to AGORA models using a pre-calculated alignment between the two databases. KEGG Ortholog abundances provided from a shotgun metagenomic dataset or another method can be directly linked to KEGG reactions.

Additionally, users can specify custom additions, subtractions, or modifications to any of the MIMOSA2 model templates. These can be provided at the gene, reaction, or taxon level. Example modification files are available from the MIMOSA2 server webpage and documentation. The ability to add or remove a particular reaction is useful to allow for assessing the impact of errors or likely incorrect annotations in the reference database, as we have observed in previous analyses with MIMOSA version 1 (see Chapter 3, metabolism of *trans*-4-hydroxyproline and 5AV). Full details on the implementation of each of the metabolic model construction options are provided in Appendix D.

5.3.3 *Core algorithm and identification of species-metabolite contributors*

After importing input data and constructing a community metabolic model, MIMOSA2 next calculates metabolic potential scores and fits a model relating those scores to the relevant metabolite measurements.

Metabolic potential scores are by default calculated using the same formula as the original MIMOSA (Noecker et al., 2016), as a linear combination of the abundances of genes predicted to contribute to synthesis or production of a metabolite multiplied by their expected effects given reaction stoichiometry in the model. Next, MIMOSA2 fits a linear regression model relating these scores to the measured metabolite concentrations. Optionally, MIMOSA2 uses a greedy algorithm to refine the direction and/or presence of reactions in the network, based on whether removing a reaction in a given species improves the fit with metabolite measurements.

Once this fitted model relating metabolic potential scores to metabolites has been obtained, MIMOSA2 uses it to calculate the share of metabolite variation attributable to each species, analogously to our previous approach for calculating taxonomic contributors based on metabolic fluxes (Chapter 4). Metabolites are identified as putatively microbe-controlled if the overall model meets a threshold of explained variation. For these metabolites, microbial taxa with the largest contributions to model variation (defined by the same metric as previously used to define gold standard flux-based contributions) are then identified as potential contributors. This metric prioritizes taxa whose estimated metabolic potential is abundant, variable, and correlated with the true metabolite concentrations. It is in contrast with the taxonomic contributor metric used in MIMOSA version 1, which was strictly based on correlations of a taxon's metabolic potential across samples.

Importantly, each of these steps is modular and can be easily substituted in the future with alternative approaches. For instance, community metabolic potential is currently calculated from the community metabolic model using a gene abundance-based scoring approach, but in the future metabolic fluxes could instead be estimated using flux balance analysis (Varma and Palsson, 1994).

5.3.4 *Output results and visualizations*

MIMOSA2 provides three main types of results: the constructed community metabolic network model, the statistics on model fit between metabolic potential scores and metabolite concentrations for each metabolite, and the identified contributions to variation for each taxon and metabolite. The workflow also automatically generates a set of plots summarizing the model fit and contribution results, which are displayed interactively when run using the Shiny server app.

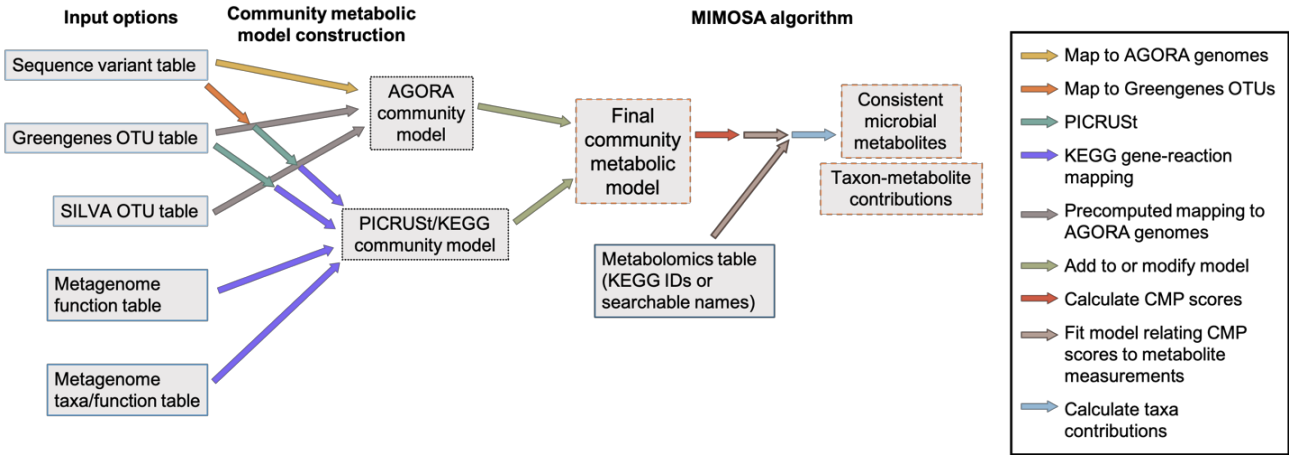

Figure 5.1. **Summary of the MIMOSA2 analysis pipeline.**

Datasets are shown as gray boxes: input data (solid blue outlines), community metabolic model templates (dotted gray outline), or analysis products (orange dashed outlines). Analysis steps are shown as colored arrows, with the specific function described in the table on the right. Briefly, input microbiome taxonomic abundance data is linked to either the AGORA collection of genome-scale metabolic reconstructions or to the KEGG metabolic model template, and the resulting community metabolic model can also be edited. The metabolic model is then used to calculate community metabolic potential scores for each metabolite, which are then compared with metabolomics data. The resulting model fit is used to calculate the primary taxonomic contributors to variation in each metabolite that was successfully predicted based on microbial metabolic potential.

5.4 RESULTS AND EXAMPLES

5.4.1 *Validating and comparing MIMOSA2 with simulated datasets*

We first applied MIMOSA2 to two disparate simulation datasets, allowing us to compare the results with the true set of key microbial contributors to variation in each metabolite, as calculated based on fluxes. We also compared the performance and results of MIMOSA2 with those obtained using several alternative approaches: the original MIMOSA pipeline (Noecker et al., 2016), a microbe-metabolite Spearman correlation analysis, and a modified correlation analysis where significant species-metabolite correlations are only retained if the species is identified as possessing relevant reactions for modifying the metabolite (Appendix D).

The two simulation datasets were previously generated using dynamic Flux Balance Analysis and AGORA model reconstructions (Magnúsdóttir et al., 2016) (see Chapter 4). Dataset 1 consists of communities with varying compositions of 10 representative gut species, and 3% variation in nutrient inflow contents across samples. Dataset 2 is more diverse and variable, consisting of 57 samples whose initial compositions were designed to emulate Human Microbiome Project gut samples (Huttenhower et al., 2012), along with 1% variation in nutrient inflow concentrations. These compositions were determined by aligning HMP 16S rRNA sequencing variants against genomes linked to the AGORA model collection, which resulted in 131 species unevenly distributed across the dataset (see section 4.4.6). Both datasets consisted of species and metabolite concentrations at the final time point of a 144-hour dynamic co-culture simulation.

MIMOSA2 analysis of these datasets (using the unconstrained community metabolic models from which each set of simulations was originally generated) correctly identified most metabolites as consistent with microbial potential, and recovered many key species contributors with high predictive value. In Dataset 1, only 27 of 78 metabolites vary predominantly due to

microbial metabolism (as opposed to environmental variation). MIMOSA2 identified 24 of these with only 2 false positives (92% precision). Key taxonomic contributors to metabolite variation were identified with higher sensitivity, specificity and precision by MIMOSA2 than any alternative approach (Figure 5.2A-B).

In Dataset 2, nearly all simulated metabolites are controlled mainly by microbial metabolism (204 of 222). MIMOSA2 identified 99 of these metabolites as microbially consistent, with no false positives. In terms of specific microbial contributors, MIMOSA2 identified true contributors with higher precision, 81%, than any of the other three methods (56% for MIMOSA version 1, 54% for correlation with reaction presence, and 36.8% for correlation alone), although its sensitivity was consistently lower than correlation analysis (26% compared to 40%), and this difference held across a range of thresholds (precision-recall curve shown in Figure 5.2C). These results are unsurprising in light of MIMOSA2's approximate metabolic model, which acts as a filter for spurious associations, but fails to detect true mechanisms not well described by the model.

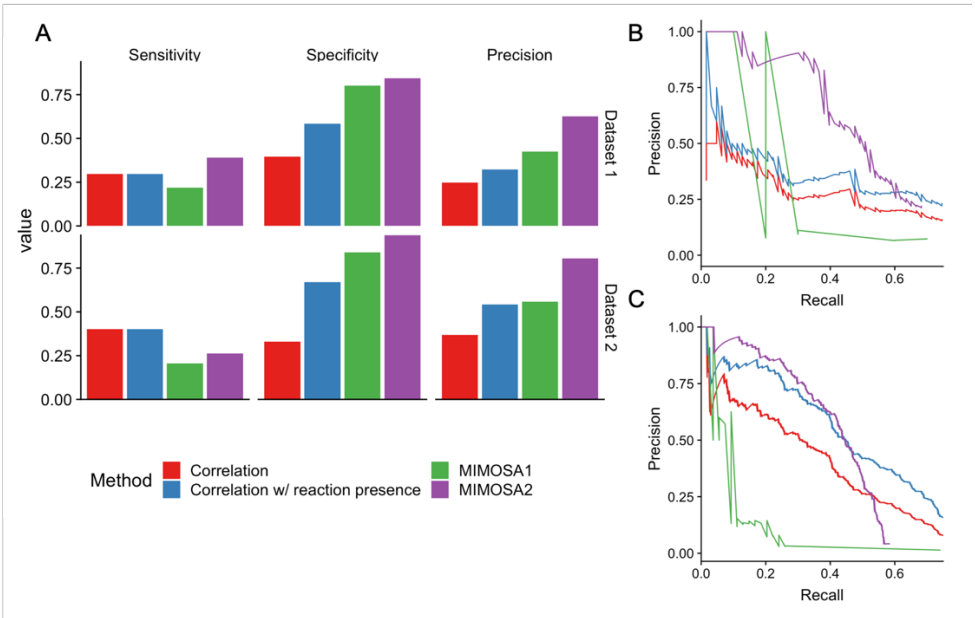

Figure 5.2. **Performance of MIMOSA2 on simulated data.**

A) Sensitivity (recall), specificity, and precision of MIMOSA2 analysis for recovering true key microbial contributors to metabolite variation from two simulated datasets, compared with

alternative approaches. Standard thresholds were used to designate contributors for each method (Correlation: 1% q-value, MIMOSA1: 5% q-value, MIMOSA 2: 10% contribution). **B-C)** Precision-recall curves for identifying key contributors for MIMOSA2 and the same three alternatives, for Dataset 1 **(B)** and Dataset 2 **(C).**

5.4.2 *Example analysis of a microbiome-metabolome mouse study*

We next applied MIMOSA2 to a microbiome-metabolome dataset describing fecal samples from 41 genetically divergent mice (Snijders et al., 2016) (full details in Appendix D). Animals were all fed the same corn-based chow, which was also assayed with metabolomics. The processed fecal dataset consisted of measurements of 175 Greengenes 97% OTUs and 110 metabolites, of which 52 (47%) were also present in the diet. We had previously observed that MIMOSA version 1 was less likely to identify metabolites present in the chow as consistent with microbial metabolism (see chapter 3).

We performed a MIMOSA2 analysis using the KEGG community metabolic network template. First, we compared the resulting set of putative microbial metabolites with those identified by the three other methods described above (MIMOSA version 1, Spearman correlation, and Spearman correlation with gene presence), and also with the set present in the diet (Figure 5.3A). MIMOSA2 identified similar metabolites as the original version of MIMOSA, but these only overlapped partially with the set of metabolites with significant OTU correlations. MIMOSA2 identified a larger share of dietary metabolites than the original version of MIMOSA, indicating that its model fitting approach may be better able to separate environmental and microbial effects.

We then examined the putative taxonomic contributions to metabolite variation obtained by MIMOSA2 (Figure 5.3B). Several larger contributions were attributed taxa in the *Lachnospiraceae* family, especially to an OTU identified as *Ruminococcus gnavus* (262633). MIMOSA2 also found significant contributions to amino acid metabolism from *Akkermansia*

municiphila (OTU 182176), possibly related to its known capability of producing free amino acids from host mucins (Ottman et al., 2017).

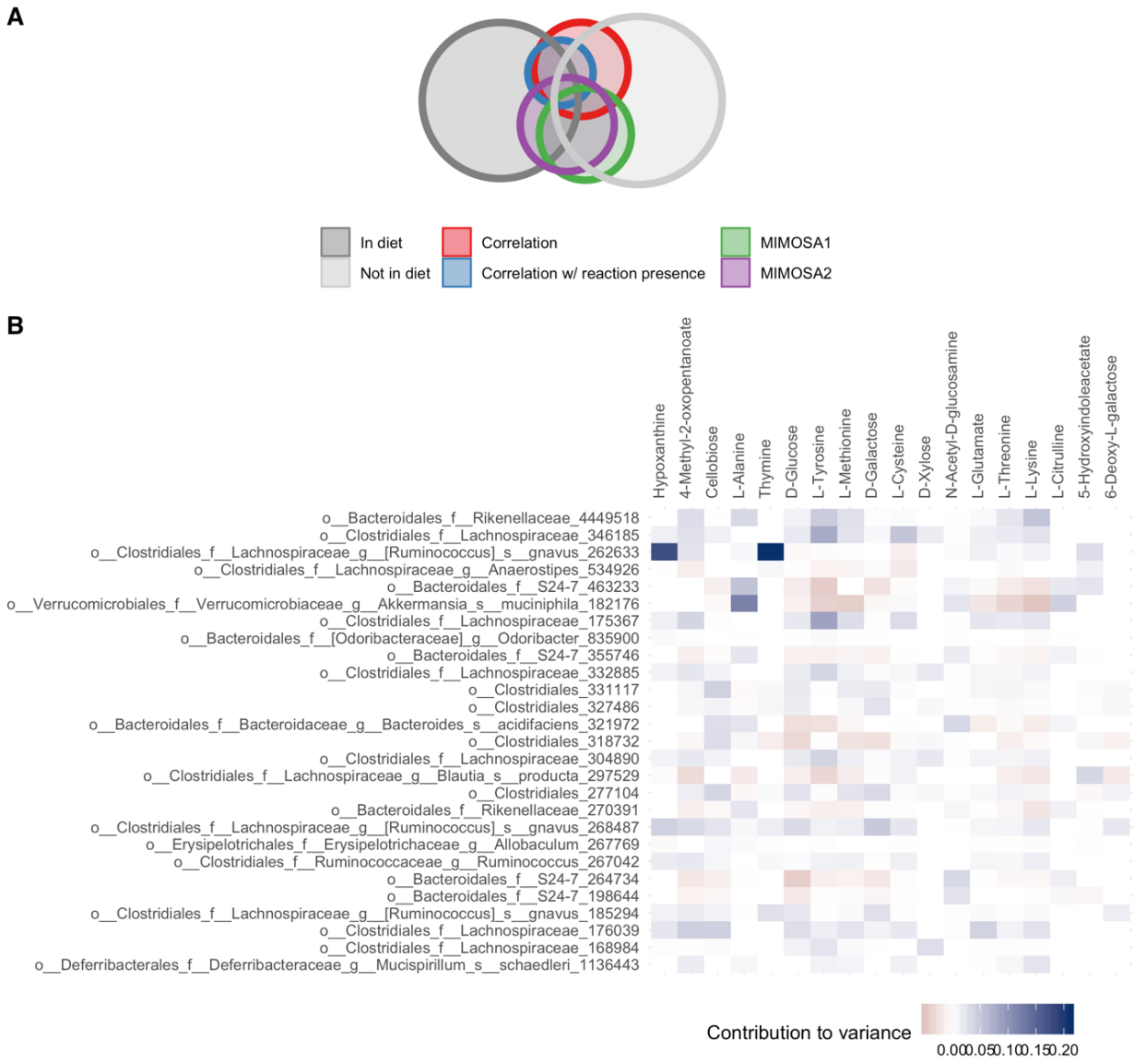

Figure 5.3. **Analysis of a mouse fecal microbiome-metabolome dataset with MIMOSA2.** **A)** Euler plot comparison of putative microbial metabolites identified by MIMOSA2 compared with other methods and compared with the set of metabolites detected in the mouse chow. The area of each circle scales with the number of metabolites in each set, and the overlap between circles approximately represents the portion of each set shared between the two. The smallest circle (for correlated pairs with a known linking reaction) represents 9 metabolites. **B)** Putative microbial contributions to metabolite variation identified by MIMOSA2. Each square in

the grid indicates the estimated relative contribution to variation to a putative microbial metabolite from an individual Greengenes OTU. OTUs with a greater than 2% relative contribution to any metabolite are shown.

5.5 CONCLUSIONS

MIMOSA2 is a software framework for generating and evaluating mechanistic metabolic hypotheses from microbiome-metabolome datasets, by determining whether metabolite levels across a population are consistent with the estimated effects of microbial community members. MIMOSA2 is compatible with many formats of microbiome-metabolome data and can provide evidence of metabolic mechanisms from complex microbiomes measured in their natural context. This type of analysis will be useful for assessing possible microbial causes of differing metabolite phenotypes between health and disease, as well as evaluating the metabolic consequences of microbiome shifts across lifestyle or environmental factors.

Chapter 6. CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, I describe efforts to develop bioinformatic tools that utilize modeling to infer and assess metabolic mechanisms from large-scale microbiome omics data. I sought to develop generalizable frameworks capable of generating reasonable hypotheses and informing the interpretation of any collection of observations about microbial communities and their metabolic environments. Moreover, I used modeling tools to describe and quantify the limitations of microbiome-metabolome association studies in achieving this purpose, and to propose evaluation frameworks to support continued development and improvement of these efforts.

First, I introduced an exploratory approach to assess whether metabolite variation across a set of microbial communities, as measured by metabolomics, is consistent with the expected differences in community metabolic activity, as calculated from composition using an approximate metabolic model. Applying this method to analyze several paired microbiome-metabolome datasets, we found that in fact, even this approximate model is predictive of a large share of metabolite shifts. This finding highlights the impact of community assembly on the metabolic phenotype of a microbiome: in other words, answering why some communities produce different levels of a metabolite can be understood in large part to be answering why those communities support different levels of organisms capable of producing that metabolite.

I then described the application of this framework, MIMOSA, in two collaborative case studies, demonstrating that it provides useful results for answering multiple types of research questions. In both cases, MIMOSA uncovered links between the two databases that were not identified by a more standard association analysis and manual interpretation. These case studies demonstrated that modeling tools could provide biological insights even if they do not produce completely accurate estimates of all reaction fluxes in all taxa. However, with the use of these

results to prioritize subsequent hypotheses and experiments, it was also clear that further evaluation of the predictive power of this approach was necessary.

To better assess and improve our ability to infer metabolic function from microbiome-metabolome datasets, I developed a definition and platform to define gold standards and evaluate the identification of microbe-metabolite links from these datasets. I proposed that a common objective of microbiome-metabolome studies is to identify the subset of microbial taxa whose metabolic activity is responsible for observed differences in metabolites between disparate communities. By identifying these true contributors from simulated data generated using a dynamic constraint-based flux balance analysis pipeline, I then found that the most common and intuitive analyses can be highly ineffective at addressing this objective, but that this outcome varies greatly as a function of the microbes, metabolites, environment, and dataset in question. This effort was a new application of the extensive resources and tools in constraint-based metabolic modeling: to generate independent simulation for methods development, validation, and benchmarking. The need for better validation of analysis methods and more tools for doing so has recently been recognized in bioinformatics more broadly (Lotterhos et al., 2018). Ideally, these results and the general approach to model evaluation will provide context for the interpretation of microbiome-metabolome association studies and contribute to further statistical efforts to more accurately infer the links between species and metabolites.

In that vein, I finally introduced a new software framework, MIMOSA2, informed by the applications thus far and the ability to benchmark against gold standards. MIMOSA2 not only uses a more effective algorithm to identify key links between microbes and metabolites, but also includes improvements in computability, accessibility, and modularity, which will enable further improvements and refinements in each of the individual components of the pipeline.

While these efforts have advanced the ability to understand what microbiome-metabolome assays can reveal about metabolic mechanisms, developing tools to obtain a full picture of those mechanisms will depend on addressing several remaining structural limitations. First, any predictive model based on the metabolic capacities described in reference databases is fundamentally limited by the quality of said databases. Roughly 10-50% of reads in a human metagenomic study typically cannot be mapped to any known organism (Pasolli et al., 2019), and even in the most well-studied microbial genomes have a large share of genes of unknown function (Qin et al., 2010). Moreover, the vast majority of chemicals detected by mass spectrometry from microbiome samples cannot be identified (Wissenbach et al., 2016). This category likely includes many complex and significant secondary metabolites and natural products (Olivon et al., 2017), which are also therefore rarely included in genome-scale metabolic reconstructions or pathway databases like KEGG (Kim et al., 2016). Reducing these unknown fractions of species, genes, reactions, and metabolites is challenging in part because of the difficulty of growing and studying many microbial taxa in the lab—genomes from uncultured clades have more genes of unknown function (among bacteria, 27% of genes from uncultured taxa vs. 19% from cultured strains) (Lloyd et al., 2018). Additionally, microbial metabolism is finely tuned and regulated, with ongoing discoveries of new mechanisms of metabolic regulation even in the most well-studied microbial species (Fuhrer et al., 2017; Li et al., 2018).

However, fully documenting the human-associated microbial and metabolic diversity and assigning roles to all genes of unknown function will not necessarily lead to clear paths towards manipulating microbiomes or explaining every microbiome-related metabolic phenotype. There are also many areas in need of improvement in the construction and application of microbiome metabolic models, including reducing the need for manual curation of genome-scale metabolic reconstructions (Machado et al., 2018) and accounting for spatial constraints and host physiology

(Bauer et al., 2017). Better validation of community metabolic models can also be achieved through the collection of high-quality metabolomics from designed combinatorial communities as well as the application of stable isotope probing technologies to measure metabolic fluxes of individual community members *in vivo* (Ayayee et al., 2015; Berry et al., 2013; Kraft et al., 2014; Kurczy et al., 2016).

Despite these challenges, the importance of modeling microbiome metabolism is especially clear in light of the fact that the metabolic activities of a microbial strain in monoculture do not necessarily recapitulate their activities in a naturally occurring complex community setting (Fiegna et al., 2015; Guo and Boedicker, 2016; Medlock et al., 2018). Determining whether a metabolic process of interest occurs in a natural setting as expected based on experimental and computational models is crucial not just for the development of microbiome therapies, but also more broadly for greater understanding of the ecological and evolutionary processes that shape microbial communities and their functions. One particularly notable area for future applications of modeling approaches is the prediction of a gut community's metabolic response to specific dietary and xenobiotic compounds (Haiser and Turnbaugh, 2013; Maier et al., 2018; Reese and Carmody, 2018), which has sizable implications for nutrition, pharmacology, and health in general.

Ultimately, the pervasive, diverse, and interwoven nature of microbial communities makes any attempt to predict and manipulate them a tall order. Nevertheless, even preliminary and approximate approaches to doing so, like those described here, have already led to new knowledge and potential improvements in medicine. With continuing growth in this area, both the limits and benefits of our understanding of the activities of our microbial counterparts are, to a large degree, still unknown.

REFERENCES

Adams, J.B., Audhya, T., McDonough-Means, S., Rubin, R.A., Quig, D., Geis, E., Gehn, E., Loresto, M., Mitchell, J., Atwood, S., et al. (2011). Nutritional and metabolic status of children with autism vs. neurotypical children, and the association with autism severity. Nutrition & Metabolism *8*, 34.

Agus, A., Planchais, J., and Sokol, H. (2018). Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. Cell Host & Microbe *23*, 716–724.

Aldred, S., Moore, K.M., Fitzgerald, M., and Waring, R.H. (2003). Plasma amino acid levels in children with autism and their families. J Autism Dev Disord *33*, 93–97.

Alivisatos, A.P., Blaser, M.J., Brodie, E.L., Chun, M., Dangl, J.L., Donohue, T.J., Dorrestein, P.C., Gilbert, J.A., Green, J.L., Jansson, J.K., et al. (2015). A unified initiative to harness Earth's microbiomes. Science *350*, 507–508.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A., et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. MSystems *2*, e00191-16.

Antharam, V.C., McEwen, D.C., Garrett, T.J., Dossey, A.T., Li, E.C., Kozlov, A.N., Mesbah, Z., and Wang, G.P. (2016). An Integrated Metabolomic and Microbiome Analysis Identified Specific Gut Microbiota Associated with Fecal Cholesterol and Coprostanol in Clostridium difficile Infection. PLoS ONE *11*, e0148824.

van der Ark, K.C.H., van Heck, R.G.A., Martins Dos Santos, V.A.P., Belzer, C., and de Vos, W.M. (2017). More than just a gut feeling: constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. Microbiome *5*.

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. Nature Biotechnology *36*, 566–569.

Ayayee, P.A., Jones, S.C., and Sabree, Z.L. (2015). Can (13)C stable isotope analysis uncover essential amino acid provisioning by termite-associated gut microbes? PeerJ *3*, e1218.

Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. PLOS Computational Biology *13*, e1005544.

Beaumont, M., Goodrich, J.K., Jackson, M.A., Yet, I., Davenport, E.R., Vieira-Silva, S., Debelius, J., Pallister, T., Mangino, M., Raes, J., et al. (2016). Heritable components of the human fecal microbiome are associated with visceral fat. Genome Biology *17*, 189.

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

Benson, A.K. (2016). The gut microbiome—an emerging complex trait. Nature Genetics *48*, 1301–1302.

Berg, R. (1996). The indigenous gastrointestinal microflora. Trends in Microbiology *4*, 430–435.

Berry, D., Stecher, B., Schintlmeister, A., Reichert, J., Brugiroux, S., Wild, B., Wanek, W., Richter, A., Rauch, I., Decker, T., et al. (2013). Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. PNAS *110*, 4720–4725.

Biggs, M.B., Medlock, G.L., Moutinho, T.J., Lees, H.J., Swann, J.R., Kolling, G.L., and Papin, J.A. (2016). Systems-level metabolism of the altered Schaedler flora, a complete gut microbiota. ISME J.

Bisanz, J.E., Seney, S., McMillan, A., Vongsa, R., Koenig, D., Wong, L., Dvoracek, B., Gloor, G.B., Sumarah, M., Ford, B., et al. (2014). A Systems Biology Approach Investigating the Effect of Probiotics on the Vaginal Microbiome and Host Responses in a Double Blind, Placebo-Controlled Clinical Trial of Post-Menopausal Women. PLoS ONE *9*, e104511.

Blekhman, R., Goodrich, J.K., Huang, K., Sun, Q., Bukowski, R., Bell, J.T., Spector, T.D., Keinan, A., Ley, R.E., Gevers, D., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. Genome Biology *16*.

Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature.

Buffington, S.A., Di Prisco, G.V., Auchtung, T.A., Ajami, N.J., Petrosino, J.F., and Costa-Mattioli, M. (2016). Microbial Reconstitution Reverses Maternal Diet-Induced Social and Synaptic Deficits in Offspring. Cell *165*, 1762–1775.

Burge, M.N. (1988). Fungi in biological control systems (Manchester, UK ; New York : New York, NY, USA: Manchester University Press ; Distributed exclusively in the USA and Canada by St. Martin's Press).

Burgin, A.J., Yang, W.H., Hamilton, S.K., and Silver, W.L. (2011). Beyond carbon and nitrogen: how the microbial energy economy couples elemental cycles in diverse ecosystems. Frontiers in Ecology and the Environment *9*, 44–52.

Cai, J., Zhang, J., Tian, Y., Zhang, L., Hatzakis, E., Krausz, K.W., Smith, P.B., Gonzalez, F.J., and Patterson, A.D. (2017). Orthogonal Comparison of GC–MS and ^{1}H NMR Spectroscopy for Short Chain Fatty Acid Quantitation. Analytical Chemistry *89*, 7900–7906.

Califf, K.J., Schwarzberg-Lipson, K., Garg, N., Gibbons, S.M., Caporaso, J.G., Slots, J., Cohen, C., Dorrestein, P.C., and Kelley, S.T. (2017). Multi-omics Analysis of Periodontal Pocket Microbial Communities Pre- and Posttreatment. MSystems *2*, e00016-17.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods *13*, 581–583.

Campo, A.M., Bodea, S., Hamer, H.A., Marks, J.A., Haiser, H.J., Turnbaugh, P.J., and Balskus, E.P. (2015). Characterization and Detection of a Widely Distributed Gene Cluster That Predicts Anaerobic Choline Utilization by Human Gut Bacteria. MBio *6*, e00042-15.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics *26*, 266–267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. Nature Methods *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal *6*, 1621–1624.

Carr, R., and Borenstein, E. (2014). Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. PLoS ONE *9*, e105776.

Casero, D., Gill, K., Sridharan, V., Koturbash, I., Nelson, G., Hauer-Jensen, M., Boerma, M., Braun, J., and Cheema, A.K. (2017). Space-type radiation induces multimodal responses in the mouse gut microbiome and metabolome. Microbiome *5*.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl. Acids Res. *42*, D459–D471.

Casteleyn, C., Rekecki, A., Van Der Aa, A., Simoens, P., and Van Den Broeck, W. (2010). Surface area assessment of the murine intestinal tract as a prerequisite for oral dose translation from mouse to man. Laboratory Animals *44*, 176–183.

Chaidez, V., Hansen, R.L., and Hertz-Picciotto, I. (2014). Gastrointestinal Problems in Children with Autism, Developmental Delays or Typical Development. Journal of Autism and Developmental Disorders *44*, 1117–1127.

Chan, S.H.J., Simons, M.N., and Maranas, C.D. (2017). SteadyCom: Predicting microbial abundances while ensuring community stability. PLOS Computational Biology *13*, e1005539.

Chiu, H.-C., Levy, R., and Borenstein, E. (2014). Emergent Biosynthetic Capacity in Simple Microbial Communities. PLoS Comput Biol *10*, e1003695.

Chong, J., and Xia, J. (2017). Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites *7*, 62.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., and Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Research.

Choo, J.M., Kanno, T., Zain, N.M.M., Leong, L.E.X., Abell, G.C.J., Keeble, J.E., Bruce, K.D., Mason, A.J., and Rogers, G.B. (2017). Divergent Relationships between Fecal Microbiota and Metabolome following Distinct Antibiotic-Induced Disruptions. MSphere *2*, e00005-17.

Clemente, J.C., Ursell, L.K., Parfrey, L.W., and Knight, R. (2012). The Impact of the Gut Microbiota on Human Health: An Integrative View. Cell *148*, 1258–1270.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research *37*, D141–D145.

Collaborative Cross Consortium (2012). The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. Genetics *190*, 389–401.

Collins, W.D., Bitz, C.M., Blackmon, M.L., Bonan, G.B., Bretherton, C.S., Carton, J.A., Chang, P., Doney, S.C., Hack, J.J., Henderson, T.B., et al. (2006). The Community Climate System Model Version 3 (CCSM3). Journal of Climate *19*, 2122–2143.

Connolly, A.J., Angeli, G.Z., Chandrasekharan, S., Claver, C.F., Cook, K., Ivezic, Z., Jones, R.L., Krughoff, K.S., Peng, E.-H., Peterson, J., et al. (2014). An end-to-end simulation framework for the Large Synoptic Survey Telescope. G.Z. Angeli, and P. Dierickx, eds. p. 915014.

Coretti, L., Cristiano, C., Florio, E., Scala, G., Lama, A., Keller, S., Cuomo, M., Russo, R., Pero, R., Paciello, O., et al. (2017). Sex-related alterations of gut microbiota composition in the BTBR mouse model of autism spectrum disorder. Scientific Reports *7*.

Coretti, L., Paparo, L., Riccio, M.P., Amato, F., Cuomo, M., Natale, A., Borrelli, L., Corrado, G., Comegna, M., Buommino, E., et al. (2018). Gut Microbiota Features in Young Children With Autism Spectrum Disorders. Frontiers in Microbiology *9*.

Cowan, T.E., Palmnäs, M.S.A., Yang, J., Bomhof, M.R., Ardell, K.L., Reimer, R.A., Vogel, H.J., and Shearer, J. (2014). Chronic coffee consumption in the diet-induced obese rat: impact on gut microbiota and serum metabolomics. The Journal of Nutritional Biochemistry *25*, 489–495.

Cox, L.M., Yamanishi, S., Sohn, J., Alekseyenko, A.V., Leung, J.M., Cho, I., Kim, S.G., Li, H., Gao, Z., Mahana, D., et al. (2014). Altering the Intestinal Microbiota during a Critical Developmental Window Has Lasting Metabolic Consequences. Cell *158*, 705–721.

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., and Markley, J.L. (2008). Metabolite identification via the Madison Metabolomics Consortium Database. Nature Biotechnology *26*, 162–164.

Dabney, A., and Storey, J.D. (2015). qvalue: Q-value estimation for false discovery rate control.

Daniel, H., Gholami, A.M., Berry, D., Desmarchelier, C., Hahne, H., Loh, G., Mondot, S., Lepage, P., Rothballer, M., Walker, A., et al. (2014). High-fat diet alters gut microbiota physiology in mice. ISME J *8*, 295–308.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., and Alm, E.J. (2014). Host lifestyle affects human microbiota on daily timescales. Genome Biology *15*, R89.

De Angelis, M., Piccolo, M., Vannini, L., Siragusa, S., De Giacomo, A., Serrazzanetti, D.I., Cristofori, F., Guerzoni, M.E., Gobbetti, M., and Francavilla, R. (2013). Fecal Microbiota and Metabolome of Children with Autism and Pervasive Developmental Disorder Not Otherwise Specified. PLoS ONE *8*, e76993.

De Filippis, F., Vannini, L., La Storia, A., Laghi, L., Piombino, P., Stellato, G., Serrazanetti, D.I., Gozzi, G., Turroni, S., Ferrocino, I., et al. (2014). The same microbiota and a potentially discriminant metabolome in the saliva of omnivore, ovo-lacto-vegetarian and Vegan individuals. PLoS ONE *9*, e112373.

De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I.B., La Storia, A., Laghi, L., Serrazanetti, D.I., Di Cagno, R., Ferrocino, I., Lazzi, C., et al. (2015). High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut gutjnl-2015-309957.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

DiGiulio, D.B., Callahan, B.J., McMurdie, P.J., Costello, E.K., Lyell, D.J., Robaczewska, A., Sun, C.L., Goltsman, D.S.A., Wong, R.J., Shaw, G., et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences 201502875.

Dodd, D., Spitzer, M.H., Van Treuren, W., Merrill, B.D., Hryckowian, A.J., Higginbottom, S.K., Le, A., Cowan, T.M., Nolan, G.P., Fischbach, M.A., et al. (2017). A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. Nature.

Doledec, S., and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology *31*, 277–294.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proceedings of the National Academy of Sciences *107*, 11971–11975.

Dominguez-Bello, M.G., De Jesus-Laboy, K.M., Shen, N., Cox, L.M., Amir, A., Gonzalez, A., Bokulich, N.A., Song, S.J., Hoashi, M., Rivera-Vinas, J.I., et al. (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat Med *advance online publication*.

Donaldson, G.P., Ladinsky, M.S., Yu, K.B., Sanders, J.G., Yoo, B.B., Chou, W.-C., Conner, M.E., Earl, A.M., Knight, R., Bjorkman, P.J., et al. (2018). Gut microbiota utilize immunoglobulin A for mucosal colonization. Science *360*, 795–800.

Donia, M.S., and Fischbach, M.A. (2015). Small molecules from the human microbiota. Science *349*, 1254766.

Donohoe, D.R., Garge, N., Zhang, X., Sun, W., O'Connell, T.M., Bunger, M.K., and Bultman, S.J. (2011). The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. Cell Metabolism *13*, 517–526.

Drost, H.-G., and Paszkowski, J. (2017). Biomartr: genomic data retrieval with R. Bioinformatics btw821.

Dumas, M.-E., Barton, R.H., Toye, A., Cloarec, O., Blancher, C., Rothwell, A., Fearnside, J., Tatoud, R., Blanc, V., Lindon, J.C., et al. (2006). Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. Proceedings of the National Academy of Sciences *103*, 12511–12516.

Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., and Alm, E.J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications *8*.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

Edlund, A., Yang, Y., Yooseph, S., Hall, A.P., Nguyen, D.D., Dorrestein, P.C., Nelson, K.E., He, X., Lux, R., Shi, W., et al. (2015). Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. ISME J.

Erbilgin, O., Bowen, B.P., Kosina, S.M., Jenkins, S., Lau, R.K., and Northen, T.R. (2017). Dynamic substrate preferences predict metabolic properties of a simple microbial consortium. BMC Bioinformatics *18*.

Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS ONE *7*, e49138.

Evans, C., Dunstan, H.R., Rothkirch, T., Roberts, T.K., Reichelt, K.L., Cosford, R., Deed, G., Ellis, L.B., and Sparkes, D.L. (2008). Altered amino acid excretion in children with autism. Nutritional Neuroscience *11*, 9–17.

Faith, J.J., McNulty, N.P., Rey, F.E., and Gordon, J.I. (2011). Predicting a Human Gut Microbiota's Response to Diet in Gnotobiotic Mice. Science *333*, 101–104.

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., et al. (2016). Population-level analysis of gut microbiome variation. Science *352*, 560–564.

Ferreyra, J.A., Wu, K.J., Hryckowian, A.J., Bouley, D.M., Weimer, B.C., and Sonnenburg, J.L. (2014). Gut Microbiota-Produced Succinate Promotes C. difficile Infection after Antibiotic Treatment or Motility Disturbance. Cell Host & Microbe *16*, 770–777.

Fiegna, F., Moreno-Letelier, A., Bell, T., and Barraclough, T.G. (2015). Evolution of species interactions determines microbial community productivity in new environments. ISME J *9*, 1235–1245.

Flores, G.E., Caporaso, J., Henley, J.B., Rideout, J., Domogala, D., Chase, J., Leff, J.W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., et al. (2014). Temporal variability is a personalized feature of the human microbiome. Genome Biology *15*, 531.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Krogh Pedersen, H., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature *528*, 262–266.

Franzosa, E.A., Morgan, X.C., Segata, N., Waldron, L., Reyes, J., Earl, A.M., Giannoukos, G., Boylan, M.R., Ciulla, D., Gevers, D., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. Proceedings of the National Academy of Sciences *111*, E2329–E2338.

Franzosa, E.A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X.C., and Huttenhower, C. (2015). Sequencing and beyond: integrating molecular "omics" for microbial community profiling. Nat Rev Micro *13*, 360–372.

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., et al. (2018a). Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods *15*, 962–968.

Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2018b). Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology.

Fuentes, S., Rossen, N.G., van der Spek, M.J., Hartman, J.H., Huuskonen, L., Korpela, K., Salojärvi, J., Aalvink, S., de Vos, W.M., D'Haens, G.R., et al. (2017). Microbial shifts and signatures of long-term remission in ulcerative colitis after faecal microbiota transplantation. The ISME Journal *11*, 1877–1889.

Fuhrer, T., Zampieri, M., Sévin, D.C., Sauer, U., and Zamboni, N. (2017). Genomewide landscape of gene–metabolome associations in *Escherichia coli*. Molecular Systems Biology *13*, 907.

Garg, N., Wang, M., Hyde, E., da Silva, R.R., Melnik, A.V., Protsyuk, I., Bouslimani, A., Lim, Y.W., Wong, R., Humphrey, G., et al. (2017). Three-Dimensional Microbiome and Metabolome Cartography of a Diseased Human Lung. Cell Host & Microbe.

Garza, D.R., van Verk, M.C., Huynen, M.A., and Dutilh, B.E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. Nature Microbiology.

Gentile, C.L., and Weir, T.L. (2018). The gut microbiota at the intersection of diet and human health. Science *362*, 776–780.

Gibbons, S.M., Kearney, S.M., Smillie, C.S., and Alm, E.J. (2017). Two dynamic regimes in the human gut microbiome. PLOS Computational Biology *13*, e1005364.

Gibson, M.K., Wang, B., Ahmadi, S., Burnham, C.-A.D., Tarr, P.I., Warner, B.B., and Dantas, G. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nature Microbiology 16024.

Gilbert, J.A., Quinn, R.A., Debelius, J., Xu, Z.Z., Morton, J., Garg, N., Jansson, J.K., Dorrestein, P.C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. Nature *535*, 94–103.

Gondalia, S.V., Palombo, E.A., Knowles, S.R., Cox, S.B., Meyer, D., and Austin, D.W. (2012). Molecular Characterisation of Gastrointestinal Microbiota of Children With Autism (With and Without Gastrointestinal Dysfunction) and Their Neurotypical Siblings: GI microbiota of children with autism. Autism Research *5*, 419–427.

Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods *15*, 796–798.

Goodman, A.L., Kallstrom, G., Faith, J.J., Reyes, A., Moore, A., Dantas, G., and Gordon, J.I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proceedings of the National Academy of Sciences *108*, 6252–6257.

Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R., and Ley, R.E. (2014). Conducting a Microbiome Study. Cell *158*, 250–262.

Goodrich, J.K., Davenport, E.R., Beaumont, M., Jackson, M.A., Knight, R., Ober, C., Spector, T.D., Bell, J.T., Clark, A.G., and Ley, R.E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host & Microbe *19*, 731–743.

Gotham, K., Risi, S., Pickles, A., and Lord, C. (2007). The Autism Diagnostic Observation Schedule: Revised Algorithms for Improved Diagnostic Validity. Journal of Autism and Developmental Disorders *37*, 613–627.

Greenblum, S., Turnbaugh, P.J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. PNAS *109*, 594–599.

Greenblum, S., Chiu, H.-C., Levy, R., Carr, R., and Borenstein, E. (2013). Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. Current Opinion in Biotechnology *24*, 810–820.

Greenblum, S., Carr, R., and Borenstein, E. (2015). Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. Cell *160*, 583–594.

Grimaldi, R., Gibson, G.R., Vulevic, J., Giallourou, N., Castro-Mejía, J.L., Hansen, L.H., Leigh Gibson, E., Nielsen, D.S., and Costabile, A. (2018). A prebiotic intervention study in children with autism spectrum disorders (ASDs). Microbiome *6*.

Guo, X., and Boedicker, J.Q. (2016). The Contribution of High-Order Metabolic Interactions to the Global Activity of a Four-Species Microbial Community. PLOS Comput Biol *12*, e1005079.

Haiser, H.J., and Turnbaugh, P.J. (2013). Developing a metagenomic view of xenobiotic metabolism. Pharmacological Research *69*, 21–31.

Hale, V.L., Jeraldo, P., Chen, J., Mundy, M., Yao, J., Priya, S., Keeney, G., Lyke, K., Ridlon, J., White, B.A., et al. (2018). Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. Genome Medicine *10*.

Hallmayer, J. (2011). Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. Archives of General Psychiatry *68*, 1095.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. *5*, R245-249.

Hazen, T.C., Dubinsky, E.A., DeSantis, T.Z., Andersen, G.L., Piceno, Y.M., Singh, N., Jansson, J.K., Probst, A., Borglin, S.E., Fortney, J.L., et al. (2010). Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. Science *330*, 204–208.

He, B., Nohara, K., Ajami, N.J., Michalek, R.D., Tian, X., Wong, M., Losee-Olson, S.H., Petrosino, J.F., Yoo, S.-H., Shimomura, K., et al. (2015). Transmissible microbial and metabolomic remodeling by soluble dietary fiber improves metabolic homeostasis. Scientific Reports *5*, 10604.

Heinken, A., and Thiele, I. (2015a). Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. Gut Microbes *6*, 120–130.

Heinken, A., and Thiele, I. (2015b). Anoxic conditions promote species-specific mutualism between gut microbes in silico. Appl. Environ. Microbiol. AEM.00101-15.

Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., de Beaufort, C., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nature Microbiology *2*, 16180.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotech *28*, 977–982.

Hiller, K., Hangebrauk, J., Jäger, C., Spura, J., Schreiber, K., and Schomburg, D. (2009). MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. Analytical Chemistry *81*, 3429–3439.

Holland, H.D. (2006). The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society B: Biological Sciences *361*, 903–915.

Holmes, K.K., Chen, K.C.S., Lipinski, C.M., and Eschenbach, D.A. (1985). Vaginal Redox Potential in Bacterial Vaginosis (Nonspecific Vaginitis). Journal of Infectious Diseases *152*, 379–382.

Holzhütter, H.-G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks: Flux minimization. European Journal of Biochemistry *271*, 2905–2922.

Hoyles, L., Jiménez-Pranteda, M.L., Chilloux, J., Brial, F., Myridakis, A., Aranias, T., Magnan, C., Gibson, G.R., Sanderson, J.D., Nicholson, J.K., et al. (2018). Metabolic retroconversion of trimethylamine N-oxide and the gut microbiota. Microbiome *6*.

Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., et al. (2013). Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders. Cell *155*, 1451–1463.

Hua, C., Tian, J., Tian, P., Cong, R., Luo, Y., Geng, Y., Tao, S., Ni, Y., and Zhao, R. (2017). Feeding a High Concentration Diet Induces Unhealthy Alterations in the Composition and Metabolism of Ruminal Microbiota and Host Response in a Goat Model. Frontiers in Microbiology *8*.

Huang, Y.Y., Martínez-del Campo, A., and Balskus, E.P. (2018). Anaerobic 4-hydroxyproline utilization: Discovery of a new glycyl radical enzyme in the human gut microbiome uncovers a widespread microbial metabolic activity. Gut Microbes 1–16.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., et al. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

iHMP Research Network Consortium (2014). The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. Cell Host & Microbe *16*, 276–289.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., et al. (2007). Multiple High-Throughput Analyses Monitor the Response of E. coli to Perturbations. Science *316*, 593–597.

Iwai, S., Weinmaier, T., Schmidt, B.L., Albertson, D.G., Poloso, N.J., Dabbagh, K., and DeSantis, T.Z. (2016). Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. PLOS ONE *11*, e0166104.

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., and Schmitt-Kopplin, P. (2009). Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease. PLoS ONE *4*, e6386.

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, A., Steinhauser, D., Selbig, J., and Willmitzer, L. (2010). Metabolomic and transcriptomic stress response of Escherichia coli. Molecular Systems Biology *6*, n/a-n/a.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research *28*, 27–30.

Kang, D.-W., Park, J.G., Ilhan, Z.E., Wallstrom, G., LaBaer, J., Adams, J.B., and Krajmalnik-Brown, R. (2013). Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. PLOS ONE *8*, e68322.

Kang, D.-W., Adams, J.B., Gregory, A.C., Borody, T., Chittick, L., Fasano, A., Khoruts, A., Geis, E., Maldonado, J., McDonough-Means, S., et al. (2017). Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome *5*.

Kang, D.-W., Ilhan, Z.E., Isern, N.G., Hoyt, D.W., Howsmon, D.P., Shaffer, M., Lozupone, C.A., Hahn, J., Adams, J.B., and Krajmalnik-Brown, R. (2018). Differences in fecal microbial metabolites and microbiota of children with autism spectrum disorders. Anaerobe *49*, 121–131.

Kasubuchi, M., Hasegawa, S., Hiramatsu, T., Ichimura, A., and Kimura, I. (2015). Dietary Gut Microbial Metabolites, Short-chain Fatty Acids, and Host Metabolic Regulation. Nutrients *7*, 2839–2849.

Kelly, C.J., Zheng, L., Campbell, E.L., Saeedi, B., Scholz, C.C., Bayless, A.J., Wilson, K.E., Glover, L.E., Kominsky, D.J., Magnuson, A., et al. (2015). Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function. Cell Host & Microbe *17*, 662–671.

Kendall, A.I. (1909). SOME OBSERVATIONS ON THE STUDY OF THE INTESTINAL BACTERIA. Journal of Biological Chemistry *6*, 499–507.

Keren, N., Konikoff, F.M., Paitan, Y., Gabay, G., Reshef, L., Naftali, T., and Gophna, U. (2015). Interactions between the intestinal microbiota and bile acids in gallstones patients: Bile acid and microbiota in gallstones patients. Environmental Microbiology Reports *7*, 874–880.

Kešnerová, L., Mars, R.A.T., Ellegaard, K.M., Troilo, M., Sauer, U., and Engel, P. (2017). Disentangling metabolic functions of bacteria in the honey bee gut. PLOS Biology *15*, e2003467.

Kim, H.U., Charusanti, P., Lee, S.Y., and Weber, T. (2016). Metabolic engineering with systems biology tools to optimize production of prokaryotic secondary metabolites. Natural Product Reports *33*, 933–941.

Kim, S., Kim, H., Yim, Y.S., Ha, S., Atarashi, K., Tan, T.G., Longman, R.S., Honda, K., Littman, D.R., Choi, G.B., et al. (2017). Maternal gut bacteria promote neurodevelopmental abnormalities in mouse offspring. Nature *549*, 528–532.

Kim, Y.-M., Schmidt, B.J., Kidwai, A.S., Jones, M.B., Deatherage Kaiser, B.L., Brewer, H.M., Mitchell, H.D., Palsson, B.O., McDermott, J.E., Heffron, F., et al. (2013). Salmonella modulates metabolism during growth under conditions that induce expression of virulence genes. Molecular BioSystems *9*, 1522.

Kim, Y.-M., Nowack, S., Olsen, M.T., Becraft, E.D., Wood, J.M., Thiel, V., Klapper, I., Kühl, M., Fredrickson, J.K., Bryant, D.A., et al. (2015). Diel metabolomics analysis of a hot spring chlorophototrophic microbial mat leads to new hypotheses of community member metabolisms. Front. Microbiol *6*, 209.

Kind, T., Wohlgemuth, G., Lee, D.Y., Lu, Y., Palazoglu, M., Shahbaz, S., and Fiehn, O. (2009). FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. Analytical Chemistry *81*, 10038–10048.

Klingenberg, H., and Meinicke, P. (2017). How to normalize metatranscriptomic count data for differential expression analysis. PeerJ *5*, e3859.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., et al. (2018). Best practices for analysing microbiomes. Nature Reviews Microbiology *16*, 410–422.

Koeth, R.A., Wang, Z., Levison, B.S., Buffa, J.A., Org, E., Sheehy, B.T., Britt, E.B., Fu, X., Wu, Y., Li, L., et al. (2013). Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nature Medicine *19*, 576–585.

Kraft, B., Tegetmeyer, H.E., Sharma, R., Klotz, M.G., Ferdelman, T.G., Hettich, R.L., Geelhoed, J.S., and Strous, M. (2014). The environmental controls that govern the end product of bacterial nitrate respiration. Science *345*, 676–679.

Kurczy, M.E., Forsberg, E.M., Thorgersen, M.P., Poole, F.L., Benton, H.P., Ivanisevic, J., Tran, M.L., Wall, J.D., Elias, D.A., Adams, M.W.W., et al. (2016). Global Isotope Metabolomics Reveals Adaptive Strategies for Nitrogen Assimilation. ACS Chemical Biology *11*, 1677–1685.

Kushak, R.I., Winter, H.S., Buie, T.M., Cox, S.B., Phillips, C.D., and Ward, N.L. (2017). Analysis of the Duodenal Microbiome in Autistic Individuals: Association With Carbohydrate Digestion. Journal of Pediatric Gastroenterology and Nutrition *64*, e110–e116.

Lane, N. (2015). The unseen world: reflections on Leeuwenhoek (1677) “Concerning little animals.” Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140344–20140344.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotech *31*, 814–821.

Larsen, P.E., and Dai, Y. (2015). Metabolome of human gut microbiome is predictive of host dysbiosis. GigaScience *4*.

Larsen, P.E., Collart, F.R., Field, D., Meyer, F., Keegan, K.P., Henry, C.S., McGrath, J., Quinn, J., and Gilbert, J.A. (2011). Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. Microbial Informatics and Experimentation *1*, 4.

Lau, J.T., Whelan, F.J., Herath, I., Lee, C.H., Collins, S.M., Bercik, P., and Surette, M.G. (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Medicine *8*.

LeBlanc, J.G., Milani, C., de Giori, G.S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Current Opinion in Biotechnology *24*, 160–168.

Levin, B.J., Huang, Y.Y., Peck, S.C., Wei, Y., Martínez-del Campo, A., Marks, J.A., Franzosa, E.A., Huttenhower, C., and Balskus, E.P. (2017). A prominent glycyl radical enzyme in human gut microbiomes metabolizes *trans* -4-hydroxy-l-proline. Science *355*, eaai8386.

Li, S.H.-J., Li, Z., Park, J.O., King, C.G., Rabinowitz, J.D., Wingreen, N.S., and Gitai, Z. (2018). Escherichia coli translation strategies differ across carbon, nitrogen and phosphorus limitation conditions. Nature Microbiology *3*, 939–947.

Lin, Z., Ye, W., Zu, X., Xie, H., Li, H., Li, Y., and Zhang, W. (2018). Integrative metabolic and microbial profiling on patients with Spleen-yang-deficiency syndrome. Scientific Reports *8*.

Liu, F., Horton-Sparks, K., Hull, V., Li, R.W., and Martínez-Cerdeño, V. (2018). The valproic acid rat model of autism presents with gut bacterial dysbiosis similar to that in human autism. Molecular Autism *9*.

Liu, L.-K., Becker, D.F., and Tanner, J.J. (2017). Structure, function, and mechanism of proline utilization A (PutA). Archives of Biochemistry and Biophysics *632*, 142–157.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. MSystems *3*.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature.

Lotterhos, K.E., Moore, J.H., and Stapleton, A.E. (2018). Analysis validation has been neglected in the Age of Reproducibility. PLOS Biology *16*, e3000070.

Louis, P., Hold, G.L., and Flint, H.J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Micro *12*, 661–672.

Lu, K., Abo, R.P., Schlieper, K.A., Graffam, M.E., Levine, S., Wishnok, J.S., Swenberg, J.A., Tannenbaum, S.R., and Fox, J.G. (2014). Arsenic Exposure Perturbs the Gut Microbiome and Its Metabolic Profile in Mice: An Integrated Metagenomics and Metabolomics Analysis. Environmental Health Perspectives.

Lyall, K., Schmidt, R.J., and Hertz-Picciotto, I. (2014). Maternal lifestyle and environmental risk factors for autism spectrum disorders. International Journal of Epidemiology *43*, 443–464.

Machado, D., and Herrgård, M. (2014). Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Comput Biol *10*, e1003580.

Machado, D., Andrejev, S., Tramontano, M., and Patil, K.R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research *46*, 7542–7553.

Macklaim, J.M., Fernandes, A.D., Di Bella, J.M., Hammond, J.-A., Reid, G., and Gloor, G.B. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome *1*, 12.

Magnúsdóttir, S., and Thiele, I. (2018). Modeling metabolism of the human gut microbiome. Current Opinion in Biotechnology *51*, 90–96.

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., et al. (2016). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature Biotechnology.

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E.E., Brochado, A.R., Fernandez, K.C., Dose, H., Mori, H., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. Nature *555*, 623–628.

Maier, T.V., Lucio, M., Lee, L.H., VerBerkmoes, N.C., Brislawn, C.J., Bernhardt, J., Lamendella, R., McDermott, J.E., Bergeron, N., Heinzmann, S.S., et al. (2017). Impact of Dietary Resistant Starch on the Human Gut Microbiome, Metaproteome, and Metabolome. MBio *8*, e01343-17.

Manor, O., and Borenstein, E. (2015). MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biology *16*.

Manor, O., and Borenstein, E. (2017). Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host & Microbe *21*, 254–267.

Manor, O., Levy, R., and Borenstein, E. (2014). Mapping the Inner Workings of the Microbiome: Genomic- and Metagenomic-Based Study of Metabolism and Metabolic Interactions in the Human Microbiome. Cell Metabolism *20*, 742–752.

Mao, S.-Y., Huo, W.-J., and Zhu, W.-Y. (2014). Microbiome-metabolome analysis reveals unhealthy alterations in the composition and metabolism of ruminal microbiota with increasing dietary grain in a goat model. Environ. Microbiol.

Marcobal, A., Kashyap, P.C., Nelson, T.A., Aronov, P.A., Donia, M.S., Spormann, A., Fischbach, M.A., and Sonnenburg, J.L. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. ISME J *7*, 1933–1943.

Mardinoglu, A., Shoaie, S., Bergentall, M., Ghaffari, P., Zhang, C., Larsson, E., Bäckhed, F., and Nielsen, J. (2015). The gut microbiota modulates host amino acid and glutathione metabolism in mice. Mol Syst Biol *11*.

Markley, J.L., Brüschweiler, R., Edison, A.S., Eghbalnia, H.R., Powers, R., Raftery, D., and Wishart, D.S. (2017). The future of NMR-based metabolomics. Current Opinion in Biotechnology *43*, 34–40.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research *40*, D115–D122.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal *6*, 610–618.

McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018a). American Gut: an Open Platform for Citizen Science Microbiome Research. MSystems *3*.

McDonald, L.C., Gerding, D.N., Johnson, S., Bakken, J.S., Carroll, K.C., Coffin, S.E., Dubberke, E.R., Garey, K.W., Gould, C.V., Kelly, C., et al. (2018b). Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clinical Infectious Diseases *66*, e1–e48.

McGuire, K.L., and Treseder, K.K. (2010). Microbial communities and their relevance for ecosystem models: Decomposition as a case study. Soil Biology and Biochemistry *42*, 529–535.

McHardy, I.H., Goudarzi, M., Tong, M., Ruegger, P.M., Schwager, E., Weger, J.R., Graeber, T.G., Sonnenburg, J.L., Horvath, S., Huttenhower, C., et al. (2013). Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome *1*, 17.

McNally, C.P., and Borenstein, E. (2018). Metabolic model-based analysis of the emergence of bacterial cross-feeding via extensive gene loss. BMC Systems Biology *12*.

McNulty, N.P., Wu, M., Erickson, A.R., Pan, C., Erickson, B.K., Martens, E.C., Pudlo, N.A., Muegge, B.D., Henrissat, B., Hettich, R.L., et al. (2013). Effects of Diet on Resource Utilization by a Model Human Gut Microbiota Containing Bacteroides cellulosilyticus WH2, a Symbiont with an Extensive Glycobiome. PLoS Biology *11*, e1001637.

Medlock, G.L., Carey, M.A., McDuffie, D.G., Mundy, M.B., Giallourou, N., Swann, J.R., Kolling, G.L., and Papin, J.A. (2018). Inferring Metabolic Mechanisms of Interaction within a Defined Gut Microbiota. Cell Systems *7*, 245-257.e7.

Mee, M.T., Collins, J.J., Church, G.M., and Wang, H.H. (2014). Syntrophic exchange in synthetic microbial communities. Proceedings of the National Academy of Sciences *111*, E2149–E2156.

Melnik, A.V., da Silva, R.R., Hyde, E.R., Aksenov, A.A., Vargas, F., Bouslimani, A., Protsyuk, I., Jarmusch, A.K., Tripathi, A., Alexandrov, T., et al. (2017). Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples. Analytical Chemistry *89*, 7549–7559.

Metges, C.C. (2000). Contribution of microbial amino acids to amino acid homeostasis of the host. J. Nutr. *130*, 1857S-64S.

Miller, G.E., Engen, P.A., Gillevet, P.M., Shaikh, M., Sikaroodi, M., Forsyth, C.B., Mutlu, E., and Keshavarzian, A. (2016). Lower Neighborhood Socioeconomic Status Associated with Reduced Diversity of the Colonic Microbiota in Healthy Adults. PLOS ONE *11*, e0148952.

Milo, R. (2002). Network Motifs: Simple Building Blocks of Complex Networks. Science *298*, 824–827.

Ming, X., Stein, T.P., Barnes, V., Rhodes, N., and Guo, L. (2012). Metabolic Perturbance in Autism Spectrum Disorders: A Metabolomics Study. Journal of Proteome Research *11*, 5856–5862.

Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology *13*, R79.

Mussap, M., Noto, A., and Fanos, V. (2016). Metabolomics of autism spectrum disorders: early insights regarding mammalian-microbial cometabolites. Expert Review of Molecular Diagnostics *16*, 869–881.

Nayak, R.R., and Turnbaugh, P.J. (2016). Mirror, mirror on the wall: which microbiomes will help heal them all? BMC Medicine *14*.

Nguyen, T.L.A., Vieira-Silva, S., Liston, A., and Raes, J. (2015). How informative is the mouse for human gut microbiota research? Disease Models and Mechanisms *8*, 1–16.

Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., Young, V.B., Jansson, J.K., Fredricks, D.N., and Borenstein, E. (2016). Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation. MSystems *1*, e00013-15.

Noecker, C., Chiu, H.-C., McNally, C.P., and Borenstein, E. (2019). Defining and Evaluating Microbial Contributions to Metabolite Variation in Microbiome-Metabolome Association Studies. BioRxiv.

Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., Daníelsdóttir, A.D., Krecke, M., Merten, D., Haraldsdóttir, H.S., et al. (2018). The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Research gky992–gky992.

Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., Daníelsdóttir, A.D., Krecke, M., Merten, D., Haraldsdóttir, H.S., et al. (2019). The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Research *47*, D614–D624.

O'Keefe, S.J.D. (2016). Diet, microorganisms and their metabolites and colon cancer. Nature Reviews Gastroenterology & Hepatology *13*, 691–706.

Olivon, F., Allard, P.-M., Koval, A., Righi, D., Genta-Jouve, G., Neyts, J., Apel, C., Pannecouque, C., Nothias, L.-F., Cachet, X., et al. (2017). Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. ACS Chemical Biology.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986). Microbial Ecology and Evolution: A Ribosomal RNA Approach. Annual Review of Microbiology *40*, 337–365.

Org, E., Blum, Y., Kasela, S., Mehrabian, M., Kuusisto, J., Kangas, A.J., Soininen, P., Wang, Z., Ala-Korpela, M., Hazen, S.L., et al. (2017). Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the METSIM cohort. Genome Biology *18*.

Ottman, N., Geerlings, S.Y., Aalvink, S., de Vos, W.M., and Belzer, C. (2017). Action and function of Akkermansia muciniphila in microbiome ecology, health and disease. Best Practice & Research Clinical Gastroenterology *31*, 637–642.

Park, E., Cohen, I., Gonzalez, M., Castellano, M.R., Flory, M., Jenkins, E.C., Brown, W.T., and Schuller-Levis, G. (2017). Is Taurine a Biomarker in Autistic Spectrum Disorder? In Taurine 10, D.-H. Lee, S.W. Schaffer, E. Park, and H.W. Kim, eds. (Dordrecht: Springer Netherlands), pp. 3–16.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell.

Pedersen, H.K., Gudmundsdottir, V., Nielsen, H.B., Hyotylainen, T., Nielsen, T., Jensen, B.A.H., Forslund, K., Hildebrand, F., Prifti, E., Falony, G., et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. Nature *advance online publication*.

Pedersen, H.K., Forslund, S.K., Gudmundsdottir, V., Petersen, A.Ø., Hildebrand, F., Hyötyläinen, T., Nielsen, T., Hansen, T., Bork, P., Ehrlich, S.D., et al. (2018). A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. Nature Protocols.

Pérez-Cobas, A.E., Gosalbes, M.J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., Otto, W., Rojo, D., Bargiela, R., Bergen, M. von, et al. (2013). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut *62*, 1591–1601.

Perry, R.J., Peng, L., Barry, N.A., Cline, G.W., Zhang, D., Cardone, R.L., Petersen, K.F., Kibbey, R.G., Goodman, A.L., and Shulman, G.I. (2016). Acetate mediates a microbiome–brain–β-cell axis to promote metabolic syndrome. Nature *534*, 213–217.

Price, N.D., Magis, A.T., Earls, J.C., Glusman, G., Levy, R., Lausted, C., McDonald, D.T., Kusebauch, U., Moss, C.L., Zhou, Y., et al. (2017). A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nature Biotechnology.

Qi, J., You, T., Li, J., Pan, T., Xiang, L., Han, Y., and Zhu, L. (2018). Circulating trimethylamine N-oxide and the risk of cardiovascular diseases: a systematic review and meta-analysis of 11 prospective cohort studies. Journal of Cellular and Molecular Medicine *22*, 185–194.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. *41*, D590–D596.

Randolph, T.W., Zhao, S., Copeland, W., Hullar, M., and Shojaie, A. (2015). Kernel-Penalized Regression for Analysis of Microbiome Data. ArXiv:1511.00297 [Stat].

Rath, S., Heidrich, B., Pieper, D.H., and Vital, M. (2017). Uncovering the trimethylamine-producing bacteria of the human gut microbiota. Microbiome *5*.

Reese, A.T., and Carmody, R.N. (2018). Thinking outside the cereal box: non-carbohydrate routes for dietary manipulation of the gut microbiota. Applied and Environmental Microbiology.

Reigstad, C.S., Salmonson, C.E., Rainey, J.F., Szurszewski, J.H., Linden, D.R., Sonnenburg, J.L., Farrugia, G., and Kashyap, P.C. (2014). Gut microbes promote colonic serotonin production through an effect of short-chain fatty acids on enterochromaffin cells. The FASEB Journal.

Rettedal, E.A., Gumpert, H., and Sommer, M.O.A. (2014). Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. Nature Communications *5*.

Rezola, A., Pey, J., Tobalina, L., Rubio, Á., Beasley, J.E., and Planes, F.J. (2014). Advances in network-based metabolic pathway analysis and gene expression data integration. Brief Bioinform bbu009.

Rivière, A., Gagnon, M., Weckx, S., Roy, D., and De Vuyst, L. (2015). Mutual Cross-Feeding Interactions between Bifidobacterium longum subsp. longum NCC2705 and Eubacterium rectale ATCC 33656 Explain the Bifidogenic and Butyrogenic Effects of Arabinoxylan Oligosaccharides. Applied and Environmental Microbiology *81*, 7767–7781.

Roberts, A.B., Gu, X., Buffa, J.A., Hurd, A.G., Wang, Z., Zhu, W., Gupta, N., Skye, S.M., Cody, D.B., Levison, B.S., et al. (2018). Development of a gut microbe–targeted nonlethal therapeutic to inhibit thrombosis potential. Nature Medicine *24*, 1407–1417.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

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

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature.

Sajitz-Hermstein, M., Töpfer, N., Kleessen, S., Fernie, A.R., and Nikoloski, Z. (2016). iReMet-flux: constraint-based approach for integrating relative metabolite levels into a stoichiometric metabolic models. Bioinformatics *32*, i755–i762.

Sandler, R.H., Finegold, S.M., Bolte, E.R., Buchanan, C.P., Maxwell, A.P., Väisänen, M.-L., Nelson, M.N., and Wexler, H.M. (2000). Short-Term Benefit From Oral Vancomycin Treatment of Regressive-Onset Autism. Journal of Child Neurology *15*, 429–435.

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols *6*, 1290–1307.

Schlesinger, E.B., and Metchnikoff, E. (1908). The prolongation of life (New York: New York, Putnam).

Schneider, C.K., Melmed, R.D., Barstow, L.E., Enriquez, F.J., Ranger-Moore, J., and Ostrem, J.A. (2006). Oral Human Immunoglobulin for Children with Autism and Gastrointestinal Dysfunction: A Prospective, Open-Label Study. Journal of Autism and Developmental Disorders *36*, 1053–1064.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLOS Biology *14*, e1002533.

Sgritta, M., Dooling, S.W., Buffington, S.A., Momin, E.N., Francis, M.B., Britton, R.A., and Costa-Mattioli, M. (2019). Mechanisms Underlying Microbial-Mediated Changes in Social Behavior in Mouse Models of Autism Spectrum Disorder. Neuron *101*, 246-259.e6.

Shaffer, M., Armstrong, A.J.S., Phelan, V.V., Reisdorph, N., and Lozupone, C.A. (2017). Microbiome and metabolome data integration provides insight into health and disease. Translational Research.

Shankar, V., Homer, D., Rigsbee, L., Khamis, H.J., Michail, S., Raymer, M., Reo, N.V., and Paliy, O. (2015). The networks of human gut microbe–metabolite associations are different between health and irritable bowel syndrome. ISME J *9*, 1899–1903.

Shapley, L.S. (1953). 17. A Value for n-Person Games. In Contributions to the Theory of Games (AM-28), Volume II, H.W. Kuhn, and A.W. Tucker, eds. (Princeton: Princeton University Press), p.

Sharon, I., and Banfield, J.F. (2013). Genomes from Metagenomics. Science *342*, 1057–1058.

Sharon, G., Garg, N., Debelius, J., Knight, R., Dorrestein, P.C., and Mazmanian, S.K. (2014). Specialized Metabolites from the Microbiome in Health and Disease. Cell Metabolism *20*, 719–730.

Shen, X., Miao, J., Wan, Q., Wang, S., Li, M., Pu, F., Wang, G., Qian, W., Yu, Q., Marotta, F., et al. (2018). Possible correlation between gut microbiota and immunity among healthy middle-aged and elderly people in southwest China. Gut Pathogens *10*.

Shi, W., Moon, C., Leahy, S., Kang, D., Froula, J., Kittelmann, S., Fan, C., Deutsch, S., Gagic, D., Seedorf, H., et al. (2014). Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res. gr.168245.113.

Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters, T., Juste, C., Rizkalla, S., Chilloux, J., et al. (2015). Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. Cell Metabolism *22*, 320–331.

Sinha, R., Ahn, J., Sampson, J.N., Shi, J., Yu, G., Xiong, X., Hayes, R.B., and Goedert, J.J. (2016). Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. PLOS ONE *11*, e0152126.

Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., and Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. Ther Drug Monit *27*, 747–751.

Smith, M.I., Yatsunenko, T., Manary, M.J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A.L., Rich, S.S., Concannon, P., Mychaleckyj, J.C., et al. (2013a). Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor. Science *339*, 548–554.

Smith, P.M., Howitt, M.R., Panikov, N., Michaud, M., Gallini, C.A., Bohlooly-Y, M., Glickman, J.N., and Garrett, W.S. (2013b). The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. Science *341*, 569–573.

Smith, R., Mathis, A.D., Ventura, D., and Prince, J.T. (2014). Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. BMC Bioinformatics *15*, S9.

Snijders, A.M., Langley, S.A., Kim, Y.-M., Brislawn, C.J., Noecker, C., Zink, E.M., Fansler, S.J., Casey, C.P., Miller, D.R., Huang, Y., et al. (2016). Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. Nature Microbiology *2*, 16221.

Son, J.S., Zheng, L.J., Rowehl, L.M., Tian, X., Zhang, Y., Zhu, W., Litcher-Kelly, L., Gadow, K.D., Gathungu, G., Robertson, C.E., et al. (2015). Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. PLOS ONE *10*, e0137725.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J., Buhler, J.D., and Gordon, J.I. (2005). Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont. Science *307*, 1955–1959.

Spanogiannopoulos, P., Bess, E.N., Carmody, R.N., and Turnbaugh, P.J. (2016). The microbial pharmacists within us: a metagenomic view of xenobiotic metabolism. Nat Rev Micro *advance online publication*.

Sridharan, G.V., Choi, K., Klemashevich, C., Wu, C., Prabakaran, D., Pan, L.B., Steinmeyer, S., Mueller, C., Yousofshahi, M., Alaniz, R.C., et al. (2014). Prediction and quantification of bioactive microbiota metabolites in the mouse gut. Nat Commun *5*.

Srinivasan, S., Morgan, M.T., Fiedler, T.L., Djukovic, D., Hoffman, N.G., Raftery, D., Marrazzo, J.M., and Fredricks, D.N. (2015). Metabolic Signatures of Bacterial Vaginosis. MBio *6*, e00204-15.

Stagaman, K., Cepon-Robins, T.J., Liebert, M.A., Gildner, T.E., Urlacher, S.S., Madimenos, F.C., Guillemin, K., Snodgrass, J.J., Sugiyama, L.S., and Bohannan, B.J.M. (2018). Market Integration Predicts Human Gut Microbiome Attributes across a Gradient of Economic Development. MSystems *3*.

Stefka, A.T., Feehley, T., Tripathi, P., Qiu, J., McCoy, K., Mazmanian, S.K., Tjota, M.Y., Seo, G.-Y., Cao, S., Theriault, B.R., et al. (2014). Commensal bacteria protect against food allergen sensitization. PNAS *111*, 13145–13150.

Stegen, J.C., Bottos, E.M., and Jansson, J.K. (2018). A unified conceptual framework for prediction and control of microbiomes. Current Opinion in Microbiology *44*, 20–27.

Stewart, C.J., Mansbach, J.M., Wong, M.C., Ajami, N.J., Petrosino, J.F., Camargo, C.A., and Hasegawa, K. (2017). Associations of Nasopharyngeal Metabolome and Microbiome with Severity among Infants with Bronchiolitis. A Multiomic Analysis. American Journal of Respiratory and Critical Care Medicine *196*, 882–891.

Strati, F., Cavalieri, D., Albanese, D., De Felice, C., Donati, C., Hayek, J., Jousson, O., Leoncini, S., Renzi, D., Calabrò, A., et al. (2017). New evidences on the altered gut microbiota in autism spectrum disorders. Microbiome *5*.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., et al. (2007). LMSD: LIPID MAPS structure database. Nucleic Acids Research *35*, D527–D532.

Suez, J., Zmora, N., Zilberman-Schapira, G., Mor, U., Dori-Bachash, M., Bashiardes, S., Zur, M., Regev-Lehavi, D., Ben-Zeev Brik, R., Federici, S., et al. (2018). Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. Cell *174*, 1406-1423.e16.

Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H., Beger, R., Daykin, C.A., Fan, T.W.-M., Fiehn, O., Goodacre, R., Griffin, J.L., et al. (2007). Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics *3*, 211–221.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., and the UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.

Tabouy, L., Getselter, D., Ziv, O., Karpuj, M., Tabouy, T., Lukic, I., Maayouf, R., Werbner, N., Ben-Amram, H., Nuriel-Ohayon, M., et al. (2018). Dysbiosis of microbiome and probiotic treatment in a genetic model of autism spectrum disorders. Brain, Behavior, and Immunity *73*, 310–319.

Taxis, T.M., Wolff, S., Gregg, S.J., Minton, N.O., Zhang, C., Dai, J., Schnabel, R.D., Taylor, J.F., Kerley, M.S., Pires, J.C., et al. (2015). The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. Nucleic Acids Research gkv973.

de Theije, C.G.M., Wopereis, H., Ramadan, M., van Eijndthoven, T., Lambert, J., Knol, J., Garssen, J., Kraneveld, A.D., and Oozeer, R. (2014). Altered gut microbiota and activity in a murine model of autism spectrum disorders. Brain, Behavior, and Immunity *37*, 197–206.

Theriot, C.M., Koenigsknecht, M.J., Carlson Jr, P.E., Hatton, G.E., Nelson, A.M., Li, B., Huffnagle, G.B., Z. Li, J., and Young, V.B. (2014). Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. Nat Commun *5*.

Tikhonov, M., Leach, R.W., and Wingreen, N.S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. The ISME Journal *9*, 68–80.

Tong, M., McHardy, I., Ruegger, P., Goudarzi, M., Kashyap, P.C., Haritunians, T., Li, X., Graeber, T.G., Schwager, E., Huttenhower, C., et al. (2014). Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. ISME J.

Tramontano, M., Andrejev, S., Pruteanu, M., Klünemann, M., Kuhn, M., Galardini, M., Jouhten, P., Zelezniak, A., Zeller, G., Bork, P., et al. (2018). Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. Nature Microbiology.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods *12*, 902–903.

Tu, W.-J., Chen, H., and He, J. (2012). Application of LC-MS/MS analysis of plasma amino acids profiles in children with autism. Journal of Clinical Biochemistry and Nutrition.

Turnbaugh, P.J., and Gordon, J.I. (2008). An Invitation to the Marriage of Metagenomics and Metabolomics. Cell *134*, 708–713.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027–1131.

Unterseher, M., Jumpponen, A., öPik, M., Tedersoo, L., Moora, M., Dormann, C.F., and Schnittler, M. (2011). Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology: COMMUNITY ECOLOGY IN FUNGAL METAGENOMICS. Molecular Ecology *20*, 275–285.

Ursell, L.K., Haiser, H.J., Van Treuren, W., Garg, N., Reddivari, L., Vanamala, J., Dorrestein, P.C., Turnbaugh, P.J., and Knight, R. (2014). The Intestinal Metabolome: An Intersection Between Microbiota and Host. Gastroenterology *146*, 1470–1476.

Vandeputte, D., Falony, G., Vieira-Silva, S., Wang, J., Sailer, M., Theis, S., Verbeke, K., and Raes, J. (2017). Prebiotic inulin-type fructans induce specific changes in the human gut microbiota. Gut *66*, 1968–1974.

Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M., Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US Immigration Westernizes the Human Gut Microbiome. Cell *175*, 962-972.e10.

Varma, A., and Palsson, B.O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Bio/Technology *12*, 994–998.

Vijayakumar, S., Conway, M., Lió, P., and Angione, C. (2017). Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. Briefings in Bioinformatics.

Waldor, M.K., Tyson, G., Borenstein, E., Ochman, H., Moeller, A., Finlay, B.B., Kong, H.H., Gordon, J.I., Nelson, K.E., Dabbagh, K., et al. (2015). Where Next for Microbiome Research? PLOS Biology *13*, e1002050.

Walker, A., Pfitzner, B., Neschen, S., Kahle, M., Harir, M., Lucio, M., Moritz, F., Tziotis, D., Witting, M., Rothballer, M., et al. (2014a). Distinct signatures of host–microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. ISME J.

Walker, A., Lucio, M., Pfitzner, B., Scheerer, M.F., Neschen, S., de Angelis, M.H., Hartmann, A., and Schmitt-Kopplin, P. (2014b). Importance of Sulfur-Containing Metabolites in Discriminating Fecal Extracts between Normal and Type-2 Diabetic Mice. Journal of Proteome Research *13*, 4220–4231.

Walsh, A.M., Crispie, F., Kilcawley, K., O'Sullivan, O., O'Sullivan, M.G., Claesson, M.J., and Cotter, P.D. (2016). Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. MSystems *1*, e00052-16.

Walter, J., and Ley, R. (2011). The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. Annual Review of Microbiology *65*, 411–429.

Walters, W., Hyde, E.R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J.A., Jansson, J.K., Caporaso, J.G., Fuhrman, J.A., et al. (2016). Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. MSystems *1*.

Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotech *34*, 828–837.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., DuGar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.-M., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature *472*, 57–63.

Weir, T.L., Manter, D.K., Sheflin, A.M., Barnett, B.A., Heuberger, A.L., and Ryan, E.P. (2013). Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. PLoS ONE *8*, e70803.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal *10*, 1669–1681.

Welsh, C.E., Miller, D.R., Manly, K.F., Wang, J., McMillan, L., Morahan, G., Mott, R., Iraqi, F.A., Threadgill, D.W., and de Villena, F.P.-M. (2012). Status and access to the Collaborative Cross population. Mammalian Genome *23*, 706–712.

Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biology *18*.

Werhli, A.V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics *22*, 2523–2531.

Whitt, D.D., and Demoss, R.D. (1975). Effect of Microflora on the Free Amino Acid Distribution in Various Regions of the Mouse Gastrointestinal Tract. Appl Microbiol *30*, 609–615.

Wikoff, W.R., Anfora, A.T., Liu, J., Schultz, P.G., Lesley, S.A., Peters, E.C., and Siuzdak, G. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. PNAS *106*, 3698–3703.

Williams, B.L., Hornig, M., Buie, T., Bauman, M.L., Cho Paik, M., Wick, I., Bennett, A., Jabado, O., Hirschberg, D.L., and Lipkin, W.I. (2011). Impaired Carbohydrate Digestion and Transport and Mucosal Dysbiosis in the Intestines of Children with Autism and Gastrointestinal Disturbances. PLoS ONE *6*, e24585.

Williams, C.F., Walton, G.E., Jiang, L., Plummer, S., Garaiova, I., and Gibson, G.R. (2015). Comparative Analysis of Intestinal Tract Models. Annual Review of Food Science and Technology *6*, 329–350.

Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., et al. (2009). HMDB: a knowledgebase for the human metabolome. Nucleic Acids Research *37*, D603–D610.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al. (2018). HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research *46*, D608–D617.

Wissenbach, D.K., Oliphant, K., Rolle-Kampczyk, U., Yen, S., Höke, H., Baumann, S., Haange, S.B., Verdu, E.F., Allen-Vercoe, E., and von Bergen, M. (2016). Optimization of metabolomics of defined in vitro gut microbial ecosystems. International Journal of Medical Microbiology *306*, 280–289.

Wlodarska, M., Luo, C., Kolde, R., d'Hennezel, E., Annand, J.W., Heim, C.E., Krastel, P., Schmitt, E.K., Omar, A.S., Creasey, E.A., et al. (2017). Indoleacrylic Acid Produced by Commensal Peptostreptococcus Species Suppresses Inflammation. Cell Host & Microbe *22*, 25-37.e6.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences *74*, 5088–5090.

Woese, C.R., Fox, G.E., and Pechman, K.R. (1977). Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. International Journal of Systematic and Evolutionary Microbiology *27*, 44–57.

Wolrath, H., Forsum, U., Larsson, P.G., and Boren, H. (2001). Analysis of Bacterial Vaginosis-Related Amines in Vaginal Fluid by Gas Chromatography and Mass Spectrometry. Journal of Clinical Microbiology *39*, 4026–4031.

Wu, G.D., Compher, C., Chen, E.Z., Smith, S.A., Shah, R.D., Bittinger, K., Chehoud, C., Albenberg, L.G., Nessel, L., Gilroy, E., et al. (2016). Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. Gut *65*, 63–72.

Yano, J.M., Yu, K., Donaldson, G.P., Shastri, G.G., Ann, P., Ma, L., Nagler, C.R., Ismagilov, R.F., Mazmanian, S.K., and Hsiao, E.Y. (2015). Indigenous Bacteria from the Gut Microbiota Regulate Host Serotonin Biosynthesis. Cell *161*, 264–276.

Yassour, M., Lim, M.Y., Yun, H.S., Tickle, T.L., Sung, J., Song, Y.-M., Lee, K., Franzosa, E.A., Morgan, X.C., Gevers, D., et al. (2016). Sub-clinical detection of gut microbial biomarkers of obesity and type 2 diabetes. Genome Medicine *8*, 17.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

Yen, S., McDonald, J.A.K., Schroeter, K., Oliphant, K., Sokolenko, S., Blondeel, E.J.M., Allen-Vercoe, E., and Aucoin, M.G. (2015). Metabolomic Analysis of Human Fecal Microbiota: A Comparison of Feces-Derived Communities and Defined Mixed Communities. J. Proteome Res. *14*, 1472–1482.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized nutrition by prediction of glycemic responses. Cell *163*, 1079–1094.

Zelante, T., Iannitti, R.G., Cunha, C., De Luca, A., Giovannini, G., Pieraccini, G., Zecchi, R., D'Angelo, C., Massi-Benedetti, C., Fallarino, F., et al. (2013). Tryptophan Catabolites from Microbiota Engage Aryl Hydrocarbon Receptor and Balance Mucosal Reactivity via Interleukin-22. Immunity *39*, 372–385.

Zhang, Y., Zhao, F., Deng, Y., Zhao, Y., and Ren, H. (2015). Metagenomic and metabolomic analysis of the toxic effects of trichloroacetamide-induced gut microbiome and urine metabolome perturbations in mice. Journal of Proteome Research 150122053921003.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science *352*, 565–569.

Zhou, B., Wang, J., and Ressom, H.W. (2012). MetaboSearch: Tool for Mass-Based Metabolite Identification Using Multiple Databases. PLoS ONE *7*, e40096.

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

Zierer, J., Jackson, M.A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., Mohney, R.P., Small, K.S., Bell, J.T., Steves, C.J., et al. (2018). The fecal metabolome as a functional readout of the gut microbiome. Nature Genetics *50*, 790–795.

Chapter 7. APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

7.1 SUPPLEMENTARY TABLES

Table 7.1. Study design and metabolomic data description for all analyzed datasets.

Dataset	Organism and site	Study type	Data used	# samples	# subjects	# metabolites	# KEGG metabolites	# analyzed metabolites	Ref
1	Human, vagina	Case-control (BV)	16S qPCR, metabolomics	67 (42/25/0)*	59 (40/23/0)*	279	197	82	(22)
2	Human, vagina	Case-control (BV)[a]	16S sequencing, targeted metabolomics	70 (40/20/10)*	70 (40/20/10)*	101	96	58	(22)
3	Mouse, gut	Perturbation (antibiotics)	16S sequencing, metabolomics	29		480	297	116	(23)
4	Human, gut (fecal)	Case-control (CD)[b]	Shotgun sequencing, metabolomics	12		1113	136	31	(15, 40)
E. coli	*E. coli* cultures	Perturbation (stress)[c]	Microarray, metabolomics	28		196	87	42	(44)

** (BV+/BV-/ intermediate)*
a *Bacterial vaginosis with intermediate samples*
b *Twins discordant and concordant for Crohn's disease*
c *E. coli cultures treated with heat, cold, oxidative stress, glucose-lactose shift*

Table 7.2. Reference genomes used to infer genomic content for strains measured by qPCR in Dataset 1.

Species	Taxon ID	Genome Used for Analysis
Gardnerella vaginalis	2702	Gardnerella vaginalis 409-05
Lactobacillus crispatus	47770	Lactobacillus crispatus 214-1
Lactobacillus jensenii	109790	Lactobacillus jensenii JV-V16
Lactobacillus iners	147802	Lactobacillus iners AB-1
BVAB1	186802_1	None
BVAB2	186802_2	None
BVAB3	186802_3	Clostridiales genomosp. BVAB3 UPII9-5
Megasphaera sp. type 1	906_1	Megasphaera genomosp. type_1 str. 28L incomplete assembly
Atopobium vaginae	82135	Atopobium vaginae PB189-T1-4
Sneathia/Leptotrichia	168808	None
Eggerthella sp. type 1	84111_1	Eggerthella sp. 1_3_56FAA
Prevotella timonensis	386414	Prevotella timonensis CRIS 5C-B1
Prevotella buccalis	28127	Prevotella buccalis ATCC 35310
Prevotella amnii	419005	Prevotella amnii CRIS 21A-A

7.2 SUPPLEMENTARY RESULTS: PREDICTING VARIATION IN METABOLITES ABUNDANCES IN A SIMPLE MONO-CULTURE SYSTEM IDENTIFIES KNOWN MECHANISMS REGULATING METABOLITE VARIATION.

We examined how our metabolic-model based approach performs in a simple mono-culture system with extremely well-characterized metabolic features, as a way to provide a baseline for the boundaries of such an approach. Specifically, we applied our framework to a dataset of transcriptomic (microarray) and metabolomic abundances from *E. coli* culture samples before and after various perturbations (Jozefczuk et al., 2010). As in the more complex systems, we found that a substantial portion of the measured metabolites (26.2%; 11 out of 42) are well-predicted (as defined above) at a 1% false discovery rate (Figure 7.2A). These well-predicted metabolites include the TCA metabolite succinate, ethanolamine, tyramine, the fatty acid 2-oxobutanoate, GABA (4-aminobutanoate), and 6 amino acids (Figure 7.2B-G). The key gene contributors for these metabolites coincided with known metabolic regulatory mechanisms. For instance, ethanolamine shifts are predicted by expression changes in the periplasmic glycerophosphodiester phosphodiesterase gene, which is regulated by an operon responsive to phosphodiester availability. Similarly, succinate is predicted by variation in isocitrate lyase expression, which catalyzes the first and main limiting step of the glyoxylate cycle and is repressed during growth on glucose. Tyramine is likewise predicted by expression of the first step of an alternative degradation pathway. Amino acids were generally predicted by a balance between synthesis pathways and depletion by tRNA synthetases. These examples confirm that well-predicted metabolites and the genes that contribute to their predictions are concordant with our knowledge of metabolic network control and perturbation responses in *E. coli*.

7.3 SUPPLEMENTARY FIGURES

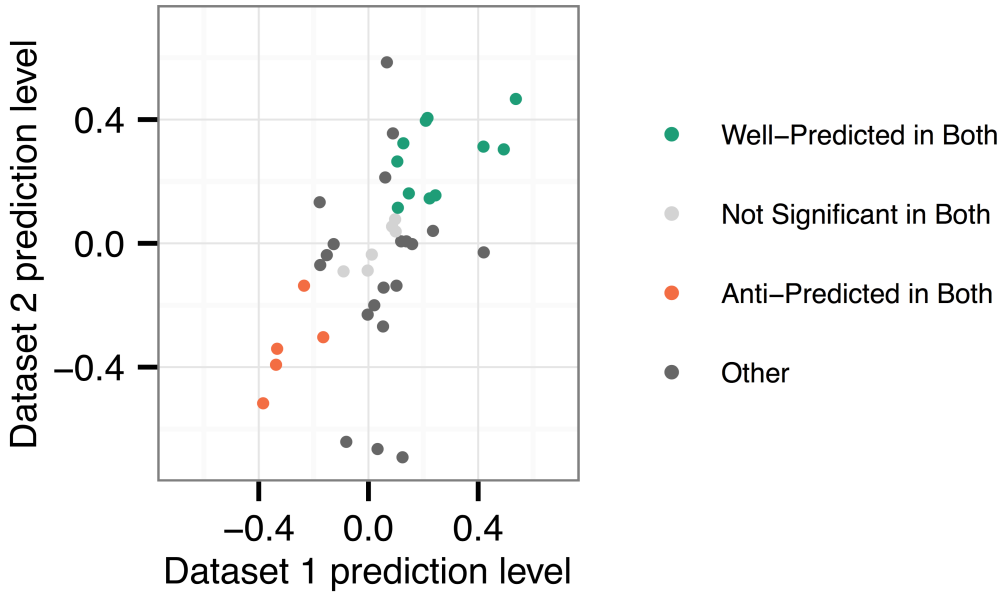

Figure 7.1. **Comparison of metabolite predictions between vaginal data sets 1 and 2.** Metabolites assayed and analyzed in both data sets are plotted, with the *x* axis representing the prediction level of the metabolite in data set 1 (measured as the Spearman correlation between pairwise differences in calculated CMP scores and pairwise differences in measured metabolite abundances) and the *y* axis representing the prediction level of that metabolite in data set 2. The color of each point denotes the agreement between the prediction type (well-predicted, anti-predicted, or neither) across the two data sets.

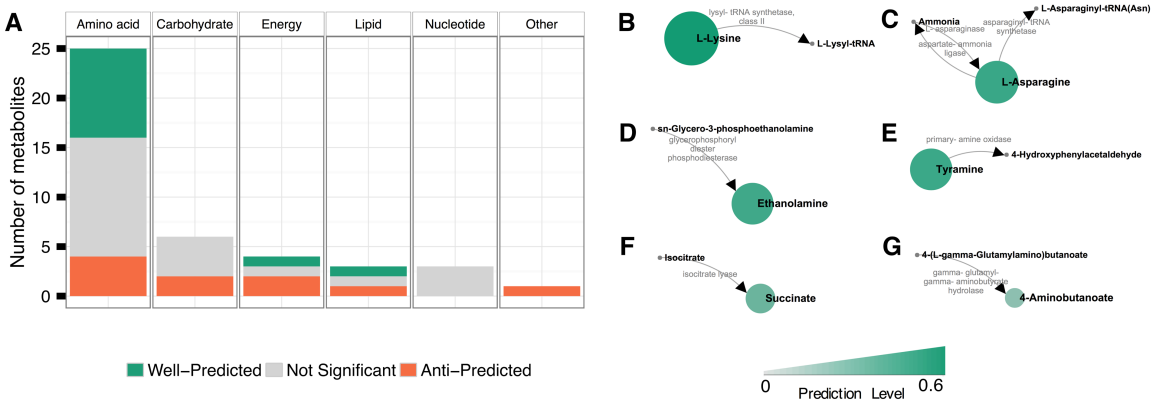

Figure 7.2. **Metabolite predictability across metabolic categories using *E. coli* monoculture gene expression data.**

(A) Prediction outcomes for metabolites assayed and analyzed in *E. coli*cultures treated to cause assorted perturbations. Bars show the number of metabolites of each prediction type in each KEGG category (see Figure 2.2 for comparison). **(B to G)** Visualization of key species and reaction contributors for metabolites of interest. The size of each node represents the magnitude of the correlation between variation in CMP scores and variation in measured metabolite abundances. Edges denote key reaction contributors.

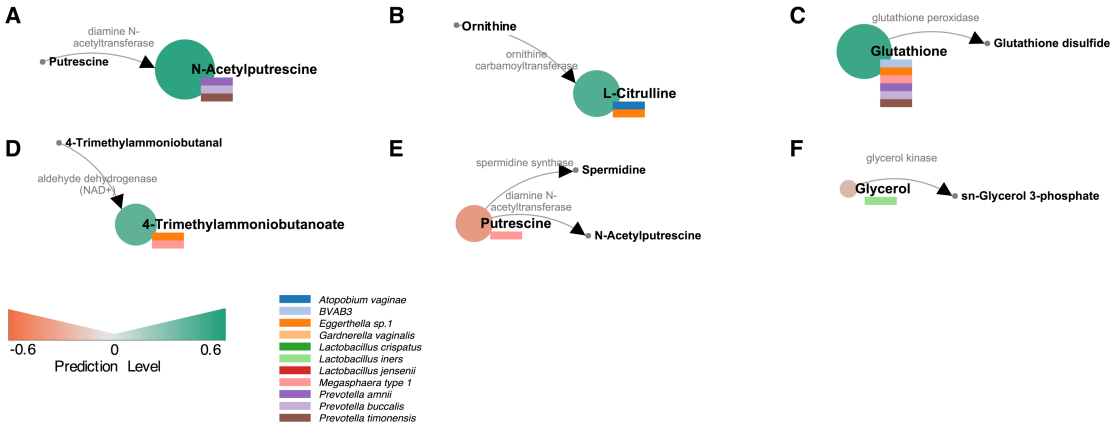

Figure 7.3. **Examples of well-predicted metabolites in data set 1 and their key species and gene contributors.**

The size of each node represents the magnitude of the correlation between variation in CMP scores and variation in measured metabolite abundances for a given metabolite, while the color indicates whether this correlation is positive or negative (i.e., well-predicted or anti-predicted).

Edges denote key reaction contributors. Colored rectangular bars denote the key species contributors for the given metabolite.

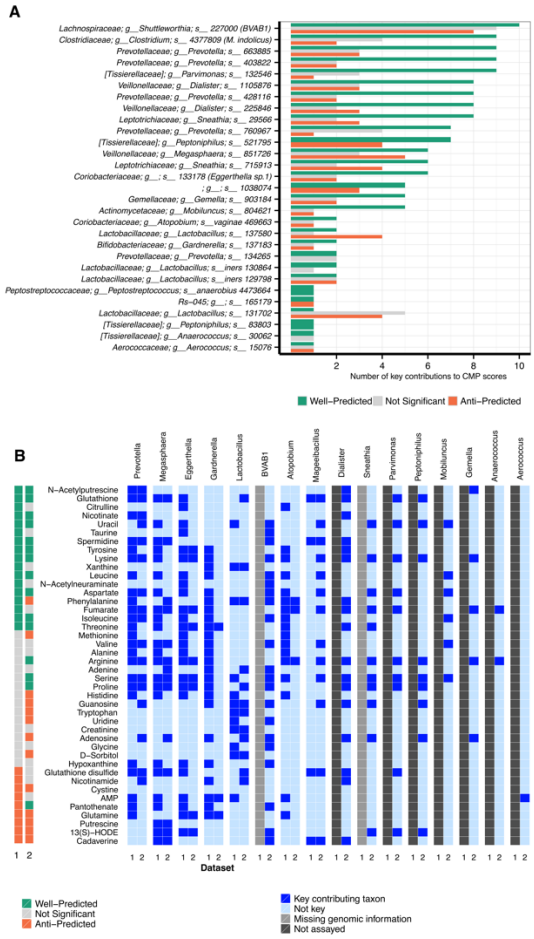

Figure 7.4. **Key species contributors to metabolites in data set 2 and variation in species contribution across data sets 1 and 2.**

(A) Each OTU that participated in the calculation of CMP scores in data set 2 is shown along the *y* axis. The *x* axis indicates the numbers of well-predicted and anti-predicted metabolites (as well as those with nonsignificant prediction) for which that OTU was a key contributor. OTUs matched to species assayed in Dataset 1 with conflicting taxonomy (Materials and Methods) are shown in parentheses. Compare also with Figure 2.3. **(B)** A comparison between key species contributors (at the genus level) of each metabolite across data sets 1 and 2. The two left columns indicate whether the metabolite was well-predicted, anti-predicted, or neither in each of the two data sets. A bright blue tile indicates that at least one species or OTU in a given genus was identified as a key contributor to predictions for the metabolite in question. Note that several genera whose abundances were measured in Dataset 2 were either not assayed in Dataset 1 (in

which taxonomic composition was assayed using 16S rRNA gene qPCR) or did not have genomic information available at the time.

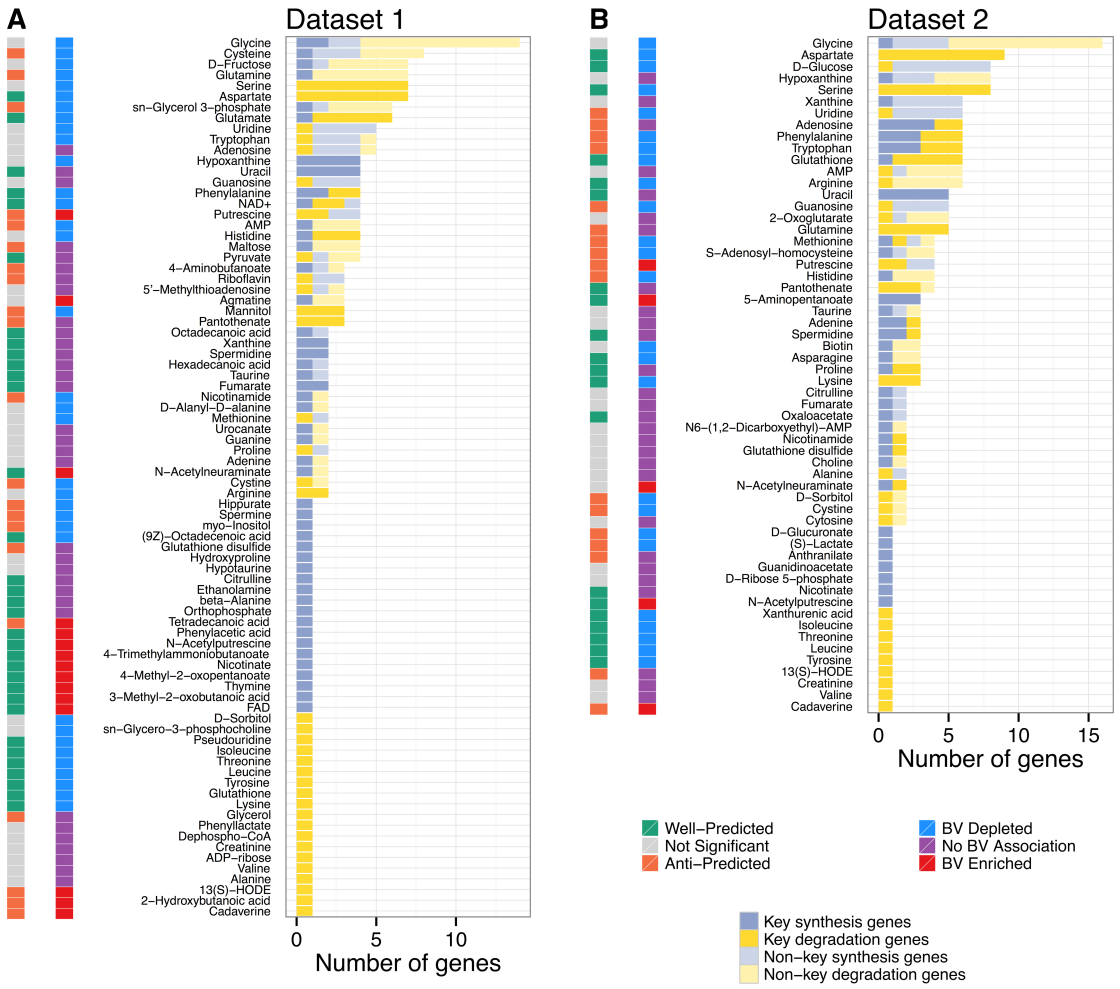

Figure 7.5. **Key gene contributors of metabolites in data sets 1 (A) and 2 (B).** Each metabolite analyzed in each data set is listed, along with the number of genes contributing to its CMP scores (*x*-axis), divided into the numbers of key genes coding for synthesis enzymes, key genes coding for degradation enzymes, and non-key genes in the same two categories. The left columns indicate whether each metabolite was enriched or depleted in BV samples and whether it was well-predicted, anti-predicted, or neither. Metabolites are ordered by the total number of relevant genes and their association with BV.

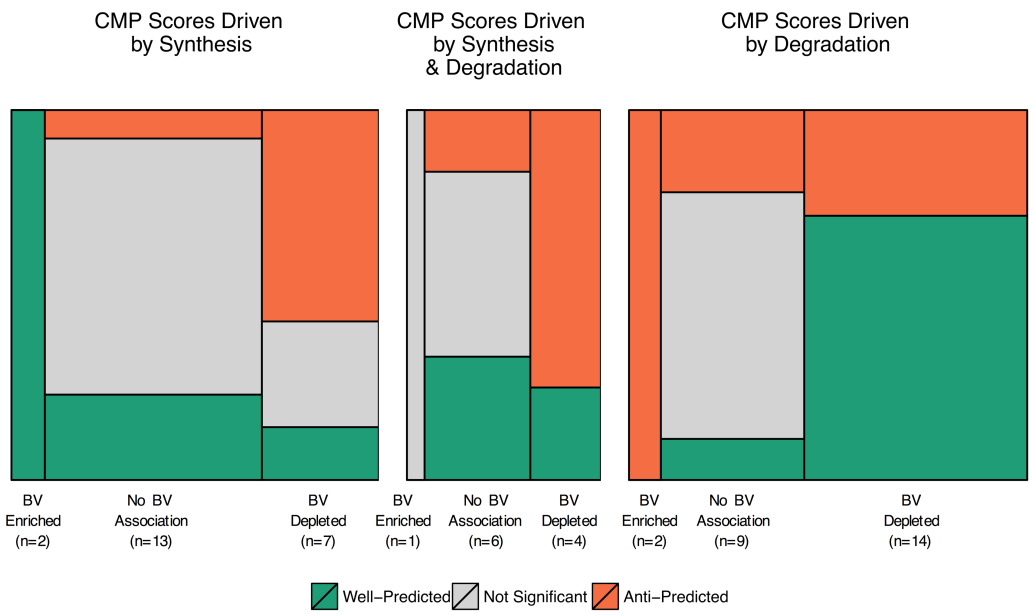

Figure 7.6. **Trends in metabolite predictability in data set 2.**

Area plots depict the numbers of metabolites whose CMP scores are driven by synthesis, by degradation, or by both in relation to their association with the host state and their predictability. The visualization used is similar to that shown in 2.4.

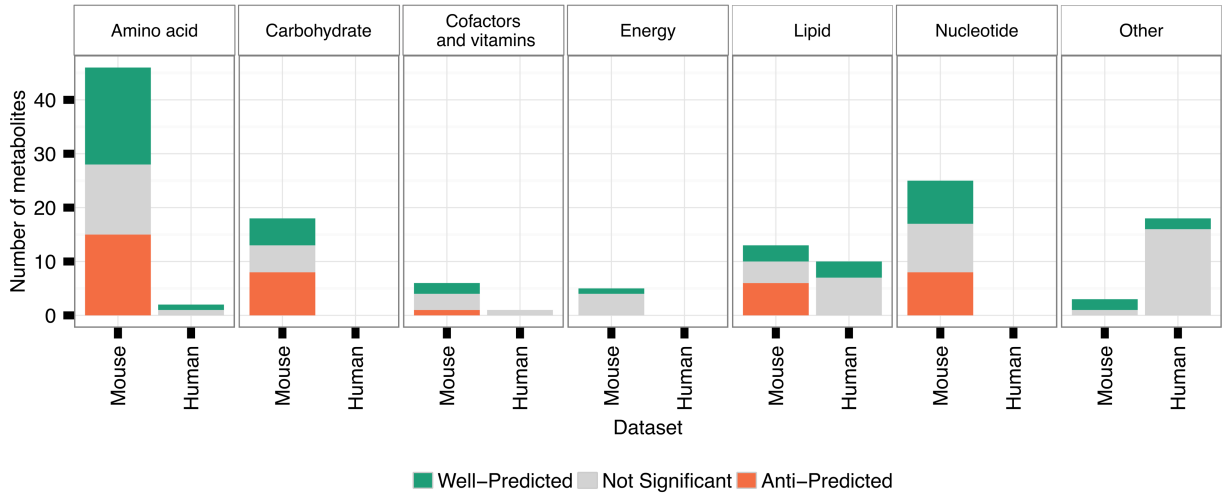

Figure 7.7. **Metabolite predictability across metabolic categories in the gut microbiome.** The visualization used is similar to that shown in Figure 2.2, with the left bar corresponding to predictions in data set 3 and the right bar corresponding to data set 4.

Chapter 8. APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

8.1 DATA COLLECTION METHODS FOR CASE STUDY 1

8.1.1 *Mouse husbandry and faecal sample collection*

Mice were obtained from the Systems Genetics Core Facility at the University of North Carolina (UNC) (Welsh et al., 2012). Before their relocation to UNC, CC lines were generated and bred at Tel Aviv University in Israel, Geniad in Australia and Oak Ridge National Laboratory in the USA. All studies were performed on young adult mice (age 9–15 weeks). For each of 30 strains, two males and two females were housed separately and maintained on PicoLab Rodent Diet 20 (5053). Mice from different strains were always housed in different cages. We observed a subtle change in microbial composition in samples collected 16 h after a cage change compared to <2 h. However, to collect sufficient mouse faecal material for combined microbiome and metabolomic analysis, all faecal samples were consistently collected from each cage, avoiding areas clearly contaminated with urine, 16 h after cage change at 2, 4, 6 and 8 weeks after arrival at Lawrence Berkeley National Laboratory (LBNL). All animal procedures were approved by the UNC Chapel Hill or LBNL Institutional Animal Care and Use Committees.

Faecal samples were stored at −80 °C for downstream metabolite and microbial analyses. Faecal samples from different strains were collected in the same way to avoid collection and storage biases. Genotyping data for CC mice were obtained from UNC (http://csbio.unc.edu/CCstatus/index.py).

Faecal samples were collected from a different cohort of genetically identical young adult mice at UNC Chapel Hill (maintained on Labdiet Prolab 3500) to determine the effect of environment on the faecal microbiome and metabolome. Faecal samples were then manually

homogenized on ice with a micropestle, 0.25 g was used for DNA isolation, 0.05 g for metabolite extraction and the remainder stored at −80 °C.

8.1.2 *Microbiome analyses*

Genomic DNA was extracted from 0.25 g of the homogenized faecal samples using the PowerSoil DNA Isolation Kit (http://www.mobio.com/) according to the manufacturer's instructions. PCR amplification of the V4 region of the 16S rRNA gene was performed using the protocol developed by the Earth Microbiome Project (http://press.igsb.anl.gov/earthmicrobiome/empstandard-protocols/16s/) and described in ref. (Caporaso et al., 2012) using updated primers described in ref. (Walters et al., 2016). Amplicons were sequenced on an Illumina MiSeq using the 150 base pair (bp) MiSeq Reagent Kit v2 (http://www.illumina.com/) according to the manufacturer's instructions.

QIIME 1.9.1 was used to join, quality filter and demultiplex libraries from three MiSeq runs (Caporaso et al., 2010b). *vsearch* 1.1.3 was used to dereplicate, sort by abundance, remove single reads and then to cluster at 97% similarity. *vsearch* was also used to check these clusters for chimaeras and construct an abundance table by mapping labelled reads to chimaera-checked clusters (Edgar, 2010; Edgar et al., 2011; Rognes et al., 2016).

8.1.3 *Extraction of metabolites from faecal homogenates*

Metabolites were extracted from mouse faecal samples using a methanol/sonication method (Walker et al., 2014b). Briefly, portions of the homogenized samples were weighed and extracted with cold (−20 °C) methanol proportionally (1 ml solvent added per 100 mg homogenate) in a microcentrifuge tube. The average weight of the homogenized faecal samples was 69.3 ± 26.3 mg (mean ± standard deviation, s.d.) and the methanol extracts contained the same theoretical concentration of metabolites. A 100 µl volume of each methanol extract was transferred

to glass vials and dried in a speed-vac concentrator (Labconco CentriVap Benchtop Vacuum Concentrator). Dried metabolite extracts were chemically derivatized using a modified version of the protocol used to create FiehnLib (Kind et al., 2009).

An Agilent GC 7890A coupled with a single quadrupole MSD 5975C (Agilent Technologies) was used and the samples were blocked and analysed in random order for each experiment. An HP-5MS column (30 m × 0.25 mm × 0.25 μm; Agilent Technologies) was used for untargeted metabolomics analyses. The sample injection mode was splitless and 1 μl of each sample was injected. The injection port temperature was held at 250 °C throughout the analysis. The GC oven was held at 60 °C for 1 min after injection and the temperature was then increased to 325 °C by 10 °C min^{-1}, followed by a 5 min hold at 325 °C (Kim et al., 2013). The helium gas flow rates for each experiment were determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid and were in the range of 0.45–0.5 ml min^{-1}. Data were collected over the mass range 50–550 *m/z*. A mixture of fatty acid methyl esters (FAMEs) (C8–C28) was analysed once per day together with the samples for retention index alignment purposes during subsequent data analysis.

GC–MS raw data files were processed using the Metabolite Detector software, version 2.5 beta (Hiller et al., 2009). Briefly, Agilent .D files were converted to netCDF format using Agilent Chemstation, followed by conversion to binary files using Metabolite Detector. Retention indices (RIs) of detected metabolites were calculated based on analysis of the FAMEs mixture, followed by their chromatographic alignment across all analyses after deconvolution. Metabolites were initially identified by matching experimental spectra to an augmented version of FiehnLib (Kind et al., 2009) (that is, the Agilent Fiehn Metabolomics Retention Time Locked (RTL) Library, containing spectra and validated retention indices for over 700 metabolites), using a Metabolite Detector match probability threshold of 0.6 (combined retention index and spectral probability).

All metabolite identifications were manually validated to reduce deconvolution errors during automated data-processing and to eliminate false identifications. We propose that this approach results in a metabolite identification confidence of Level 1.5 (Level 1 is highest, Level 4 is lowest), according to the guidelines recommended by the Metabolomics Standards Initiative Chemical Analysis Working Group of the Metabolomics Society (Sumner et al., 2007). The library used to identify metabolites was generated by an external laboratory, but this library contains both retention indices and mass spectra from analyses of authentic chemical standards and our analyses were performed using methods identical to those used to create the library. The NIST 14 GC–MS library was also used to cross-validate the spectral matching scores obtained using the Agilent library and to provide identifications of unmatched metabolites (Level 2 identifications). The three most abundant fragment ions in the spectra of each identified metabolite were automatically determined by Metabolite Detector and their summed abundances were integrated across the GC elution profile; fragment ions due to trimethylsilylation (that is, *m/z* 73 and 147) were excluded from the determination of metabolite abundance. A matrix of identified metabolites, unidentified metabolite features (characterized by mass spectra and retention indices and assigned as 'unknown'; Level 4 identifications) and their abundances was created for subsequent data analysis. Features resulting from GC column bleeding were removed from the data matrices before further data processing and analysis.

8.2 Data collection methods for Case Study 2

8.2.1 *Human fecal samples*

For all animal experiments, Arizona State University (ASU) shared human fecal samples with California Institute of Technology (Caltech) with a Material Transfer Agreement and approval to share de-identified data by the Institutional Review Board (IRB) at ASU (ASU IRB

protocol #1206007979, Caltech IRB protocol # 15-0569). Human fecal samples were previously collected from typically developing children and children with autism spectrum disorders (ASD) at ASU (Kang et al., 2013). All fecal samples and their metadata including gastrointestinal (GI)- and ASD-relevant clinical data were de-identified before being shared with Caltech.

8.2.2 *Mouse husbandry*

All animal husbandry and experiments were approved by the Caltech's Institutional Animal Care and Use Committee (IACUC protocol #1645). Throughout the study, colonized animals were maintained in autoclaved microisolator cages with autoclaved water and chow (Laboratory Autoclavable Rodent Diet - 5010, LabDiet, St. Louis, MO).

8.2.3 *Mouse Colonization*

Germ-free (GF) C57BL/6J weanlings (3-4 weeks of age) from the Mazmanian laboratory colony were colonized with fecal samples from human donors. Human fecal samples were collected by the Krajmalnik-Brown laboratory at the Arizona State University as part of a previous study (Kang et al., 2013), and kept at -80 °C. Aliquots of 16 donor samples were sent to Caltech and used for colonization. To that end, frozen aliquots were thawed in an anaerobic chamber and resuspended in two volumes of reduced sodium bicarbonate solution (final concentration 5 %). Subsequently, samples were vigorously vortexed and spun down. Supernatants were then used to colonize GF mice by a single gavage (100 ul / mouse; Instech, PA, USA). Colonized mice (4-6 females and 2-3 males per donor) were then allowed to rest for 3 weeks, and were subsequently mated according to donor. Pregnant dams were single-housed at E15.5-17.5, and offspring were weaned at 3 weeks of age. At weaning, different litters born within up to a week of each other were combined and housed in groups of 4-5 male or female mice per cage and used for subsequent

analyses. Cages were assigned to either behavior testing or for tissue collection. Behavior testing started at 6 weeks of age, while tissue was collected at P45.

8.2.4 *Mouse fecal sample collection and microbial DNA extraction*

Frozen mouse fecal samples were shipped overnight on dry ice to ASU and stored in -80°C until DNA extraction. Human feces that were used as donor samples for the mouse experiments were also shipped back to ASU in order to be processed for microbial DNA extraction and next-generation sequencing together with mouse fecal samples. At ASU, microbial genomic DNA was extracted from fecal samples using the PowerSoil® DNA Isolation Kit (Mobio Carlsbard, CA) with a modification based on the manufacturer protocol. Quality and quantity of genomic DNA was verified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technology, Rockland, DE).

8.2.5 *Microbiome analysis via shotgun metagenomic sequencing*

A miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems) was used for library generation. DNA extracts were normalized to 5 ng total input per sample in an Echo 550 acoustic liquid handling robot (Labcyte Inc). A Mosquito HTS liquid-handling robot (TTP Labtech Inc was used for 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions carried out using). Sequencing adapters were based on the iTru protocol (Glenn et al. 2016), in which short universal adapter stubs are ligated first and then sample-specific barcoded sequences added in a subsequent PCR step. Amplified and barcoded libraries were then quantified by the PicoGreen assay and pooled in approximately equimolar ratios before being sequenced on an Illumina HiSeq 4000 instrument to >30X coverage.

The metagenomic data was processed using the Oecophylla pipeline (https://oecophylla.readthedocs.io). The raw reads were examined with FastQC and low quality

sequences and unwanted reads were filtered with Bowtie2 v0.1. The taxonomic composition was profiled using the default parameters of MetaPhlAn2 (Truong et al., 2015) through the Oecophylla pipeline. The functional gene pathway was profiled using the default settings of HUMAnN2 (Franzosa et al., 2018a) through the Oecophylla pipeline. HUMAnN2 uses the UniRef90, MetaCyc and MinPath databases along with MetaPhlAn2 and ChocoPhlAn pangenome databases to characterize the pathways and genes in sequences. Gene family abundance, pathway abundance, and pathway coverage of each sample were generated from HUMAnN2; we used the gene family abundance output biom table for analysis, and the stratified gene family abundance table for MIMOSA analysis.

8.2.6 *Metabolomics analysis*

De-identified colon contents and serum samples were collected and flash-frozen at P45 without any buffers, and were shipped to the Dept. of Energy Pacific Northwest National Laboratory for metabolomic analysis by NMR and GC-MS. In colon contents, a total of 122 metabolites were identified by GC-MS (out of a total of 246 detected), and 67 metabolites were detected and identified by NMR. In serum, a total of 130 metabolites were identified by GC-MS (out of a total of 255 detected).

8.2.7 *GC-MS sample preparation and analysis*

Metabolites were extracted from murine colon contents and plasma samples using methanol (Deroussent et al., 2011; Snijders et al., 2016). Feces were homogenized and weighed, and chilled methanol (-20°C) was added proportionally to the colon content sample (1 mL to 100 mg). Glass beads were added and the suspension was agitated and sonicated to extract metabolites. Supernatant was collected after centrifugation (15,000 g × 5 min at 4°C) and 100 µL of each methanol layer was transferred to a new clean vial and subsequently dried under a speed-vacuum

concentrator. 50 µL of serum samples were thawed and 200 µL of chilled methanol was added to denature proteins. Supernatants were collected after centrifugation (15,000 g × 5 min at 4°C). All the samples were then dried completely and stored at -70°C freezer until the instrumental analysis. Prior to analysis, the stored extracts were completely dried under speed-vacuum to remove moisture and were subsequently derivatized chemically, by methoxyamination and trimethylsilyation (TMS), as reported previously (Snijders et al., 2016). Briefly, methoxyamine (20 µL of a 30 mg mL-1 stock in pyridine) was added to each sample, followed by incubation at 37°C with shaking for 90 min. subsequently, N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (80 µL) was added to each vial, and incubated at 37°C with shaking for 30 minutes. Samples were then allowed to cool to room temperature and were analyzed on the same day. Metabolites were resolved by gas chromatography using a HP-5MS column (30 m × 0.25 mm × 0.25 µm; Agilent Technologies). Samples (1 µL) were injected in splitless mode, and the helium gas flow rate was determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid (Agilent Technologies, Santa Clara, CA). The injection port temperature was held at 250°C throughout the analysis. The GC oven was held at 60°C for 1 min after injection, and the temperature was then increased to 325°C by 10°C/min, followed by a 10 min hold at 325°C. The transfer line between GC and MS was maintained at 280°C. All the MS data were collected over the mass range of 50-550 m/z under standard electron impact (EI) ionization mode at 70 eV of ionization energy. GC-MS raw data files were processed using the Metabolite Detector software, version 2.5 beta (Hiller et al., 2009). Retention indices (RI) of detected metabolites were calculated based on the analysis of the FAMEs mixture (C8-C28), followed by their chromatographic alignment across all analyses after deconvolution. Metabolites were initially identified by matching experimental spectra to a PNNL augmented version of Agilent GC-MS metabolomics Library, containing spectra and validated

retention indices for over 850 metabolites. Subsequently, any unknown peaks were matched to the NIST14 GC-MS library. All metabolite identifications and quantification ions were validated and confirmed to reduce deconvolution errors during automated data-processing and to eliminate false identifications.

8.2.8 *Proton NMR Metabolomics*

A global metabolomics approach was used to obtain assignment and quantitation of metabolites via nuclear magnetic resonance (1H NMR). The one-dimensional (1D) 1H NMR spectra of all samples were collected in accordance with standard Chenomx (Edmonton, Alberta, Canada) sample preparation and data collection guidelines (Weijle, 2006). Fecal extract samples were diluted by 10% (v/v) spike of a National Institute of Standards and Technology calibrated reference solution (100% D2O, 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d6 (DSS), and 0.1% sodium azide). All NMR spectra were collected using a Varian Direct Drive 600 MHz NMR spectrometer equipped with a 5 mm triple-resonance salt-tolerant cold probe. The 1D 1H spectra were collected following standard Chenomx data collection guidelines (Weljie et al., 2006), employing a 1D NOESY presaturation (TNNOESY) experiment with 65536 complex points and at least 512 scans at 298 K.A presaturation delay of 1.5 s was used to optimize water suppression. The 1D 1H NMR spectra of all samples were processed, assigned, and analyzed by using Chenomx NMR Suite 8.1 (Chenomx Inc.) with quantification based on spectral intensities relative to the internal standard. Candidate metabolites present in each of the complex mixture were determined by matching the chemical shift, J-coupling, and intensity information of experimental NMR signals against the NMR signals of standard metabolites in the Chenomx library which include metabolites from the HMDB database.

Chapter 9. APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

9.1 SUPPLEMENTARY RESULTS

9.1.1 *Simulated species responses to media variation partially recapitulate experimental results*

We ran several sets of simulations with the same set of initial species compositions but we maintained the same set of 61 initial species compositions but with small amounts of stochastic noise added to the nutrient inflow, sampling inflow concentrations for each compound in each simulation from a normal distribution with a mean equal to the compound's original inflow rate and a standard deviation set to a particular fraction of the mean (Methods). We evaluated whether the differences in simulated species growth across simulations with large media fluctuations (8-10%) recapitulated the experimental observations of (Faith et al., 2011), finding agreement on several trends. First, Faith *et al.* observed that casein abundance (presumably, due to limiting amino acid or nitrogen concentrations) was strongly associated with the total community biomass. In our simulations, the total amount of amino acids is indeed positively associated with total biomass (Spearman rho=0.25, p=0.005). We fit linear regression models of total biomass across all simulations based on inflow metabolite levels, and found that concentrations of both amino acids and carbohydrates explain most of the variation in this quantity in our simulations (adjusted R^2 of 0.97), with other nutrients adding only negligible effects. Faith *et al.* further observed that while most species increase their growth rate in the presence of higher protein, *E.rectale*, *D. piger* and *M. formatexigens* are negatively associated with the amount of casein in the mouse diet. In our simulations, a model predicting growth rate based on amino acid levels explains no variation in *E. rectale* and only 8% in *M. formatexigens*, in contrast to other taxa, for which amino acids explained up to 54% of growth rate variation. Additionally, the initial growth rates for all species were

positively correlated ($p < 0.1$) with an average of 2.7 different amino acid compounds (range 1 to 6), while *E. rectale* and *M. formatexigens* were each only correlated with a single one, *L*-cysteinylglycine. Lastly, our simulations also recapitulate differences in carbohydrate use, with Faith *et al.* observing preferential expansion of *B. ovatus* and *B. thetaiotaomicron* on a high-starch diet compared to a high-sugar diet. In our simulations, the growth rate of all species was associated with the amount of available simple sugars, but only *B. ovatus*, *B. thetaiotaomicron*, and *E. rectale* were significantly correlated with the quantity of starch in the inflow (Spearman rho coefficients of 0.5, 0.53, 0.53 respectively, all $p < 10^{-9}$). These results indicate that our simulation framework successfully encapsulates some, though not all, of the nutrient limitations shaping the growth dynamics of this model community.

9.1.2 *Analysis of an alternative definition of contribution values based on flux rates*

Our contribution value metric attributes metabolite variance to each species depending on its cumulative metabolite uptake or secretion over the entire simulation, rather than its arrived-at steady-state metabolite flux at the time of "sampling". To assess the impacts of this choice, we calculated contribution values using an alternative definition based solely on steady-state fluxes. Specifically, we calculated the contribution of each species to the metabolite flux at the final time point of a simulation run for 144 hours and for 1440 hours. Under this definition, steady-state contribution values explain the variation in metabolite flux rate at the time of sampling, rather than the accumulated variation in metabolite concentrations (cumulative contribution values). We compared these alternative steady-state contributions with the original set of cumulative contribution values at both time points, finding that they are highly similar. In our original dataset of simulations run for 144 hours, the Pearson correlation between steady-state and cumulative contribution values for each metabolite was on average 0.99 (minimum of 0.75). Only 6 of the 520 analyzed species-metabolite pairs differ in contributor status between the two definitions: 4 pairs

are key cumulative contributors but not steady-state contributors, and 2 pairs are the reverse. The AUC for detection of steady-state contributors is 0.710 (compared with 0.717 for cumulative contributors). These differences also naturally recede further for simulations run for longer durations: in a dataset of simulations run for 1440 hours, the average metabolite-level correlation between steady-state and cumulative contribution values was 0.999 (minimum 0.97). These results indicate that for these simulations, historical differences in species composition and metabolic activity are not a major factor in the observed discrepancy between species-metabolite correlations and true key contributors to metabolic variation.

9.1.3 *Analysis of an alternative definition of key taxon-metabolite pairs*

For most analyses, we defined the key taxonomic contributors for a particular metabolite as those species with the highest positive contribution values, or those that are responsible for the observed pattern of variation in a metabolite. However, an alternative goal could be to detect all microbes that substantially impact levels of a given metabolite across samples, regardless of whether their effects are ultimately reflected in the observed concentrations. To this end, we defined *key player* species as those with a contribution value, either positive or negative, greater in magnitude than 20% of the total contribution magnitude. This resulted in 91 species-metabolite key player pairs, including 65 of the previously defined 'positive' key contributor pairs but also 26 players with negative contributions, and these were distributed similarly across metabolites and species (Figure 9.5, panels A-B). Examining how well these key players were detected by a correlation-based analysis, we found similar performance to those reported above for key contributors (Figure 9.5, panels C-G), including a comparable positive predictive value (31.9%) and AUC (0.73).

9.1.4 *Effects of simulation length and Vmax parameter on correlation results*

We assessed the sensitivity of our correlation results to the parameters used in our simulations. Specifically, we evaluated the effect of the duration of simulations on our results. We ran additional simulations for 5,760 time points (or 1,440 hours), and calculated contribution values and correlation coefficients at 22 intermediate time points starting at 36 hours (Figure 9.8). Species compositions and metabolite concentrations became increasingly less variable with longer simulation time, converging towards similar steady states dominated in abundance by 5 of the 10 species (Figure 9.8A-B). Correspondingly, the number of key contributors decreased with increasing simulation length, from 121 contributors across all 52 analyzed metabolites at 36 hours, to 75 at 1,440 hours (Figure 9.8B-C). The number of significantly correlated species-metabolite pairs, however, increased from 179 to 375 over the same datasets, detecting contributors with higher sensitivity but lower specificity (Figure 9.8D). Ultimately, the AUC and positive predictive value both decrease slightly with increasing simulation length, with the AUC shifting from 0.67 to 0.73 and positive predictive value from 39.7% to 18.4% (Figure 9.8E). This transition occurs sharply initially before reaching an inflection point and beginning to stabilize around 144 hours, the length of time chosen for our main analysis.

We also generated additional datasets with the same initial species compositions but with widely varied values for the universal *Vmax* parameter, which was set to 20 in the main set of analyses. Changing this parameter had very minimal impact on both the simulation abundance profiles and the results of correlation analysis (Figure 9.9). The AUC for the identification of key contributors was not associated with the value of the *Vmax* parameter, and only ranged from 0.70 to 0.72.

9.1.5 *Features distinguishing true key contributors from false positives among correlated pairs*

We constructed additional regression models to assess whether there are features that can distinguish true key contributors from false positives among all correlated species-metabolite pairs. We fit regression models to similarly assess whether species and/or metabolite identity are indicative of whether a correlated species-metabolite pair represents a true or false positive relationship. We found that species identity ($p = 0.047$), but not metabolite identity, was predictive of key contributor status among correlated pairs. This is unsurprising given that the number of key contributions from each species varied widely, while all metabolites have at least one key contributor.

9.1.6 *Additional effects of inflow fluctuations on contribution and correlation profiles*

We assessed whether the addition of external metabolite fluctuations impacted the profile of species contributing to each metabolite. For most metabolites (28 out of 52, including 12 out of 14 non-inflow metabolites), the top microbial contributor did not change across all levels of fluctuation. However, for many inflow metabolites, the large external fluctuations can result in a switch in contribution values. In these cases, activity by a microbe that contributed to variation in a constant-inflow setting instead has a mitigating impact, resulting in a negative contribution. Of the 65 key contributors to variation in inflow metabolites in the original dataset, 34 (52%) of them have a negative contribution value in at least one simulation run with external fluctuations. This observation highlights that our definition of key contributors is context-dependent, identifying the entities primarily responsible for the observed variation.

We also examined whether the detection of the 14 variable metabolites not present in the nutrient inflow was affected by random fluctuations in inflow metabolites. Variation in 8 of these metabolites was significantly positively correlated with variation in the surrounding inflow

(Spearman rho, $p < 0.01$), suggesting that their synthesis fluxes were affected by changes in microbial growth or nutrient usage that resulted from environmental shifts. Correlation analysis tended to identify key contributors for these metabolites with slightly higher specificity and lower sensitivity as inflow fluctuations increased (Figure 9.10).

9.2 SUPPLEMENTARY FIGURES

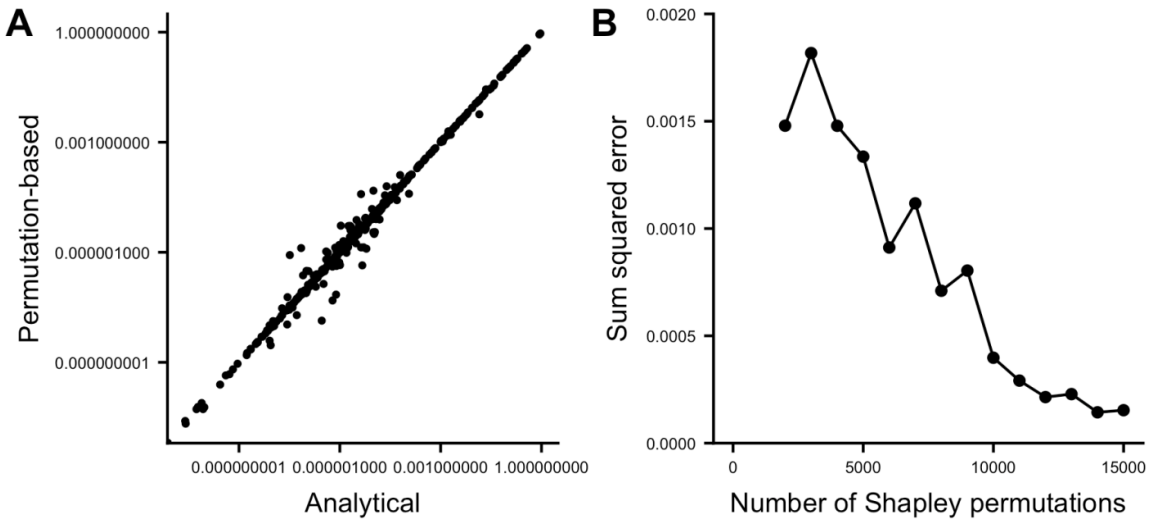

Figure 9.1. **Shapley values are equivalent to analytically calculated variance contributions.**

(A) Plot of contribution values calculated analytically versus those obtained from a Shapley value-based permutation analysis using 15,000 species orderings (see Methods), for all 52 analyzed metabolites in our simulated dataset. Axes are on a $\log_{10}$ scale. **(B)** Plot of the total sum of squared error between Shapley values calculated using permuted species subsets and our analytically calculated variance contributions, for all metabolites. With increasing numbers of permutations and therefore increasingly precise contribution estimates, the difference between these values approaches 0.

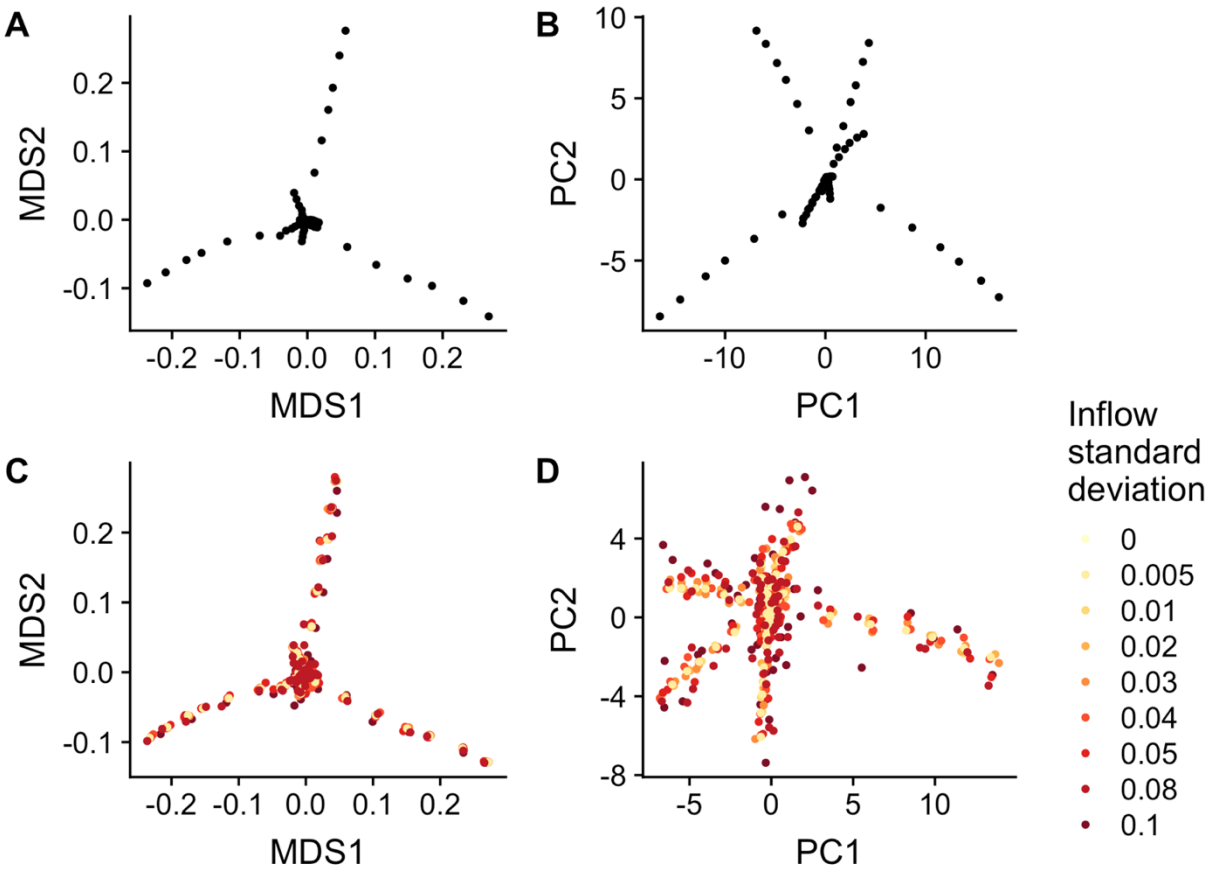

Figure 9.2. **Ordination plots of species and metabolite abundances in simulated datasets.** **(A)** Non-metric multidimensional scaling plots of species composition across the 61 original simulation runs, using Bray-Curtis dissimilarity. **(B)** Principal components analysis of metabolite concentrations across the 61 original simulation runs. **(C-D)** The same plots as **(A)** and **(B)**, but including all simulation runs with environmental fluctuations in the nutrient inflow.

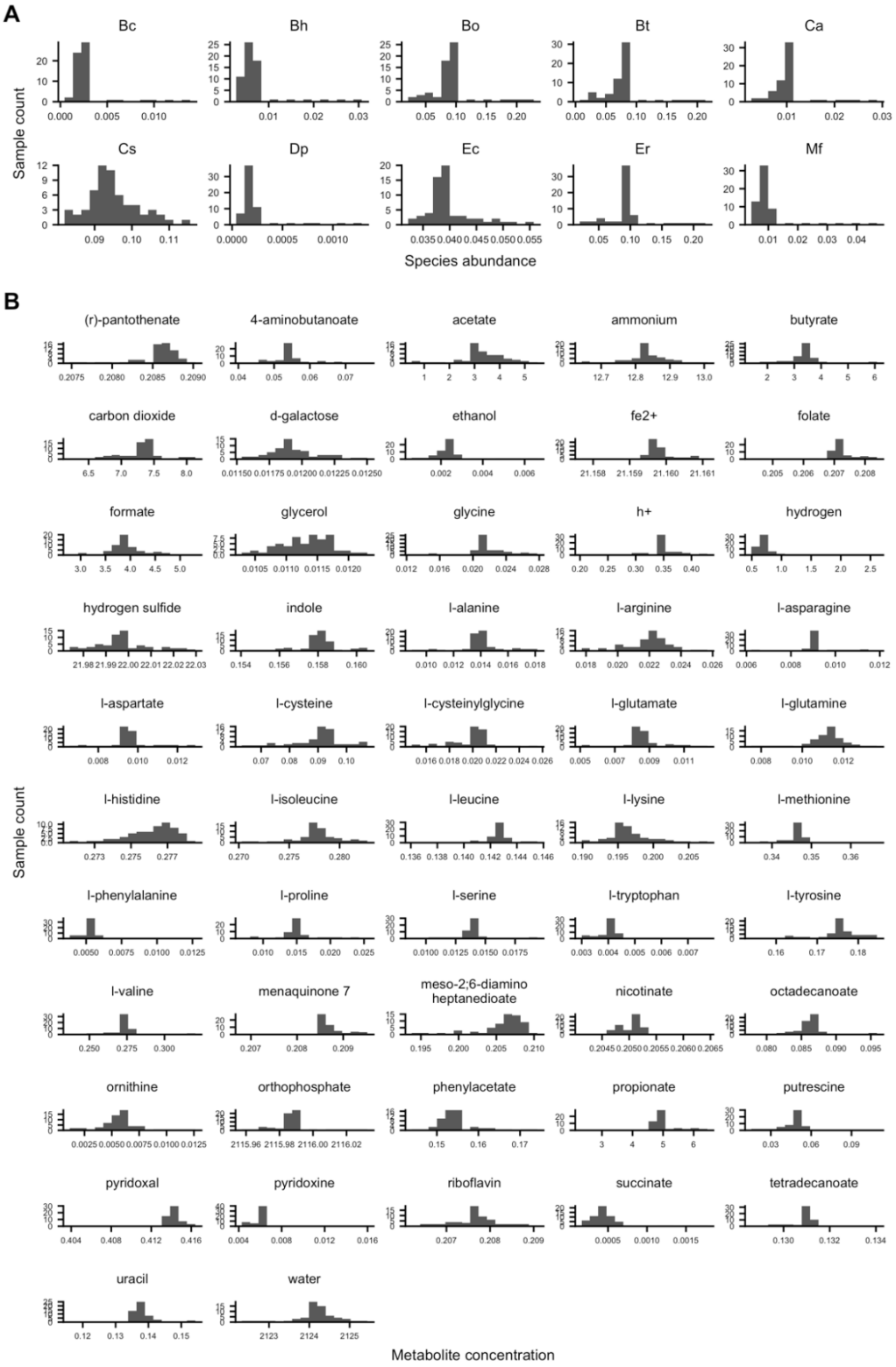

Figure 9.3. **Distributions of species and metabolite abundances.**
Each panel shows a histogram of abundances for a single species (A) or a single variable metabolite (B) across all 61 simulation runs.

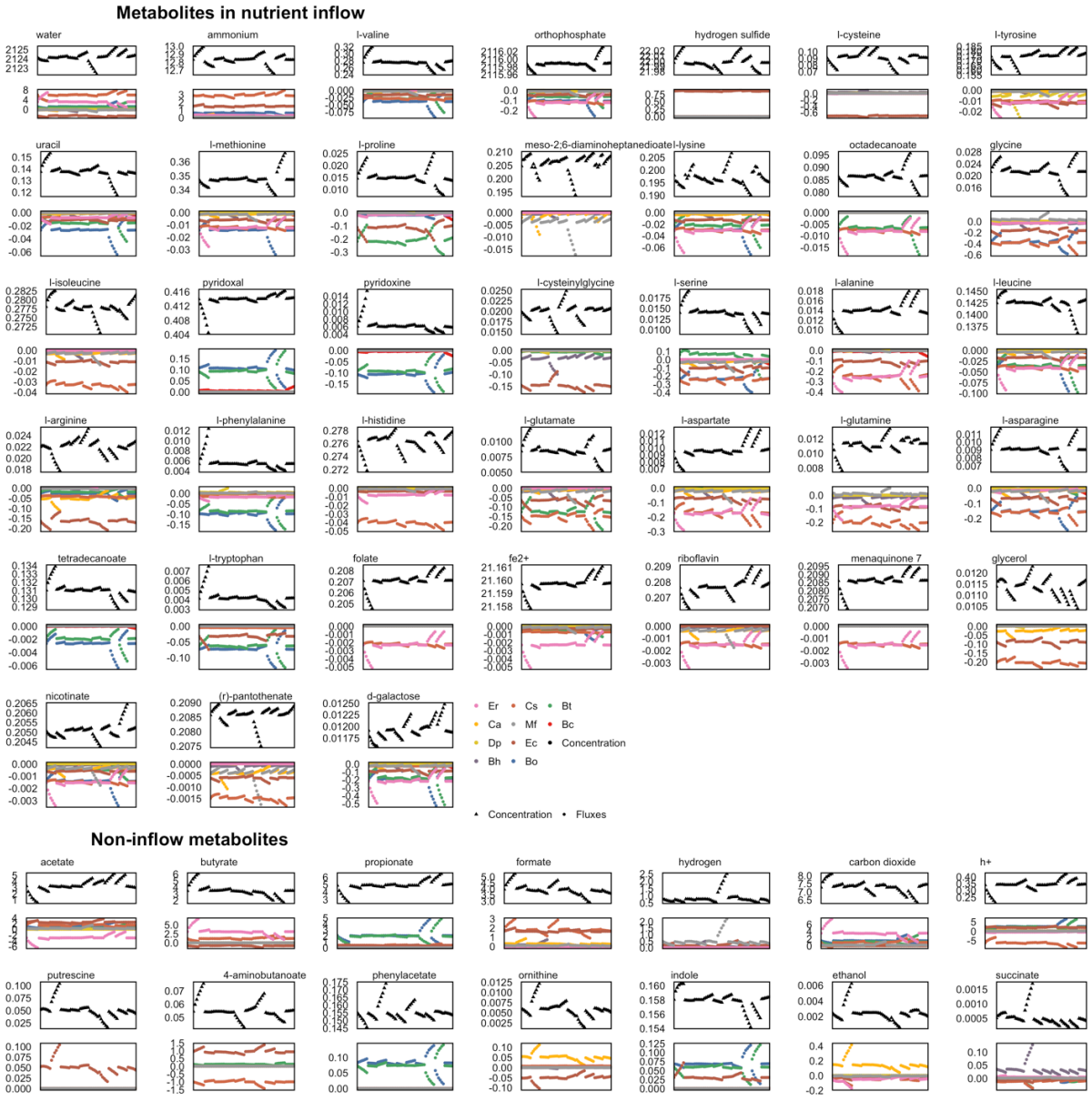

Figure 9.4. **Cumulative uptake and secretion fluxes for all species and all metabolites, across all 61 simulations.**

For all analyzed metabolites, an upper panel shows the total cumulative secretion or uptake of that metabolite by each species across all 61 simulation runs. A lower panel shows the corresponding environmental concentration at the final time point. Each plot shows fluxes for a single metabolite, with those found in the nutrient inflow in the upper section and microbially-produced metabolites below. Metabolites are ordered by their total variance. Simulations are ordered on the x-axis in the same ordering as in Figures 1 and 2.

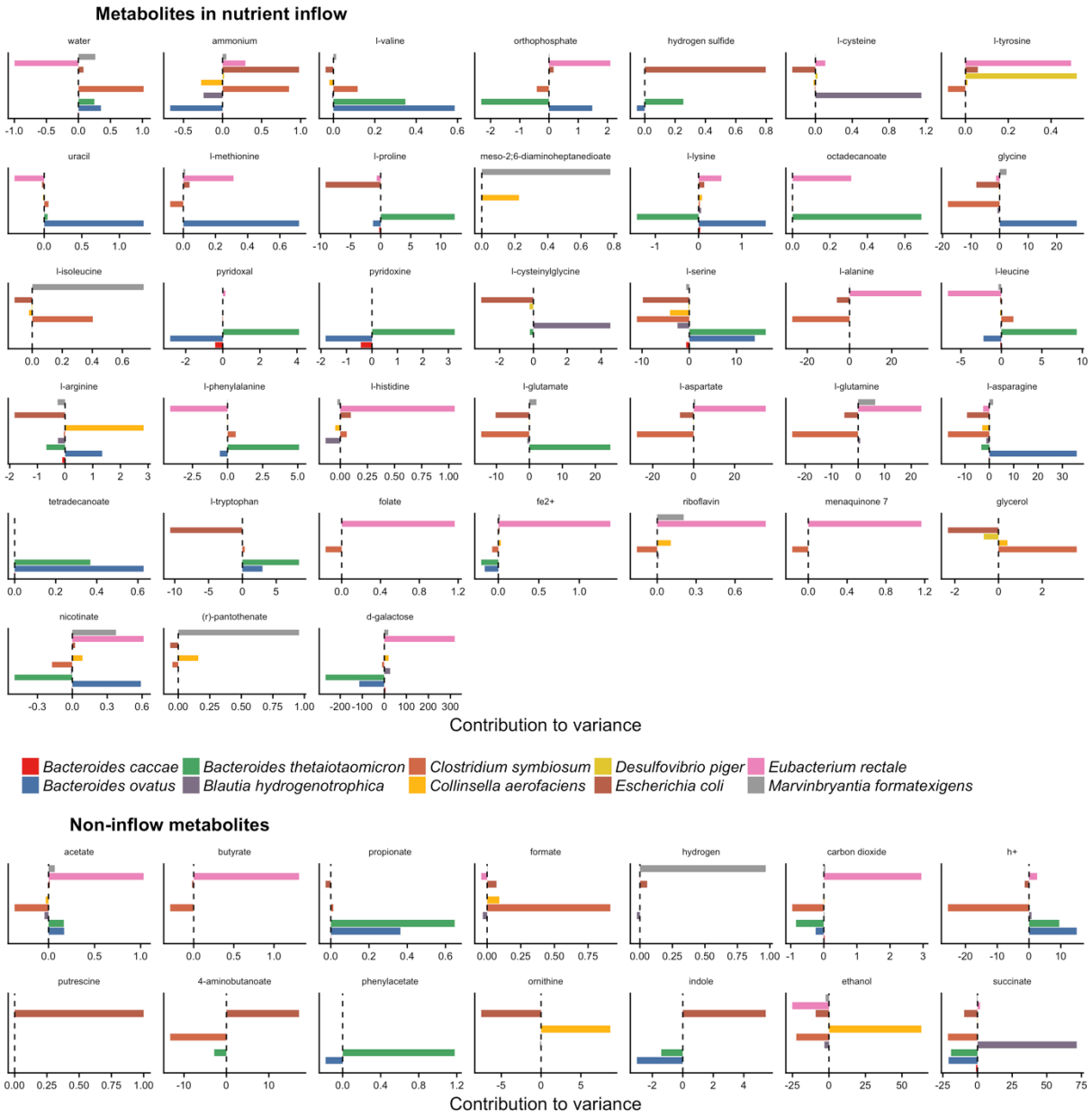

Figure 9.5. **Variance contribution profiles for all metabolites.**

Each plot shows contribution values for a single metabolite, with those found in the nutrient inflow in the upper section and microbially-produced metabolites below. Metabolites are ordered by their total variance. The relative contribution values, $\hat{c}_i$, are plotted on the x-axis.

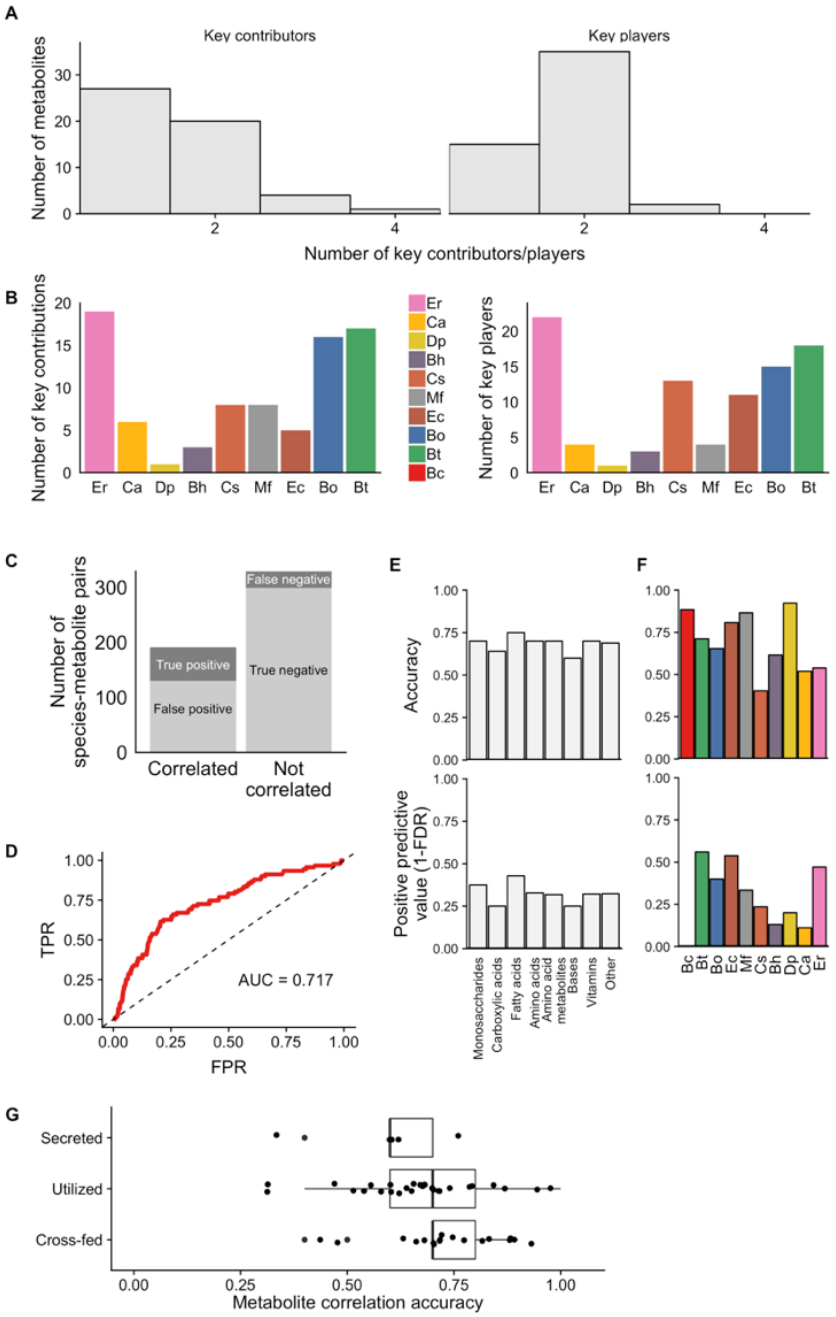

Figure 9.6. **Key contributors and key players driving metabolite variance have similar properties and correlation results.**

(A) Histograms of the number of key contributor and key player species for all 52 analyzed metabolites. **(B)** Number of key contributor and key player relationships for each species. Full species names can be found in Figure 2. **(C-G)** Correlation results for identification of key players, as shown in Figures 3 and 4 for key contributors. **(C)** The number of species-metabolites pairs that were significantly correlated (left bar) or not correlated (right bar) and its correspondence with true species-metabolite key players. **(D)** Receiver

operating characteristic (ROC) plot, showing the ability of absolute Spearman correlation values to classify key players among all species-metabolite pairs. **(E-F)** Accuracy and positive predictive value of Spearman correlation analysis for detecting true key players across metabolite classes (Panel E) and for each of the 10 species (Panel F). **(G)** As in Figure 4, correlation-based analysis detected key players equally accurately regardless of whether a metabolite is secreted, utilized, or cross-fed by the species. Each point represents the accuracy of correlation-based analysis for a single metabolite across its comparisons with all 10 species.

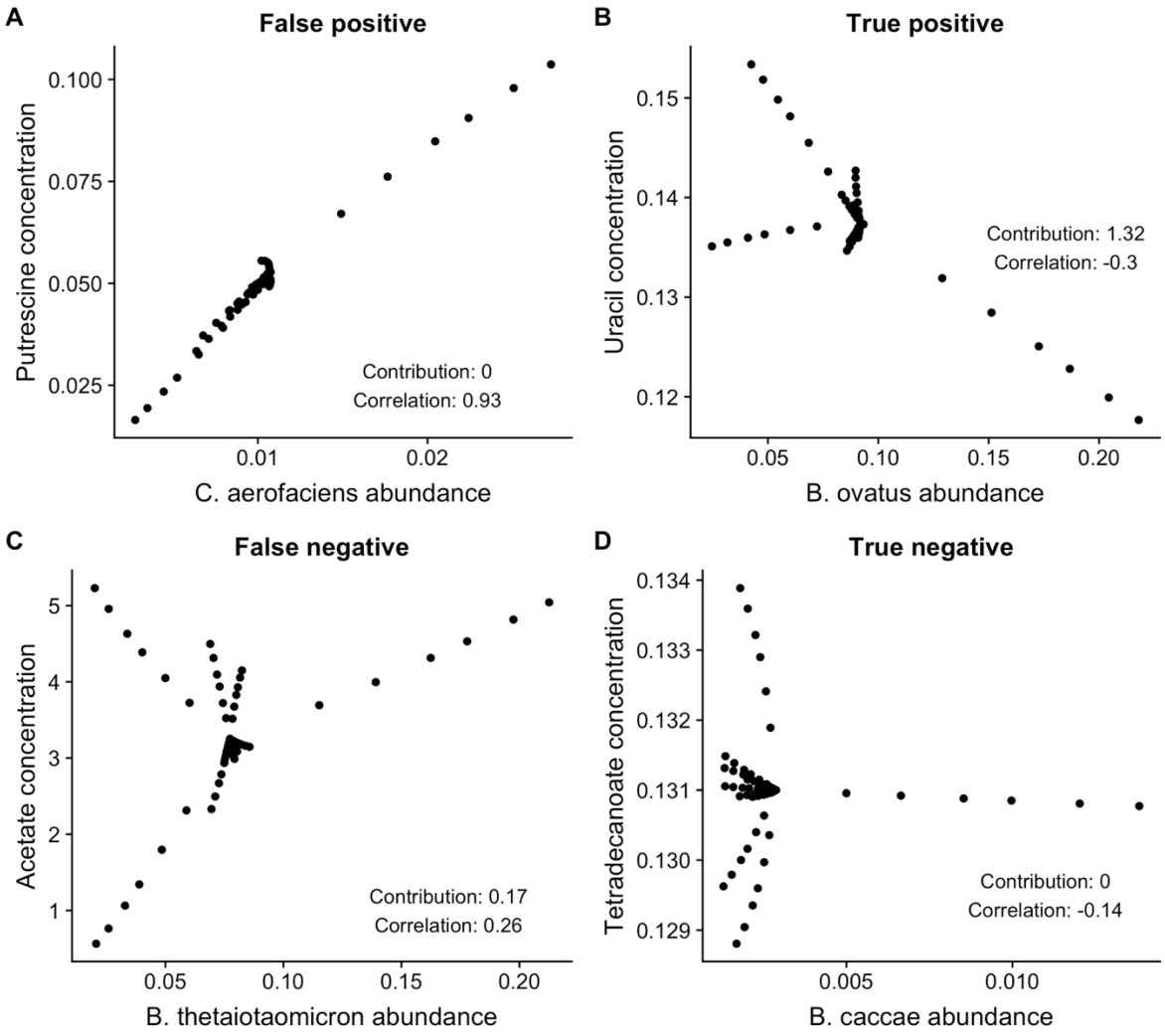

Figure 9.7. **Examples of species-metabolite correlation outcomes.**

Each panel plots the concentration of one of the example metabolites shown in Figure 2 against the abundance of a key contributor or non-contributor species, with annotations of the corresponding correlation and contribution values.

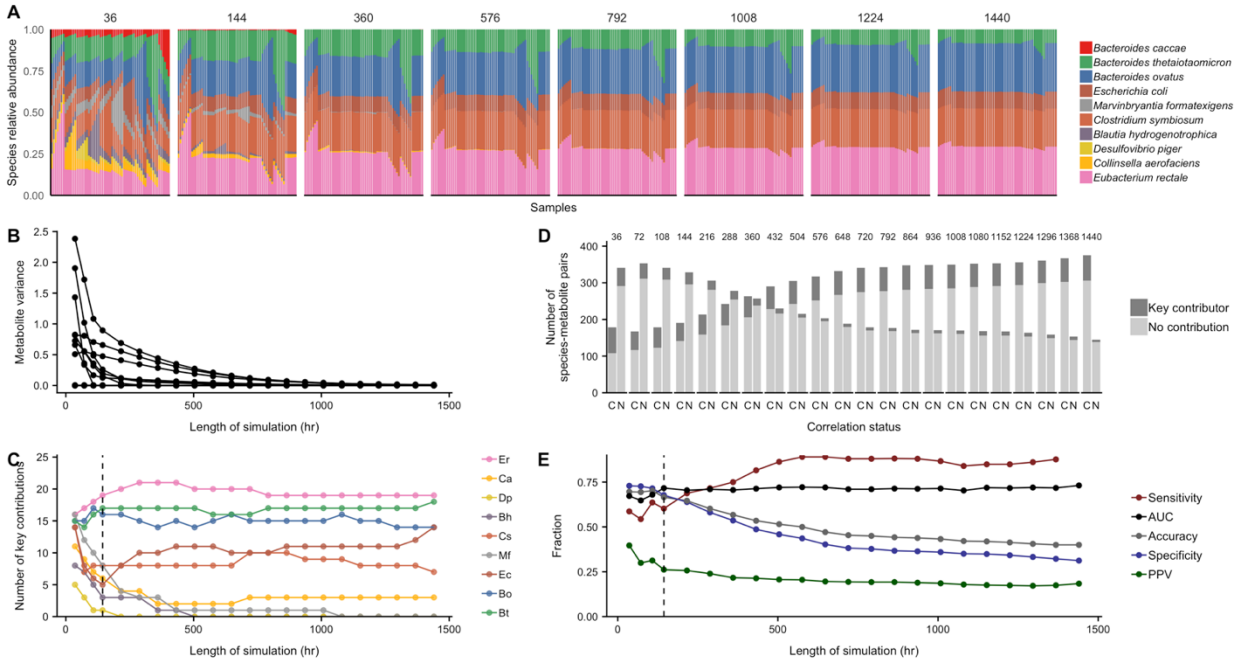

Figure 9.8. **Effects of simulation duration on contribution profiles and correlation efficacy.**

(A) Species abundances after simulation runs of increasing length. Longer simulations increasingly converge towards a consistent profile. **(B)** Metabolite variance decreases with increasing simulation duration. Each line represents the total variance in concentration of a metabolite across the 61 simulations. **(C)** Each line represents the number of key contributions by each species across simulation datasets of increasing duration. The total number of key contributors decreases with increasing length. The length of 144 hours described in the main results is indicated with a dotted line. **(D)** Bar plots of correlation and contribution outcomes with increasing simulation duration, with the "C" labeled bar indicating the number of correlated species-metabolite pairs and the "N" indicating the number of non-correlated pairs. Datasets generated from longer simulations display more significant correlations and fewer key contributors. **(E)** Overall shifts in prediction metrics for correlation analysis with increasing simulation duration. AUC and predictive value are largely constant.

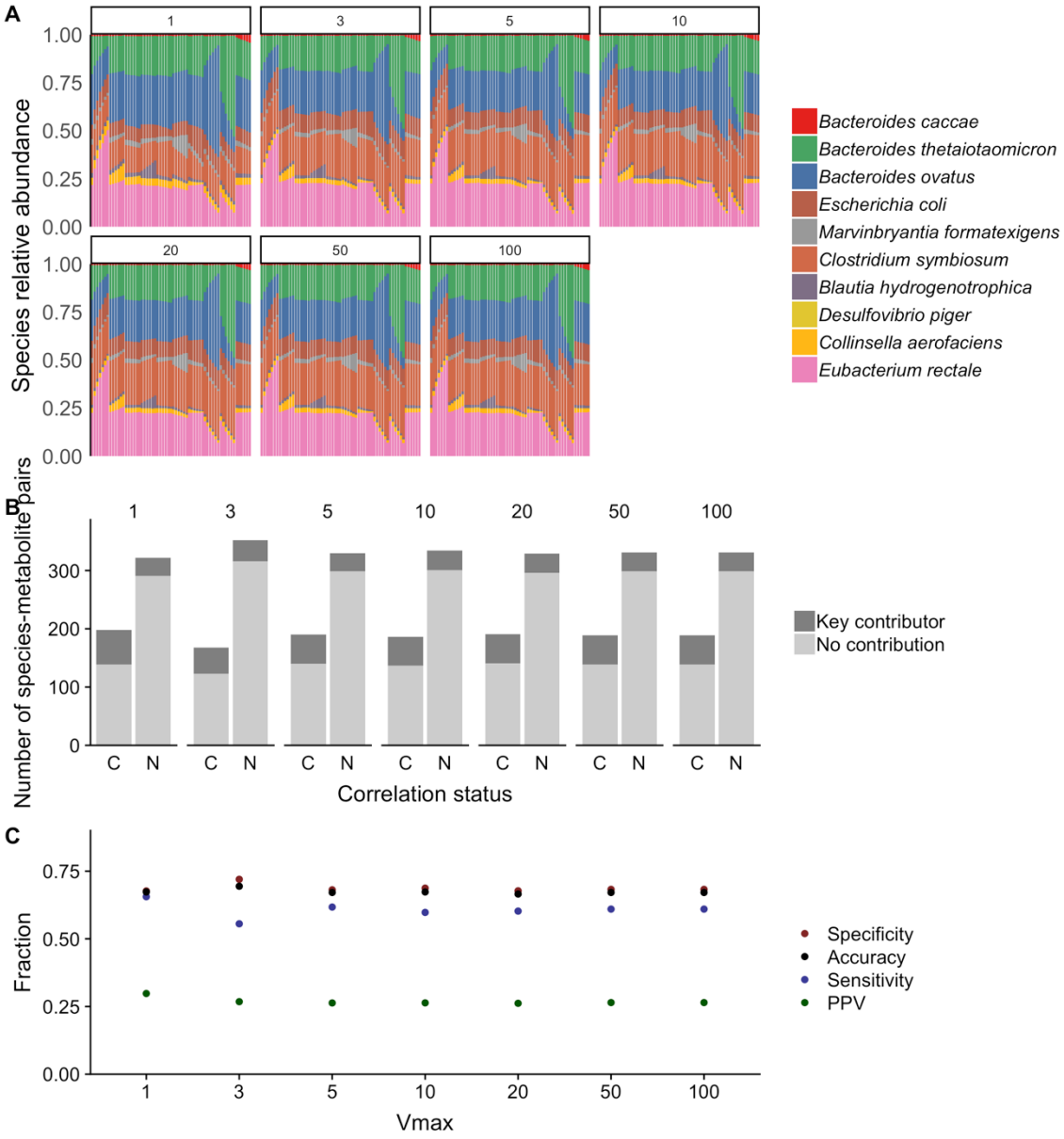

Figure 9.9. **Effects of *V_max* parameter on simulation and correlation results.** **(A)** Species compositions generated using different values of the parameter are nearly identical. **(B)** Bar plots of correlation and contribution outcomes from simulations with varying values of *V_max*, with the "C" labeled bar indicating the number of correlated species-metabolite pairs and the "N" indicating the number of non-correlated pairs. **(C)** Overall prediction metrics for correlation analysis are largely constant across simulations generated with different *V_max* values.

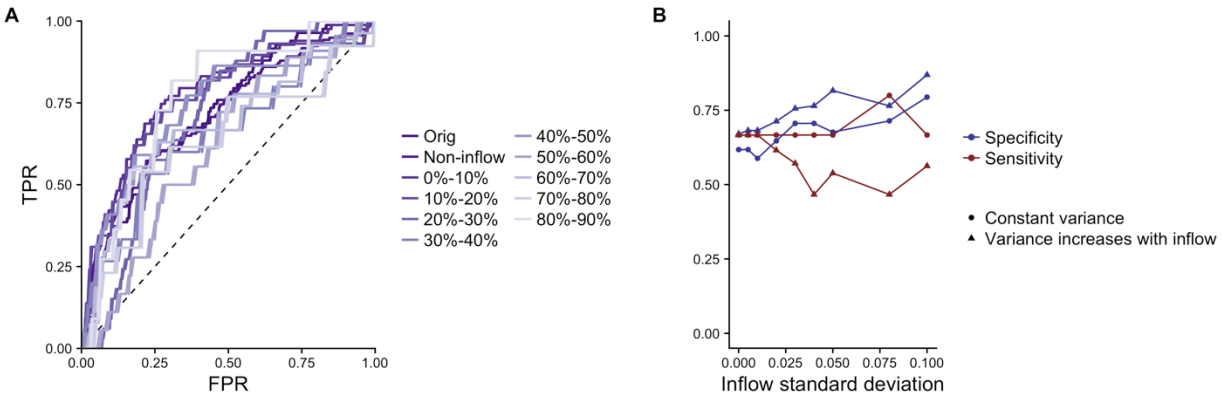

Figure 9.10. **Effects of environmental metabolite variation on correlation analysis. (A) Environmental fluctuations do not significantly affect overall correlation performance.** ROC curves are shown for sets of metabolites with increasing environmental contribution. None of the levels of environmental contribution had a significantly different area under the curve, based on 95% confidence intervals calculated using bootstrap resampling with 500 replicates. **(B) The sensitivity and specificity of correlation analysis to detect key microbial contributors to non-inflow metabolites are affected by variation in metabolic inflow.** Each point represents the specificity, sensitivity, or positive predictive value of the 14 analyzed non-inflow metabolites in a dataset of 61 simulations. The percent standard deviation (coefficient of variation) in inflow metabolite concentrations for each set of simulations is plotted on the *x*-axis.

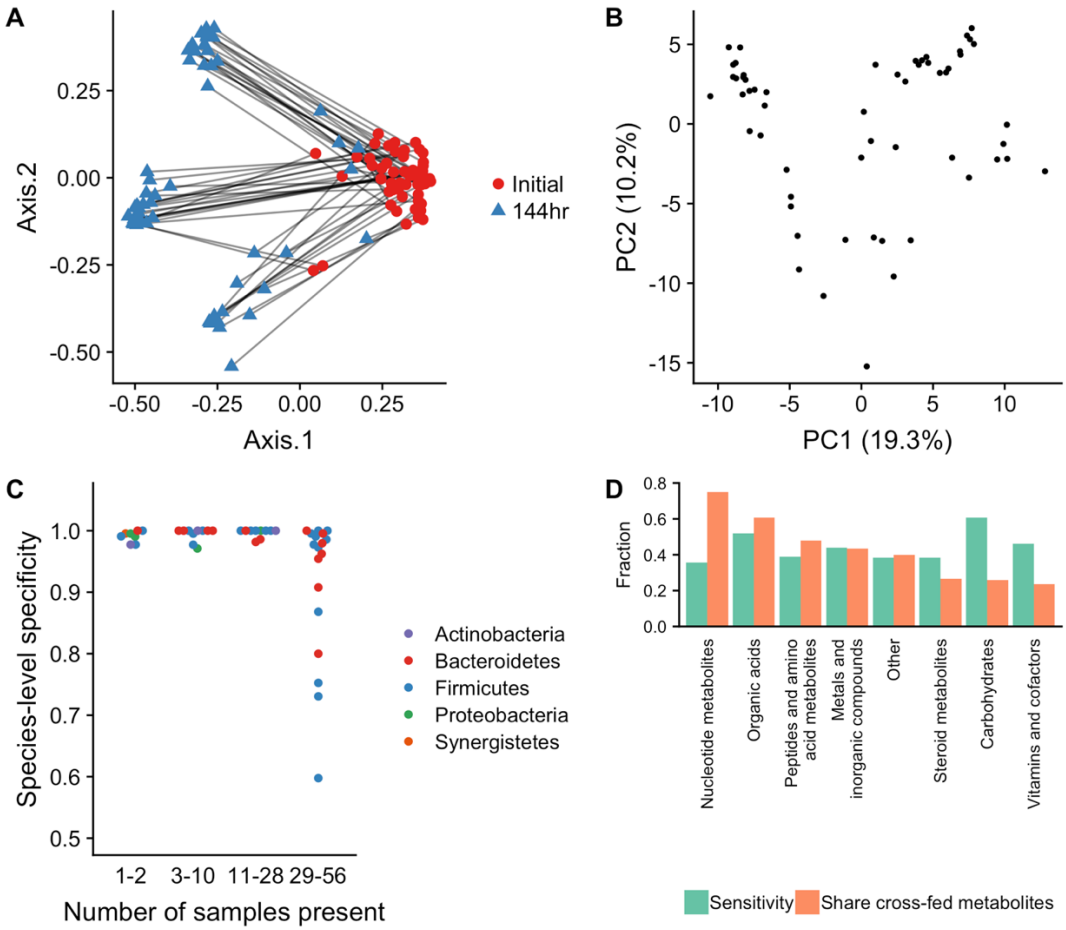

Figure 9.11. **Simulations of HMP-based microbiota.**

(A) Progression of HMP-based simulations. A principal coordinates analysis of the species compositions of the 57 HMP-based simulations at their initial and final time points, using the Bray-Curtis dissimilarity metric. Initial compositions tended to become dominated by a limited number of fast-growing species, leading to distinct subgroups. **(B) Metabolite variation across HMP-based simulations.** A principal component analysis of the metabolite concentration data at the final simulation time point. **(C) The specificity of species-metabolite correlation analysis is associated with species prevalence.** Each point represents a species with at least one key contribution to metabolite variation. The *x*-axis categorizes species into quartiles based on the number of samples in which they appear. Species that are present in a wider subset of the dataset have a higher rate of false positive correlations (lower specificity). **(D) The sensitivity of species-metabolite correlation analysis is related to metabolite class and cross-feeding status.** Green bars represent the overall sensitivity of identification of key contributor species-metabolite pairs within that category. Orange bars represent the share of metabolites in that category that are both synthesized and utilized by community members (cross-fed).

Chapter 10. APPENDIX D: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

10.1 SUPPLEMENTARY METHODS AND IMPLEMENTATION

10.1.1 *16S rRNA pre-processing and mapping*

NCBI Taxonomy IDs provided in Table S5 of (Magnúsdóttir et al., 2016) were used to download associated RNA genes for each model from NCBI RefSeq and GenBank, using the *biomartR* package in R (Drost and Paszkowski, 2017).

To establish a mapping between Greengenes OTUs and the AGORA genome collection, the representative set of sequences for Greengenes 13_8 99% OTUs were aligned optimally to the downloaded AGORA RNA database using *vsearch* 2.8.1 and returning the single best alignment for each OTU. The same alignment was performed for SILVA v132 99% OTUs to find their closest match in both the Greengenes database (to construct a KEGG-based model), and the AGORA RNA database (to construct an AGORA-based model).

10.1.2 *Metabolic network construction*

For KEGG-based metabolic models, MIMOSA2 uses a generic KEGG model template constructed using 3 files from the KEGG database, following (Noecker et al., 2016): *reaction_mapformula.lst* (streamlined set of core metabolic KOs, pathways, reactions, and metabolites), *reaction_ko.list* (KO-reaction links), and *reaction* (reaction annotations). The currently used version is from the February 2018 KEGG release.

To generate a set of reactions for each Greengenes 99% OTU, the generic KEGG model described above was merged with the list of KOs inferred to be present in each OTU according to

PICRUSt (Langille et al., 2013). The copy number for each reaction is also normalized according to the 16S rRNA copy number inferred by PICRUSt for each OTU.

To use the AGORA genome-scale metabolic reconstructions, we first downloaded the AGORA model collection version 1.0.2 from the Virtual Metabolic Human database (Noronha et al., 2018), including the set of 818 unconstrained models as well as the set constrained based on an average European diet. We converted the models to a format usable by MIMOSA2 using the *R.matlab* package. A pre-processed reaction file was generated from the stoichiometry matrix and bound constraints for each model. A normalized copy number was included for each reaction, which was assumed to be 1 divided by the number of 16S rRNA genes found in the relevant reference genome. Metabolite IDs were mapped to KEGG compounds using the mappings provided with each model.

10.1.3 *Reaction directionality inference algorithm*

When using a metabolic network model template involving many reversible reactions (such as the AGORA model collection), MIMOSA2 can optionally infer directionality of major reactions using a greedy algorithm. For each analyzed metabolite (ordered by coefficient of variation), it assesses whether model fit is greatly improved (25% increase in variance explained) by removing any single reaction direction in any single taxon. If so, the taxon-specific reaction with the largest improvement is removed from the network model and metabolic potential scores are recalculated. This process is repeated as long as the model continues to improve and a minimum number of reactions (3) are still present for each metabolite.

10.1.4 *Simulation data analysis*

To apply MIMOSA2, the original AGORA network was used as the basis for the community metabolic network model. This model was also used as the basis for running MIMOSA

version 1 and to filter significant correlations based on whether each species was capable of modifying each metabolite. All reactions were assumed to proceed in a forward direction. True microbial metabolites were identified as those for which the scaled contribution to variation from nutrient inflow fluxes was less than 90%. True contributors to metabolite variation were defined as any taxon with a scaled positive variance contribution greater than 10% (see section 4.3). Putative microbial metabolites were inferred by MIMOSA2 if the metabolic potential model explained greater than 10% of metabolite variation, and putative taxonomic contributors were those whose contributions to the model were greater than 10%. Precision, recall, and ROC curve statistics were calculated using the *pROC* package in R (Robin et al., 2011).

10.1.5 *MIMOSA2 analysis of Snijders et al. 2016 dataset*

This dataset was collected and processed using the same methods described in Chapter 3 and Appendix B. We limited the analysis to samples from mice on the autoclaved LabDiet Prolab 3500 (Diet 2). The input to the MIMOSA2 analysis was the closed-reference table of Greengenes OTU abundances produced by *vsearch*, and the metabolomics dataset of raw peak areas identified with KEGG compound IDs. The KEGG-based metabolic network template was used for the community metabolic model, with the option to refine reaction directionality when fitting the model. The same thresholds as for the simulation datasets were used to infer microbial metabolites and key contributors, except that taxonomic contributors were identified as those with a contribution greater than 5%.

VITA

Cecilia Noecker grew up in St. Paul, Minnesota, and attended St. Olaf College for her bachelor's degree, majoring in biology with a concentration in statistics. There, she performed research in both the statistics and biology departments and also completed a research project in mathematical modeling of HIV infection at the University of Tennessee-Knoxville. Upon graduating, she spent a year as a Fulbright fellow at the National Institute of Public Health in Cuernavaca, Mexico, studying associations between the severity of dengue virus infection and human genetic variation. In the PhD program in Genome Sciences at the University of Washington, advised by Elhanan Borenstein, she developed and applied tools for the analysis and interpretation of microbiome omics data. She is also an active contributor to efforts in science outreach, education, and advocacy for equity and inclusion, working with the Pacific Science Center, the Carpentries, and Women in Genome Sciences, among other organizations.