Summer 6-15-2022

Deciphering Taxa-function Relationships in Population-level Studies of Human Gut Microbiomes

Quang P. Nguyen
Dartmouth College, quang.p.nguyen.gr@dartmouth.edu

DECIPHERING TAXA-FUNCTION RELATIONSHIPS IN POPULATION-LEVEL STUDIES OF HUMAN GUT MICROBIOMES

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Quantitative Biomedical Sciences

by

Quang P. M. Nguyen

DARTMOUTH COLLEGE

Hanover, New Hampshire

May 2022

Examining Committee:

H. Robert Frost, Ph.D., Co-Chair

Anne G. Hoen, Ph.D., Co-Chair

Margaret R. Karagas, Ph.D.

Brock C. Christensen, Ph.D.

Levi Waldron, Ph.D.

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

The human gut microbiome is a complex and dynamic ecosystem, featuring a multitude of microbes all interacting with their hosts in an elaborate manner. Even though this exchange is often mediated through microbial metabolic and functional outputs, such as the production of certain metabolites, environmental exposures and host lifestyle are highly influential in shaping the presence of microbial species irrespective of their individual roles. As such, a comprehensive understanding of the microbiome requires researchers to examine the relationship between taxonomic abundance and function simultaneously. Assessing microbial contributions to important ecosystem services can enable identification of robust functions supported by a variety of species, or to identify important keystone taxa that are associated with a disease-causing biochemical pathway. The primary objective of this thesis is to assess different approaches for investigating the taxa-function relationship and evaluate its value in providing unique biological insights. First, we leveraged densely collected multi-omics data from the New Hampshire Birth Cohort Study to identify genus-metabolite pairs that are core to infant gut microbiomes. Second, we developed a novel statistical method that enable integrating taxa-function relationships in epidemiological studies. Third, we assessed microbial phenotypic traits as a potential source for defnining interpretable and human-centric microbiome functions.

Preface

This thesis would not have been written if it were not for my undergraduate advisor, Dr. Larissa Williams. Even though I had been studying science for while, I was not really introduced to the conduct and methodology of scientific research until I took the advanced molecular biology course at Bates College instructed by Dr. Williams. The methods-focused approach to biology showed me how much I enjoyed the act of doing science (despite some pretty unsavory grades in the standard biology courses). I came to understand that scientific research is an act of love driven by curiosity and excitement. There is nothing else as thrilling as being able to tackle a difficult question armed with a set of tools you know by heart. Dr. Williams' mentorship was the major reason why I applied to graduate school in the first place, which was the first step towards completion of this work.

I would be remiss to not thank my graduate advisors, Dr. Anne Hoen and Dr. H. Robert Frost, without whom I would have been completely lost. I am amazed at how you are able to tolerate me even after three mental breakdowns (one for each manuscript, like clockwork). I am eternally grateful for your mentorship and guidance. Your patience and dedication to independent research has allowed me to confidently explore a diverse array topics, even if the end result is an eclectic collection of papers that is more akin to looking at a random cabinet at an antique store than anything found at Williams Sonoma.

This thesis was also supported by the wonderful members of the Hoen and Frost

labs. Thank you to Dr. Weston Viles and Dr. Jie Zhou for providing expert statistical help whenever I need them. I am thankful for Dr. Yasmin Kamal, Dr. Modupe Coker, and Courtney Schiebout for all the feedback and enthusiasm on my work at lab meetings. Finally, and most importantly, I would not have survived without Erika Dade's steadfast and expert grasp on all things data, and my lab mate Dr. Rebecca Lebeaux's insightful recommendations. I really appreciate Becky taking the time out of her day to bounce ideas, collaborate, and share our excitement (and the occasional apprehension) for microbiome research.

I also want to thank my committee members Dr. Margaret Karagas and Dr. Brock Christensen for all of their support and guidance for the past years, who have never said no to anything I asked (even if they are last minute requests). I am also appreciative of Dr. Levi Waldron for agreeing to serve on my dissertation committee.

I am thankful for all the assistance from researchers within and oustide of Dartmouth. Special thanks to members of the Dartmouth microbiome research group Dr. Juliette Madan, Dr. Modupe Coker, Dr. Hannah Laue, Dr. Thomas Palys, and Yuka Moroishi; as well as my collaborators Dr. Susan Sumner, Dr. Susan McRitche, and Dr. Wimal Pathmasiri from the UNC Nutrition Institute, and Dr. Hilary Morrison from the Marine Biological Laboratory. I am thankful for all the staff and participants of the New Hampshire Birth Cohort Study for their efforts.

I would like to acknowledge the funding sources that have made this research possible. My work is supported by grants from the NIH (grants R21CA253408, P20GM130454, and R01LM012723) and from the EPA (grants EPA RD-83544201).

So much of this thesis is also owed to the wonderful QBS community. I want to give thanks to Dr. Diane Gilbert-Diamond, Dr. James O'Malley, Dr. Krissy Griffin, Rosemary White, Shaniqua Jones, and so many others who have helped create a supportive, fun, and wonderful community. Special love for the 62 Sachem

crew, Yuka Moroishi and Diana Song (and Keke Song) for 3 years of wonderful co-habitating. Always grateful for my life-long friends Catherine Pollack, Becky Lebeaux and Despina Karalis. I also cannot forget my co-workers at the COVID Tracking Project: Kara Schechtman, Michal Mart, Nicki Camberg, Brandon Park, and Rebecca Glassman, all of whom gave me an amazing community at the darkest of times.

My final dedications are for my family. I would not have been here at this point without my parents, my aunts, and my cousins. They have been incredibly supportive, and are my biggest cheerleaders. I would also like to thank the Ziegler family, who have allowed me to eat all the soda bread in their house for a whole year. Lastly, my most heartfelt thanks to my partner, Katie Ziegler, for all her support in the past years. I constantly annoy her by asking about the difference between "genera" and "genus" (and still getting it wrong), yet she has shown nothing but patience and love. I do not have enough words to express how much you mean to me, and how much your presence have given me strength to move forward.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Section 1.1

The human gut microbiome and population health

1.1.1. Overview of the human gut microbiome

The human microbiome is the collection of microorganisims (which includes bacteria, protozoa, archaea, fungi, viruses, and their genes) that participates in a symbiotic co-existence with their hosts [289]. It is difficult to study human microbiomes outside the host due to difficulties in being able to culture the majority of organisms [298]. However, advances in sequencing technologies have allowed researchers to glimpse the inner workings of these complex communities via their genomes [345]. Researchers found that different body sites harbor unique environmental determinants that give rise to distinct groups of microbes [59]. For example, in the oral microbiome, oral surfaces have different surface receptors [102], promoting only microbes with specific adhesins that are complementary [1]. This results in differences such as *Streptococcus mitis* bv.2 species being represented in the tongue dorsum but not even detected in the related lateral tongue surface [1]. Even though the degree of diversity across body sites extend across the entire tree of life, most studies so-far have focused on profiling bacteria, which exists in high-abundance and is relatively easy to profile [41].

The gut environment specifically promotes a consortium of microbes that varies across the digestive track [185, 70]. Most microbiome research has been focused on the colon via fecal samples [278], where it has been estimated to contain the highest microbial density recorded in any habitat [267]. The gut microbiome is initially acquired at birth via maternal transfer [266, 15, 333]. The community matures over time, increasing in diversity and reaching an adult-like state in around 2-3 years of age, where it is characterized mainly by members of the Firmicutes, Actinobacteria, and Bacteroidetes phyla, with *Bacteroides*, *Faecalibacterium*, and *Bifidobacterium* as the most abundant genera [142, 200]. Many of the species identified to be in the gut cannot be found in other habitats, suggesting a strong co-evolutionary relationship with human hosts [164].

Various environmental factors can shape the composition of the gut microbiome. In early life, the mode of delivery and breastfeeding status are significant modifiers of composition. Infants who were born via vaginal delivery have increased abundances of *Bacteroides*, *Pectobacterium*, and *Bifidobacterium* genera, while those born via Cesarian section have decreased diversity and higher propensity to be colonized by *Staphylococcus* and members of the *Clostridum* cluster [183, 141, 274]. Breastfeeding is associated with lower levels of *Escherichia coli*, *Tyzzerella nexilis*, and *Roseburia intestinalis* while on the other hand promoting the coloization of various *Bifidobacterium* species such as *B. breve* and *B. dentium* which harborsspecific genes that help in digesting complex oligosaccharides [274, 291]. Among adults, diet is of particular interest. Studies have shown that the Western diet, high in saturated and trans fats while low in mono and polyunsaturated fats, is associated with decreased abundance in *Bifidobacterium*, *Eubacterium*, and *Lactobacillus* genera [325]. Other extrinsic and intrinsic factors such as smoking status [27] and alcohol consumption [74] also contribute to microbiome modulation (reviewed in [263]). This demonstrates that the microbiome is likely to be highly variable in composition as it is sensitive to changes in host physiology.

1.1.2. Outcomes associated with changes in gut microbiome composition

Observed shifts in gut microbiome composition are often associated with adverse health outcomes. As such, there is a great interest in epidemiological applications, where the microbiome can be identified as either a marker or as a causal agent in human disease [90]. Observational studies have linked changes in the gut microflora to various metabolic and infectious diseases. This is because host inflammation responses can be linked to the microbiome's role in mediating immune function (reviewed in [302]). Short-chain fatty acids (SCFAs), such as acetate, propionate, and butyrate, are metabolic products of microbiota digestion from dietary fiber and resistant starches. These compounds bind to G-protein-coupled receptor 43 expressed in immune cells, an effect that allows for immune responses to resolve post-infection, thereby preventing persistent inflammation. As such, the gut microbiome is found to be associated with inflammation-related diseases such as colorectal cancer [49, 336] and inflammatory bowel disease [101, 93, 174]. There are many other conditions (reviewed in [52]) that can be linked to changes in the gut microbiome, such as *Clostridium difficile* infection [306] and obesity [288, 9].

Additionally, the gut microbiome is also linked to other conditions that are not localized in the intestinal tract. The gut-brain axis refers to how residential gut microbes are involved in regulating host cognition, mood, and behavior [207]. There are established links, for example, between the gut microbiome and neurological conditions such as autisum spectrum disorder (ASD), where individuals with ASD have different gut microbial signatures (an increase in several mucosa-associated Clostridiales) compared to neurotypical controls [179]. Another instance of the gut microbiome's overarching impact on human physiology is the gut-lung axis [83], where the crosstalk between gut and lung communities is linked to chronic and acute respiratory conditions. For example, a study from the Canadian Healthy Infant Longitudinal Development cohort identified decreases in *Lachnospira*, *Veillonella*, *Faecalibacterium*, and *Rothia* genera among children at risk of asthma attacks [11].

1.1.3. Challenges in determining potential microbial biomarkers

Despite the wealth of microbiome association studies, there still exist considerable challenges in identifying consistent microbial markers that have meaningful associations with health outcomes [76]. In a meta-analysis of 10 studies for inflammatory bowel disease and obesity [300], no individual microbe was consistently associated with subject case status. Even though false discoveries are expected since population-level studies are indeed only exploratory and hypothesis generating, the practice of interpreting differentially abundant taxa lists using post-hoc literature searches makes results unreliable and biased. As a result, validating these markers proves to be arduous as it is impossible to disentangle which hypotheses are more contextually probable to follow-up in expensive *in vitro* or *in vivo* laboratory experiments.

Statistical and computational difficulties remain one of the major hurdles towards this replication issue [165, 169]. A mainstay of research in the field is the usage of high-throughput sequencing technologies to profile microbial communities taxonomically and genomically. Various steps within the sample processing protocol such as storage, DNA extraction, and library preparation, can contribute to differences in results between studies [55]. Contamination is also a big issue [65], especially in environments where there is low total microbial load [81], thereby producing erroneous results. Additionally, bioinformatic analyses and use of different databases may also cause divergences in observations [206]. For example, when 16S rRNA gene amplicon sequencing is used for taxonomic profiling, choosing to cluster amplicon sequence variants (ASVs) to operational taxonomic units (OTUs) using a sequence similarity threshold can yield radically different observed community profiles [50, 206].

In addition to issues pertaining to converting raw sequence data to named microbes, statistical analyses used to identify relevant candidates also suffer from inconsistencies. Microbiome taxonomic data, like other sequencing data sets, is compositional [104, 249] and constrained by the total number of reads (or library size). This constraint induces spurious negative correlations between variables, whereby changes in true absolute counts might not be accurately reflected at the observed relative scale [172, 212]. Microbiome data is sparse, containing a mixture of both structural zeroes (true absence of a taxon), and technical zeroes (abundance below the limit of detection) [136, 271]. This makes it difficult to distinguish between low-abundance and absent taxa, especially when studies differ considerably in sequencing quality and depth. Finally, microbiome data is also high-dimensional, where a typical data set contains from hundreds of species to thousands of sequence variants resulting in a high multiple testing burden. All of the challenges above contribute to methods that have to make difficult trade-offs between power, type I error control, and effect size estimations. As a consequence, researchers are faced with a complex landscape of available methods, which have been shown to produce different results on the same data sets [219].

In addition to technical challenges, the gut microbiome itself is also dynamic with a great degree of intra-individual variability. Even though the adult gut microbiome composition is stable over longer time scales [59], studies have also shown that shifts in composition occur on a day-to-day basis, impacted by host diet and lifestyle [63, 64]. However, the magnitude of shifts are small when compared to between host variability.

In the original human microbiome project (HMP) paper, researchers estimated that for signature genera that are supposed to be habitat-specific, their presence in samples collected from respective environments can be as low as 17% (highest 84%) [59]. In another study of gut microbiomes from a Chinese cohort of healthy individuals (N = 120), around 90 (±16) species-level taxa (assayed using full length 16S rRNA gene sequencing) were shared between individuals from the total of 1,235 identified species [332]. This effect is consistent even when looking at the strain level [175]. For context, humans share around 99.5% of their genomes [125]. The degree of microbiome personalization is significant enough to warrant initial exploration of forensic applications [89].

Section 1.2 Approaching microbiome research from a mechanistic perspective

1.2.1. The functional microbiome

The end goal of medical and epidemiological research on the gut microbiome is to ascertain why selected microbes exist in a given environment, and how they can affect human physiology to cause or mediate disease. The questions of "why" and "how" usually underline discussions surrounding microbial function [144], the answers to which can allow for the design of therapeutics and interventions [75]. As such, researchers are interested in moving beyond looking at the microbiome from a purely taxonomic perspective to a functional one [116].

One of the most important and well-known roles of the gut microbiome involves the fermentation of foods into metabolites that can be absorbed by its host. Studies have shown that the microbiome is involved in the digestion of all three macronutrient sources from the human diet: carbohydrates, lipids, and proteins (reviewed in [228]).

For carbohydrates, even though the human gut can hydrolyze and absorb certain sugars such as glucose, fructose, and lactose in the proximal gastrointestinal (GI) tract, the complexity of bonds that exist between monosaccharide units in dietary starches (especially plant polysaccharides) means that the majority of carbohydrates pass through to the distal GI where it is digested by very well-equipped microbes [319]. The species *Bacteroides thetaiotaomicron* alone has 260 glycoside hydrolases in its genome [329], well beyond that of native human enzymes. The by-product of carbohydrate fermentation are SCFAs (most notably acetate, proprionate, and butyrate) [182], which are compounds that not only have inherent nutritional value but also play a role in maintaining the gut epithelial barrier. Even though proteins and fats are not as central to microbiome function as carbohydrates, certain important compounds have been shown to be linked to gut communities (reviewed in [208]). In a study by Backhed et al., germ-free mice were fed significantly more chow compared to wild-type controls yet had 42% less body fat [14]. Surprisingly, when a single species from the microbiome of wild-type mice was transplanted (also *Bacteroides thetaiotaomicron*), these formerly germ-free mice were able to partially recover their capacity to produce body fat, thereby suggesting a relationship between gut microbes and the process of adiposity. This provides further evidence for a link between the gut microbiome and obesity [288].

The microbiome is also implicated with immune programming. Even though the specific signalling pathways are still under active research, this process is mediated through commensal secreted metabolites or surface-associated antigens that interact directly with host immune cells (reviewed in [21]). For example, murine studies have shown that the innate immune receptor Toll-like receptor 5 (TLR5) selects for certain microbes during the neonatal period by serving as a sensor for bacterial flagellin [96]. During early life, in order to accommodate the initial colonization process, host immune response is limited and certain types of cell activity is suppressed. Microbes can directly induce this by secreting molecules such as sphingolipids that inhibit the induction of invariant natural killer T cells (iNKT) [8]. Even though this process results in an increased propensity of being infected, it also suppresses the host's inflammatory response, preventing excessive behaviors that might cause harmful outcomes such as necrotizing enterocolitis [221]. This crosstalk continues throughout life, where the microbiome participates in a co-operative relationship that helps maintain intestinal homeostatsis (reviewed in [341]).

Overall, the gut microbiome participates in extensive biochemical pathways that help maintain homeostasis and promote human health. The list of potential targets grows as more studies demonstrate new and exciting host-microbe intractions. It is clear that profiling microbial function can help get at the questions of "why" certain microbes exist, as well as "how" they can lead to positive or negative health outcomes.

1.2.2. Approaches to characterize function in epidemiological studies

Advances in high-throughput molecular technologies have allowed researchers to comprehensively profile different functional components of the microbiome [90]. Although powerful, each meta'omic method faces different challenges in accomplishing their intended goals, such as complex sampling preparation strategies or limited resolution and annotation. Here, we give a short description of major microbiome profiling technologies currently in use and the types of biological insight they can provide.

Metagenomics. Metagenomics refers to untargeted DNA sequencing of the entire gene content of a sample via "shotgun" shearing of fragments [248]. In terms of taxonomic profiling, metagenomic approaches can provide species to strain level resolution [286], as well as being able to detect non-bacterial organisms such as archaea and viruses. In terms of function, researchers can estimate the abundance of certain gene families in the entire community, which can be used downstream to infer entire pathways and even predict structural variants [140]. Finally, assembly-based approaches can categorize sequence fragments into putative genomes, thereby enabling the discovery of novel strains and genes [239]. However, since this is a DNA-based approach, gene family copy numbers only represent functional potential rather than true outputs. This is further complicated by the fact that it is challenging to discern which bacteria are "alive" since DNA is a stable molecule [248]. As such, inferences about microbiome function drawn from metaegenomic data sets are limited, and difficult to trace back to specific microbes. Finally, databases are woefully biased towards easily identifiable strains, whereby most genes are defined as unmapped, making annotating specific functions difficult.

Metatranscriptomics and Metaproteomics. Metatranscriptomics and metaproteopmics refer to profiling the entire transcript (i.e. RNA) and protein content of a sample [91]. Both techniques measure downstream products of gene abundances, therefore they are better representations of the total amount of functional information that is "active" within a community. Metatranscriptomics involve nucleic acid sequencing similar to metagenomics (in fact, with proper tagging, one can simultaneously sequence both the transcriptional and genomic content of a sample), while metaproteomics requires separation and quantification using a combination of liquid chromatography (LC) and mass spectrometry (MS). Both metatranscriptomics and metaproteomics can be mapped back to sequences obtained from original metagenomics results via sequence translation. This allows for powerful multi-omics approaches that can identify how a function reservoir is activated and expressed downstream. For example, in a paired metagenomic-metatranscriptomic analysis of healthy gut microbiomes, researchers found that even though there is a reservoir of genes coding for biosynthesis of amino acids, low transcriptional activity suggests that this function is under-expressed [92]. This is consistent with the fact that these procedures are energetically unfavorable [228]. Despite such benefits, there are various challenges in the technical aspect of these profiling approaches. For metatranscriptomics, not only are mRNAs highly unstable, they are dwarfed in abundance by rRNA in the total RNA pool, requiring specialized techniques to enrich for them while also removing human contaminants. For metaproteomics, the process of protein purification and extraction is demanding due to the complexity of the environment, requiring more biomass as well as special sample preparations to reach the degree of depth obtained by nucleic acid sequencing technologies [160, 261, 293].

Metabolomics. Metabolomics refers to the direct quantification of metabolites and other small molecules. It is different from metatranscriptomics and metaproteomics in the fact that there is no convenient map to sequence information [91], making direct integration with taxonomic or gene abundances from metagenomic sequencing more challenging. However, metabolomics reflects the layer of microbiome function that is closest to the host-microbiome interface, as measured metabolites interact directly with host receptors or participate in collaborative metabolic pathways [277]. Integrative multi-omics data sets featuring metabolomics have shown important links between microbes, their metabolic outputs, and human disease. For example, a study identified that changes in the microbiome have been implicated in the production of trimethylamine N-oxide (TMAO), a compound associated with cardiovascular disease, from l-carnitine (commonly found in red meat) [303]. One large benefit of metabolomics is that its sample preparation requirements are not as demanding as that of metaproteomics [93, 293]. However, each molecule has different properties and chemical structures, which means that some compounds are easier to measure than others, creating biases in which type of features gets measured [277]. Finally, metabolomic profiles are highly variable and sensitive to perturbations such as food consumption prior to measurement [118]. As such, it is suggested that metabolite flux might be a more meaningful measure of microbiome function rather than cross-sectional measures of concentration.

1.2.3. Obstacles in a function-based approach

A common misconception is that microbiome functional profiles are relatively more stable across individuals compared to their taxonomic counterparts [59]. This means that analyses focusing on community-wide functional 'omics analyses might produce more consistent and validatable results. However, the degree of comparative stability is difficult to ascertain due to differences in the scale of comparison across taxonomy and function [156]. In fact, when gene families are considered instead of pathways, the degree of stability decreased significantly [124]. Additionally, empirical studies have also shown that using pathway abundances are not significantly better at classifying patients disease status [330].

Two major challenges exist for function-driven microbiome analyses (reviewed in [116]). First, the large number of available molecular technologies mean that there is considerable choice in what constitutes as "function" in a certain context. In other words, depending on the research question, researchers have to make a decision on the analytical unit of microbial function, be it gene family abundance, pathway presence-absence, or concentrations of groups of metabolites. For example, in an analysis of strain-specific functional adaptation of the infant gut microbiome from the DIABIMMUNE cohort [291], the authors were interested in microbial capacity to digest human milk oligosaccharides (HMO) as a core ecosystem function. However, HMO metabolism is not encoded as a single pathway in frequently used databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) or Gene Ontology (GO), but rather as a group of 30 genes identified via literature review. As such, defining functions based on the research question of interest can provide meaningful interpretations on the services that microbes confer. Second, there is a considerable amount of "microbial dark matter" that hampers the characterization of functional dynamics [131]. For example, in a study involving metatranscriptomic profiling, around 9% of differentially abundant transcripts had unknown function [115]. Results can be misleading if these limitations are not acknowledged or improved upon.

Section 1.3

An integrative approach to incorporate both structure and function

1.3.1. Importance of taxa-function relationships

A holistic understanding of microbiome-related processes requires a conception of the relationship between taxonomic compositions and their functional profiles, termed the taxa-function relationship [156, 116]. Unfortunately, limited numbers of studies have explored taxonomic drivers of functional shifts beyond anecdotal searches in the literature [188], where it is more common to study them independently. This gap is problematic because even though drivers of the microbiome's impact on host health are its functional outputs, modulation can only occur at the taxonomic level. Niche differentiation driven by abiotic factors, such as the availability of nutrients, shape community assembly [237]. As such, any attempt to design restorative interventions or understand environmental perturbations requires the ability to pinpoint exact groups of taxon relevant to the functional processes of interest [318].

The taxa-function relationship is also integral to the complex ecological processes that exist within the microbiome. The plasticity of certain functions to external perturbations can be attributed to redundancies in the number of contributing strains [299, 215]. This idea of "robustness" [84] can be used as a proxy to diagnose community-wide or function-specific dysbiosis [295]. This also helps explain the underlying mechanisms as to why healthy microbiomes are found to be generally more “diverse”, fitting with the prevailing ecological theory linking biodiversity and ecosystem functioning in microbial systems [284]. On the other hand, if a core function is only contributed by a single taxon, targeted therapies can be designed to improve growth conditions or to provide the necessary probiotic supplements.

Finally, taxa-function relationships can also enhance quality of existing population-level studies. Providing the relevant context that can increase interpretability, while also helping researchers distinguish possible false positives when considering targets for validation.

1.3.2. Challenges and opportunities

Many studies have attempted to systematically tackle taxon-function integration in microbiome studies [188, 295, 84, 226]. However, they face challenges regarding assumptions, limitations in resolution, and biases in reference databases. For example, FishTaco [188], a tool to estimate species contributions to functional shifts, assumes that the functional content of contributing microbes is consistent at the species level, ignoring strain-level variation. Other ecological processes such as inter-taxa interactions and horizontal gene transfer can also radically change the taxa-function landscape, whereby removals or additions of certain contributing species might affect other seemingly unrelated members. Additionally, there are difficulties in defining relevant functions. Even though pathway annotations from reference databases such as KEGG are informative, it is still difficult to interpret long lists of pathways identified as differentially abundant.

However, despite such drawbacks, there are numerous opportunities. Utilizing different multi-omic approaches can provide additional insight into “active” functions whose products are persistent in the gut environment [130]. This is particularly important considering that microbes might harbor genes but can choose not to express them depending on environmental or ecological conditions [199]. Defining functions in terms of ecosystem roles such as traits [308] can allow for more interpretable results that are relevant to the condition of interest. The goal would be to look for context specific functions such as in the aforementioned study by Vatanen et al. [291] where multiple related gene clusters are simultaneously evaluated.

Chapter 2

Associations between the gut microbiome and metabolome in early life

This work was published on August 28th, 2021 as a *Research Article* at *BMC Microbiology.* The final article can be found here:

Nguyen, Q.P., Karagas, M.R., Madan, J.C., Dade, E., Palys, T.J., Morrison, G.H., Pathmasiri, W.W., McRitche, S., Sumner, S.J., Frost, H.R., Hoen, A.G. Associations between the gut microbiome and metabolome in early life. BMC Microbiol 21, 238 (2021). https://doi.org/10.1186/s12866-021-02282-3

Section 2.1 Abstract

Background: The infant intestinal microbiome plays an important role in metabolism and immune development with impacts on lifelong health. The linkage between the taxonomic composition of the microbiome and its metabolic phenotype is undefined and complicated by redundancies in the taxon-function relationship within microbial communities. To inform a more mechanistic understanding of the relationship between the microbiome and health, we performed an integrative statistical and machine learning-based analysis of microbe taxonomic structure and metabolic function in order to characterize the taxa-function relationship in early life.

Results: Stool samples collected from infants enrolled in the New Hampshire Birth Cohort Study (NHBCS) at approximately 6-weeks (N = 158) and 12-months (N = 282) of age were profiled using targeted and untargeted nuclear magnetic resonance (NMR) spectroscopy as well as DNA sequencing of the V4-V5 hypervariable region from the bacterial 16S rRNA gene. There was significant inter-omic concordance based on Procrustes analysis (6 weeks: p = 0.056; 12 months: p = 0.001), however this association was no longer significant when accounting for phylogenetic relationships using generalized UniFrac distance metric (6 weeks: p = 0.376; 12 months: p = 0.069). Sparse canonical correlation analysis showed significant correlation, as well as identifying sets of microbe/metabolites driving microbiome-metabolome relatedness. Performance of machine learning models varied across different metabolites, with support vector machines (radial basis function kernel) being the consistently top ranked model. However, predictive R^2 values demonstrated poor predictive performance across all models assessed (avg: -5.06% – 6 weeks; -3.7% – 12 months). Conversely, the Spearman correlation metric was higher (avg: 0.344 – 6 weeks; 0.265 – 12 months). This demonstrated that taxonomic relative abundance was not predictive of metabolite concentrations.

Conclusions: Our results suggest a degree of overall association between taxonomic profiles and metabolite concentrations. However, lack of predictive capacity for stool metabolic signatures reflects, in part, the possible role of functional redundancy in defining the taxa-function relationship in early life as well as the bidirectional nature

of the microbiome-metabolome association. Our results provide evidence in favor of a multi-omic approach for microbiome studies, especially those focused on health outcomes.

Section 2.2
Background

The human gut microbiome is a complex and diverse system of microorganisms that co-inhabit the intestinal lumen and play a crucial role in modulating human health and disease [270, 229]. The development of the microbiota in early life is a sensitive process akin to primary ecological succession [145], and therefore both reliant on, and vulnerable to, external perturbations. Studies have linked microbiome alterations to long-term health consequences, including risk of obesity [273], type I diabetes [147], and inflammatory bowel disease [10]. As such, there is a need to understand how the microbiome participates in the multifactorial pathways leading to long-term disease outcomes. One key to this open question lies in the currently undefined relationship between the taxonomic structure of the microbiome and its metabolic phenotype. Previous studies addressing this question have mainly focused on DNA-based profiling of microbial functional potential, which, due to complicated regulatory mechanisms at the cellular level beyond the genome, is not equivalent to the microbiota's realized functional landscape [116].

There exists a bidirectional association between the metabolome and the microbiome in the gut [287, 86]. These low molecular weight molecules can either be nutrients that shape the composition of the microbiome [228] or important byproducts of host-microbe nutrient co-metabolism that help regulate host metabolic homeostasis [114, 168, 224]. For example, members of the Clostridium clusters can produce and increase serum levels of branched chain amino acids, which are markers for insulin resistance and diabetes [233, 220]. However, studies suggest that the fecal metabolome specifically can be used as a readout of gut microbe metabolic functions. Zierer et al. [343] showed in a large cohort of adult females (n = 786) from the TwinsUK study that around 60% of the fecal metabolome is associated with microbial composition, where on average, 67% of variance in the metabolome can be explained by the microbiome.

Recent studies have integrated the metabolome in microbiome analyses of health outcomes, most notably Lloyd et al. [174] from the integrative Human Microbiome Project. However, these studies have mostly focused on adult populations with specific metabolic disease etiologies such as inflammatory bowel disease. Only a limited number of studies [301, 13, 275, 335, 36, 143], have simultaneously profiled the gut microbiome and metabolome from infant stool samples. These studies have preliminarily established that metabolomic profiles shift as taxonomic abundances change between subject case/control status [301, 275, 335, 117]. Specifically, Ayeni et al. (n = 48) [13] and Kisuse et al. (n = 35) [143] demonstrated that inter-sample distances calculated using metabolite abundances are correlated with those calculated from taxonomic profiles using Mantel tests across African and Asian cohorts. However, studies to date have either focused on preterm infants [301, 275, 335] or had small sample sizes (less than 50) [13, 36, 143]. We identified a major gap in defining microbiome-metabolome relatedness among infants from a population-based cohort capturing both early in infancy and near the first birthday, with regards to determining the strength of association and to identify key contributors to the overall concordance.

Here, we present our study investigating associations between microbe abundances assayed with 16S rRNA sequencing and metabolomic profiles measured with ^{1}H NMR spectroscopy in a cohort of infants representing a rural general population from the

New Hampshire Birth Cohort Study (NHBCS). This is a unique epidemiological cohort with multi-omic profiling of infant stool samples at multiple time points accompanied with rich sociodemographic, dietary and health outcomes data [183]. Our study utilizes predictive modeling, multivariate correlation methods and distance-based approaches to characterize the dynamic relationship between the gut microbiome and the gut metabolome in early life.

Section 2.3

Results

The overall workflow and subject selection process are described in Fig 2.1. Primary analyses were performed on paired microbiome-metabolome data sets on samples collected at approximately 6 weeks ($N = 158$ samples) and 12 months ($N = 282$ samples) of age (65 subjects have paired samples collected at both time points). In order to take advantage of the most samples from this study, each time point was analyzed separately with sensitivity analyses performed on sample pairs. As such, the sample size N will thereafter represent the number of samples found in each time point rather than the number of unique infants. After processing and filtering, we evaluated a final taxonomic dataset of 48 genera in 6 weeks samples and 72 genera in 12 months samples. Metabolomic profiles were represented as 208 unique untargeted spectral bins and a concentration-fitting method [309] was used to acquire specific relative concentrations of 36 targeted metabolites. These metabolites were chosen based on a literature search (Supplementary Note 4) for compounds known to be associated with commensal gut microbes. Results presented here will primarily feature the targeted dataset, with accompanying figures and tables for the untargeted data set in the supplemental section. Fig 2.1 shows the overall workflow including the sample selection process. In summary, we performed three analyses: First, an overall concordance analysis using ordinations with ecological distances; second, a parametric multivariate correlation approach with a variable selection component to quantify the overall correlation and determine important factors that contribute to the overall microbiome-metabolite association; third, a predictive analysis approach to see if taxonomic abundance alone can accurately predict the concentrations of specific metabolites.

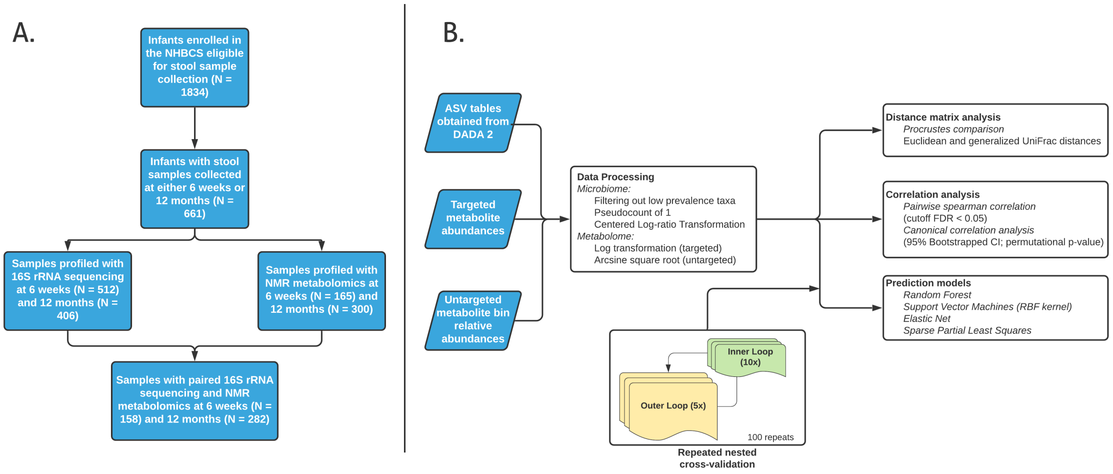
A.
Infants enrolled in the NHBCS eligible for stool sample collection (N = 1834)
Infants with stool samples collected at either 6 weeks or 12 months (N = 661)
Samples profiled with 16S rRNA sequencing at 6 weeks (N = 512) and 12 months (N = 406)
Samples profiled with NMR metabolomics at 6 weeks (N = 165) and 12 months (N = 300)
Samples with paired 16S rRNA sequencing and NMR metabolomics at 6 weeks (N = 158) and 12 months (N = 282)
B.
ASV tables obtained from DADA 2
Targeted metabolite abundances
Untargeted metabolite bin relative abundances
Data Processing
Microbiome:
Filtering out low prevalence taxa
Pseudocount of 1
Centered Log-ratio Transformation
Metabolome:
Log transformation (targeted)
Arcsine square root (untargeted)
Distance matrix analysis
Procrustes comparison
Euclidean and generalized UniFrac distances
Correlation analysis
Pairwise spearman correlation
(cutoff FDR < 0.05)
Canonical correlation analysis
(95% Bootstrapped CI; permutational p-value)
Prediction models
Random Forest
Support Vector Machines (RBF kernel)
Elastic Net
Sparse Partial Least Squares
Inner Loop (10x)
Outer Loop (5x)
100 repeats
Repeated nested cross-validation

Figure 2.1: Overview of the analysis. Panel A describes the subject selection workflow and panel B describes the analytic pipeline.

2.3.1. Study population

Study subject characteristics are summarized in Table 2.1 separately for both subjects providing samples at 6-week of age ($N = 158$) and 12-months of age ($N = 282$). Characteristic of our population, most infants are White (97.5% among subjects contributing a 6-week sample; 95.4% among subjects contributing a 12-month sample), delivered vaginally (6 weeks samples: 72.2%; 12 months samples: 70.9%) and did not have any systemic antibiotic exposure during initial hospitalization following birth (6 weeks samples: 95.6%; 12 months samples 97.2%). The average birth weight was also similar across subjects irrespective of the sample time point, 3370 g ($\pm$ 480) for infants contributing 6-week samples and 3430 g (± 528) for infants contributing 12-month samples. Similarly, the average gestational age was 39.1 weeks (± 1.59) (6-week samples) and 39 weeks (± 1.7) (12-month samples). At the time of 6-week sample collection, 62% of infants had been exclusively breastfed while by the time of 12-month sample collection, 59.2% of infants received breast milk supplemented with formula, however a large minority (35.1%) remained exclusively breastfed. There were more male than female infants in the cohort (53.8% male among infants contributing a 6-week sample; 56.4% male among infants contributing a 12-month sample). Maternal smoking during pregnancy was rare (6-week samples: 7%; 12-month samples: 5%).

2.3.2. Inter-omic sample distance comparison suggests overall concordance between data sets

Global concordance between the microbiome and the metabolome was observed across both time points and metabolomic data sets (Fig 2.2A, Fig B.1A) when analyzed using a symmetric Procrustes test with samples ordinated by Euclidean distances Materials and Methods. It is noted that the p-value at 6 weeks for the targeted data set ($p = 0.057$) was only trending close to significant at the 0.05 level.

Since the Procrustes test was performed on principal coordinate (PCoA) ordinations of sample distances, the result is sensitive to the choice of dissimilarity metric. In addition to standard Euclidean distances, the generalized UniFrac (gUniFrac) metric was also leveraged to account for phylogeny in calculating differences between samples. With gUniFrac, the association was not significant at either time points for the targeted data set only (Fig 2.2B), while the untargeted data set still maintained strong concordance (6 weeks samples – $p = 0.001$; 12 months samples – $p = 0.006$; Fig B.1B).

Table 2.1: Selected characteristics of subjects contributing samples at 6 weeks ($N = 158$) and at 12 months of age ($N = 282$)

	6 weeks (N = 158)	12 months (N = 282)
Birthweight (grams)		
Mean (Standard Deviation)	3370 (480)	3430 (528)
Median [Minimum, Maximum]	3430 [1910, 4710]	3450 [1320, 4660]
Missing	2 (1.3%)	4 (1.4%)
Sex		
Male	85 (53.8%)	159 (56.4%)
Female	73 (46.2%)	123 (43.6%)
Feeding mode until time of sample collection		
Unknown	6 (3.8%)	7 (2.5%)
Exclusively breastfed	98 (62%)	99 (35.1%)
Exclusively formula fed	13 (8.2%)	9 (3.2%)
Mixed	41 (25.9%)	167 (59.2%)
Delivery Mode		
Vaginal	114 (72.2%)	200 (70.9%)
Cesarean	44 (27.8%)	82 (29.1%)
Gestational Age (Weeks)		
Mean (SD)	39.1 (1.59)	39 (1.70)
Median [Minimum, Maximum]	39.1 [33.4, 43.0]	39.1 [29.1, 42.0]
Post-delivery infant systemic antibiotic exposure		
No	151 (95.6%)	274 (97.2%)
Yes	7 (4.4%)	8 (2.8%)
Maternal smoking during pregnancy		
No	143 (90.5%)	262 (92.9%)
Yes	11 (7.0%)	14 (5.0%)
Missing	4 (2.5%)	6 (2.1%)
Infant Race		
Other	4 (2.5%)	13 (4.6%)
White	154 (97.5%)	269 (95.4%)

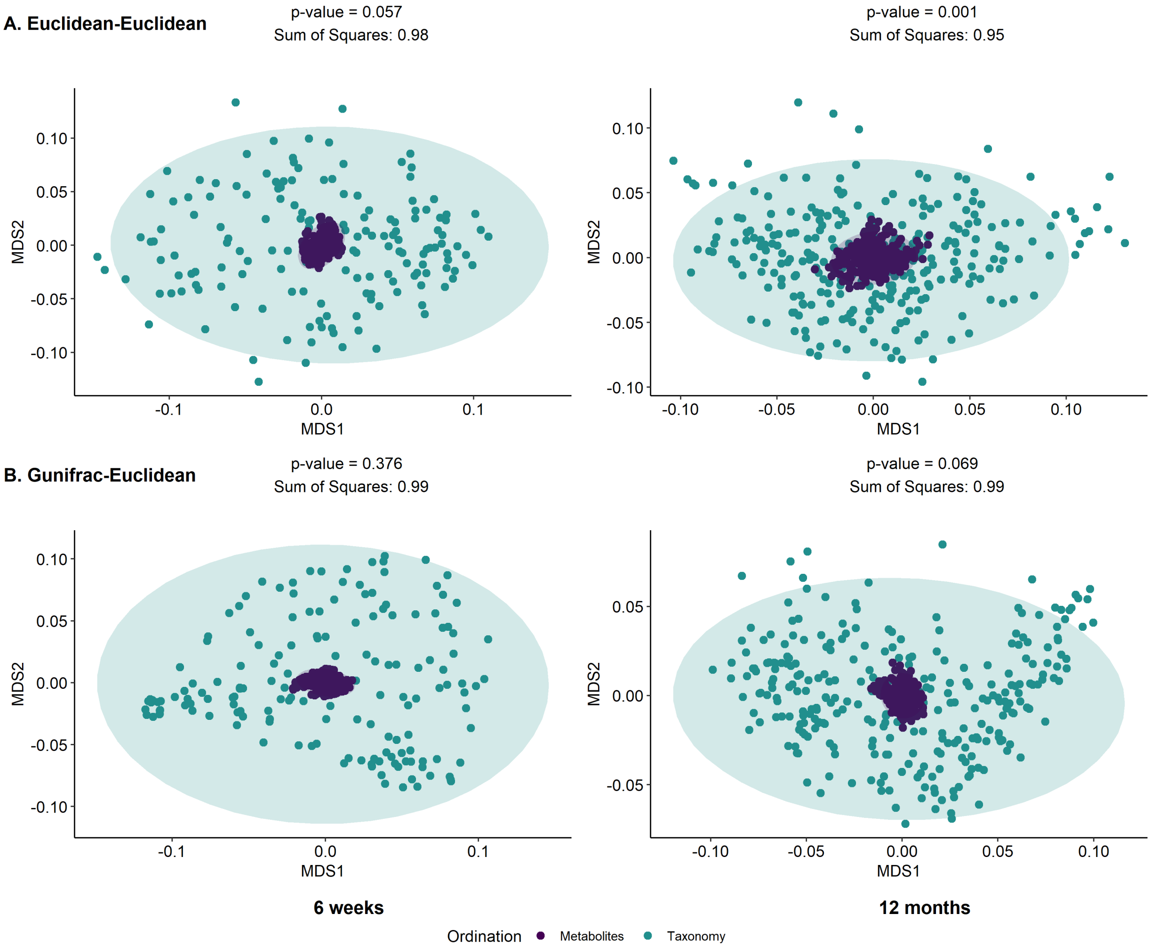

Figure 2.2: Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and taxonomic relative abundances for 6 weeks (left panels) ($N = 158$) and 12 months (right panels) ($N = 262$). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of log-transformed metabolite profiles. Bottom panel presents analyses based on gUniFrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of log-transformed metabolite profiles. There were significant associations between the microbiome and the metabolome (both targeted and untargeted) when utilizing Euclidean distances, however this association goes away when the gUniFrac distance was employed for the targeted metabolites only.

2.3.3. Sparse canonical correlation analyses reveal the core set of microbe-metabolite groups driving inter-omic relatedness

Given broad concordance between the gut microbiome and metabolome from sample distance analyses, we employed sparse canonical correlation analysis (sCCA) and pairwise Spearman rank correlation to ascertain the strength of association as well as to identify core microbes and metabolites driving this relationship (Materials and Methods).

The majority of taxa (65% of total genera at 6-weeks and 80% at 12-months) and metabolites (100% of total metabolites at 6-weeks and 80% at 12-months) were part of significant (FDR threshold 0.05) Spearman pairwise comparisons (Supplementary Note 1). This demonstrated a high level of congruence, where most microbes are involved in metabolic processes captured in the stool metabolome. This was also reflected in the significant multivariate correlation (permutation $p < 0.001$). However, at 6 weeks (correlation: 0.606 [0.61 – 0.73]), the degree of concordance was slightly higher than at 12 months (correlation: 0.52 [0.431 - 0.646]) but this difference was not significant due to overlapping confidence intervals. The canonical correlation was slightly higher in the untargeted data set (6 weeks: 0.636 [0.621 – 0.733]; 12 months: 0.49 [0.475 – 0.702]), however the difference between time points was similar (Fig B.2, Supplementary Note 2).

Using sCCA, we identified a core set of microbes and metabolites that are major contributors to the multivariate correlation (Fig 2.3 right panels; Supplementary Note 2). Selected microbes (in both the targeted and untargeted data set) belonged to the Firmicutes, Actinobacteria and Proteobacteria phyla, as those are the most commonly found phyla in the infant gut [183, 15]. However, previously established dominant genera such as *Bifidobacterium*, *Bacteroides* and *Lactobacillus* were not consistently selected across both time points. In the targeted data set *Bifidobacterium* was selected only at 6 weeks and *Lactobacillus* was only selected at 12 months. Most notably, more microbes were selected at 12 months compared to 6 weeks in the targeted data set, however in the untargeted data set this pattern was reversed (Fig B.2, right panels). In terms of selected metabolites, the majority of the selected metabolites in the targeted data set were amino acids (Supplementary Note 1), with some short chain fatty acids (SCFAs) selected at the 6-week time point.

2.3.4. Microbial community structure is weakly predictive of stool metabolite relative concentrations

In order to determine how well the fecal metabolome acts as a functional representation of the gut microbiome, we fitted metabolite-specific prediction models based on taxonomic profiles. Chosen models include random forest (RF), elastic net (EN), support vector machines with radial basis kernel (SVM-RBF) and sparse partial least squares (SPLS), all of which had previously been shown to work well with microbiome-associated learning tasks [342]. Evaluation was based on predicted R^2 and Spearman correlation coefficient (SCC) as measured using 100 repeats of 5-fold nested cross validation (Materials and Methods).

Predictive performance was more dependent on the metabolite being predicted than by choice of model (Fig 2.4, Supplementary Note 3, Supplementary Note 1). Looking at predictive R^2 (Fig 2.4A), the average posterior mean performance across all models and metabolites was negative for both time points (-5.6% at 6 weeks; -3.07% at 12 months), which indicated that for most prediction tasks the fitted model was less predictive than a naive, intercept only model. At 6 weeks 22.2% of metabolites had models that perform significantly better than the null (lower bound of 95% credible interval larger than 0) while at 12 months 38.9% of metabolites fit the classification. However, even among such metabolites, the maximum R^2 is only 11.8% at 6 weeks and 8.7% at 12 months. Conversely, SCC values were higher in comparison (cross-metabolite avg.: 0.339 at 6 weeks and 0.249 at 12 months) (Fig 2.4B, Supplementary Note 3). At 6 weeks, 83% of metabolites were significantly more performant than the null, while at 12 months all metabolites were selected. Using a more stringent cutoff as used by Mallick et al. [187], the majority of metabolites at 6 weeks (69.4%

of total metabolites) still remained as well predicted while conversely at 12 months only 38.9% (of total metabolites) were predictable.

Results from the untargeted analysis showed higher performance values for both evaluation metrics (Supplementary Note 3). Specifically, 56.7% of metabolites bins at 6 weeks and 42.7% of bins at 12 months had R^2 values significantly higher than 0. However, under SCC, while 57% of metabolite bins at 6 weeks had SCC values significantly larger than 0.3 cutoff, only 28.8% of metabolite bins at 12 months fit this criterion. Despite better performance, the overall average values were still low, suggesting that across the entire metabolome few metabolites were highly predictable.

Despite weak predictive performance values, we were still interested in determining a model that stands out as the most appropriate for our prediction task. Aggregating performance across metabolites stratified by model for both evaluation metrics (Fig 2.5, top panel), it can be observed that the average performances were similar (Supplementary Note 3), for which no semi-targeted analyses performed better on average than the naive model under R^2. This is further illustrated when model performance was aggregated by rank using Borda scores (Fig 2.5, bottom panel). A higher score indicated that a model was selected as the top choice more times than others, where an even score distribution across models corroborated the suggestion that no model was best across all prediction tasks. That said, SVM-RBF seemed to be the highest scoring model, particularly for the 6-week time point. The untargeted analysis also found similar results (Fig B.3).

2.3.5. Sensitivity analyses

We performed both Procrustes and correlation analyses on a data set restricted to the 65 subjects with paired samples collected at both time points (6 weeks and 12 months). Each time point was analyzed separately as in our main analysis. In the targeted data set, significant Procrustes concordance was observed at 12 months (p-value = 0.003) but not at 6 weeks (p-value = 0.106). This association was no longer significant when considering taxonomic ordination using the gUniFrac distance metric (6 weeks). Surprisingly, in the untargeted data set, no association was observed across both time points and choice of distance metric (Fig B.5, Fig B.6). In the canonical correlation analyses, significance was only observed in the targeted data set at 6 weeks only (6 weeks: permutation p-value = 0.044; 12 months: permutation p-value = 0.388). Even though most correlations were not significantly different from the permuted null, the canonical correlation coefficient is higher at 6 weeks compared to 12 months in both the targeted (6 weeks: 0.676 [0.661 – 0.765]; 12 months: 0.52 [0.484 – 0.663]), and untargeted (6 weeks: 0.703 [0.685 – 0.788]; 12 months: 0.444 [0.52 – 0.705]) data sets (Fig B.7, Fig B.8).

Furthermore, to ascertain the uncertainty of model choice, we evaluated all selected modelling approaches with simulated data sets based on bootstrapped resampling of taxonomic relative abundances (Fig B.4). For the first simulation scenario, models were assessed against generated metabolite concentrations under different degrees of model saturation (number of taxa associated with the outcome) and association strength (signal to noise ratio). As expected, model performance asymptotically reached perfect prediction with increasing signal strength and model saturation, which demonstrated that prediction models were able to capture predictive associations should they arise even in sparse microbiome data sets. Most notably, simulation performance differed more by signal-to-noise ratio than model saturation, which indicated that the strength of association plays a larger role in the observed weak predictive performance than the number of taxa involved. Surprisingly, we obtained very similar results to our real data values under our lowest simulation setting (model saturation = 5%; signal-to-noise ratio 0.5). As such, it can be suggested that the lack of predictability is due to weak coupling rather than model choice.

Section 2.4

Discussion

In this study, we provide a descriptive and hypothesis generating analysis of the relationship between fecal microbial taxonomic abundances and metabolite concentrations with multi-omic profiling via paired targeted sequencing of the 16S rRNA gene and ^{1}H NMR metabolomics at multiple time points. Ecological, statistical and machine learning approaches were applied to provide a multi-faceted view of this association. To our knowledge, this study is one of the few comprehensive investigations addressing the microbiome/metabolome interface in a large general population cohort of infants.

2.4.1. The microbiome is significantly correlated but weakly predictive of the metabolome

Overall global concordance was found from three independent methods (Procrustes analysis, sCCA and univariate Spearman correlation), consistent with previous studies on both infant [13, 143] and adult populations [174, 193]. This overall effect was found at both time points, suggesting there coupling exists throughout infancy despite high levels of both inter- and intra-individual variability in taxonomic compositions [15].

Although our analyses demonstrated significant multivariate and univariate correlation between the microbiome and the metabolome, most metabolites were not predictable when evaluated across multiple machine learning models. Even among the small number of metabolites that are significantly predictable compared to the null, the maximum performance values were still low for both the untargeted and targeted analyses. When compared to a recent study performing metabolite predictions from taxonomic abundances using an adult cohort [187], both the number of well-predicted metabolites and the average performance values were much lower, even when using similar evaluation criterion and cut offs. It is unlikely that model choice was driving the lack of predictability, since all chosen methods had been shown to be suited for microbiome-associated prediction tasks [342, 232] as well as covering both non-linear and linear associations. This is further evidenced in our sensitivity analyses, where non-parametric simulations demonstrated that low predictability across both evaluation metrics was driven by low signal-to-noise ratio rather than model choice or number of taxa driving the association.

These results can be attributed to the limitations of our study design. We utilized partial 16S rRNA sequencing instead of whole genome shotgun sequencing. This limits our taxonomic resolution to the Genus level for most of the analysis [132]. Since bacterial functions relevant to human metabolism are likely to be strain specific [338, 175], we hypothesized that aggregating to Genus level might dilute the direct effects, where different strains within the same Genus might have opposite impacts on the abundance of a certain metabolite. This would result in a lack of predictability as the same feature would contain elements that both increase and decrease the values of the outcome of interest.

However, we can potentially attribute overall performance to other ecological processes. A likely candidate is functional redundancy, an aspect ubiquitous in microbial communities [176], plays an important role in this weak coupling. Functional redundancy is the ecological phenomena that multiple species representing a spectra of taxonomic groups can perform similar roles [176, 295], and is usually a marker for ecosystem resilience [7]. Under this paradigm, the loss of a certain metabolite producing taxa would not impact the abundance of that metabolite as different taxa can complement the functional role, complicating taxa to metabolite predictions. This is evidenced when inter-omic associations is no longer significant in Procrustes analyses when phylogenetic relatedness was adjusted using the gUniFrac distance metric. Since gUniFrac adjusts for phylogeny by weighting the differences in proportions of each taxa across two samples by the branch length from constructed evolutionary trees [48], the absence of an association suggests that samples with similar metabolic profiles might be numerically comparable (cluster together under Euclidean distances) but with evolutionarily divergent taxonomic compositions. This is further supported by our supplementary PICRUSt2 analyses, where we found for most pathways no single genera dominate functional contribution (Fig B.10). Functional redundancy is also consistent with previous research in human associated microbiomes [178].

2.4.2. Taxa and metabolites selected to be core to the microbiome-metabolome correlation reveal the importance of amino acid metabolism

Taxa and metabolites with non-zero loading coefficients in sCCA analyses were identified as factors driving this overall correlation. The sCCA procedure utilized a L1-penalized matrix decomposition of the cross-product matrix akin to a LASSO regression problem [316], which means that variables were selected based on their importance to the overall covariance between taxa and metabolite abundances.

At six weeks, two short chain fatty acids (SCFAs), butyrate and propionate, were selected as core to the microbiome-metabolome interface. SCFAs (which includes compounds such as isobutyrate, and acetate) are important metabolites obtained primarily from colonic microbial fermentation of carbohydrates that escape digestion in the small intestines [210]. Butyrate is an energy source for colonocytes [161] as well as participating in the maintenance of the gut epithelial barrier through mucin production [236]. Similarly, propionate is part of the gluconeogenesis pathway in liver hepatocyte cells, which is core to lipid and energy metabolism in liver [68]. Most importantly, SCFAs participate in immune programming in early life, where the reduction in SCFA producing bacteria is associated with inflammatory bowel disease [61, 265].

SCFA production in early life is linked to the *Bifidobacterium* and *Bacteroides* catabolism of human milk oligosaccharides (HMO) [128, 158, 189], which explains the selection of the *Bifidobacterium* genera at 6 weeks where infants are exclusively on a milk-based diet. This is further supported in our supplementary PICRUSt2 analysis, where predicted pathways whose abundance significantly correlate with butyrate concentrations were those associated with breakdown of sugars into butanoate (Fig B.9). The genera breakdown of those functions features prominently *Bacteroides*, *Bifidobacterium*, *Lachnoclostridium*, *Flavonifractor*, and *Clostridium* sensu stricto 1 genera (Fig B.10). This demonstrates that at 6 weeks, infant microbiome-metabolome interaction is primarily concerned with breakdown complex sugars into SCFAs, cementing it's functional role in microbiome development [274].

Surprisingly, the selected *Bifidobacterium* genus is negatively correlated with butyrate abundance. We hypothesized that this might be due the complex cross-feeding relationship that exist between *Bifidobacterium* and butyrate-producing taxa [258]. On one hand, some *Bifidobacterium* species can be completely commensal, producing secondary metabolites such as acetate that assist in the growth of butyrate producing species. On the other hand, other *Bifidobacterium* strains such as *B. longum* LMG 11047 and *B. adolescentis* can compete for the same substrates as butyrate producing species [203]. The selection of the negative association between *Bifidobacterium* and butyrate suggests that butyrate-suppressing *Bifidobacterium* strains might be more important in our infant samples.

However, the most selected metabolites in sCCA analyses are amino acids (7 out of 10 metabolites selected at 6 weeks were amino acids). Prior studies have shown that the microbiota participate in regulating host amino acid homeostasis by acting as both producers and utilizers [220]. The most common amino acid fermenters in the human gut include those from the *Clostridia* class [62]. Our results further support this as most selected microbes with positive correlation with amino acids are of the *Eisenbergniella*, *Flavonifractor*, *Ruminococcaceae UCG-004*, *Oscillibacter* and *Ruminiclostridium* genera under *Clostridia*. This is further seen in our supplemental PICRUSt2 analyses, where predicted abundance of isoleucine and methionine biosynthesis pathways are significantly correlated with observed concentrations (Fig B.9).

Aside from being fermenters, microbes can also either directly utilize amino acids and incorporate them into protein synthesis, or catabolize them as an energy source, producing secondary metabolites. Even though the process of amino acid catabolism for energy alone is not energetically efficient [228], it produces secondary metabolites such as aforementioned SCFAs, which are important molecules in the metabolic interactions between the microbiota and the host. However, amongst selected microbes whose abundance are negatively correlated with amino acid concentrations (hence, suggestive of catabolism), we do not observe corresponding positive correlation with selected SCFAs. We hypothesized that this might be due to the fact that bacterial concentrations are higher in distal parts of the intestine [86, 220] where nutrient availability is low. This lack of available carbohydrates might incentivize microbes to conserve energy by directly incorporating free amino acids rather than metabolizing them. On the other hand, prior studies suggested that microbial amino acid catabolism is compartment specific and occurs in more proximal regions [62, 182]. However, our study design is limited to cross-sectional metabolomic profiling, which limits the possibility of detecting SCFAs that are rapidly produced and absorbed.

2.4.3. The microbiome is more tightly coupled with the metabolome in early infancy

Results suggest some level of significant difference in microbiome-metabolome coupling across development. Canonical correlation, while not significantly different, were lower at 12 months than at 6 weeks, suggesting a time-varying effect. When looking at predictability, we observed a higher number of well predicted metabolites at 6 weeks compared to 12 months. Among those selected as well predicted metabolites, the average performance values (both R^2 and SCC) where higher. This is also replicated in the global untargeted data set. Furthermore, in our supplementary PICRUSt2 analyses, there exists a higher number of significantly correlated predicted pathway abundance to observed metabolite concentrations (Fig B.11), indicating increased metabolic coupling between the microbiome and the metabolome at 6 weeks compared to at 12 months.

There are various factors that can contribute to the difference in microbiome-metabolome coupling between infants at 6 weeks and 12 months. First, there exists substantive differences in dietary patterns for those included in our analysis. The majority of infants at 6 weeks (62%) were exclusively breastfed, while that number is markedly less (35%) at 12 months, at which time infants are also consuming complimentary solid family foods. This transition in diet to solid foods have been shown to induce a change in the gut microbiome composition and diversity due to increased amounts of fiber and protein [205, 157], which might favor certain microbes over others. Such changes in diet, particularly the cessation of breastmilk intake, also contributed towards the development of infant gut microbiomes towards a more “adult like” state [15, 205]. We hypothesized that earlier in life when infants are only consuming a limited type of food (predominantly breast milk or formula), the microbiome participates more actively in host-microbiome co-metabolic activity as infants are more reliant on microbes to breakdown complex nutrients (reviewed in [202]). Conversely, at one year of age where the microbiome has matured, this relationship is not as strongly coupled as a larger share of the metabolome comes from host-produced metabolites.

However, as analyses were conducted within each timepoint independently with little subject overlap, further investigations are required to make more conclusive statements about the potential time-varying effect of microbiome-metabolome coupling. Particularly, aside from differences in diet, factors such as differences in antibiotic exposure [57] and maternal covariates [180] might result in differences between time points. In future studies we hope to examine this factor using samples across multiple time points for the same infants.

2.4.4. Limitations

This study has various limitations. First, we utilized partial 16S rRNA gene sequencing instead of shotgun whole genome sequencing, which limits our taxonomic resolution to the genus level for most of the analysis [338]. We hypothesized this lack of resolution contribute to overall lack of predictability, as well as limiting the interpretability of variables selected by the sCCA process as species and strain level differences can result in completely separate metabolic contributions [175]. For example, we cannot disentangle the different *Bifidobacterium* strains that might compete with butyrate producing taxa and generating the negative correlation between measured SCFAs and *Bifidobacterium* abundance.

Second, our cohort includes only infants from the NHBCS, a population-based cohort reflecting mostly rural and White demographics of northern New England in the United States. While this increases confidence in the internal validity of our study, this homogeneity in race and geography limits the generalizability of our results to other populations.

Third, our study is a cross-sectional survey if microbiome-metabolome relationships at two different time points. This means that we cannot capture associations relating to metabolites that are highly produced and consumed. This means that the metabolites selected might not be representative of the intricate relationship between the microbiome and the metabolome. This interpretation is further limited by the lack of annotation for our untargeted metabolite bins, which cannot be compensated by the small number of metabolites selected for the targeted analyses.

Finally, each time point was analyzed independently with only 65 subjects with samples in both time points. As such, this limits the ability to explore the differences in coupling across the first year of life.

Section 2.5 Conclusion

In conclusion, we conducted one of the first large-scale multi-omics analysis of the microbiome-metabolome relationship using samples from a large birth cohort study at 2 time points (6 weeks and 12 months). Although we found global concordance between the microbiome and the metabolome, the inter-omic concordance is weak, where bacterial abundances at the genus level cannot accurately predict metabolite concentrations. We hypothesized that this might be due to functionally relevant diversity at the strain level, as well as the impact of functional redundancy on the contribution of each microbe to metabolite abundances. Additionally, we were able to identify metabolites and microbes driving the overall correlation. Results pointed to support the importance of SCFA metabolism particularly at 6 weeks, as well as the role of amino acid metabolism, either as a source of SCFA and energy in the absence of carbohydrates, or as a general mechanism for microbes to save energy as they incorporate amino acids around their environment. Finally, our analysis suggests preliminary evidence that the degree of microbiome-metabolome coupling changes across time, being much more integrated at six weeks compared to one year.

We conclude that although the metabolome is a functional output of the microbiome, there exists massive challenges in being able to trace specific microbial contributions to host-microbe metabolism due to the complexity of factors such as functional redundancy and strain level variability. As such, we recommend studies to profile both the microbiome and the metabolome, as aspects of microbial metabolic contributions cannot be found solely through one omic data set. This is particularly important in settings where it is important to have a mechanistic understanding of the role of microbes such as developing of microbiome therapies [163].

Section 2.6 Materials and Methods

2.6.1. Study population

Subjects for this study were from the New Hampshire Birth Cohort Study (NHBCS) who provided infant stool samples at 6-weeks and 12-months after birth. These two timepoints are chosen as each correspond to routine maternal postpartum visit, allowing sample collection with minimal participant burden. Furthermore, at both time points, infant feeding patterns are comparatively more well established. As described in previous studies [183, 180], NHBCS is a prospective study of mother-infant dyads in New Hampshire, USA. Participants eligible are pregnant women between the ages of 18 and 45 years old, currently receiving routine prenatal care at one of the study clinics, consuming water out of a private well with no intention to move prior to delivery. The Center for the Protection of Human Subjects at Dartmouth provided institutional review board approval. All methods were carried out in accordance with guidelines. Written informed consent was obtained for participation from all subjects for themselves and their children. Comprehensive sociodemographic, exposure and outcome data such as infant feeding method, delivery mode, maternal smoking status, etc. were collected for all participants through surveys, medical records and telephone interviews conducted during pregnancy, about 6 weeks postpartum, and updated every 4 months up until first year of age and every 6 months thereafter.

2.6.2. Collection of infant stool samples

Infant stool samples were collected at 6-weeks and 12-months. Stool samples were provided in diapers and stored by subjects in their home freezer ($-20°$C) until they were able to return it to the study site. Stool was thawed at 4°C so that it could be aliquoted into cryotubes. Stools collected for 16S rRNA gene sequencing were aliquoted (range 350-850 mg) into 3ml RNAlater and homogenized before storing at $-80°$C. Stools collected for metabolomic analysis were aliquoted (1-2 grams) into 15ml centrifuge tubes before storing at $-80°$C.

2.6.3. Taxonomic profiling using 16S rRNA targeted gene sequencing

RNAlater stool samples were thawed and DNA was extracted using the Zymo Fecal DNA extraction kit (Cat #D6010, Zymo Research, Irvine, CA), according to the manufacturer's instructions. For each sample extraction, $400\mu l$ RNAlater stool slurry (50–100 mg of stool) was used to isolate DNA. Extractions were performed in batches of multiple samples and included a composite RNAlater stool positive control and a RNAlater negative control. Lysis was performed using $750\mu l$ Lysis Buffer in ZR BashingBead™ Lysis Tubes (0.5 mm beads), mixed and then shaken on a Disruptor Genie for 6 min. Eluted DNA was quantified on a Qubit™ fluorometer using the Qubit™ dsDNA BR Assay. Average coefficient of variation of DNA yields ($\mu g/\mu l$) for composite RNAlater stool positive controls was 28%. No DNA was ever detectable in negative control elutions. Concentrations of DNA samples used for 16S rRNA gene sequencing range from 1 $ng/\mu l$ to 25 $ng/\mu l$.

The V4-V5 hypervariable region of bacterial 16S rRNA gene was sequenced at Marine Biological Laboratory in Woods Hole, MA, using standard Illumina MiSeq amplicon approach (paired end sequenced between 518F and 926R) [222, 121]. As described previously [183, 180], 16S rDNA V4-V5 amplicons were generated from purified genomic DNA samples using fusion primers. The use of forward primers containing one of eight five-nucleotide barcodes between the Illumina-specific bridge and sequencing primer regions and the 16S-specific region and a single reverse primer containing 1 of 12 Illumina indices enables 96 samples per lane multiplexing. Amplifications were done in triplicate with one negative control for internal quality control at MBL. We used qPCR (Kapa Biosystems) to quantify the amplicon pool, and one Illumina MiSeq 500 cycle paired end run to sequence each pool of 96 libraries. We demultiplex and divided datasets using Illumina MiSeq reporter and a custom Python script. Demultiplexed reads derived from Illumina sequencing were denoised and quality filtered using DADA2 (v. 1.12.1) [40] in R (v. 3.6.1) [251]. Illumina adapter sequences were removed prior using cutadapt (v. 1.18). We utilized DADA2's filterAndTrim function to remove reads either containing a quality score of 2 or lower (minQ = 2) or with expected errors [77] of 2 (maxEE = c(2,2)) or higher. Post filtering, we obtained an average of 119,800 reads per sample for 6-week samples and 120,480 reads per samples for 12-month samples. On average, we 74.7% of reads were kept for 6-week samples and 76.3% of reads were kept for 12-month samples. We then use the RDP classifier implemented natively in the DADA2 R package with SILVA database (v. 128) to profile the taxonomy of identified amplicon sequence variants (ASVs).

2.6.4. Metabolomics profiling using untargeted and targeted ^{1}H NMR

^{1}H NMR metabolomics was performed in collaboration with the NIH Eastern Regional Comprehensive Metabolomics Resource Core (RCMRC) at UNC Chapel Hill. De-identified stool aliquots were shipped on dry ice and immediately stored at -80°C for metabolomics analysis. Samples were thawed and $\sim$150mg of stool samples were transferred to MagNA Lyser tubes after recording the weight. Samples were then homogenized with 50% acetonitrile in water by using the Omni Bead Disruptor (Omni International, GA, USA). Homogenized samples were centrifuged at 16000 rcf and the supernatant was separated into another tube. An aliquot (1000 μL, 100 mg equivalent of fecal mass) was transferred into an Eppendorf tube and lyophilized overnight. The dried extract was reconstituted in 700 μL of NMR master mix (containing 0.2M phosphate buffer, 0.5 mM DSS-d6 (internal standard), and 0.2% sodium azide (preventing bacterial growth)), vortexed on a multi tube vortexer at speed 5 for 2 min and centrifuged at 16000 rcf for 5 min. A 600 μl aliquot of the supernatant was transferred into pre-labeled 5mm NMR tubes. Additionally, study pooled quality control (QC) samples (created from randomly selected study samples) and batch pooled QC samples were generated from supernatants of study samples and aliquots of supernatants were dried and reconstituted similar to study samples described above and used for QC purposes.

^{1}H NMR spectra of feces extracts were acquired on a Bruker 700 MHz NMR spectrometer using a 5 mm cryogenically cooled ATMA inverse probe and ambient temperature of 25°C. A 1D NOESY presaturation pulse sequence (noesygppr1d [18, 69], [recycle delay, RD]-90°-t1-90°-tm-90°-acquire free induction decay (FID)]) was used for data acquisition. For each sample, 64 transients were collected into 64k data points using a spectral width of 12.02 ppm, 2 s relaxation delay, 10 ms mixing time, and an acquisition time of 3.899 s per FID. The water resonance was suppressed using resonance irradiation during the relaxation delay and mixing time. NMR spectra were processed using TopSpin 3.5 software (Bruker-Biospin, Germany). Spectra were zero filled, and Fourier transformed after exponential multiplication with line broadening factor of 0.5. Quality control measures included review of each NMR spectrum for line shape and width, phase and baseline of spectra, and tight clustering of QC samples in Principal Component Analysis [37]. NMR bin data (0.49-9.0 ppm) were generated (untargeted data) excluding water (4.73-4.85 ppm) using intelligent bucket integration of 0.04 ppm bucket width with 50% looseness using ACD Spectrus Processor (ACD Labs Inc, Toronto, Canada).The integrals of each bin were normalized to the total spectral intensity of each spectrum and transferred to analysis software. This resulted in a collection of spectral bins with bin-specific relative abundances, which will be called the untargeted data. In addition, relative concentration of library-matched metabolites (selected from the literature implicated to be important in host-microbe relationships - Supplementary Note 4) was determined by using Chenomx NMR Suite 8.4 Professional software [309].This data set will be called the targeted data set.

2.6.5. Software and tools

All analyses were performed using the R programming language (v. 3.6.3) [251] and associated packages. All data wrangling steps were performed using `phyloseq` [196], `plyr` and `tidyverse` packages [310], as well as the compositions package [290] for log-ratio transformations. All figures were generated using the `ggplot2` [311], `cowplot` [313], `viridis` [98] and `pheatmap` [146] packages. Additionally, the `tidymodels` [152] suite of packages was utilized to assist in all modelling tasks. Specific packages used for modelling will be enumerated below. All scripts as well as intermediary analysis objects are available on GitHub with all dependencies and their versions (`https://github.com/qpmnguyen/infant_metabolome_microbiome`).

2.6.6. Data transformation and normalization

For microbiome data, we retained all ASVs present in at least 10% of samples [187] and added one pseudocount to all cells [153]. We then subsequently aggregated all ASVs to the genus taxonomic level [342] and converted data to relative proportions using total read counts by sample to account for differential sequencing depth. We further filtered out taxa with mean relative proportion $< 0.005\%$ [30]. This filtration step resulted in 46 genera for 6-week samples and 72 genera for 12-month samples. To address the compositional problem induced by a sum to one constraint, we apply the centered log ratio transformation (CLR), which is often used to remove such constraints in microbiome data sets [4]. The CLR transformation is favored compared to other statistically equivalent log-ratio transformations due to its scale invariant property and ease of interpretation [104].

For metabolomic data sets, we employed different transformations to approximate homoscedasticity depending on the data type (targeted vs untargeted). For targeted metabolites, we performed a $\log(x+1)$ transformation while for untargeted metabolites we utilized the arcsine square root transformation which has been previously used for transforming composition metabolomics data sets [187].

2.6.7. Distance matrix analyses

Principal coordinates analysis (PCoA) was performed using the `pcoa` function from the `ape` package in R [230] with sample distance matrices. The PCoA procedure seeks to represent high dimensional multivariate data sets in lower dimensions through eigen decomposition of the doubly centered distance matrix. PCoA allows the usage of non-Euclidean distances between samples such as ecological indices, which makes it a preferable method for sample ordination compared to principal component analysis (PCA).

We constructed Euclidean distance matrices for both metabolic and taxonomic profiles post data transformation described in the previous section. Additionally, gUniFrac distances (alpha = 0.5) [48] were considered for taxonomic data using the implementation provided in the package `MiSPU` [323]. gUniFrac requires a phylogenetic tree, of which an approximate maximum likelihood phylogenetic tree was constructed with representative ASV sequences using `FastTree` (v 2.1) [244]. Multiple sequence alignment was performed using the `AlignSeqs` function from the `DECIPHER` package in R [320] and trees were midpoint rooted using `phytools` [255]. Since multiple sequence alignment is not conserved under filtering and aggregation of ASVs, gUniFrac distance calculations were performed with pre-filtered ASV-level abundances normalized to relative abundances.

The first two axes of constructed ordinations were then compared using a symmetric Procrustes procedure implemented in the protest function in the `vegan` package [227]. Procrustes superimposes two ordinations by translating and rotating the coordinates, which preserves the general structure of the data. This method performs a superimposition fit between two data sets minimizing the sum-of-squared differences (m^2), which describes the degree of concordance between the two configurations normalized to unit variance. Significance is obtained by testing against the permuted null using a permutation test. This method was shown to have more power while also limiting type I error compared to the traditional Mantel test in ecological analysis [238]. Significance was determined using a permutation test on the sum of squared differences with 999 permutations [37].

2.6.8. Sparse canonical correlation and Spearman correlation analyses

Sparse canonical correlation analysis (sCCA) was performed to identify strongly associated metabolite-microbe groups. sCCA seeks to find linear combinations of variables from each dataset that maximizes the correlation with each other while simultaneously thresholding variable specific weights to induce sparsity and performing variable selection. The correlation coefficient in the first canonical variate quantifies the overall degree of multivariate associations. As such, sCCA is a popular method in integrating multi-omics datasets with the ability to select more biologically relevant sets of features compared to traditional ecological methods such as co-inertia analysis [42]. In this study, we use the sCCA implementation in the package `PMA` in R [316] which uses a novel penalized matrix decomposition procedure to achieve sparsity [317]. We tune hyperparameters using a permutation approach in the `CCA.permute` function (`nperms` = 50) prior to fitting the final model. We obtained the correlation coefficients as a measure of overall correlation between the two data sets and calculated a bootstrapped 95% confidence interval (`nboot` = 5000). Additionally, we tested for significance using a permutation test (`nperm` = 1000) for α at the 0.05 level. In order to keep the structure of the data set across different permutations, we use the function `randomizeMatrix` from the package `picante` in R [138] using the richness null model, which randomizes community abundances within samples to maintain sample species richness.

Pairwise Spearman correlations were also performed using the cor function in R. Hypothesis testing was done using cor.test, with multiple hypothesis testing correction using the Benjamini-Hochberg procedure [24] using `p.adjust`. An FDR value of 0.05 is used as cutoff for significance pairwise correlations. Visualization was done using `pheatmap` package in R.

2.6.9. Predictive modelling and evaluation

We choose candidate models based on previous research utilizing supervised learning with microbiome associated prediction tasks [342, 232, 187]. Specifically, we chose random forest (RF) [33], support vector machine with radial basis function kernel (SVM-RBF) [32], elastic net (EN) [344] and sparse partial least squares (SPLS) [54], which have all been shown to perform with high-dimensional predictors. These models also support linear and non-linear associations between the microbiome and the outcome of interest. Model fitting, parameter tuning, and evaluation were done using `caret` package in R [314]. Parallel processing was performed using the `doParallel` [60] and `parallel` packages.

We evaluate prediction performance by performing 100 repeats of 10-fold nested cross validation, whereby within each training fold is a separate 5-fold cross-validation procedure done to perform hyperparameter selection when appropriate with parameter grids modelled after Pasolli et al. [232]. For RF, we set the number of trees to be 500, and the number of features used in each decision tree to be the square root of the number of the original features. For SVM-RBF, we tuned across a grid for the regularization parameter C (values $2^{(-5)}, 2^{(-3)}, \ldots, 2^{15}$) and the kernel width parameter γ (values $2^{(-15)}, 2^{(-13)}, \ldots, 2^{3}$). For EN, we tuned over a grid of the regularization parameter λ and the L_1 to L_2 penalty ratio α, where for each α value (spaced by 0.1) between 0 (equivalent to a LASSO model) and 1 (equivalent to a ridge regression model), we evaluate 100 lambda values chosen by the `glmnet` procedure. For SPLS, we kept the concavity parameter κ constant at 0.5 while tuning the number of components K (values $1, 2, \ldots, 10$) and the thresholding parameter η (values $0.1, 0.2, \ldots, 0.9$).

We utilize standard regression evaluation metrics include predictive R^2 and Spearman correlation coefficient (SCC). These statistics were chosen due to their ability to capture two different aspects of the regression task. Predictive R^2 captures the predicted residual sum of squares (PRESS) normalized by the total sum of squares, which can be thought of as PRESS values for a naive, intercept only model. On the other hand, SCC quantifies the monotonic association between true and predicted values, providing perspective as to whether the predicted values can capture the overall trend of the outcome. Prior to evaluation, all metabolites were back transformed to their original scale. In order to perform comparisons between models across time points and metabolites as well as ascertaining the uncertainty of each evaluation metric, a Bayesian approach as presented in [23]. Specifically, a generalized Bayesian hierarchical linear model (with identity link and gaussian standard error) in the following form was fitted for each metabolite:

$$\text{Evaluation Statistic} \sim \text{Model} + (1|\text{repeat}) + (1|\text{repeat : fold})$$

This model assumes that the distribution of the evaluation statistic as a linear function of model assignment, with random intercepts varying among repeats and for folds within each repeat. Models were fitted using implementation in the R package `tidyposterior` [151] using default weakly informative priors as described in the `rstanarm` package [35]. Using this model, a predictive posterior mean and 95% credible interval can be generated. The posterior mean is then used to rank the best performing model for each metabolite according to the evaluation metric of interest. Ranks are then aggregated using the Borda method [173] to generate Borda scores. In detail, for each metabolite, 4 points are added to the top ranked model, 3 points to the second ranked model and so on. The model with the highest total points for each metric is the most performant model aggregated across all prediction tasks.

2.6.10. Simulation design

Simulations were performed to examine the behavior of models under known association/null settings in order to validate findings.

For the first simulation scenario, a linear association between genus-level taxonomic abundance and log transformed metabolite concentrations were simulated. The predictor matrix were bootstrapped resamples of the community matrix post data processing. β coefficient values were sampled from the standard normal distribution $\mathcal{N}(0, 1)$ values for each genus would have a probability p (0.05, 0.1, 0.5, 0.95) of being 0 which determines the sparsity of the coefficients (or the level of model saturation). We generate metabolite outcome values Y following the model $Y = \beta_0 + \mathbf{X}\beta + \epsilon$ where X is the $n \times p$ simulated taxonomic predictor matrix, β is the $p \times 1$ previously defined coefficient vector, $\epsilon \sim \mathcal{N}(\mu = 0, \sigma = \sigma_\epsilon)$ is the standard normal noise vector.

Similar to Xiao et al. 2018 [327] and Shi et al. 2016 [269], we set all $\beta_0 = 6/\sqrt{10}$ and $\sigma_\epsilon = (\sigma(\beta_0 + \mathbf{X}\beta))/\text{SNR}$ where signal-to-noise ratio (SNR) are set at 0.5, 0.7, 3, 5 to simulate both situations where noise is higher than signal and vice versa. For each simulation setting, 100 data sets were generated. For the second simulation scenario, null models were assessed through a permutation procedure using the `picante` package in R as described earlier. A total of 500 permutations was performed for each model.

To evaluate the predictive capacity of models for each simulation scenario, each data set was split into a train and test set (80% train; 20% test). Within each training set, a 10-fold cross validation procedure was employed to tune any hyperparameters. Similar evaluation metrics were assessed as described in the model fitting section.

2.6.11. Metagenomic prediction with PICRUSt2

We conducted a PICRUSt2 (version 2.3.0_b) [73] analysis to investigate the potential relationship between the functional metagenome (obtained via in sillico predictions) and measurements of associated metabolites. We performed this analysis for metabolites obtained in the targeted data set. The PICRUSt2 pipeline was performed on pre-filtered ASV sequences and abundance tables using default settings. Snakemake was used to construct the computational pipeline [204].

After obtaining predicted MetaCyc pathway abundances [45], for each metabolite, we selected a subset of the pathways where the metabolite is a known product (accessed via MetaCyc SmartTables; 6-week samples: `https://metacyc.org/group?id=biocyc13-50254-3822215614`, 12-month samples: `https://metacyc.org/group?id=biocyc13-50254-3822215614`) and performed Spearman correlation analysis with the measured metabolite abundances. For each significant correlation (significance level defined as q-values below 0.05 following the Benjamini-Hochberg procedure [24]), we profiled the relative contributions of the top five Genera. Relative contribution is calculated as total abundance of a pathway assigned to that Genus across all samples divided by the total abundance of the pathway across all samples. Additionally, pairwise Spearman correlation between all identified pathway abundances and targeted metabolite concentrations was also performed. Significance is defined similarly as FDR adjusted q-values below 0.05.

Section 2.7

Availability of data and materials

The 16S rRNA gene sequencing datasets used in this study are stored in the National Center for Biotechnology Information (NCBI) Sequence Read Archive: `http://www.ncbi.nlm.nih.gov/sra` under accession number PRJNA296814. The raw and processed metabolomics data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, `https://www.metabolomicsworkbench.org` where it has been assigned Project ID PR001146. The data can be accessed directly via its project DOI: `https://doi.org/10.21228/M8K69N`. All intermediary analysis objects and scripts are available on GitHub.

Section 2.8

Acknowledgements

We thank the participants of the New Hampshire Birth Cohort Study and to the study staff, without whom this research would not be possible. We would also like to thank all members of the Microbiome research group at Dartmouth College and our collaborators at the University of North Carolina and the Marine Biological Laboratory for all advice and additional assistance in preparing this manuscript.

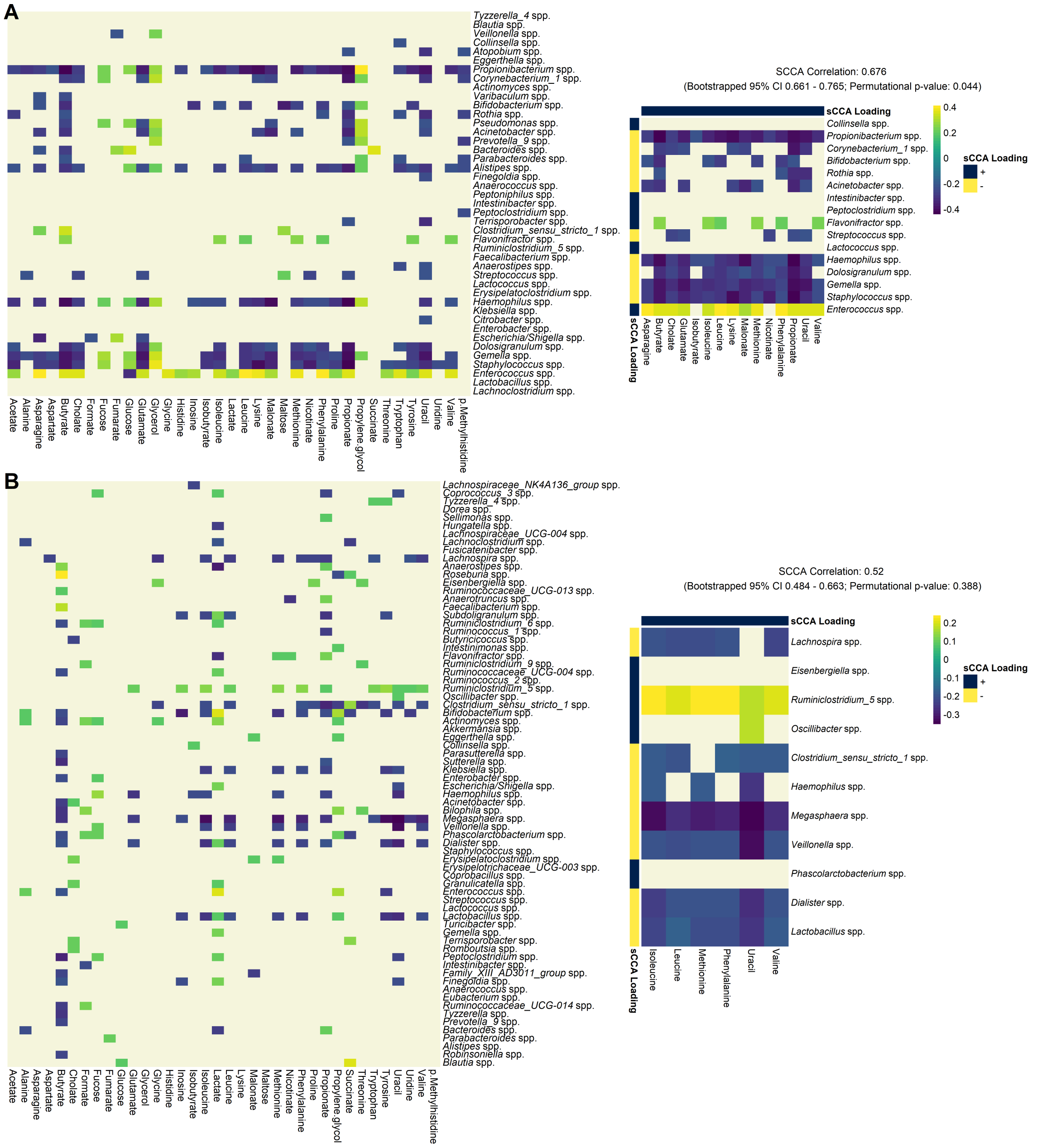

Figure 2.3: Pairwise Spearman correlation of concentration-fitted metabolites and genus-level taxonomic abundances for 6-weeks (panel A, $N = 158$) and 12-months (panel B, $N = 282$) infants. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (false discovery rate (FDR) controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sparse CCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval and permutation p-value are also reported. Significant microbiome-metabolome correlation was observed at both time points, however no significant difference was found between the time points.

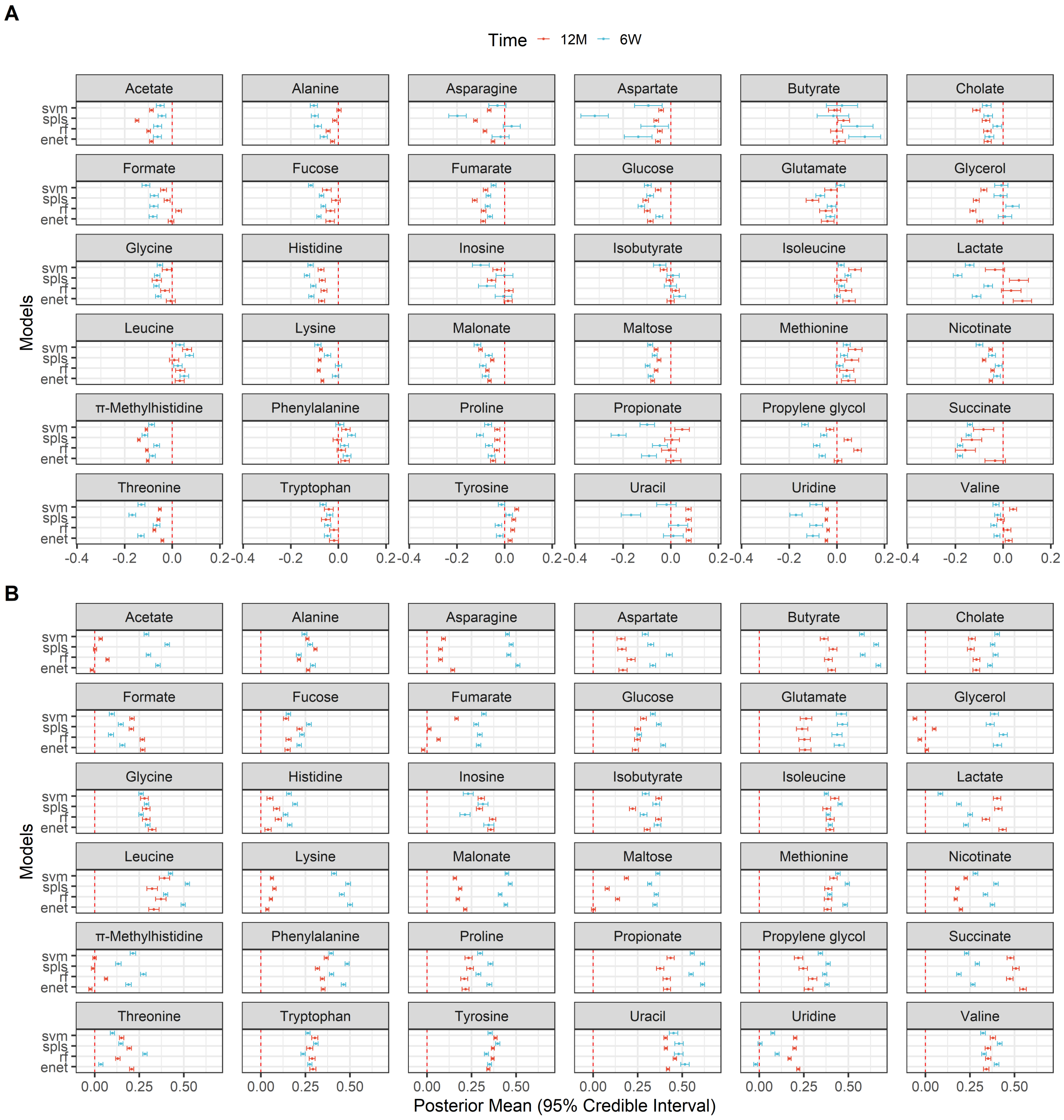

Figure 2.4: Forest plots of each prediction performance metric (R^2 - Panel A, Spearman correlation - Panel B) for each time point (6 weeks ($N = 158$), 12 months ($N = 282$)) across all 36 metabolites and 4 machine learning models. 95% credible interval and predictive posterior means were generated using Bayesian modelling of the evaluation statistic (Materials and Methods) after 100 repeats of 5-fold nested cross validation. Red vertical lines indicate a value of 0 for the evaluation metric (equivalent to null model). Metabolites were classified as predictable if the null value did not lie within the estimated 95% credible interval. For most metabolites, predictive performance was not significantly better than null models.

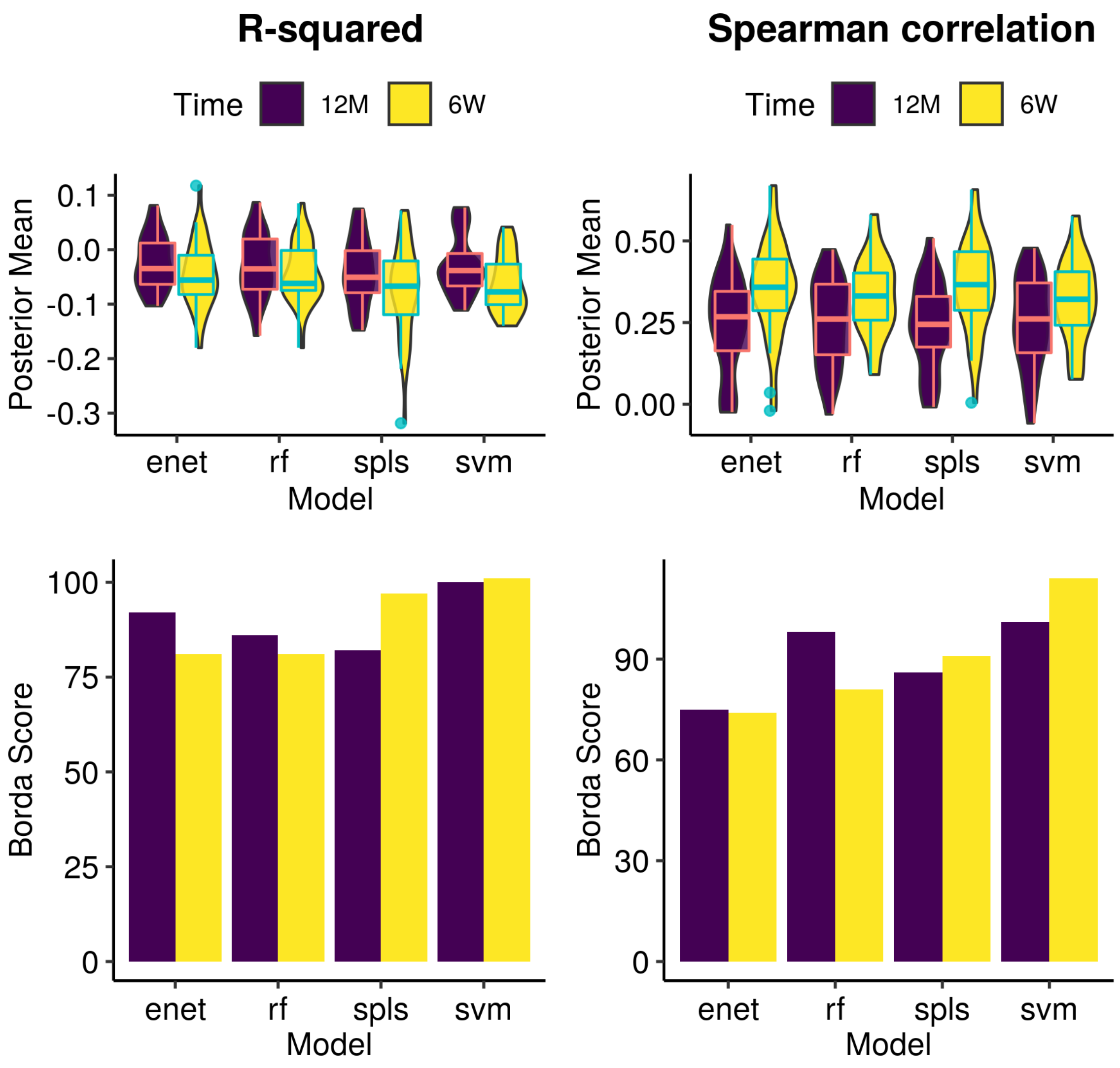

Figure 2.5: Comparative analysis predictive model performance across all metabolites in the targeted dataset for both 6-weeks ($N = 158$) and 12-months ($N = 282$) time points. Top panel shows superimposed boxplots and violin plots of the distribution of predictive posterior mean for each evaluation metric across all 36 metabolites. Bottom panels show aggregated model rankings for all metabolites using R^2 (left) and Spearman correlation (right) using Borda scores (Materials and Methods). Higher scores indicate that a model was consistently selected as a better performing. Relatively similar Borda scores and cross-metabolite average predictive performances indicate that no model was clearly the most performant. However, support vector machines (with radial basis function kernel) was highest scoring model.

Chapter 3

CBEA: Competitive balances for taxonomic enrichment analysis

This chapter was accepted for publication on May 5th, 2022 as a *Research Article* at *PLOS Computational Biology* and is currently in press. The latest pre-print of the article can be found here:

Nguyen, Q.P., Hoen, A.G., Frost, H.R. CBEA: Competitive balances for taxonomic enrichment analysis. bioRxiv. `https://doi.org/10.1101/2021.09.07.459294`

Section 3.1 Abstract

Background: Research in human-associated microbiomes often involves the analysis of taxonomic count tables generated via high-throughput sequencing. It is difficult to apply statistical tools as the data is high-dimensional, sparse, and compositional. An approachable way to alleviate high-dimensionality and sparsity is to aggregate variables into pre-defined sets. Set-based analysis is ubiquitous in the genomics literature and has demonstrable impact on improving interpretability and power of downstream analysis. Unfortunately, there is a lack of sophisticated set-based analysis methods specific to microbiome taxonomic data, where current practice often employs abundance summation as a technique for aggregation. This approach prevents comparison across sets of different sizes, does not preserve inter-sample distances, and amplifies protocol bias. Here, we attempt to fill this gap with a new single-sample taxon enrichment method that uses a novel log-ratio formulation based on the competitive null hypothesis commonly used in the enrichment analysis literature.

Methods: Our approach, titled competitive balances for taxonomic enrichment analysis (CBEA), generates sample-specific enrichment scores as the scaled log-ratio of the subcomposition defined by taxa within a set and the subcomposition defined by its complement. We provide sample-level significance testing by estimating an empirical null distribution of our test statistic with valid p-values.

Results: Herein, we demonstrate, using both real data applications and simulations, that CBEA controls for type I error, even under high sparsity and high inter-taxa correlation scenarios. Additionally, CBEA provides informative scores that can be inputs to downstream analyses such as prediction tasks.

Section 3.2

Background

The microbiome is the collection of microorganisms (bacteria, protozoa, archaea, fungi, and viruses) which co-exist with their host. Previous research has shown that changes in the composition of the human gut microbiome are associated with important health outcomes such as inflammatory bowel disease [245], type II diabetes [268], and obesity [9]. To understand the central role of the microbiome in human health, researchers have relied on high-throughput sequencing methods, either by targeting a specific representative gene (i.e. amplicon sequencing) or by profiling all the genomic content of the sample (i.e. whole-genome shotgun sequencing) [52]. Raw sequencing data is then processed through a variety of bioinformatic pipelines [40, 286], yielding various data products, one of which are taxonomic tables which can be used to study associations between members of the microbiome and an exposure or outcome of interest.

However, there are unique challenges in the analysis of these taxonomic count tables [166, 165]. The data is sparse, high-dimensional, and likely compositional [104, 166, 165]. Even though these problems are challenging, a very approachable solution is to use set-based analysis methods, also termed gene set testing in the genomics literature [139, 105]. Aggregated variables can be less sparse, and testing on a smaller number of features can reduce the multiple-testing burden. As such, gene set testing methods have been shown to increase power and reproducibility of genomic analyses. Furthermore, through the usage of functionally informative sets defined *a priori* based on historical experiments (for example MSigDB [276], and Gene Ontology [12]), gene set analysis also allows for more biologically informative interpretations.

A diverse set of methods have already been developed in this field. Traditional methods utilize the hypergeometric distribution to test for the overrepresentation of a gene set using a candidate list of genes screened based on a marginal model [105]. Unfortunately, these approaches are sensitive to the differential expression test as well as the chosen threshold when trying to select genes for the candidate list. Aggregate score methods, which are generally preferred [126], instead assign a score for each gene set based on gene-specific statistics such as z-scores or fold change. Of these approaches, methods such as GSEA [276] perform a test for each gene set at the population level, summarizing information across all samples. Conversely, methods such as GSVA [111] and VAM [95], generate enrichment scores at the sample level and are more akin to a transformation. In addition to being able to screen for enriched sets per sample, this strategy also allows for the flexible incorporation of different downstream analyses, such as fitting prediction models, or performing dimension reduction.

In microbiome research, even when no explicit enrichment analysis is performed, researchers often aggregate taxa to higher Linnean classification levels such as genus, family, or phylum. However, there is limited research done to extend existing set-based methods to microbiome relative abundance data. Some software suites, such as `MicrobiomeAnalyst`, do offer tools to perform enrichment testing with curated taxon sets [53]. However, the approach used in `MicrobiomeAnalyst` is a form of overrepresentation analysis at the population level and therefore similarly sensitive to the differential abundance approach used and p-value threshold. One of the primary challenges for adapting gene set analysis to the microbiome context is the compositional nature of the data. Sequencing technologies constrain the total number of reads, and samples are expected to have the same number of reads instead of DNA content [250, 249]. However, different samples still yield arbitrarily different total read counts [104, 212], suggesting the use of normalization methods to allow for proper comparison of feature abundances across samples. However, microbiome data sets do not follow certain assumptions that enable the cross-application of methods from similar fields (such as RNA-seq) [249, 250]. For example, DESeq2's `estimateSizeFactors` [177] assumes that the majority of genes acts as housekeeping genes with constant expression levels across samples. As such, practitioners often rely on total sum normalization to transform count data into relative proportions that sum to one [307]. Some studies have provided emprical performance evaluations supporting this normalization schema [194]. Since this approach imposes a sum constraint on the data, post normalization microbiome data sets are compositional [104], which means that the abundance of any taxon can only be interpreted relative to another. Under this scenario, log-ratio based approaches from the compositional data analysis (CoDA) literature [6] are motivated to address this issue.

Unfortunately, the standard practice for aggregating variables using element-wise summations (referred to as amalgamations in the CoDA literature), does not adequately address the compositional issue [195]. First, inter-sample Aitchison distances computed on original parts are not preserved after amalgamation [79]. This means that cluster analyses might show different results depending on the level of amalgamation and differs from the those computed from original variables. Second, amalgamations do not allow for comparison between sets of different sizes within the same experimental condition since larger sets will have larger means and variances. Third, considering that each taxa has specific measurement biases [195], an amalgamation based approach would make the bias of the amalgamated variable dependent on the relative abundance of the its constituents. In other words, if taxon 1 has abundance A_1 and bias B_1, while taxon 2 has abundance A_2 and bias B_2, then the bias of the aggregate variable (for example, a genera) is $(A_1B_1 + A_2B_2)/(A_1 + A_2)$ (see Appendix 1. from McLaren et al. [195]). This means that bias invariant approaches (such as analyses of ratios) would no longer be invariant when applied to amalgamated variables as bias now varies across samples. The alternative would be to multiply the proportions rather than to sum them [79].

Here, we present a taxon-set testing method for microbiome relative abundance data that addresses the aforementioned issues. Our approach generates enrichment scores at the sample level similar to GSVA [111] and VAM [95]. We leverage the concept of the Q_1 competitive hypothesis presented in Tian et al. [283] to formulate the enrichment of a set as the compositional balance [257] of taxa within the set and remainder taxa using multiplication as the method of aggregating proportions

[79]. This well-defined null hypothesis allows us to perform significance testing with interpretable results through estimating the empirical distribution of our statistic under the null that can also account for variance inflation due to inter-taxa correlation [324].

In the following sections, we present our approach titled competitive balances for taxonomic enrichment analysis (CBEA). First, we present the step-by-step formulation of CBEA and discuss its statistical properties. Second, we detail our evaluation strategy using both real data and parametric simulations, and the methods we are comparing against. Third, we present results on enrichment testing using CBEA for single samples as well as at the population level. Fourth, we show the performance of CBEA in downstream disease prediction. Finally, we discuss our results and the limitations of our method. An R package implementation of CBEA can be installed via Bioconductor. The development version can be found on GitHub (www.github.com/qpmnguyen/CBEA).

Section 3.3

Materials and Methods

3.3.1. Competitive balances for taxonomic enrichment analysis (CBEA)

The CBEA method generates sample-specific enrichment scores for microbial sets using products of proportions [80]. Details on the computational implementation of CBEA can be found in the Supplementary Materials. The CBEA method takes two inputs:

- **X**: n by p matrix of positive proportions for p taxa and n samples measured through either targeted sequencing (such as of the 16S rRNA gene) or whole genome shotgun sequencing. Usually **X** is generated from standard taxonomic profiling pipelines such as DADA2 [40] for 16S rRNA sequencing, or

MetaPhlAn2 [286] for whole genome shotgun sequencing. CBEA does not accept X matrices with zeroes since it invalidates the log-ratio transformation. Users can generate a dense matrix X using a method of choice, however by default mode CBEA will add a pseudocount of 10^{-5} if zeroes are detected in the matrix.

- **A**: p by m indicator matrix annotating the membership of each taxon p to m sets of interest. These sets can be Linnean taxonomic classifications annotated using databases such as SILVA [247], or those based on more functionally driven categories such as tropism or ecosystem roles ($A_{i,j} = 1$ indicates that microbe i belongs to set j).

The CBEA method generates one output:

- **E**: n by m matrix indicating the enrichment score of m pre-defined sets identified in **A** across n samples.

The procedure is as follows:

(a) **Compute the CBEA statistic**: Let $\mathbf{M}$ be a n by m matrix of CBEA scores. Let $\mathbf{M}_{i,k}$ be CBEA score for set k and sample i:

$$\mathbf{M}_{i,k} = \sqrt{\frac{\sum_k A_{ik}(p - \sum_k A_{ik})}{p}} \ln \left(\frac{g(\mathbf{X}_{i,j} | \mathbf{A}_{j,k} = 1)}{g(\mathbf{X}_{i,j} | \mathbf{A}_{j,k} \neq 1))} \right) \tag{3.1}$$

where $g(.)$ is the geometric mean. This represents the ratio of the geometric mean of the relative abundance of taxa assigned to set k and remainder taxa.

(b) **Estimate the empirical null distribution** Enrichment scores represent the test statistic for the Q_1 null hypothesis H_o that relative abundances in $\mathbf{X}$ of members of set k are not enriched compared to those not in set k. Since the distribution of CBEA under the null vary depending on data characteristics (Fig 3.1), an empirical null distribution will be estimated from data.

- **Compute the CBEA statistic on permuted and un-permuted $\mathbf{X}$**. Let $\mathbf{X}_{perm}$ be the column permuted relative abundance matrix, and $\mathbf{M}_{perm}$ be the corresponding CBEA scores generated from $\mathbf{X}_{perm}$. Similarly, we have $\mathbf{M}_{unperm}$ be CBEA scores generated from $\mathbf{X}$.
- **Estimate correlation-adjusted empirical distribution for each set**. For each set, a fit a parametric distribution to both $\mathbf{M}_{perm}$ and $\mathbf{M}_{unperm}$. The location measure estimated from $\mathbf{M}_{perm}$ and the spread measure estimated from $\mathbf{M}_{unperm}$ will be combined as the correlation-adjusted empirical null distribution $\mathbf{P}_{emp}$ for each set. Two available options are the normal distribution and the mixture normal distribution. For the normal distribution, parameters were estimated using the method of maximum likelihood implemented in the `fitdistr` package [66]. For the mixture normal distribution, parameters were estimated using an expectation-maximization algorithm implemented in the `mixtools` package [22].

(c) **Calculate finalized CBEA scores with respect to the empirical null**. Enrichment scores $\mathbf{E}_{i,k}$ are calculated as the cumulative distribution function (CDF) values or z-scores with respect to $\mathbf{P}_{emp}$ distribution. Raw p-values can be calculated by subtracting $\mathbf{E}$ from 1.

3.3.2. Properties of CBEA

CBEA and balances between groups of parts. The CBEA statistic is based on the multiplication-based aggregation approach used to calculate balances between groups of parts [79]. These balances are computed using the isometric log ratio (ILR) transformation [80] formula. For a given balance i splitting variables across sets R

and S, we have the balance coordinate x_i^* as:

$$x_i^* = \sqrt{\frac{rs}{r+s}} \log \left(\frac{g(X_{j|j \in R})}{g(X_{j|j \in S})} \right) \tag{3.2}$$

where r and s are the cardinalities of sets R and S respectively, $g(z)$ is the geometric mean, and X_j are values of the original predictors with indexes defined by membership in R and S.

CBEA belongs to a set of methods that seek to leverage compositional balances for the analysis of microbiome data [304, 257, 213, 272]. Unlike methods such as PhILR [272], CBEA does not present an orthonormal basis for the complete ILR transformation (such as a a sequential binary partition) [80]. Therefore, it is not a subclass of the ILR transformation and is adjacent to this approach. A similar method to CBEA would be phylofactor [304]. However, instead of performing an optimization procedure to identify interesting balances, CBEA constructs balances *a priori* using pre-defined sets, and formulates the enrichment of a set as the scaled log-ratio between the center of the subcomposition represented by microbes within the set and the center of the subcomposition represented by remainder taxa. This formulation aligns with the Q_1 null hypothesis from the gene set testing literature [283].

Estimating the null distribution. We can assume that the CBEA statistic, similar to other log-ratio based transforms, follows a normal distribution [80, 5]. However, when applying CBEA for hypothesis testing at the sample level, it is expected that the researcher would be testing a large number of hypotheses. Under the assumption that the number of truly significant hypotheses is low, Efron [78] showed that estimating the null distribution of the test statistic directly (termed the empirical null distribution) is much more preferable than using the theoretical null due to unobserved confounding effects inherently part of observational studies. As such, to perform significance testing using CBEA, we also estimated the null distribution from observed raw CBEA variables.

This assumption is also supported by preliminary simulation studies (detailed below). We simulated microbiome taxonomic count data under the global null across different data features and compute raw CBEA scores and compute kurtosis and skewness in Fig 3.1A. We found that the characteristics of the null change depending on sparsity and inter-taxa correlation. Sparsity seems to drive the distribution to be more positively skewed while inter-taxa correlation encourages platykurtic (negative kurtosis). The effect is most dramatic under both high inter-taxa correlation and sparsity. This heterogeneity further supports the decision to estimate an empirical null distribution, as suggested by Efron [78].

Additionally, the degree of kurtosis and skewness also suggests that the normal distribution itself might not be a good approximation of the null. To address this issue, we also evaluated a two-component normal mixture distribution. The goodness of fit of the mixture normal and the normal distribution using Kolmogorov-Smirnov (KS) test statistic computed on fitted normal and mixture normal distribution when fitted on CBEA scores in simulation scenarios under the global null is shown in Fig 3.1B. We can see that the mixture normal distribution is a better fit (lower KS scores) than the normal distribution across both sparsity and correlation settings.

We performed our empirical null estimation by fitting our distribution of choice and computing relevant parameters on raw CBEA scores on taxa-permuted data (equivalent to gene permutation in the gene expression literature). As such, the null distribution is characterized by scores computed on sets of equal size with randomly drawn taxa.

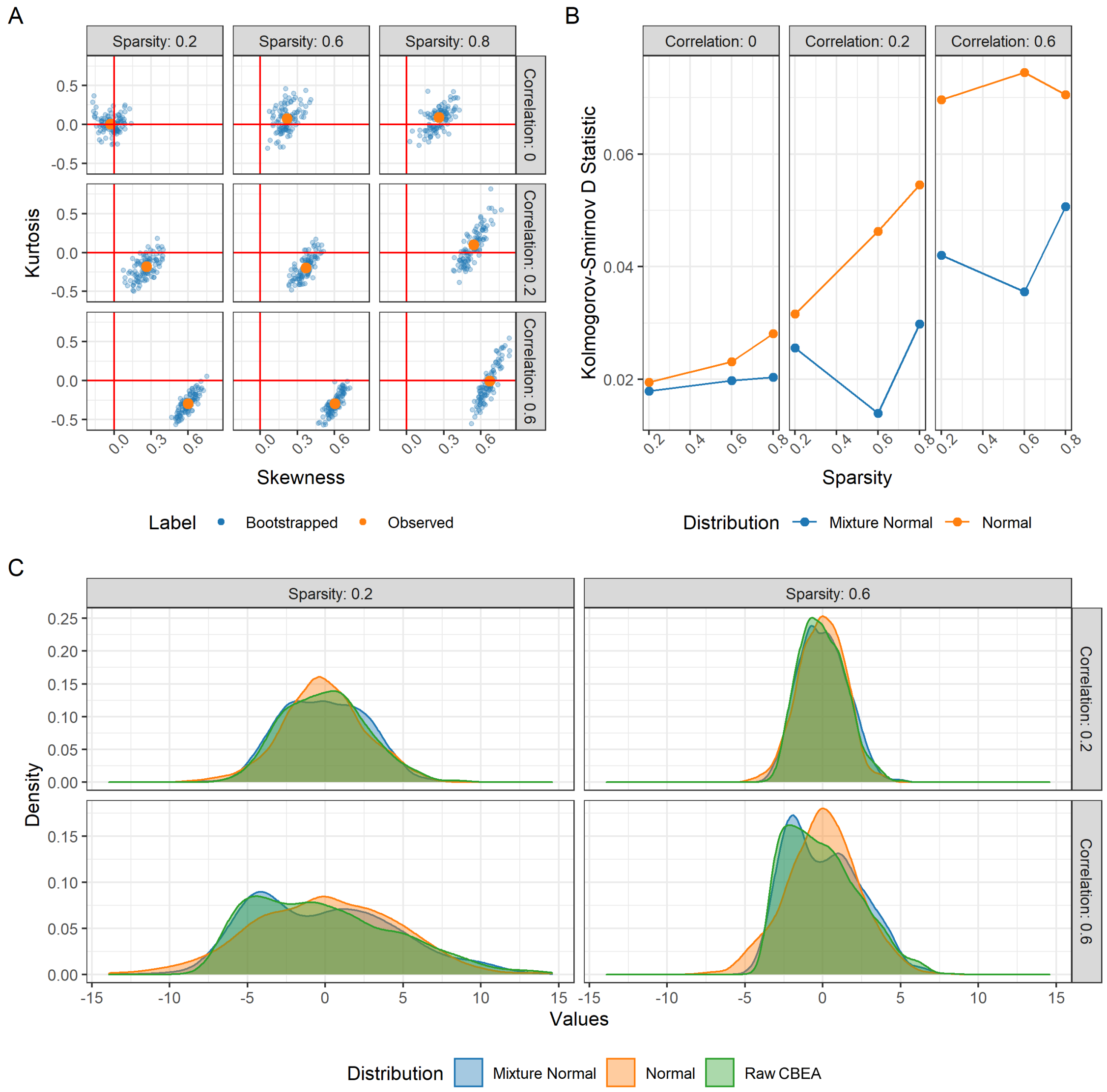

Figure 3.1: Properties of the null distribution of CBEA under the global null simulations. Panel **(B)** presents kurtosis and skewness of CBEA scores while panel **(A)** presents the goodness of fit (as Kolmogorov-Smirnov D statistic) for mixture normal and normal distributions. Panel **(C)** is a density plot of the shape of the null distribution. Results indicated the necessity of estimating an empirical null and demonstrating that the mixture distribution was the better fit compared to the basic normal.

Variance inflation due to inter-taxa correlation. When taxa within a set are highly correlated, the variance of the sample mean of taxon-wise statistics is inflated. Without loss of generalizability, for a set of taxa with taxon-specific statistics

$x_1, \ldots, x_p$ we have the variance of the mean $\bar{x}$ to be:

$$Var(\bar{x}) = \frac{1}{m^2}\left(\sum_{i=1}(\sigma_i^2) + \sum_{i<j}\rho_{ij}\sigma_i\sigma_j\right) \tag{3.3}$$

where σ_i is the standard deviation of taxon i and ρ_{ij} is the correlation between i and j. The second term of (3.3) is the correlation dependent variance component, which goes to 0 if there is no correlation. The CBEA statistic follows a similar pattern. Since the geometric mean of a set of variables is equivalent to the exponential of the arithmetic mean of their logarithms, we can re-write CBEA score for a set k with size K as follows:

$$M_{i,k} = \sqrt{\frac{K(p-K)}{K+(p-K)}}\left(\overline{\log X_{i,j|j\in K}} - \overline{\log X_{i,j|j\notin K}}\right) \tag{3.4}$$

where p is the overall number of taxa, j is the index of a taxa and K is the set of indices of taxa in set k. The CBEA statistic then looks similar to a t-statistic for difference in means of log-transformed proportions. As such, the pooled variance of CBEA is dependent on the variance inflation of both mean components $\overline{\log X_{i,j|j\in K}}$ and $\overline{\log X_{i,j|j\notin K}}$. The result of this variance inflation is inflated type I error since highly correlated sets are also detected as significantly enriched.

However, as Wu et al. [324] showed, performing column permutation to estimate the null distribution of a competitive test statistic doesn't allow for adequate capture of this variance inflation factor since the permutation procedure disrupts the natural correlation structure of the original variables. It is important to address this problem since there is strong inter-taxa correlation within the microbiome [153]. Our strategy for addressing this issue is to use the location (or mean) estimate from the column permuted raw score matrix with the spread (or variance) estimate taken from the original un-permuted scores. This still allows us to leverage the null distribution generated via column permutation while using the proper variance estimate taken from scores where the correlation structure has not been disrupted. As such, this procedure assumes that the variance of the test statistic under the alternate hypothesis is the same as that of the null. Details of the computational implementation to this estimation process can be found in the Supplementary Note 1.

However, set-based analysis is an exploratory approach that can help generate functionally informative hypotheses, and as such users might not want strict type I error control in favor of higher power. This is especially true for competitive hypotheses, where its stricter formulation compared to the self-contained approach implies that the test naturally has lower power [105, 2]. Furthermore, sets that are highly correlated compared to background can be biologically relevant. Therefore, CBEA provides an option for users to specify whether correlation adjustment is desired.

Section 3.4
Evaluation

We based our evaluation strategy on gene set testing benchmarking standards set by Geistlinger et al. [100] and utilized the same approaches whenever possible. All data sets are obtained from either the `curatedMetagenomicData` [231] and `HMP16SData` [262] R packages (2020-10-02 snapshot), or downloaded from the Qiita platform [106]. All code and data sets used for evaluation of this method is publicly available and can be found on GitHub (`www.github.com/qpmnguyen/CBEA_analysis`). Additional packages used to support this analysis includes: `tidyverse` [312], `pROC` [259], `phyloseq` [197], `mia` [85], `targets` [155].

3.4.1. Statistical significance

We evaluate the inference procedure of CBEA compared to alternate methods using two approaches: randomly sampled taxa sets and sample label permutation. These analyses were performed on the 16S rRNA gene sequencing of the oral microbiome from the Human Microbiome Project [59, 245]. This data set contains 369 samples split into two subsites: supragingival and subgingival. We processed this data set by removing all samples with total read counts less than 1000 and OTUs whose presence (at least 1 count) is in 10% of samples or less.

Sample-level inference. Due to CBEA's self-contained null hypothesis, we can perform inference at the sample level for the enrichment of a set. We evaluated this application by generating one random taxon-set of different sizes $S \in \{20, 50, 100, 150, 200\}$ across 500 iterations. Random sets can act as our estimate for type I error since this matches the CBEA null hypothesis stated in Materials and Methods, where we expect within each sample, sets of randomly drawn taxa should not be significantly enriched compared to the remainder background taxa. For this evaluation, we estimated type I error as the fraction of samples where our random set is detected as significant at a p-value threshold of 0.05 with confidence bands computed from the standard error across all iterations. Additionally, this analysis also tests whether CBEA is sensitive to different set sizes.

Population-level inference. We can perform enrichment testing at the population level by generating corresponding sample level CBEA scores and performing a two-sample test such as Welch's t-test. In order to evaluate CBEA under this context, we generated CBEA scores of sets representing genus-level annotation in above gingival data set [59, 245] and applied a t-test to test for enrichment (similar to GSVA [111]) across a randomly generated variable indicating case/control status (repeated 500 times). Type I error is estimated as the fraction of sets per iteration found to be significantly enriched with confidence bands computed from the standard error across all iterations. In addition, we also performed a random set analysis assessment, where we generated 100 sets of different set sizes $S \in \{20, 50, 100, 150, 200\}$ and evaluated the fraction of genera that were found to be differentially abundant across the original labels (supragingival versus subgingival subsite). 95% confidence intervals were computed using the Agresti-Couli approach [3].

3.4.2. Phenotype relevance

We want to evaluate whether sets found to be significantly enriched by CBEA are relevant to the research question. To perform this assessment, we relied on the gingival data set mentioned above [59, 245]. This data set was chosen because its clear biological interpretation can serve as the ground truth. Specifically, we expect aerobic microbes to be enriched in the supragingival subsite where the biofilm is exposed to the open air, while conversely anaerobic microbes thrive in the subgingival site [281]. Genus-level annotations for microbial metabolism from Beghini et al. [20] were obtained from the GitHub repository associated with Calagaro et al. [38]. For sample-level inference, we assessed power as the fraction of supragingival samples where aerobic microbes are significantly enriched. For population-level inference, power is the fraction of sets representing genus level taxonomic assignments that were significant across subsite labels.

In addition to statistical power, we also assessed phenotype relevance through evaluating whether highly ranked sets based on CBEA scores are more likely to be enriched according to the ground truth. This is represented by the area under the receiving operator curve (AUROC/AUC) scores computed on CBEA scores against true labels (similar approach was used to evaluate VAM [95]). DeLong 95% confidence intervals for AUROC [67] were obtained for each estimate.

3.4.3. Disease Prediction

CBEA scores can also be used for downstream analyses such as disease prediction tasks. We utilized two data sets for this evaluation:

(a) Whole genome sequencing of stool samples from inflammatory bowel disease (IBD) patients in the MetaHIT consortium [225]. This data set contains 396 samples from a cohort of European adults, where 195 adults were classified as having IBD (which includes patients diagnosed with either ulcerative colitis or Crohn's disease). We processed this data by removing all samples with less than 1,000 total read counts as well as any OTU that was present (with non-zero proportions) in 10% of the samples or less. Prior to model fitting, we back-transformed relative abundances into count data (to align the format with our 16S rRNA gene sequencing data set) using the provided total number of reads aligned to MetaPhlan marker genes (per sample).

(b) 16S rRNA gene sequencing of stool samples from IBD patients in the RISK cohort [101]. This data set contains 16S rRNA gene sequencing samples from a cohort of pediatric patients (ages < 17) from the RISK cohort enrolled in the United States and Canada. Of the 671 samples obtained, 500 samples belong to patients with IBD. We processed this data set by removing all samples with less than 1,000 total read counts as well as any OTU that was present (at least 1 count) in 10% of the samples or less.

We evaluate disease prediction performance by fitting a random forest model [33] using as inputs CBEA scores to classify samples of patients with IBD and healthy controls. Random forest was chosen as a baseline learner due to its flexibility as an out-of-the-box model that is easy to fit. In this instance we evaluated predictive performance of a default random forest model (without hyperparameter tuning) AUROC after 10-fold cross validation. Additionally, we utilized SMOTE to correct for class imbalances [47]. Implementation was done using the `tidymodels` suite of packages [152].

3.4.4. Comparison Methods

We benchmarked the statistical properties of CBEA against existing baseline approaches. For sample-level inference analyses, utilized the Wilcoxon rank-sum test, which non-parametrically tests the difference in mean counts between taxa from a pre-defined set and its remainder similar to CBEA. For assessments at the population level, we compared CBEA against performing a standard test for differential abundance with set-level features generated via element-wise summations instead. We chose DESeq2 [177] and corncob [191] because they represent both methods extrapolated from RNA-seq [197] and those developed specifically for microbiome data.

Since disease prediction models and rankings-based phenotype relevance analyses seek to evaluate the informativeness of CBEA scores instead of relying on computing p-values, we compared performance against other single sample based approaches from the gene set testing literature, specifically ssGSEA [17] and GSVA [111]. Additionally, for evaluating prediction, we also compared performance against a standard analysis plan where inputs are count-aggregated sets with the centered log-ratio (CLR) transformation.

Section 3.5 Results

In this section, we present results for evaluating statistical significance, phenotype relevance, and predictive performance. In addition to real data, we also evaluated models based on parametric simulations, where results can be found in the Supplemental Materials.

3.5.1. Statistical Significance

Inference at the sample level. CBEA provides significance testing at the sample level through a self-contained competitive null hypothesis. Generating random sets approximate the global null setting where within each sample, sets generated by randomly sampling taxa should not be significantly more enriched than remainder taxa.

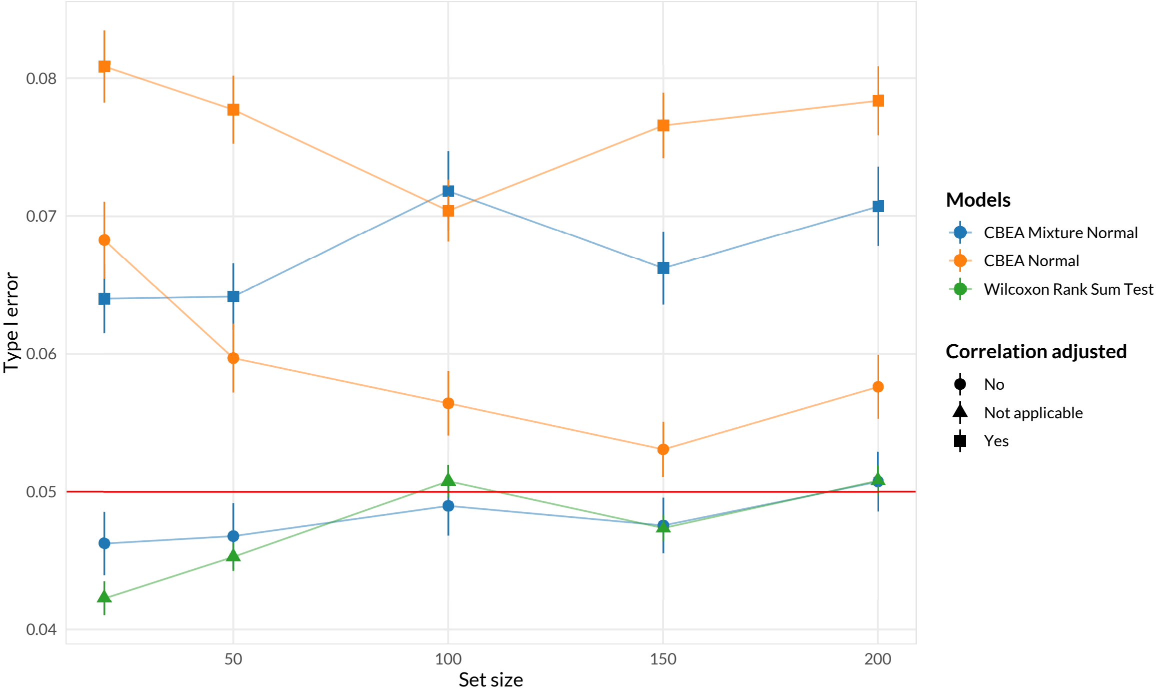

Figure 3.2: Random taxa set analyses for inference at the sample level of CBEA under different parametric assumptions compared against a Wilcoxon rank-sum test. Type I error (y-axis) was evaluated by generating random sets of different sizes (x-axis) (500 replications per size) and computing the fraction of samples where the set was found to be significantly enriched at $\alpha = 0.05$. Error bars represent the mean type I error ± sample standard error computed across 500 replications of the experiment. Only the unadjusted CBEA with the mixture normal distribution and the Wilcoxon rank sum test were able to control for type I error at 0.05. All approaches are invariant to set sizes.

Fig 3.2 demonstrates type I error of sample-level inference evaluated using the random set approach. The Wilcoxon rank sum test and unadjusted CBEA under mixture normal assumption demonstrated good type I error control at the appropriate α level. This fits with our expectations since the mixture normal distribution has a much better fit than the normal distribution especially at the tails of the empirical distribution (Fig 3.1). However, other variants of CBEA demonstrated inflated type I error, especially correlation adjusted variants compared to their unadjusted counter parts. Encouragingly, all methods demonstrate consistent performance across all set sizes, with a slight increase in type I error at the highest levels.

Interestingly, simulation results (Fig C.1) showed an opposite pattern. Adjusted approaches were good at controlling for type I error, especially under the low inter-taxa correlation values within the set (similar to generating random sets where the natural correlation structure is disrupted). In these simulations, unadjusted approaches and the Wilcoxon rank sum test had significant type I error inflation with increasing correlation. All approaches seem to be invariant to the level of data sparsity.

Inference at the population level. Similar to other single sample approaches to gene set testing such as GSVA [111], we can perform inference at the population level by utilizing a two-sample difference in means test. Here, we evaluate using CBEA scores generated under different settings with Welch's t-test in a supervised manner to assess whether a set is enriched across case/control status.

Fig 3.3 shows results for this scenario using both random sample label and random set evaluations. The random sample label approach (Fig 3.3A) provides a controlled setting where we can estimate type I error rate controlled at $\alpha = 0.05$. Across all replications, CBEA methods were able to control for type I error at the nominal threshold of 0.05, with CBEA raw scores being the most performant. Neither output types, correlation adjustment, nor distributional assumption improved performance

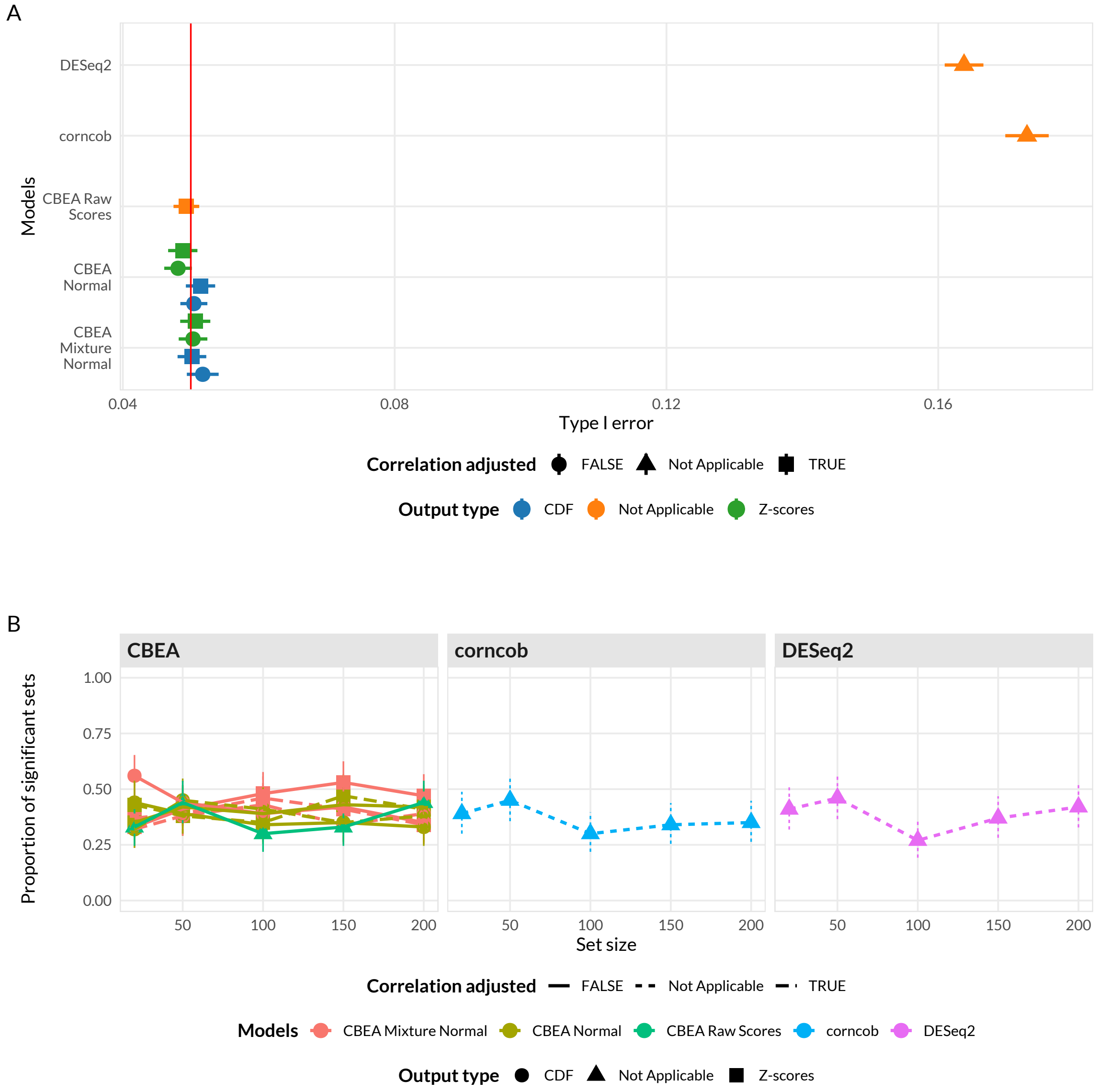

Figure 3.3: Random sample label (**A**) and random set (**B**) analyses for population level inference. (**A**) Type I error (x-axis) was estimated as the overall fraction of sets found to be enriched $\alpha = 0.05$ using randomly generated sample labels (500 permutations). Error bars represent the mean type I error ± sample standard error. (**B**) Proportion of significant sets (y-axis) using 100 randomly generated sets of different set sizes (x-axis). Confidence intervals computed using Agresti-Couli method for binomial proportions. For sample label permutation (**A**), all CBEA approaches were able to control for type I error but not for corncob and DESeq2. For random set analyses (**B**), all approaches demonstrate similar rate of accepting significant sets and were invariant to overall set size.

values. Surprisingly, DESeq2 and corncob both exhibit significantly inflated type I error.

We also assessed the impact of set-size on the inference procedure by testing for enrichment using the original sample labels but with randomly sampled sets of different sizes (Fig 3.3B). Overall we observed very similar values across CBEA as well as corncob and DESeq2, suggesting that no individual method is systematically identifying too many significant sets. Additionally, similar to analogous analyses at the sample level, no approach was significantly sensitive to changes in set sizes.

3.5.2. Phenotype Relevance

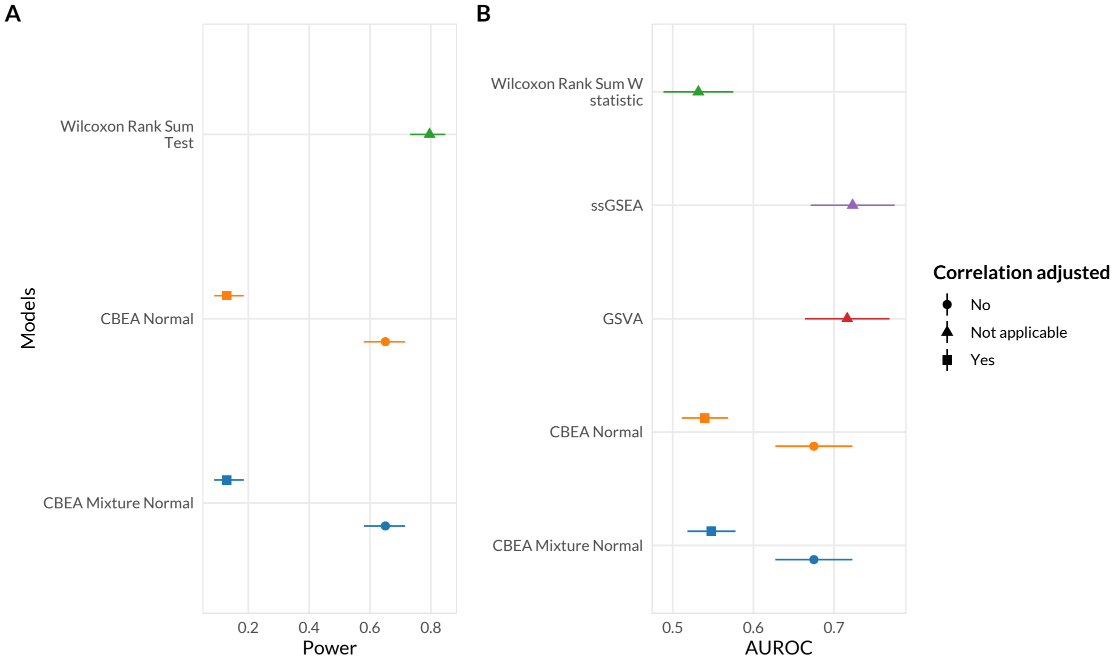

Figure 3.4: Statistical power (**A**) and score rankings (**B**) to assess phenotype relevance. (**A**) Power (x-axis) was estimated as the overall fraction of aerobic microbes found to be enriched in supragingival samples at $\alpha = 0.05$. 95% confidence intervals were computed using the Agresti-Couli approach for binomial proportions. (**B**) Score rankings were evaluated by comparing computed scores against true values using AUROC (x-axis). DeLong 95 % confidence intervals for AUROC were computed.

Inference at the sample level. In Fig 3.4, we evaluate whether sets found to be significant by CBEA are relevant to the phenotype of interest. We leveraged the gingival data set as stated in Evaluation section where we know beforehand that aerobic microbes are more likely to be enriched in supragingival subsite samples and vice versa.

We estimated statistical power using this data set as the fraction of supragingival samples where the set representing aerobic microbes were significantly enriched. We observed that adjusted CBEA approaches demonstrate much lower power compared to the Wilcoxon rank-sum test and unadjusted variants. This is surprising given the fact that in statistical significance analyses, the adjusted CBEA approach provides inflated type I error, especially if the normal distribution assumption was chosen, which indicates a mismatch in estimating the null distribution since a high type I error did not result in increased power.

We also evaluated phenotype relevance by assessing whether enriched sets according to ground truth are preferentially ranked higher using assigned continuous scores (instead of performing a hypothesis test). This aspect is captured through computing AUROC values comparing computed enrichment scores and true labels. Consistent with the previous type I error evaluation, adjusting for correlation did not improve performance, where obtained AUROC were around 0.5 and at the same level as the benchmark Wilcoxon rank sum statistic. Unadjusted methods were much better at ranking true enriched sets, however the mean AUROC values are lower than alternate single sample enrichment methods (GSVA [111] and ssGSEA [17]) even though this difference is not significant due to overlapping confidence intervals.

The above results were replicated in simulation studies where we observed that adjusted approaches were very conservative and demonstrated significantly lower power (Fig C.3) with increasing correlation even at the highest evaluated effect sizes. When assessing score rankings, the performance of CBEA was closer to ssGSEA and GSVA compared to real data evaluations, however all single sample approaches were much better than using the W statistic from the Wilcoxon Rank Sum test.

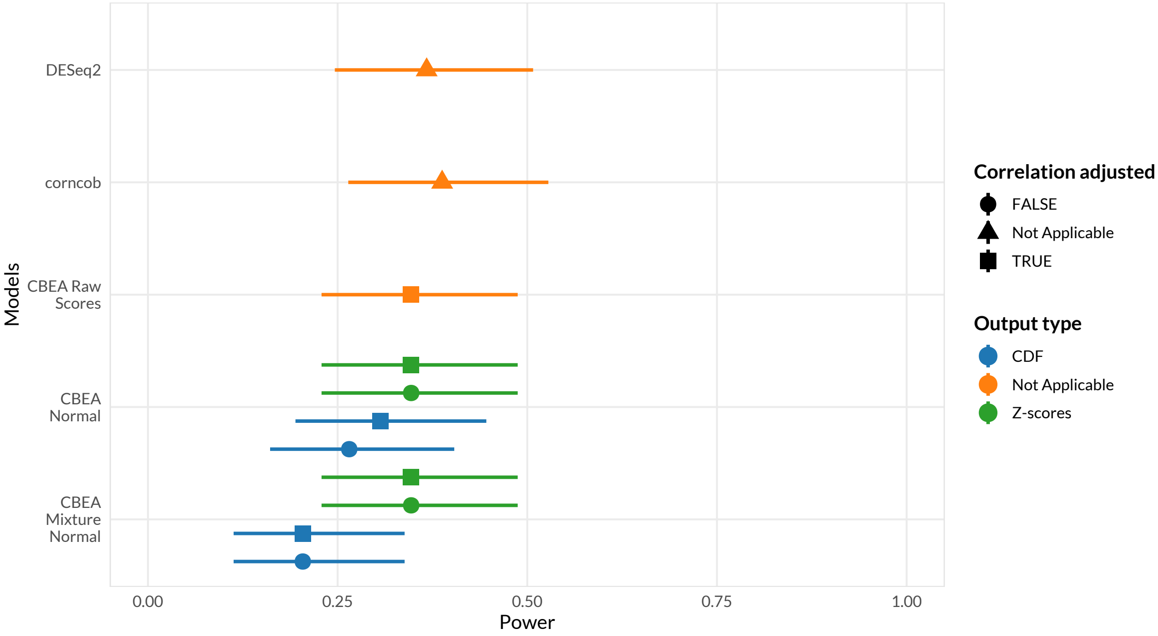

Figure 3.5: Statistical power to assess phenotype relevance of inference tasks at the population level. Power (x-axis) was estimated as the overall fraction of sets representing genera that are aerobic or anaerobic microbes found to be differentially enriched across sample type (supragingival or subgingival). 95% confidence intervals were computed using the Agresti-Couli approach for binomial proportions.

Inference at the population level. We also assessed statistical power for population level inference scenarios using a similar approach. Here, enrichment scores for sets representing all identified genera were computed, and power was estimated as the fraction of sets found to be differentially enriched across sample site labels (supragingival or subgingival). We compared these results against performing a differential abundance test of genus level features generated via sum-based approaches. Results are shown in Fig 3.5. Some CBEA variants, such as CDF outputs for the mixture normal distributional assumption, did not correctly detect as many significant sets as DESeq2 or corncob despite very close performance values. Using raw CBEA scores was best approach, however it did not exceed values obtained from DESeq2 and corncob.

3.5.3. Disease Prediction

Since CBEA can generate informative scores that can discriminate between samples with inflated counts for a set (Fig 3.4), we wanted to assess whether they can also act as useful inputs to predictive models. In this section we assessed the predictive performance of a standard baseline random forest model [33] with different single sample enrichment scoring methods as inputs (CBEA, ssGSEA, and GSVA). Additionally, we also compared predictive performance of using these scores against the a standard approach of using the centered log ratio transformation (CLR) on taxon sets aggregated via abundance summations.

We fit our model to two data sets with a similar disease classification task of discriminating patients who were diagnosed with IBD (includes both Crohn's disease and ulcerative colitis) using only microbiome taxonomic composition. The two data sets represent different microbiome sequencing aprpaoches: the Gevers et al. [101] data set uses 16S rRNA gene sequencing, while the Nielsen et al [225] data set uses whole genome shotgun sequencing.

Fig 3.6 illustrates the performance of our model with AUROC as the evaluation criteria. In the 16S rRNA data set, the best performing CBEA variant (CDF values computed from an unadjusted mixture normal distribution) outperforms both GSVA and ssGSEA but not the standard CLR approach. Interestingly, in the whole genome sequencing data set, CBEA outperforms CLR, but was similar in performance to GSVA. However, due to large confidence intervals, no method significantly out-performed other evaluated approaches. As such, these results indicate that, for a given pre-determined collection of sets, CBEA generated scores are can be informative and provide competitive performance when acting as inputs to disease predictive

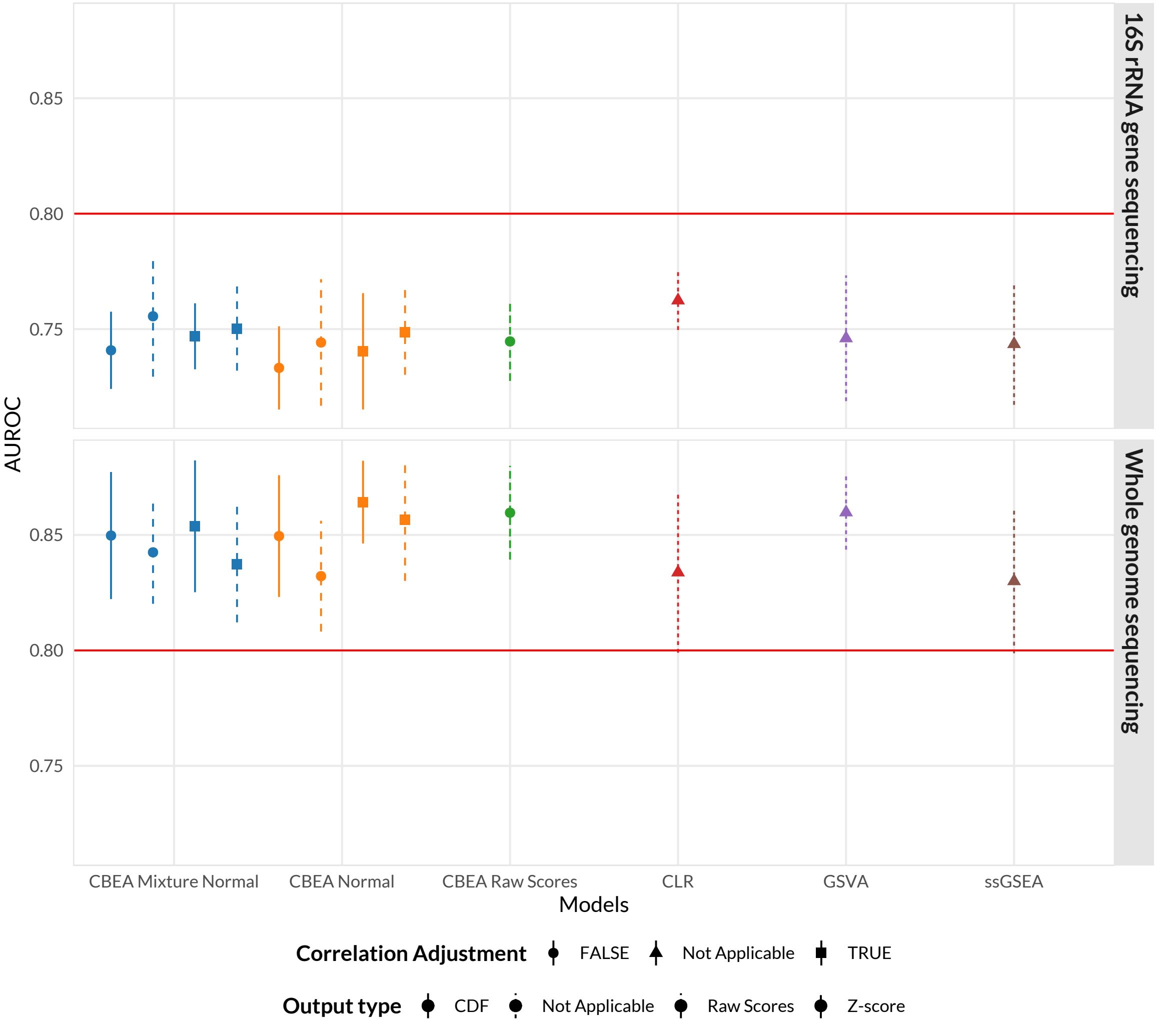

Figure 3.6: Predictive performance of a naive random forest model trained on CBEA, ssGSEA, GSVA generated scores as well as the standard CLR approach on predicting patients with inflammatory bowel disease versus controls using genus level taxonomic profiles. Data sets used span both 16S rRNA gene sequencing (Gevers et al. [101]) and whole-genome shotgun sequencing (Nielsen et al. [225]). CBEA performs better than GSVA and ssGSEA but not as well as CLR, with the exception of the whole genome sequencing data set.

models. Simulation studies (Fig C.5) showed similar results, however CBEA more consistently underperformed compared to CLR across all scenarios. Interestingly, the performance gap decreases with increasing sparsity levels and correlation.

Section 3.6

Discussion

3.6.1. Inference with CBEA

CBEA is a microbiome-specific approach to generate sample specific enrichment scores for taxonomic sets defined *a priori*. The formulation of CBEA as a comparison between taxa within the set and its complement corresponds to the competitive null hypothesis in the gene set testing literature [283]. Since this null hypothesis is self-contained per sample, this allows users perform enrichment testing at the sample level. Additionally, in combination with a difference in means test, CBEA can also test for enrichment at the population level across case/control status similar to GSVA [111].

For single-sample analyses, we demonstrated that the CBEA approach (unadjusted with mixture normal parametric assumption) was able to control for type I error at the nominal level of 0.05 under the global null (Fig 3.2) while also demonstrating adequate power (Fig 3.4). This performance is consistent across different set sizes as well as our prior distributional fit analyses (Fig 3.1), where the mixture normal displayed superior fit to the null distribution. Unfortunately, other variants of CBEA demonstrated neither good type I error control nor power. Interestingly, while the adjusted methods showed poor performance in real data evaluations (Fig 3.2), in simulation studies (Fig C.1, Fig C.3) these approaches were able to control for type I error well with the trade-off of much lower power. For the population-level inference task, CBEA also performed very well. Under the permutation global null, representing genera abundance using CBEA scores in combination with Welch's t-test controls for type I error at the correct α threshold while also keeping respectable power. Since the population level enrichment test is equivalent to a differential abundance test using set-based features, we compared the CBEA approach against using element-wise summations with corncob [191] and DESeq2 [177] to test for set-level differential abundance. We chose DESeq2 because it is an older approach from the bulk RNA-seq literature that has strong support for usage in microbiome taxonomic data [197]. Alternately, corncob is a newer method developed specifically for microbiome taxonomic data sets, where taxonomic counts are modeled directly using a beta-binomial distribution instead of relying on normalization via size factor estimation. We observed that using this approach resulted in an inflated type I error compared to all variants of CBEA (Fig 3.2), yet did not improve power (Fig 3.4). Results for CBEA approaches were replicated in simulation analyses, however for corncob and DESeq2 we observed an opposite effect: in simulation experiments, both methods show good type I error control but low power (Fig C.2, Fig C.4).

We hypothesized that the discrepancy between simulation and real data evaluations could be due to differences in our assumptions regarding the data generating process that informed our simulation schema. For the non-zero component of each taxon, we sampled from the same negative binomial distribution where designated enriched taxa were generated with inflated means (but the same dispersion). These marginals were simulated to account for block exchangable correlation within the enriched set only. This might have affected our results in two ways. First, our simulation scenario ensures that all designated non-enriched taxa are identical to each other. This is not the case for real data, where our null scenario involves randomly sampled sets that might by chance all have taxa with inflated means compared to remainder taxa. This is represented in Fig C.7, where the distribution of type I error across 500 replications is right skewed for underperforming CBEA variants, indicating that these approaches are much more sensitive compared to the Wilcoxon rank sum test or unadjusted CBEA with mixture normal distribution. Second, as described in the Background section, we did not consider taxon-specific biases that distort the observed relative abundance of taxa compared to true values [195]. In the context of sum-based aggregations, the resulting bias of the aggregated taxon-set is dependent on the relative abundances of the contributing taxa (Appendix I in [195]). Conceptually, this means that measurement error for a taxon-set is different across samples as relative abundance of contributing taxa changes, leading to issues when attempting to perform inference. As such, we expect methods like corncob or DESeq2 when performed on such sum aggregates in the presence of taxon-specific biases to have inflated type I error compared to our multiplication based approach. This also explains why conversely in simulation studies, where taxon-specific biases are absent, corncob and DESeq2 performed better.

3.6.2. Downstream analysis using predictive models

The sample-level enrichment scores generated by the CBEA method can be used in downstream analyses such as disease prediction. We evaluated whether CBEA can be used to generate set-based features for disease prediction models. We fit a basic random forest model [33] to predict continuous and binary outcomes using CBEA generated scores as inputs. Similar to our inference analysis, we compared CBEA against both ssGSEA and GSVA. Additionally, we also evaluated CBEA with the approach where counts of a set were aggregated using sums and applied the centered log-ratio transformation (CLR). This is because CLR is considered standard practice in using microbiome variables as predictors for a model [104]. Results showed that CBEA generate scores perform well across both real data and simulation scenarios. Since predictive models consider the effect of variables jointly (and in the case of random forest, consider interactions as well), good performance indicates that CBEA scores can capture joint distribution of sets, enabling both uniset and multi-set type analyses. Comparatively, CBEA generated scores outperformed other enrichment score methods (GSVA and ssGSEA), suggesting that it is more tailored for microbiome taxonomic data sets. This is consistent with our sample ranking analysis (Fig 3.4), where CBEA scores are on average more informative when used to rank samples based on their propensity to have inflated counts. However, CBEA did not outperform the CLR approach across our simulation studies, and only marginally performed better in the real data analysis with WGS data. Fortunately, in simulation studies, this performance gap between CLR and CBEA decreases with higher sparsity and correlation, especially in low effect saturation scenarios.

3.6.3. Limitations and future directions

These above results demonstrate the applicability of CBEA under different data analysis scenarios. If researchers are interested in performing inference, they can decide between an unsupervised sample level approach (i.e. screen samples for enrichment of certain characteristics) or a supervised population level approach (i.e. identifying characteristics that are differentially abundant across case/control status). For the unsupervised approach, utilizing the unadjusted CBEA with the mixture normal distribution provides a good initial starting point. In the case where researchers only want to screen samples with mean-inflated taxon sets (instead of additionally detecting taxon sets with increased correlation), they can apply the adjusted approach, which can be effective at conserving type I error even for high correlation scenarios. However, the trade off for this adjustment is power, which decreases with increasing correlation. For the supervised analysis, all CBEA variants control for type I error and provide adequate statistical power. However, using raw CBEA scores with difference-in-means test such as Welch's t-test is preferable since is the least computationally expensive (no estimation process) while still outperforming the use of a sum-based approach with a standard differential abundance test. Beyond inference, CBEA scores are flexible and can be useful for downstream analysis. We demonstrated that for a given number of set-based features, CBEA can produce informative scores that contribute to competitive performance of prediction models even in low signal-to-noise ratios with high inter-taxa correlation and sparsity. This is especially true for whole genome sequencing data sets, where CBEA outperfroms the standard approach of applying a CLR transformation. Researchers might find CBEA useful under situations of high sparsity and inter-taxa correlation, or if the property of a singular covariance matrix (a byproduct of the CLR transformation [104]) is undesired. Even though we only evaluated prediction models, researchers can benchmark their own usage of CBEA for other downstream tasks such as sample ordination. However, there are various limitations to our evaluation of CBEA. First, our simulation analysis may not capture the appropriate data-generating distributions underlying microbiome taxonomic data. There is strong evidence to suggest that our zero-inflated negative binomial distribution is representative [39], however other distributions such as the Dirichlet multinomial distribution [322] have been used in the evaluation of prior studies. More recent studies have suggested utilizing the hierarchical multinomial logistic normal distribution to model microbiome data sets [214, 181]. As such, there is space to evaluate and adapt CBEA to these different distributional assumptions that underlie the data generating process. Second, we were not able to evaluate the phenotype relevance of enrichment results as in Geistlinger et al. [100] due to limited consistent annotations for microbiome signatures in health and disease, especially those that are experimentally verified (and not just from differential abundance studies). We attempted to perform this evaluation by leveraging the gingival data set similar to [39]. However, we acknowledge that this is not a perfect solution, since the oxygen usage label of each microbe in the data set is only available at the genus level, and the difference in counts for obligate aerobes and anaerobes across the supragingival and subgingival sites might not be as clear cut. As such, results from power analyses using this data set is only relative between the comparison methods and cannot be treated as absolute measures of power or phenotype relevance. Third, fitting the mixture normal distribution to raw CBEA scores using the expectation-maximization algorithm is difficult, as the convergence rate is slow when there is high overlap between the mixtures, resulting in small mixing coefficient for one of the components and increased runtime (FigC.6) [218]. In our implementation, we attempted to account for this by increasing the maximum number of iterations and relaxed the tolerance threshold. Finally, we assumed that taxa within a set are all equally associated with the outcome. This limits our ability to evaluate the performance of CBEA when only a small number of taxa within the set is associated with the outcome, or if there are variability in effect sizes or association direction of taxa within a set.

Our evaluation also showed various drawbacks of the CBEA method itself. First, inference with CBEA at the sample level is limited, and can be affected by inter-taxa correlation if users wish to only detect mean-inflated sets. Second, for downstream analyses, CBEA might not always perform better than competing methods, especially when being used to generate inputs to predictive models. We hypothesized that this might be due to the lack of fit for the underlying null distribution in high correlation settings, especially the identifiability problem associated the estimation procedure associated with adjusting the mixture normal distribution. As such, we hope to refine the null distribution estimating procedure by either choosing a better distributional form, or to further constrain the optimization procedure of the mixture normal distribution by fixing the third and fourth moments.

In addition, CBEA itself did not consider other aspects of microbiome data. First, across all analyses, we relied on adding a pseudocount to ensure log operations are valid. Users can directly address this by incorporating model-based zero correction methods prior to modelling such as in [192] or [135]. However, in our simulation studies, sparsity seems to not have a significant impact on the overall performance of our approach. Second, CBEA also treated all taxa within the set as equally contributing to the set. Incorporation of taxa-specific weights (similar to PhILR [272]) could reduce the influence of outliers, such as rare or highly invariant taxa. Finally, even though for a given set of *a priori* annotations CBEA can generate useful summary scores, such values are limited in their utility if the annotations themselves are not meaningful. As such, curating and validating sets (similar to MSigDB [276]) based on physiological or genomic characteristics of microbes [308] or their association with human disease (in beta BugSigDB `https://bugsigdb.org/Main_Page`) can allow for incorporating functional insights into microbiome-outcome analyses.

Section 3.7

Conclusion

Gene set testing, or pathway analysis, is an important tool in the analysis of high-dimensional genomics data sets; however, limited work has been done developing set based methods specifically for microbiome relative abundance data. We introduced a new microbiome-specific method to generate set-based enrichment scores at the sample level. We demonstrated that our method can control for type I error for significance testing at the sample level, while generated scores are also valid inputs in downstream analyses, including disease prediction and differential abundance.

Section 3.8

Availability of data and materials

All data sets are available publically via `Qiita`, `HMP16SData`, and `curatedMetagenomicData` with raw sequence data available on NCBI in their respective project repositories. All analysis scripts and generated figures are available on GitHub (`https://www.github.`

com/qpmnguyen/CBEA_analysis). An implementation of this approach is available on GitHub as well as on Bioconductor (version 3.15) as CBEA.

Section 3.9

Acknowledgments

The authors thank Becky Lebeaux, Modupe Coker, Erika Dade, Jie Zhou, and Weston Viles for insightful comments and suggestions that greatly improved the paper.

Chapter 4

Evaluating trait databases for taxon set enrichment analysis

This chapter was submitted as a pre-print on bioRxiv and can be found here:

Nguyen, Q.P., Hoen, A.G., Frost, H.R. Evaluating trait-based sets for taxonomic enrichment analysis applied to human microbiome data sets. bioRxiv. https://doi.org/10.1101/2022.05.16.492155

Section 4.1

Abstract

Background Set-based pathway analysis is a powerful tool that allows researchers to summarize complex genomic variables in the form of biologically interpretable sets. Since the microbiome is characterized by a high degree of inter-individual variability in taxonomic compositions, applying enrichment methods using functionally driven taxon sets can increase both the reproducibility and interpretability of microbiome association studies. However, there is still an open question of which knowledge base to utilize for set construction. Here, we evaluate microbial trait databases, which aggregate experimentally determined microbial phenotypes, as a potential avenue for meaningful construction of taxon sets.

Methods Using publicly available microbiome sequencing data sets (both 16S rRNA gene metabarcoding and whole-genome metagenomics), we assessed these trait-based sets on two criteria: first, do they cover the diversity of microbes obtained from a typical data set, and second, do they confer additional predictive power on disease prediction tasks when assessed against measured pathway abundances and PICRUSt2 prediction.

Results Trait annotations are well annotated to a small number but most abundant taxa within the community, concordant with the concept of the core-peripheral microbiome. This pattern is consistent across all categories of traits and body-sites for whole genome sequencing data, but much more heterogenous and inconsistent in 16S rRNA metabarcoding data due to difficulties in assigning species-level traits to genus. However, trait-set features are well predictive of disease outcomes compared against predicted and measured pathway abundances. Most importantly trait-set features are more interpreable and reveal interesting insights on the relationship between microbiome, its function, and health outcomes.

Section 4.2
Introduction

Advancements in high-throughout sequencing technologies have allowed researchers to characterize the identity and functional potential of a large proportion of microorganisms in human-associated microbiomes. This has enabled efficient study of the link between health outcomes and the microbiota without reliance on currently limited culture-based approaches [154]. As such, there has been an increase in microbiome profiling studies, primarily aiming towards identifying specific microbes that are differentially abundant between groups of individuals defined by an exposure or disease state vs a control population [340]. However, such analyses face unique computational and statistical challenges [166], which includes addressing the burden of multiple testing and providing meaningful biological interpretations.

The challenge of understanding results of microbiome analyses in a broader context of biological systems mirrors that of other high-throughput data sets. One approach that has proven to be fruitful in human genomic studies is gene set testing (or pathway analysis) which focuses on analyzing the coordinated expression of groups of genes (termed gene sets or pathways) [186]. From a statistical perspective, set-based statistics are are more reproducible and have greater power compared to their gene-level counterparts [105]. The true benefit of set-based approaches, however, is the ability to incorporate *a priori* knowledge of specific cellular processes [170]. Microbiome differential abundance analyses can also benefit from set-based approaches instead of a microbe-centric approach. In addition to statistical benefits such as reduced dimensionality and sparsity [223, 148], set-based approaches are also more reflective of the underlying biology. Like genes, microbes act in concert with co-abundant partners to drive biochemical processes that interact with the host, thereby impacting health outcomes [326]. For example, when comparing patients with inflammatory bowel disease against healthy subjects, microbes thought to be disease-causing for inflammatory bowel disease were also strongly co-occurring [101], suggesting that they might jointly contribute to the microbiome-disease causal pathway instead of acting as independent factors. This is also represented in the development of therapies, where products often contain multiple strains of bacteria [25, 75]. Furthermore, organizing microbes into functionally-driven groups (also termed “guilds” [326]) is also congruent with the perspective that human microbiomes are complex ecosystems whose properties emerge from localized interactions between microbial communities representing individuals that exploit and contribute to their environment in similar ways [87].

Unfortunately, there is currently limited research in curating and evaluating appropriate microbe annotations similar to the transcriptomic literature. Repositories like the Molecular Signatures Database (MSigDB) [170] aggregate information about gene function across multiple sources, incorporating both laboratory results and computational inferences. Even though similar databases such as Disbiome [129] and MSEA [148] exist, they are usually human-centric and define microbial groups based on their potential to be pathological rather than through common biochemical roles. As such, these databases are limited in generating meaningful hypotheses linking taxonomic changes to ecosystem function especially in novel disease conditions. Trait-based analysis [308, 26, 184], with its long history in traditional macroecological studies [87, 149], is a promising approach to address this gap. Traits directly represent microbial physiological characteristics and metabolic phenotypes (for example, sulfur reduction, nitrate utilization, or gram positivity) and therefore can serve as annotations for potential ecosystem function. For 16S rRNA gene sequencing data sets, where one can only obtain taxonomic abundances, performing enrichment analysis on trait-based sets can elucidate the taxa-function relationship and identify microbial processes that are differentially active between healthy and diseased patients. For whole genome metagenomic data sets, traits still offer unique perspectives. First, traits are often sourced from the long history of laboratory experiments such as journal articles and *Bergey's Manual of Systematic Archaea and Bacteria* [285] which is different from homology-based sequence queries typically performed to profile gene family abundances. Second, traits are complex phenotypes that represent multiple molecular pathways, which means that they are more comparable to higher-order pathway annotations in hierarchical databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [134] and MetaCyc [45]. As such, utilizing traits as the source to group microbes into functional and phenotypical categories can assist in interpreting microbiome profiling studies, and generating mechanistically meaningful hypotheses that link ecosystem function and its taxa.

Even though trait-based approaches have been utilized in various studies [308, 26, 107, 149], to our knowledge there is currently no effort to formalize trait-based databases in terms of microbial sets and evaluate their utility in a typical enrichment analysis of 16s rRNA metabarcoding or metagenomic data. Here, we constructed taxon sets from pre-existing trait databases at both the species and genus level. Then, we computed the coverage of these traits across different human-associated environments and sequencing approaches. Finally, we evaluated whether trait-based set features confer predictive capacity for diseased individuals compared to measured (from whole genome sequencing data) and predicted (from PICRUSt2 [73]) pathway abundances. Finally, we identified the most important features for prediction and assessed whether they matched existing literature on the microbiome-disease relationship of interest.

Section 4.3

Material and methods

All analyses were performed in the R programming language (version 4.1.2) [251] and the Python programming language (version 3.10.4). All graphics were generated using `ggplot2` [311], `ggsci` [328], `patchwork` [234]. Tabular data manipulation was performed using `pandas` for python, and `tidyverse` [310] suite of packages for R. Additional packages utilized include: `BiocSet` [209], `taxizedb` [46], `phyloseq` [196], `TreeSummarizedExperiment` [119]. For enrichment analyses, we leveraged the CBEA [223] method (version 1.0.1) developed previously by our lab. All analyses were performed using the `snakemake` workflow [204]. All reproducible code and intermediate analysis products can be found on GitHub (https://www.github.com/qpmnguyen/microbe_set_trait).

4.3.1. Generating taxonomic sets from trait databases

We utilized pre-compiled trait databases from previous publications: Madin et al. 2020 [184] and Weissman et al. 2021 [308]. The former was chosen due to the fact that it is the most comprehensive compilation of microbial (bacteria and archaea) physiological traits based on existing sources to date. The latter is a newer database that hand curates traits specifically for human microbiomes based on Bergey's manual. Both of these databases source their trait assignments primarily from biochemical and microbiological laboratory experiments over genomic-based annotation. We focused our analyses on categorical traits, namely metabolism, gram stain, enzymatic pathways, sporulation, motility, cellular shape, and substrate utilization. We are particularly interested in traits belonging to the class of enzymatic pathways and substrate utilization as they represent functions that most directly impact the microbe-host relationship [295].

We combined both databases into a joint knowledge base and constructed sets for each available categorical trait. Additionally for the Madin et al. database, we updated data entries sourced from Genomes Online Database (GOLD) [216] due to the fact that compared to other compiled sources, GOLD is continuously updated via community submissions. We grouped all traits belonging to the same National Center for Biotechnology Information (NCBI) species-level identifier. When there are conflicts in assigning traits, we prioritized Weissman et al. over Madin et al. and GOLD due to its hand curated nature. If there are ambiguities in taxonomic assignment in the Weissman et al. source, we considered that trait to be missing. The exceptions to the above logic are enzymatic pathways and substrate utilization categories where trait values across sources for the same species are concatenated instead of reconciled. For example, if a species A has entries from multiple databases suggesting the presence of “nitrogen degredation” and “ammonia degredation”, then instead of attempting to chose the best annotation based on source we assumed that species A has the capacity to metabolize both nitrogen and ammonia.

All traits are defined at the species level via NCBI identifiers, however, due to restrictions for 16S rRNA gene sequencing data sets to resolve beyond the genus level [132], we also assigned traits to each genus based on a two-step process for each major trait category:

- A hypergeometric test is used to ascertain whether the genus is underrepresented in the database based on the total number of species assigned to that genus in NCBI Taxonomy [264] compared to our trait database. If a genus is underrepresented in our database (i.e. the proportion of species number of genera in the database is significantly less than what one would expect if one were to randomly draw species from the NCBI database), then the trait is not assigned to that genus since we do not have enough information. Specifically, we assessed $P(X \leq x)$ at $\alpha = 0.05$ where $X \sim Hypergeometric(k, N, K)$, with x as the total number of species assigned to that genus in the database with an assigned value for the trait category of interest, k as the total number of species in the database with an assigned value for the trait category, N as the total number of species in NCBI Taxonomy, and K as the total number of species assigned to the genus in NCBI Taxonomy.

- For all genera that are well represented in the database, we then assessed the proportion of species under that genus that have the trait. If over 95% of species of a given genus have the trait, then the trait is assigned to the genus.

We then defined trait-based sets using the aforementioned assignments. Each trait value with a category, e.g. “obligate anaerobic” from the category “metabolism”, is defined as a set with elements representing the species (or genus) annotated to that trait value. In the analysis stage, each identified taxon within a data set is matched to a trait based on their NCBI identifier. For 16S rRNA gene metabarcoding data sets, we matched all amplicon sequence variants (ASV) with traits belonging to the genus level NCBI identifier matched to the ASV sequence. All processed databases and resulting taxonomic sets can be found on GitHub in the analysis repository.

4.3.2. Evaluation data sets

We evaluated trait-based sets on publicly available 16S rRNA gene metabarcoding and whole-genome metagenomic data sets. For study-specific metabarcoding data sets, we obtained data directly from associated European Nucleotide Archive (ENA) repositories and re-processed raw sequence files into ASV tables using the `dada2` QIIME 2 (version 2022.2) plugin [40, 31]. Taxonomic classification was performed using a pre-trained weighted naive bayes model [29, 133] using the SILVA NR 99 database version 138 [247] available via QIIME 2. For all our metagenomic data sets, we downloaded taxonomic and pathway abundance tables directly from the `curatedMetagenomicData` R package [231] (2021-10-19 snapshot), which processed the data via the `bioBakery` [19] metagenomic data processing pipeline by the package authors. Data from the Human Microbiome Project (HMP) was obtained using the `HMP16SData` [262] (for metabarcoding data) and `curatedMetagenomicData` [231] (for metagenomic data) R packages.

To assess trait annotation coverage, we utilized data from both Phase I and II of the HMP [59] as it contains surveys for multiple human-associated environments from healthy subjects. For predictive and concordance analyses, we focused on colorectal cancer (CRC) and inflammatory bowel disease (IBD) as study conditions. Both CRC and IBD are well represented across both metabarcoding and metagenomic data sets, allowing comparisons across sequencing methodology. Furthermore, these conditions are also under active study within the microbiome literature, which improves the ability to interpret the biological significance of the results. For CRC, we utilized data from Zeller et al. [337], Feng et al. [88], Gupta et al. [108], Hannigan et al. [110], Thomas et al. [279], Vogtmann et al. [296], Wirbel et al. [315], Yachida et al. [331], and Yu et al. [336]. For IBD, we utilized data from the integrative HMP [245], Gevers et al. [101], Hall et al. [109], Ijaz et al. [123], Li et al. [167], Nielsen et al. [225], and Vich Vila et al. [294]. A detailed description of each data set and data-processing procedures is available in the Supplementary Materials.

4.3.3. Coverage analysis

In this analysis, we sought to identify how well trait databases cover the taxonomic diversity of different human-associated environments. We leveraged healthy samples from multiple body sites from Phase I and II of the HMP [59]. We quantified coverage as a per-sample measure considering both taxa absence/presence and its abundance.

- For each sample, we computed the proportion of taxa that is present (non-zero counts) assigned to at least one trait (a sample-level trait-specific richness).
- For each sample, we computed the proportion of reads assigned to taxa that is present and annotated to at least one trait (a sample-level trait-specific evenness).

In addition to coverage stratified by trait categories and body sites, we also generated category-specific and site-specific coverage values by averaging across all sites or categories, respectively.

4.3.4. Prediction analysis

We also aimed to evaluate whether trait-based features can add information for microbiome-based disease prediction compared to other data inputs. Here, we generated sample-level enrichment scores for each trait using CBEA and utilized them as inputs to a standard random forest model [33]. Model fitting was done using scikit-learn [235] where all parameters were set to default values with the exception of the total number of trees per ensemble (500) and the total number of features considered per split (equal to the square root of the total number of features). We compared model performance using trait enrichment scores against measured and PICRUSt2 predicted pathway abundances (for metagenomic and metabarcoding data sets, respectively).

Model performance was measured using the area under the receiver operating characteristic curve (AUROC) and Brier scores [34]. These metrics and associated confidence intervals were obtained by fitting and evaluating the model via a 10-fold cross-validation procedure. To obtain calibrated predictive probabilities for Brier scores, we applied Platt's method (using CalibratedClassifierCV) with 5-fold cross validation nested within the training fold and used the ensemble model to generate test set probabilities [242].

In order to identify which features are important to the disease prediction process, for each input type we re-split the entire data set into train/test splits (80% training data). We then refitted our calibrated random forest model on the training set as described above. Since our final model is an ensemble of calibrated random forest classifiers, we obtained feature importance values as the average across all calibrated cross validation folds ($N = 10$). Feature importance per random forest model is defined based on the implementation in scikit-learn as the decrease in Gini impurity when the feature is split averaged across all decision trees in a forest.

Section 4.4

Results

4.4.1. Database coverage

We computed the coverage for each trait category across each body site in the HMP data set. Fig 4.1 illustrates results for species-level trait assignment for samples profiled via whole genome metagenomics. In panel A, coverage is evaluated as the total number of taxa present per sample annotated to a trait (a measure of cross-trait richness), while in panel B coverage is the total number of reads assigned to taxa annotated to a trait (a measure of cross-trait evenness). Richness provides a general overview on how many members of a community is assigned to a trait, while evenness accounts for their relative abundances by up-weighting species that have high abundance across all samples. Overall, for any body site, at most 25% of taxa are assigned to a trait, but when considering the proportion of reads, coverage increased to more than 80%. This shows that traits are usually well annotated to the most abundant taxa. This pattern holds for samples profiled with 16S rRNA gene sequencing (Fig D.1), even though the proportions were much lower due to difficulties in aggregating species level traits to genus. For many body sites and trait category combinations, traits could not be assigned to any taxa.

We also observed heterogeneity in the annotation coverage across different body sites and trait categories. For richness, nasal cavity and vaginal body sites were the lowest in coverage, with less than 5% of taxa annotated with at least one trait across all trait categories while conversely, oral cavity sites consistently had the highest coverage under this metric. This pattern was reversed when considering coverage as the proportion of assigned reads per sample, but overall values were consistently high. Averaging coverage across body sites (Fig 4.2) also supports this observation, showing

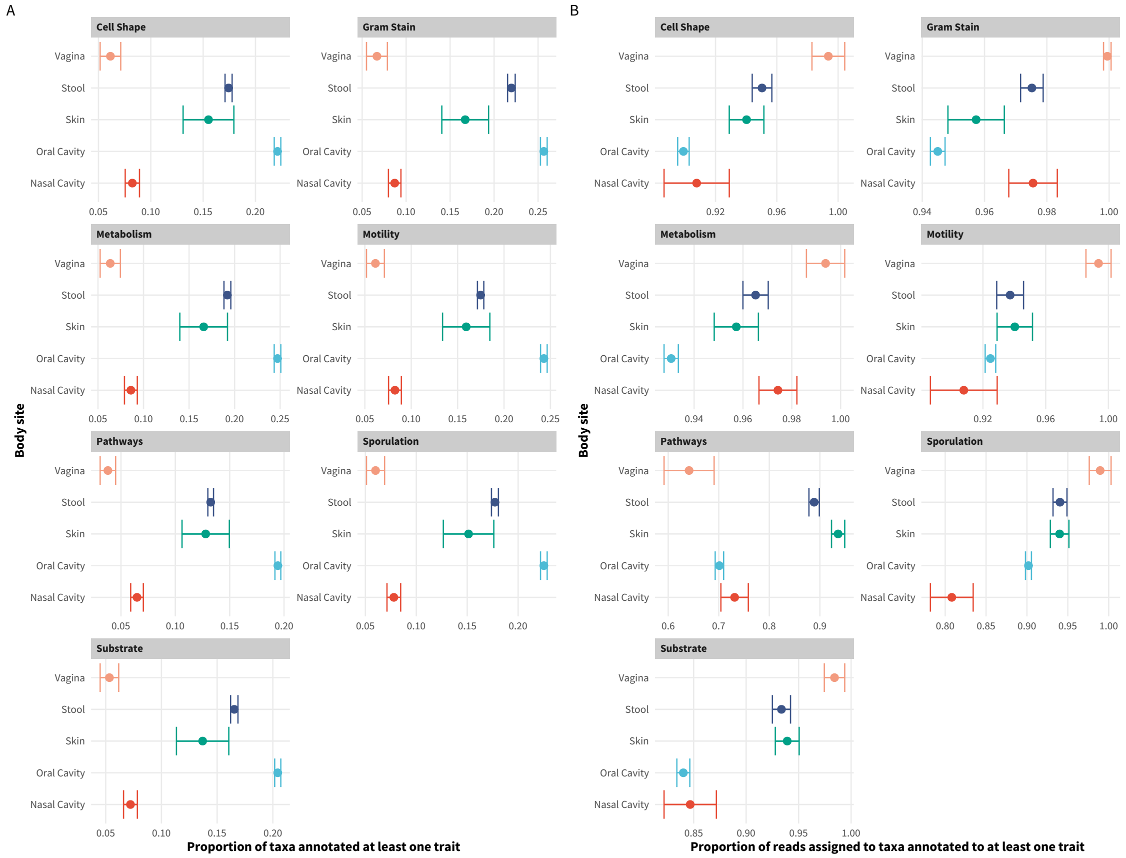

Figure 4.1: Trait annotation coverage across different body sites for the HMP data set profiled using whole genome shotgun sequencing. Panel **(A)** illustrates the proportion of present taxa per sample annotated to at least one trait. Panel **(B)** illustrates the proportion of reads assigned to taxa annotated to at least one trait which accounts for taxa relative abundances. Each plot facet represents different trait categories that were evaluated. Error bar represents the standard error of the evaluation statistic of interest across the total number of samples evaluated per body site.

overall that the proportion of reads covered are similar across all body sites despite differences in the proportion of present taxa covered by trait annotations. Similar results were observed for sites profiled with 16S rRNA gene sequencing (Fig D.1), where oral sub-sites have the highest coverage across both richness and evenness metrics but, on average, all sites were similar in coverage statistics. Surprisingly, stool samples were low in coverage across multiple categories despite being one of the well studied systems.

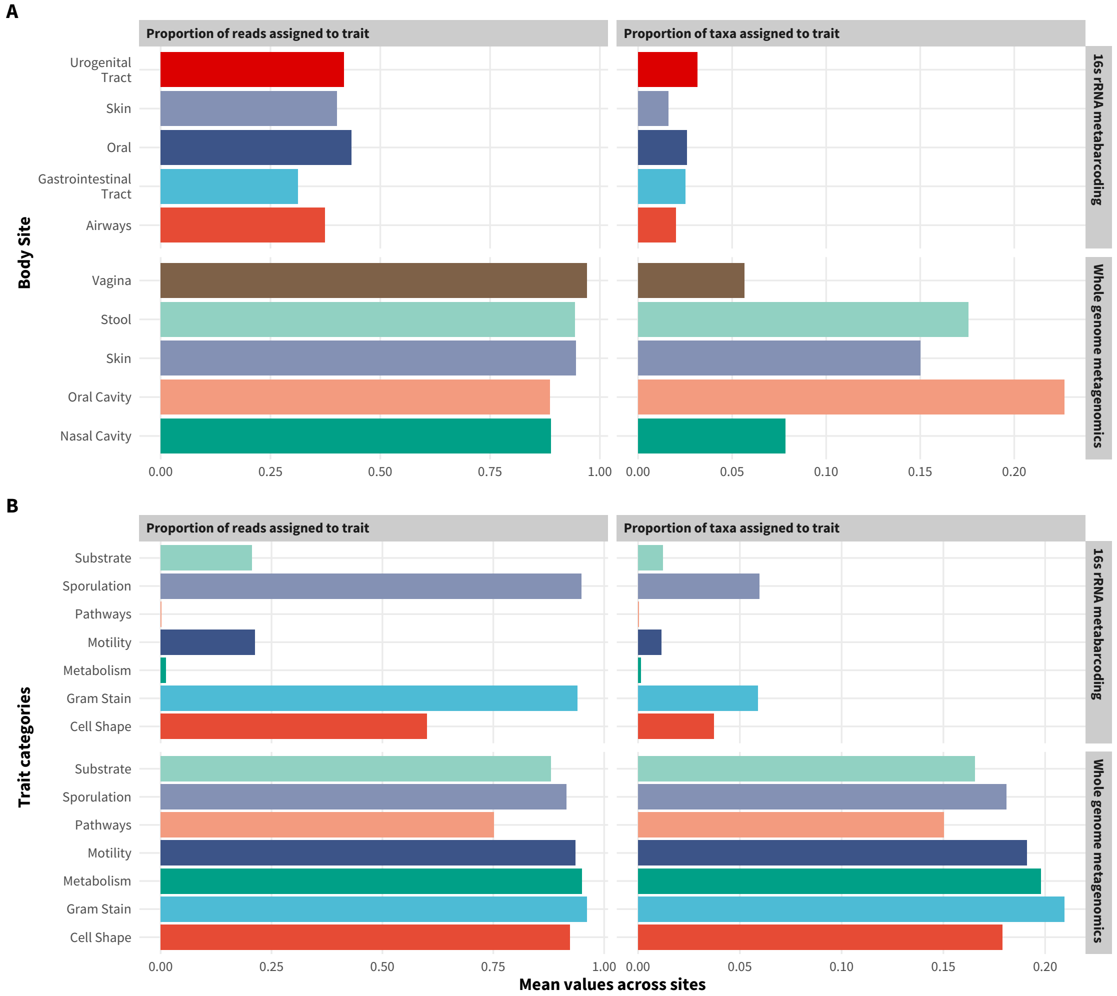

Figure 4.2: Trait annotation coverage statistics for HMP data across samples profiled with both 16S rRNA gene metabarcoding and whole genome metagenomics. Panel **(A)** illustrates coverage statistics for each body site averaged across evaluated samples and trait categories. Panel **(B)** illustrates statistics for each trait category averaged across evaluated samples and body sites.

We also stratified our coverage analyses by trait categories (Fig 4.1, Fig 4.2, Fig D.1). For samples profiled with whole genome sequencing, all trait categories are evenly covered, with about 15% - 20% of taxa were annotated to a trait of any category. However, these taxa comprise around 75% to 100% of the total reads per sample suggesting that the overall read level coverage is very high. However, in samples profiled with 16S rRNA gene sequencing, the overall coverage value across categories is low. Sporulation, substrate utilization and motility are the most covered category while pathways and metabolism has no coverage at all.

4.4.2. Predictive analysis

To determine the utility of trait-based sets, we generated enrichment scores for covered traits using CBEA [223] for evaluated data sets and compared the predictive performance of using trait-set enrichment scores as inputs compared to alternative functional-based predictors. We evaluated two disease conditions, CRC and IBD, with data sets drawn from both 16S rRNA gene metabarcoding and whole genome metagenomic profiling techniques. We fitted a calibrated random forest model to each input type and computed predictive performance as AUROC (discriminatory power) and Brier scores (probability estimates) using 10-fold cross-validation.

For the CRC prediction task, traits covered 2.7% of taxa and 27.3% of reads for the 16S rRNA gene metabarcoding data set, while for the whole genome sequencing data set, traits covered 9.1% of taxa and 87.2% of reads. For the IBD prediction task, traits covered 1.61% of taxa and 26.7% of reads for 16S rRNA gene metabarcoding data set, while for the whole genome sequencing data set, traits covered 6.6% of taxa and 91.2% of reads.

Fig 4.3 illustrates results of our model evaluations. Overall, enrichment scores for trait-sets are as good as other alternate function-based predictors at discriminating between case and control patients across both CRC and IBD conditions. Aside from pure discrimination power, models fitted on CBEA trait-set scores are also equivalent in approximating predicted probabilities. This is surprising especially for the 16S rRNA gene metabarcoding data sets, where the trait coverage is low. Even though the differences in performance is not significant, there are instances where trait-set scores perform slightly better than their pathway abundance counterparts. Since trait-features are also more descriptive, utilizing them can increase interpretation while also not sacrificing performance.

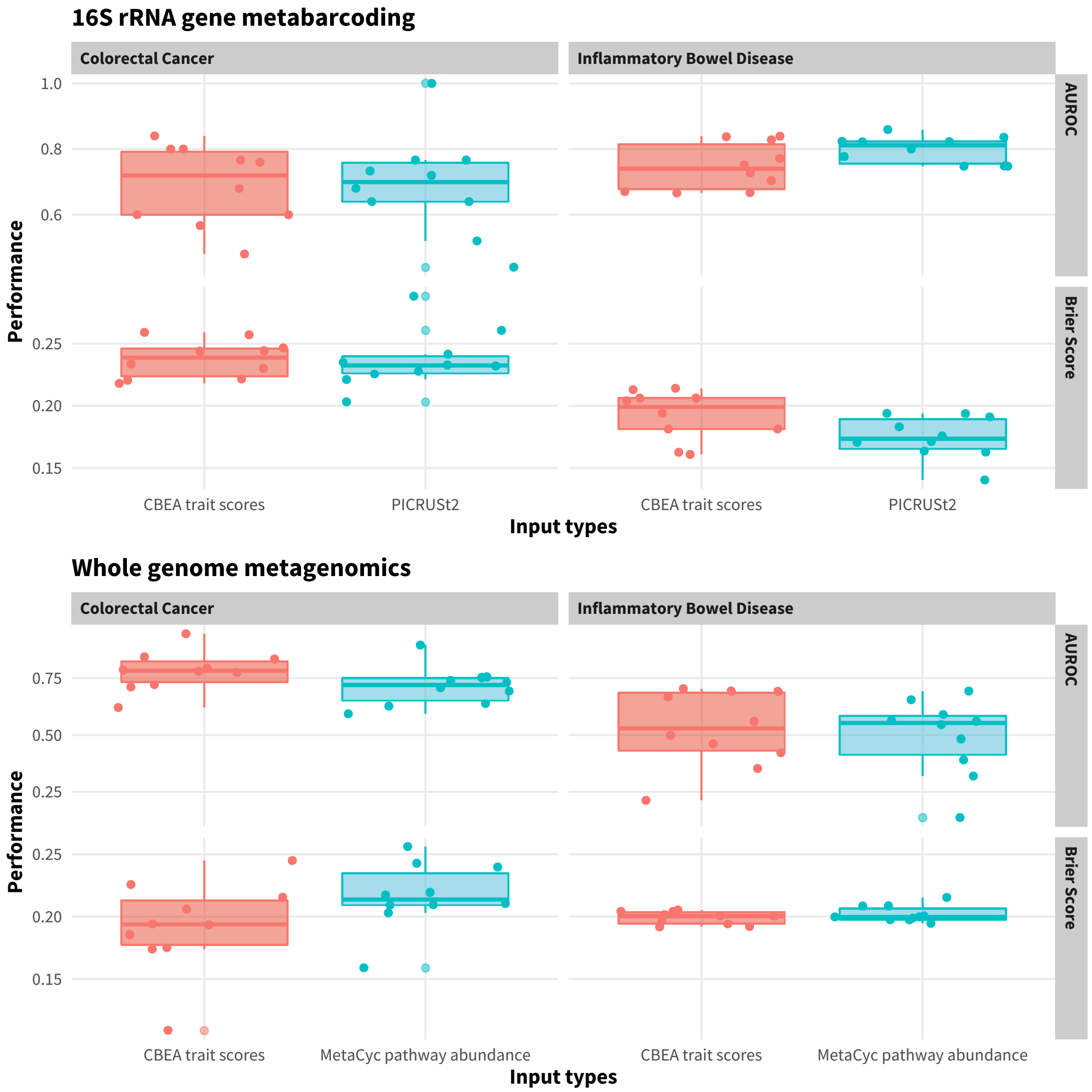

Figure 4.3: Predictive performance of a calibrated random forest model across different disease conditions, profiling technique, and performance metrics. Scores were obtained via 10-fold nested cross validation where within each training fold there is a 10-fold cross validation procedure to calibrate predicted probabilities. For each condition and data type, CBEA trait-set scores were compared against MetaCyc pathway abundances from relevant sources (measured abundances for whole genome sequencing data sets and PICRUSt2 predicted abundances for 16S rRNA gene metabarcoding data sets)

In addition to predicted performance, we also identified the top 10 features that are most important for model fitting. Since our model involves a 10-fold cross-validation procedure within the training set to calibrate predicted probabilities, top features are identified using the mean feature importance value across the 10 folds. Fig 4.4 illustrates results for whole genome metagenomic data sets while Fig D.2 illustrates results for the 16S rRNA gene metabarcoding data sets. Even though these are the top 10 features, the observed mean feature importance statistics are low, suggesting that no individual features were definitively the most important in discriminating between patient classes.

Section 4.5 Discussion

4.5.1. Traits are annotated with high coverage at the species-level

We computed the coverage of trait annotation on a typical dataset to understand the extent in which community function is captured, thereby serving as a proxy for expected confidence for an enrichment analysis performed using trait-based taxon sets. Low coverage in this case indicates that the database does not adequately capture the diversity of microbes found in the target data. This is because there might not be enough taxa present in the data set to serve as evidence for the trait. Alternately, this could also mean that the analysis is missing a majority of underlying community traits, many of which might be core to the health outcome association of interest but simply missing in the analysis. We computed coverage based on two metrics: first, a richness-like metric which computes coverage as the proportion of taxa present per sample annotated to a trait (for a given category and data set); second, an evenness-like metric that accounts for relative abundances of each annotated taxa by computing coverage as the proportion of reads per sample annotated to a trait.

When evaluated on the HMP data set (Fig 4.1), we can see that the overall

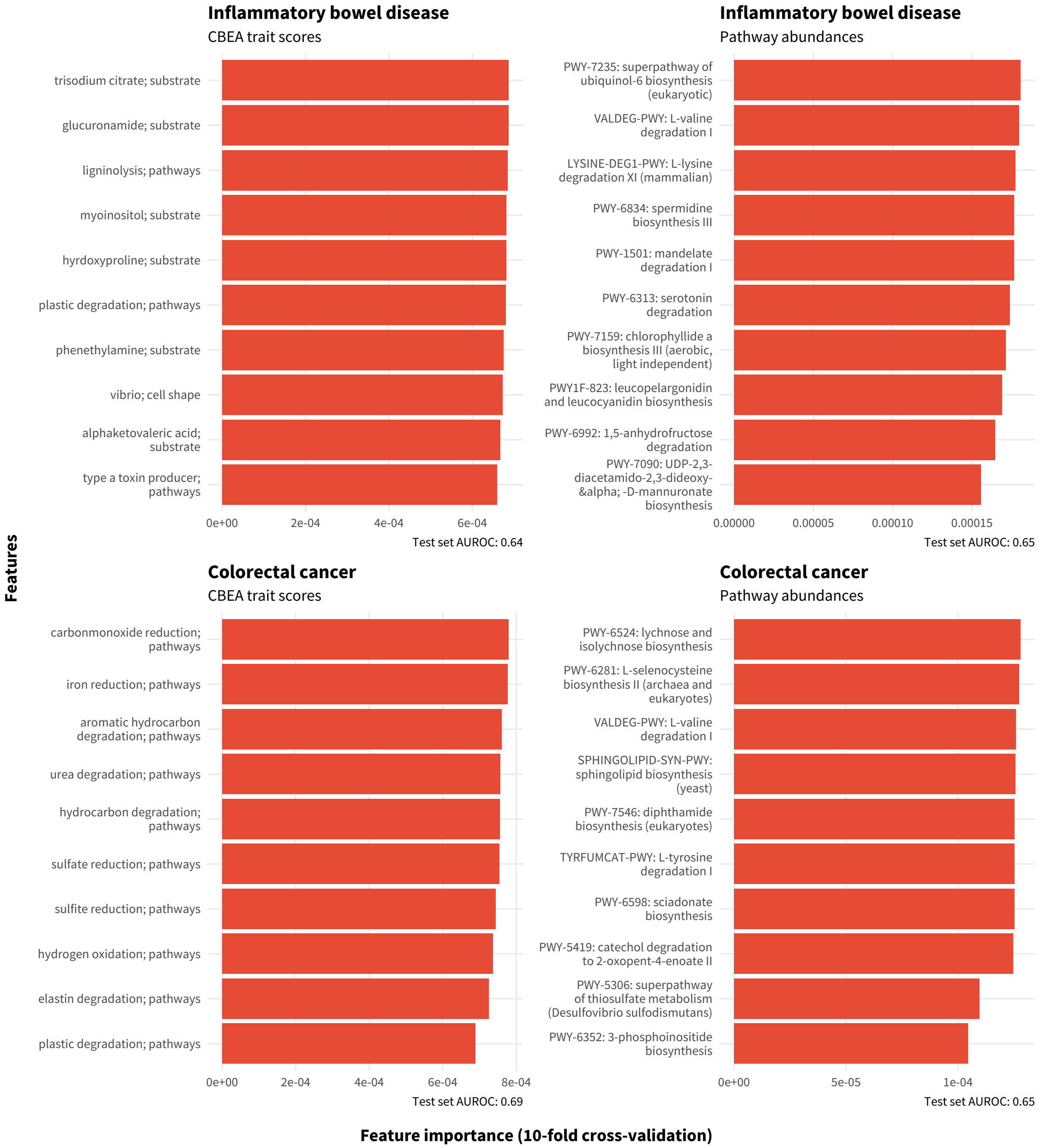

Figure 4.4: Top 10 important features based on random forest model fitted different inputs from data sets profiled with whole genome metagenomics. Features were selected from mean decrease in Gini impurity averaged across 500 decision trees and 10-fold cross-validation (nested with the training set) as implemented in `scikit-learn`. AUROC scored on a held-out test set is also presented for each input type and disease condition.

richness coverage is low (less than 25% of identified species) across all sites and data sets, particularly for nasal cavity and vaginal sub-sites. However, when considering evenness of coverage, almost all of reads were annotated to a trait. This is consistent with the observation that relative abundances of human-associated microbiomes are highly skewed [59], where a small number of species usually dominate the community. As such, even though traits might only cover a small number of taxa, they might represent the majority of community abundance. For example, Ravel et al. [253] observed that *Lactobacillus* species dominate the vaginal microbiome and, in some phylotypes, almost all reads are assigned to a single species. This shows that our trait-database has high degree of coverage across the most abundant taxon within a community, which supports utilizing these sets to perform exploratory analyses. However, low richness coverage also indicates that our database might not capture traits associated with rare taxa, which can play an important role in regulating host health [292].

Unfortunately, coverage is significantly lower for samples profiled using 16S rRNA gene metabarcoding (Fig D.1). For some trait categories, such as pathways, no traits were assigned to any taxa (Fig 4.2). We hypothesized that this is due to two issues. First, metabarcoding data sets can only resolve taxonomies at the genus level [132], while traits are usually defined at the species and strain levels. Aggregating consensus traits to the genus is difficult due to the high degree of strain and species level diversity within the microbiome [44]. Second, taxonomic assignments for metabarcoding data sets are often based amplification of a specific hyper-variable region for a marker gene (most often the 16S rRNA gene). This means that taxonomic assignment can be sensitive to the choice of region, and can be inaccurate. Furthermore, the choice of taxonomic database (e.g. Ribosomal Database Project [58], SILVA [247]) can also play a part in reducing the ability for trait annotation coverage. Differences between taxonomic paths [16] can result in certain taxa not being able to be matched to traits, whose annotations are based on NCBI identifiers.

4.5.2. Trait-set features are predictive of disease outcomes

We assessed the predictive performance of models fit on trait-set enrichment scores compared to other function-based inputs. For whole genome sequencing data sets, measured pathway abundances were utilized as a comparison point while for 16S rRNA gene sequencing data sets, predicted pathway abundances via PICRUSt2 were utilized instead. Fig 4.2 shows that across all conditions and profiling techniques, trait-set features are competitive in producing well performing models and were able to discriminate between cases and controls. Surprisingly, performance was also comparable in the 16S rRNA gene sequencing data set despite overall low coverage across both richness and evenness metrics. This demonstrates that trait-set abundances can still provide an informative approximation to functional potential similar to PICRUSt2 that can be used for exploratory and hypothesis generating purposes.

To determine which features are important for overall model performance, we extracted the top 10 features based on the mean decrease in Gini impurity. However, the overall feature importance values are not high, suggesting that no individual feature was dominant in classifying patient status. This is further supported by the fact that some nonsense features show up in the top 10 list for models fit using pathway abundances such as PWY-7235 and LYSINE-DEG-PWY, which are mammalian and eukaryotic pathways, respectively. However, the models still show respectable discriminatory power when evaluated on the test set (AUROC $\sim$ 0.7). Since random forest models can capture interactions between predictors [112], we hypothesized that the interaction between features contribute to test set performance rather than marginal effects. As such, we did not observe a high degree of feature importance scores since these measures are not designed to capture interaction effects [321].

However, despite such limitations, we were still able to recover existing knowledge about the condition of interest. For example, “sulfide reduction;pathways” was shown to be an important feature in discriminating subjects with CRC vs control subjects in Fig 4.4. This is supported by previous research showing that an increase in abundance of sulfate reducing bacteria is associated with the condition [331]. Mechanistically, this process, when using methionine or cysteine as substrates [49], generates H_2S as a product, which can stimulate CRC by inhibiting butyrate oxidation (which helps prevent the breakdown of the gut barrier) as well as promoting the generation of reactive oxygen species [190]. Another trait feature is "urea degredation;pathways", which suggests the importance of bacterial-driven urea hydrolysis process, which is one of the main sources of ammonia in the human gut [28]. Sustained exposure of colonocytes to free ammonia may contribute to the development of CRC [56], which is supported by animal experiments showing histological damage in the distal colon after long-term ammonium exposure [171].

4.5.3. Limitations and future directions

Even though our results demonstrate that utilizing trait-based sets can provide meaningful insight to microbiome data sets, there are several major challenges to widespread adoption. Although trait databases do not suffer from the same types of biases that exist in genomic reference databases [330], the reliance on curated experimental data means that traits are usually only annotated for species that are well studied and culturable. While using predictive models can help in assigning traits to a broader category of taxa [305], such automated approaches can result in misclassification of traits and increased noise in downstream analyses. Additionally, high-quality trait annotations require a time-consuming, manual curation process [308]. A source that is based on user submission such as GOLD [216] can cover a larger number of taxa and traits, but unfortunately can have erroneous and duplicated assignments due to the lack of a standardized nomenclature. There is currently a gap in producing a high-quality and diverse trait databases that are maintained and continuously updated.

In addition to issues with trait database quality, there are also problems matching the identity of taxa in a given trait database with identifiers found in references for sequence-based taxonomic profiling such as SILVA [247]. For whole genome metagenomic data sets, standard tools (such as MetaPhlan [19]) can provide NCBI identifiers at the species or strain levels. However, it is currently unclear how to aggregate or disaggregate traits if the taxonomic resolution of the observed data set is higher or lower than that of the trait database in use. This is even more difficult with metabarcoding datasets, where low taxonomic resolution makes trait-to-taxa assignments sparse and less confident.

Finally, there are also hurdles in being able to properly validate traits that are found to be significantly enriched due to a lack of ground truth data sets. While some traits can be matched to pathways directly, others involve complex coordination of multiple genetic pathways. As such, further investigation into ways to identify biological concordance between obtained results and external measurements can help improve confidence in utilizing traits for microbiome analyses.

Section 4.6
Conclusion

Set-based enrichment analysis is a useful approach for analyzing microbiome data sets since it not only reflects underlying biology but can also provide more unique perspectives of function that is linked to ecosystem services. Microbial trait databases are a promising resource to construct taxon-sets as traits represent physiological phenotypes. We demonstrated that trait-based sets have high coverage across body sites, especially for samples profiled using whole genome metagenomics. Furthermore, enrichment scores computed on such sets are also competitive in predicting case/control status compared to pathway abundances. As such, trait features found to be important in model fitting can be used to define interesting mechanistic hypotheses.

Section 4.7

Availability of data and materials

All data sets are available publically via `Qiita`, `HMP16SData`, and `curatedMetagenomicData` with raw sequence data available on NCBI in their respective project repositories. All analysis scripts and generated figures are available on GitHub (`https://www.github.com/qpmnguyen/microbe_set_trait`).

Section 4.8

Acknowledgements

The authors thank Becky Lebeaux for insightful comments that greatly improved the paper. We would also like to thank maintainers of the packages `curatedMetagenomicData`, `HMP16SData`, `phyloseq`, `QIIME2`, `Qiita`, `TreeSummarizedExperiment` for all their efforts in maintaining open source repositories for data analysis and data sharing.

Chapter 5

Conclusion

Taxa-function relationships are difficult to characterize due to the different scales in which they operate [156]. For the taxonomic layer, one can look at species, strain, or even cell states [198]. For the functional layer, it can be gene family abundance, transcript expression, or metabolite concentrations. Each degree of granularity increases the complexity of both the data collection process as well as its interpretation. However, no approach is "wrong" as each taxa-function combination can reveal unique biological knowledge. For example, even in the face of strain-level variation, an analysis of genus level taxa and metabolite abundances can show that perhaps the metabolism of certain metabolites are phylogenetically conserved, which can have various implications. Throughout this thesis, we have attempted to decipher this relationship using multiple approaches. In chapter 2, we utilized a multi-omic data set to identify strongly associated microbe-metabolite pairs. In chapter 3, we developed a statistical method to leverage pre-defined taxa-function annotations (in the form of sets) in standard epidemiological studies. In chapter 4, we evaluated an example of such a source using trait databases aggregated from the literature.

Section 5.1

Summary of findings

5.1.1. Mapping microbes to their function using multi-omics data

In chapter 2, we examined a paired metataxonomic-metabolomic data set to explore the relationship between bacterial relative abundances and metabolite concentrations. Even though multi-omics studies involving metabolomics are not new [174, 13, 143], most studies have focused on defining differences between subject case/control status, with limited exploration of the microbe-metabolite interface. Here, we characterized associations between the microbiome (profiled using 16S rRNA gene sequencing) and the metabolome (profiled using Nuclear Magnetic Resonance – NMR – techniques). The analyzed metabolomic data set contained both untargeted taxonomic bins, as well as concentrations of 36 specific metabolites. This data was generated from a cohort of healthy infants from the New Hampshire Birth Cohort Study (NHBCS) [103] with samples collected at 6-weeks and 12-months of age.

Using both Procrustes analysis and sparse canonical correlation analysis (sCCA) [317], we found that overall metabolite concentrations are concordant with genus-level taxonomic profiles. This relationship was weakly predictive, as we observed poor performance across different machine learning models using predictive R^2 as the evaluation metric. However, model outputs performed better using Spearman correlation ρ, but still lower compared to other studies using a similar performance metric [187]. Using $\rho = 0.3$ as a threshold for defining “well-predicted” metabolites [187, 217], we found that short chain fatty acids (SCFAs) such as butyrate are most predictive, consistent with our understanding of microbiome physiology [161]. Surprisingly, the degree of coupling is higher for infants at 6 weeks compared to 12 months, suggesting that in early life humans are more reliant on the microbiome for metabolic purposes.

In addition to overall patterns of associations, we also identified genera-metabolite groups that are core to the overall multivariate correlation by looking at the non-zero loading coefficients of our sCCA model. Similar to our concordance analysis, two SCFAs Butyrate and Proprionate were selected as the most important for the overall microbiome-metabolome relationship, with a surprising negative correlation with *Bifidobacterium* genera, a commonly identified producer of SCFAs [128]. We hypothesized that this is an instance of strain-level variation where some strains of *Bifidobacterium* compete with other butyrate-producing taxa [258]. Amino acids were also well-represented among selected metabolites and were negatively correlated with taxa abundances. We hypothesized that microbes are incorporating amino acids in their environment directly instead of catabolizing them due to the fact that this process is energetically inefficient [92, 228].

Our study showed that genus level microbial abundances are not sufficient to predict metabolite concentrations. However there is still a degree of overall coupling that is supported by prior work [13, 143, 343]. Additional studies with higher taxonomic resolution using whole genome metagenomics can be used to find more granular scales of association. We also provided further evidence to support the importance of microbiome-mediated butyrate catabolism in early life, while also suggesting that amino acids might play an important role. However, our study was limited by our cross-sectional design. This is because metabolite abundances are always changing, making measures of flux (or rate of change) more meaningful in finding associations [118]. Some studies have attempted to bridge this gap using genome-scale metabolic models [226], which can be fitted to observed data. However, additional studies with dense longitudinal sampling are still needed.

5.1.2. Developing novel methods to integrate taxa-function relationships in statistical analyses

In chapter 3, we developed a statistical method to test for the enrichment of groups of microbes. Gene set testing (or pathway analysis), is a commonly utilized in the genomics literature to aggregate lists of genes obtained after a differential abundance test [126, 105]. These methods have been shown to improve power, reproduciblity, and interpretability [139]. As such, set-based analysis can be a useful method to not only address some of the challenges of analyzing sequencing-based taxonomic data tables (such as sparsity) [165], but also to provide a formal statistical approach to incorporate taxa and function via sets. Here, we provided a method for set-based enrichment analysis called competitive balances for taxonomic enrichment analysis (CBEA) that is tailored to microbiome relative abundance data. CBEA generates sample-level scores in an unsupervised manner by integrating the Q_1 competitive null hypothesis [283] and compositional balances [272, 80]. Inference is performed at the sample-level through estimating an empirical null distribution that can be adjusted for variance inflation due to inter-taxa correlation.

We evaluated our model using both real and simulated data sets. First, CBEA can be used to test for enrichment at the sample level. Results indicated that our approach was able to control for type I error at the appropriate α level, however, the trade off was limited power to detect small effect sizes, especially at higher degrees of inter-taxa correlation. In addition, CBEA can also perform population-level analyses to detect sets that are differentially abundant between case/control status by combining generated scores with a difference in means test (such as Welch's t-test). Under this task, CBEA was able to control well for type I error but without having to concede as much power. Notably, CBEA produced fewer false positives compared to using a sum-based approach to aggregate taxa to sets and performing a standard differential abundance test such as `corncob` [191]. Finally, even though CBEA generated scores were informative for discriminating between healthy controls and patients with IBD, performance scores were not significantly higher than other comparison methods.

Our study illustrates an example of a statistical method that can assist in generating taxa-function hypotheses through the use of set annotations. Using CBEA, users can not only perform inference, but also use CBEA sample scores for downstream analyses such as predictive modeling, sample ordination, or network analysis. However, additional follow-up approaches are required to improve the inference procedure, as well address data sparsity beyond pseudocounts.

5.1.3. Leveraging existing microbiology knowledge to define microbial ecosystem roles

In chapter 4, we explored using an aggregated database of microbial traits defined based on laboratory experiments to curate function-based taxon sets. Traits represents microbial phenotypic characteristics and are oriented towards describing ecosystem functions given their long history in ecological research [149]. We drew on two sources: Madin et al. [184] which is a comprehensive compilation of traits across multiple different static repositories, and Weissman et al. [308], which provides a human microbiome centric annotations based on a manual curation of Bergey's manual. We constructed our sets based on categorical traits assembled, focusing on "pathways" and "substrates" as they represent traits that relate to microbes' participation in host biochemical processes.

First, we computed the coverage of traits across body sites and trait categories using the HMP data set. We found that trait coverage is low when considering the proportion of taxa covered, but much higher when accounting for their relative abundances. This suggested that traits are well annotated for the abundant taxa within each community, but less so for rare microbes. We did not identify significant differences in average coverage across body sites, but some body sites have much lower coverage in some traits compared to others. It was difficult to annotate traits for 16S rRNA gene sequencing data sets since traits are defined at the species level. This is because the databases are not well sampled enough across the tree of life to enable comprehensive profiling of all genera. Furthermore, of strain and species level variability [175] complicates the process of aggregating traits.

We then utilized CBEA (as described in chapter 3), to generate enrichment scores for trait-sets and utilized them as inputs to predictive models. We found that trait scores are as good as pathway abundances in discriminating between case and control patients, suggesting that they can be informative features. Surprisngly, performance held for the 16S rRNA gene sequencing data sets when compared against PICRUSt2 predicted pathway abundances, despite lower coverage. We also looked at the top 10 most informative features for our models and found that some traits correspond with known disease-associated biochemical pathways.

Our results demonstrated that set-based analysis can help integrate taxa and function under one unified framework for hypothesis testing. Set annotations based on traits provide an interesting avenue, given that they are sourced from laboratory experiments with descriptions that are less granular but more interpretable than MetaCyc pathways. Even though trait coverage for rare species is limited, trait-based sets are still informative in distinguishing healthy patients from those with colorectal cancer and inflammatory bowel disease, suggesting that traits represent meaningful biological processes. As such, additional studies are needed to further refine the ontology of describing traits and to provide annotations to a larger selection of microbes.

Section 5.2

Perspectives and future research

Our work has shown promising applications of leveraging a taxa-function framework in epidemiological studies. By contextualizing shifts in taxonomic abundances in terms of their function, researchers can more easily interpret lists of differentially abundant taxa and make informed choices on what to follow-up and validate in laboratory experiments. However, there are still major hurdles to overcome before an integrative framework can be confidently applied to future studies.

5.2.1. Defining microbial function

One of the most difficult aspects of investigating microbial function is the ability to identify meaningfully relevant definitions [116, 340, 144]. Specifically, the question of how to translate between definitions of genes and pathways (from KEGG or MetaCyc) to ecosystem functions that the gut microbiome delivers. A relevant example is the role of HMO metabolism [291], which is a host-relevant function that contextualizes multiple gene families, all of which would be difficult to interpret individually.

Chapter 4 of this thesis attempted to examine function under the lens of microbial traits. Traits are usually conceptualized as defined, measurable properties of organisms that link performance and contribution to core ecosystem needs [149]. While traits provide a more holistic conception of function, issues with unstandardized databases [184] and limited coverage for rare species makes it challenging to use in scenarios where they might be conferring important services.

As such, comprehensive efforts are needed to centralize and standardize the ontology of microbial function with respect to ecosystem needs. Efforts such as the ontology of microbial phenotypes (OMP) [51] have started to generate a repository of terminology to standardize description of microbial phenotypes. Researchers can further expand the types of annotations of OMP to be specific to body sites (using the uber-anatomy ontology - UBERON) or study conditions (using experimental factor ontology - EFO). KEGG terms or MetaCyc pathways can be assigned to OMP annotations, which can then be translated to specific strains using reference genomes/pangenomes or measured metatranscriptomic data. By defining a standardized vocabulary, researchers can begin to conceptualize a host-centric view of microbial function that is both standardized and context driven.

5.2.2. Leveraging multiple meta'omic technologies

While there have been a large number of microbiome multi-omic studies (see section 1.2.2), they have mostly been focused on analyzing each data-layer independently with some studies performing limited taxa-function analyses. However, as shown in Chapter 2, paired profiling of microbiome structure and molecular functions can reveal novel aspects of host-microbe interactions. As such, multi-omic data sets, especially those including multiple layers of functional profiling, can be invaluable.

Additionally, these data sets can be further leveraged in conjunction with the functional framework defined in section 5.2.1 in two ways:

- First, researchers can use these data sets directly to test for the enrichment of ecosystem-specific functional roles similar to that of Vatanen et al. [291] and HMO metabolism.
- Second, researchers can leverage the collection of these data sets to validate encoded taxa-function relationships, specifically accounting for situations where gene carriage does not directly correlate with expression [92]
- Third, they can be used to generate new core taxa-function relationships that can be disease-associated that can serve as biomarkers or as potential candidates for intervention.

5.2.3. Novel representations of taxa-function relationships

To jointly test for association between taxa-function groups and relevant exposures or disease outcomes, there is a need to identify appropriate representations that can be translated into a statistical framework. Set-based approaches, used in Chapter 3, are simple but powerful methods. Sets naturally capture categorical information such as the assignment of strains to functions. However, the definition of sets are rigid, and does not account for nuances such as uncertainties or the degree of strain presence/absence in the overall population. As such, novel numerical representations of taxa-function relationships can help account for this gap and allow for a more flexible way to encode these relationships.

One candidate would be to use weights within sets or across different sets depending on the experimental context. For example, Frost [94] curates a set of tissue-specific weights for MSigDB gene sets. Here, body-site specific weights can be computed for each functional term, or the contribution of each taxa can be weighted by its overall prevalence estimated from a large cohort such as HMP.

Network-based methods offer another approach. Bipartite networks can be used to model connections between taxonomic and functional nodes [282]. Networks also have topological features such as degree centrality that can provide extra dimensions such as being able to identify taxa that contribute to a large number of functions or vice versa. Standard network structures that allow for connections within taxonomic and functional nodes are also useful as they can account for inter-taxa correlation or dependencies between metabolites or genes.

There are also machine-learning based approaches that can provide unique encoding opportunities. Word embeddings, such as Word2Vec [201], create dense numerical vectors that can represent high-dimensional co-occurrence relationships. This application has been explored in the context of the microbiome-metabolome relationship [211]. Researchers can extend this approach to model different functional outputs, or to provide pre-trained embeddings based on a meta-analysis of microbiome-metabolome data sets.

Appendix A

List of abbreviations

NHBCS: New Hampshire Birth Cohort Study

PCoA: Principal Coordinates Analysis

NMR: Nuclear Magnetic Resonance

ASV: Amplicon Sequence Variants

OTU: Operational Taxonomic Unit

gUniFrac: Generalized Unique Fraction

sCCA/CCA: Sparse Canonical Correlation Analysis

FDR: False Discovery Rate

RF: Random Forest

ASV: Amplicon Sequence Variants

SCFA: Short chain fatty acids

EN: Elastic Net

SVM-RBF: Support Vector Machines with Radial Basis Kernel Function

SPLS: Sparse Partial Least Squares

CLR: Centered Log Ratio Transformation

KEGG: Kyoto Encyclopedia of Genes and Genomes

GOLD: Genomes OnLine Database

CBEA: Competitive Balances for taxonomic Enrichment Analysis

HMDB: Human Metabolite DataBase

HMP: Human Microbiome Project

GSVA: Gene Set Variation Analysis

GSEA: Gene Set Enrichment Analysis

ssGSEA: Single Sample Gene Set Variation Analysis

GO: Gene Ontology

CRC: Colorectal Cancer

IBD: Inflammatory Bowel Disease

MSigDB: Molecular Signatures Database

VAM: Variance-adjusted Mahalanobis

CoDA: Compositional Data Analysis

iNKT: Invariant natural killer T cells

TLR-5: Toll-like receptor 5

GI: Gastrointestinal tract

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

SCC: Spearman correlation coefficient

PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

ILR: Isometric log ratio transformation

LASSO: Least absolute shrinkage and selection operator

FID: Free induction decay

PMA: Penalized Multivariate Analysis

PCR: Polymerase Chain Reaction

QC: Quality Control

PRESS: Predicted residual error sum of squares

QC: Quality Control

DADA: Divisive Amplicon Denoising Algorithm

AUROC: Area under the receiver operating characteristic curve

QIIME: Quantitative Insights Into Microbial Ecology

MetaPhlAn: Metagenomic Phylogenetic Analysis

PhILR: Phylogenetic isometric log ratio transformation

SMOTE: Synthetic Minority Oversampling Technique

CDF: Cumulative distribution function

MSEA: Microbe-set enrichment analysis

NCBI: National Center of Biotechnology Information

LC: Liquid Chromatography

MS: Mass Spectrometry

UBERON: Uber-anatomy ontology

EFO: Experimental factor ontology

OMP: Ontology of microbial phenotypes

Appendix B

Supporting material for "Associations between the gut microbiome and metabolime in early life"

Section B.1

Supplemental Notes

B.1.1. Supplementary Note 1

Microbe-metabolite participation in significant pairwise Spearman correlation. Univariate pairwise Spearman correlations were performed to identify significant microbe-metabolite pairs. Significance was determined by the Spearman false discovery rate (FDR) threshold of 0.05 following a Benjamini-Hochberg multiple hypothesis testing procedure. At both time points a majority of genera and metabolites were significantly correlated, where at 6 weeks, 28 genera (65% of total genera) and

36 metabolites (100% of metabolites) were part of 516 significant correlations (16.6% of total pairwise comparisons) while at 12 months, 59 genera (81.9% of total genera) and 29 metabolites (80% of metabolites) were involved in 214 significant correlations (8.01% of total pairwise comparisons). This result also supported the observation that at 6 weeks the microbiome was marginally more associated with the metabolome compared to 12 months. Similar to sCCA results, untargeted data set showed a similar signal at both time points. Specifically, at 6 weeks, 37 genera (86% of genera) and 198 metabolite bins (95.1% of bins) were part of 1480 significant associations (16.5% of pairwise comparisons). Similarly, at 12 months, 67 genera (93% of genera) and 207 metabolite bins (99.5% of bins) were part of 1392 significant associations (9.2% of total pairwise comparisons).

B.1.2. Supplementary Note 2

Sparse canonical correlation analysis selects microbes and metabolites important to the inter-omic correlation. Only a small subset of metabolites and microbes were selected (27% of taxa for both time points; 16.9% of metabolites at 6 weeks and 19.4% of metabolites at 12 months). At both time points, selected taxa belong to the Firmicutes, Actinobacteria and Proteobacteria phyla with Firmicutes being the most represented (58.3% of selected taxa at 6 weeks; 70% of selected taxa at 12 months). Actinobacteria was the second most selected phylum (25% of selected taxa) at 6 weeks while at 12 months it was Proteobacteria (30% of selected taxa). For metabolites, amino acids were the most represented metabolite class (Supplementary Note 4) (60% of selected metabolites at 6 weeks, 85% of selected metabolites at 12 months). 6-week samples demonstrated a larger diversity of metabolite classes, with additional representatives from carboxylic acids group, nucleotides and short chain fatty acids (SCFA) while at 12 months, the only non-amino-acid metabolite is uracil (of the nucleotide class). Across both time points, 3 genera (Flavonifractor, Haemophilus and Acinetobacter genera) and 5 metabolites (lysine, isoleucine, leucine, uracil, phenylalanine) were consistently selected. Surprisingly, in the untargeted analysis, nearly half of taxa and metabolites were selected at 6 weeks while the number remained more similar to the targeted analysis at 12 months (6 weeks: 46% of taxa and 42% of metabolite bins; 12 months: 13.8% of taxa and 17.3% of metabolite bins). However, the taxonomic distribution of those selected taxa remained similar, with Firmicutes being the most dominating phyla (60% of selected taxa at both time points). Additionally, for both time points, the sign of the sCCA loadings for selected variables were also concordant with patterns of negative and positive correlation via univariate Spearman correlations (Figure 3, right panels). Notably, all selected metabolites contain negative loadings for both time points, with the majority of selected pairwise correlation to be negative (6 weeks: 76.6% of selected pairwise comparisons, 12 months: 60.7% of selected pairwise comparisons). This pattern is replicated in the untargeted data set as well (6 weeks: 61.3% of selected pairwise comparisons, 12 months: 70% of selected pairwise comparisons).

B.1.3. Supplementary Note 3

Prediction results. Under R^2, at 6 weeks only 8 (22.2%) metabolites (Butyrate, Glycerol, Isobutyrate, Isoleucine, Leucine, Methionine, Phenylalanine and Tyrosine) were predictable, with a mean of 4.85% and a maximum of 11.8% (Butyrate using EN). At 12 months, only 14 (38.9%) of metabolites (Butyrate, Formate, Inosine, Isobutyerate, Isoleucine, Lactate, Leucine, Methionine, Phenylalanine, Propionate, Propylene glycol, Tyrosine, Uracil, Valine) were predictable, with a mean of 4.81% and a maximum of 8.7% (Propylene Glycol using RF). When looking at the average R2 across all metabolites, performance was not good (-5.6% at 6 weeks; -3.07% at 12 months). This negative R_2 value implies that the predicted model performs worse than the naïve, intercept only model. Conversely, correlative performance was much better. At 6 weeks 26 metabolites (83%) were predictable, with a mean correlation of 0.344 and a maximum of 0.669 (Butyrate using EN). Similarly, at 12 months all 36 targeted metabolites were predictable, with a mean of 0.265 and a maximum of 0.549 (Succinate using EN). Using the SCC cutoff of 0.3 as criteria for well predicted metabolites, many metabolites at 6 weeks were still retained (25 metabolites - 69.4%). Conversely, at 12 months, only 13 metabolites remained to be well predicted (38.9%). On average, performance based on SCC was good with a mean SCC value of 0.339 at 6 weeks and 0.249 at 12 months. In the untargeted analysis looking at the entire metabolome, performance was much better for both metrics. Under R2, 116 (56.7%) of metabolite bins were predictable at a maximum of 42.7% (Bin 33 using SPLS) and a mean of 16.7% at 6 weeks while at 12 months, 94 (45.1%) of metabolite bins were ell predicted at a maximum of 22.7% (Bin 16 using SVM) and a mean of 8.19%. The overall average across all metabolites is 3.91% (6 weeks) and -0.59% (12 months). This trend was similarly observed using SCC, as all all 208 metabolites bins were predictable for both time points (using SCC = 0 as the threshold). Specifically, of the predictable metabolites at 6 weeks the maximum value was 0.687 (Bin 32 using EN) with a mean of 0.352 while at 12 months, the maximum value was 0.53 (Bin 16 using SVM) and a mean of 0.253. Using the SCC cutoff of 0.3 as above, at 6 weeks 120 metabolite bins (57% of bins) were well predicted while at 12 months only 60 (28.8% of bins) were well predicted.

B.1.4. Supplementary Note 4

Short chain fatty acids (SFCA)

- **Metabolites**: Acetate, Butyrate, Isobutyrate, Proprionate
- **HMDB Ids**: HMDB00042, HMDB00039, HMDB01873, HMDB00237
- **Examples of associated microbes**: *Faecalibacterium*, *Eubacterium*, *Rose-*

buria, *Clostridia* clusters IV and XIVa [224]

- **Potential biological functions**: Fermented in the colon from dietary complex carbohydrates, assist in regulating host immune functionality, act as substrate for cellular activities, limit growth of pathogenic species, promote integrity of the mucosal lining.

Amino acids and derivatives

- **Metabolites**: Alanine, Asparagine, Aspartate, Glutamate, Glycine, Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Proline, Threonine, Tryptophan, Tyrosine, Valine, π-Methylhistidine
- **HMDB Ids**: HMDB00161, HMDB00168, HMDB00191, HMDB00148, HMDB00123, HMDB00177, HMDB00172, HMDB00687, HMDB00182, HMDB00696, HMDB00159, HMDB00162, HMDB00167, HMDB00929, HMDB00158, HMDB00883 HMDB00479
- **Examples of associated microbes**: Proteobacteria phylum; Bacili class; *Clostridium* and *Bifidobacterium* genera [243], Lactobacilli, Enterococci, and Streptococci families [246]; *Faecalibacterium prausnitzii* species [168].
- **Potential biological functions**: Catabolized to form other end products such as SCFAs, branched chain fatty acids (BCFAs) and other compounds [228]. For example, the catabolism of methionine results in methanethiol and hydrogen sulfide [243]; catabolism of histidine can produce histamine, which can inhibit the production of pro-inflammatory cytokines as well as act as a neurotransmitter [280]; catabolism of lysine can produce cadaverine [246], which is associated with ulcerative colitis [159]; catabolism of tryptophan and tyrosine can produce tryptamine, which is a neurotransmitter involved in intestinal motility and immune function [97].

Carbohydrates

- **Metabolites**: Fucose, Glucose, Glycerol, Maltose
- **HMDB Ids**: HMDB00174, HMDB00122, HMDB00131, HMDB00163
- **Examples of associated microbes**: *Bacteroides thetaiotaomicron* (possesses 260 glycoside hydrolases in its genome [329]), *B. Fragilis*, *Ruminococcaceae* spp.
- **Potential biological functions**: Humans rely mostly on gut commensals for breaking down complex carbohydrates [228]. Bacteria also take in complex carbohydrates for additional purposes. For example, fucose participate in the fucosylation of bacterial glycans, increasing fitness for both pathogenic and commensal microbes through host mimicry; facilitate promotion of useful bacterial species and metabolites, suppress virulence genes [241]. Similarly, gut microbe can metabolize glycerol into reuterin, which is an antimicrobial multicomponent system [339].

Carboxylic and dicarboxylic acids

- **Metabolites**: Formate, Fumarate, Malonate, Succinate Lactate
- **HMDB Ids**: HMDB00142, HMDB00134, HMDB00691, HMDB00254, HMDB00190
- **Examples of associated microbes**: Bloom of Enterobacteriaceae phylum. Others include Verrucomicrobia phylum. Producers from pyruvate catabolism includes *Bacteroides* and *Clostridia* genera [120]. Specifically for lactate, *Lactobacilli*, *Lactococci*, *Streptococci*, *Leuconostoc* and *Pediococci* genera [240].
- **Potential biological functions**: Often used as terminal electron acceptors for bacterial anaerobic respiration [150] and is produced by the microbiota itself [82] and a highly competitive resource especially if induced by antibiotics [254].

The presence and activation of enzymes associated with oxidizing these acids are markers of inflammation such as in the case of formate [120]. Notably, lactate is an important component in designing lactic acid bacteria probiotics which can modulate intestinal immunity and provide a protective effect against infection [99].

Vitamins

- **Metabolites**: Nicotinate (Vitamin B3)
- **HMDB Ids**: HMDB01488
- **Examples of associated microbes**: Lactic-acid commensal bacteria such as *Bifidobacterium bifidum*, *B. longum*, *B. breve*, and *B. adolescentis* [162].
- **Potential biological functions**: Microbiome has been shown to both metabolize dietary B vitamins as well as produce them through folate metabolism [162]. It is well known that vitamin Bs are essential micronutrients that are precursors to important enzymes in humans.

Nucleosides

- **Metabolites**: Inosine, Uridine, Uracil
- **HMDB Ids**: HMDB00195, HMDB00296, HMDB00300
- **Examples of associated microbes**: *Anaerococcus*, *Peptoniphilus*, *Fusobacterium*, *Lactobacillus* genera [72].
- **Potential biological functions**: Can play an important role in immune response at the neonatal stage [297], supplementing the process of enterocyte proliferation, maturation and apoptosis of intestinal cells [260].

Alcohols

- **Metabolites**: Propylene Glycol
- **HMDB Ids**: HMDB01881
- **Examples of associated microbes**: Firmicutes and Lachnospiraceae phyla, *Dorea*, *Robinsonella* and *Roseburia* genera [252].
- **Potential biological functions**: A solvent involved in propanoate metabolism resulting in propanal (through KEGG [134]), which have been shown to be associated with inflammatory bowel disease.

Bile acids

- **Metabolites**: Cholate
- **HMDB Ids**: HMDB00619
- **Examples of associated microbes**: Inhibits Bacterioidetes and Actinobacteria phyla, expansion of Firmicutes phylum, *Blautia*, *Clostridium* and *Ruminococcus* spp. [256, 127].
- **Potential biological functions**: Bile acid are involved in absorption of fats and lipid-soluble vitamins [224], bile acid byproducts of microbial origins can bind and activate host nuclear receptors and act as endocrine signaling molecules [137, 122], which is found to be associated with cancer [334].

B.1.5. Supplementary Notes 5

Illumina MiSeq v4v5 primers used for bacteria 16S rRNA gene sequencing

- **Forward Primer (518F)**: CCAGCAGCYGCGGTAAN

- **Reverse Primers (926R)**: CCGTCAATTCNTTTRAGT CCGTCAATTTCTTTGAGT CCGTCTATTCCTTTGANT

Section B.2

Supplemental figures

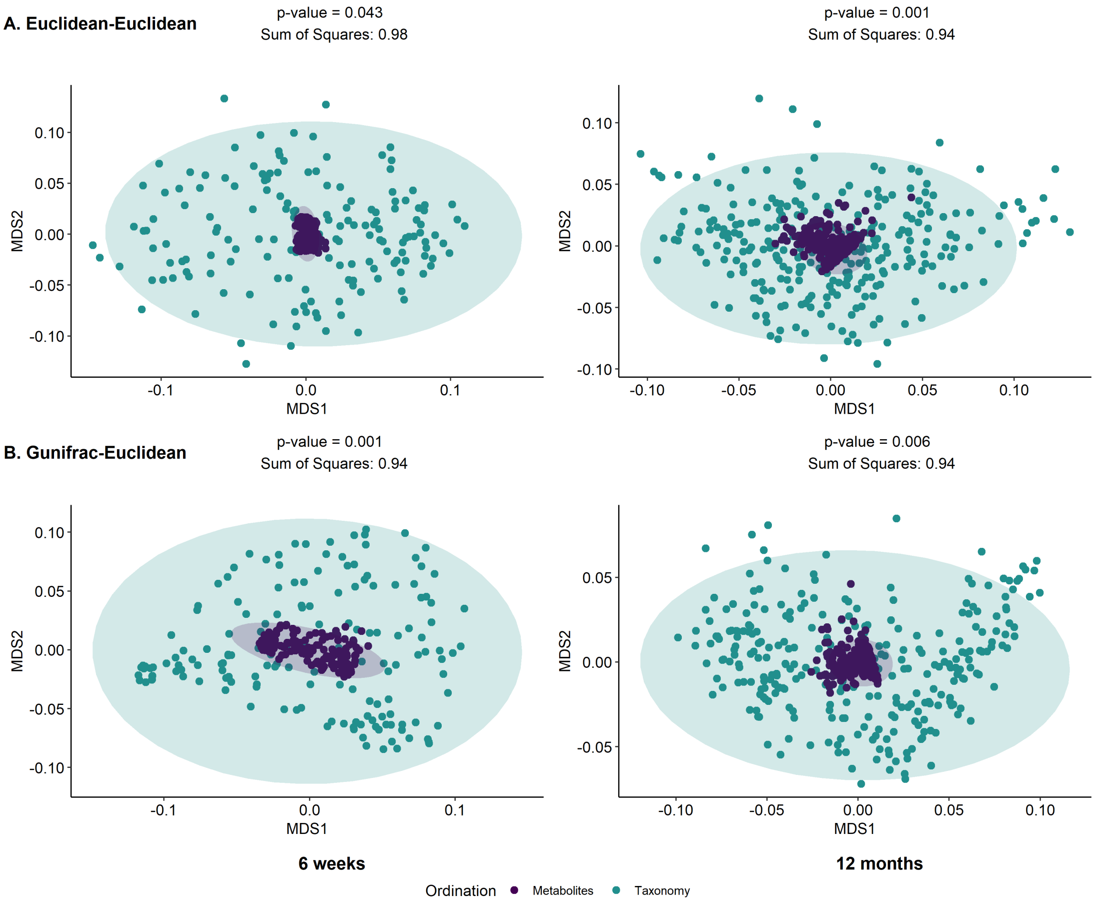

Figure B.1: Inter-omics Procrustes biplots comparing PCoA ordinations of untargeted metabolite profiles and taxonomic relative abundances for 6 weeks (left panels) (n = 158) and 12 months (right panels) (n = 262). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of arcsine square root transformed metabolite relative abundances. Bottom panel presents analyses based on generalized Unifrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of arcsine square root transformed metabolite relative abundances.

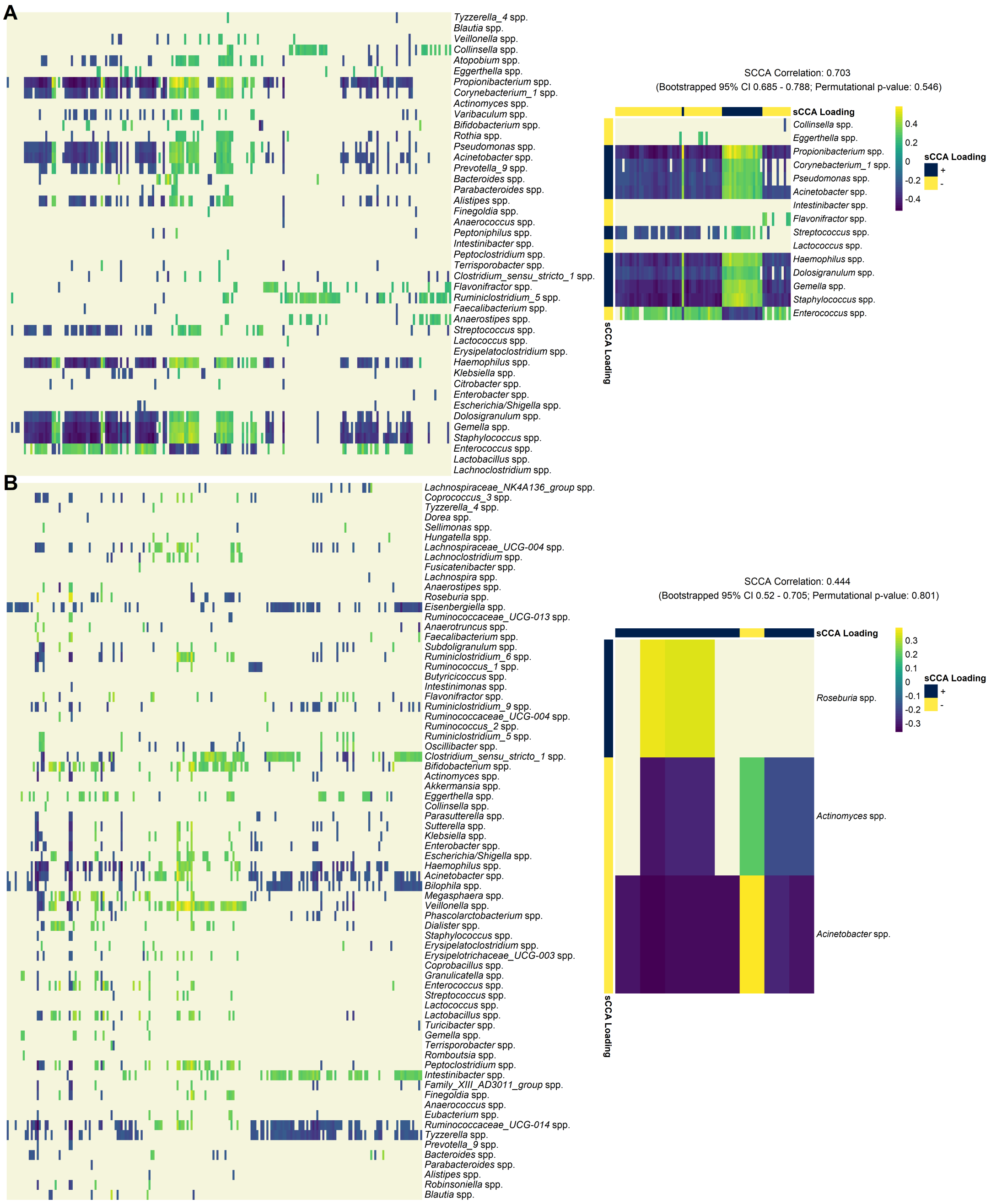

Figure B.2: Pairwise Spearman correlation of metabolite bins and genus-level taxonomic abundances for 6-weeks (panel A, N = 158) and 12-months (panel B, N = 282) infants. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (false discovery rate (FDR) controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sparse CCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval and permutation p-value are also reported.

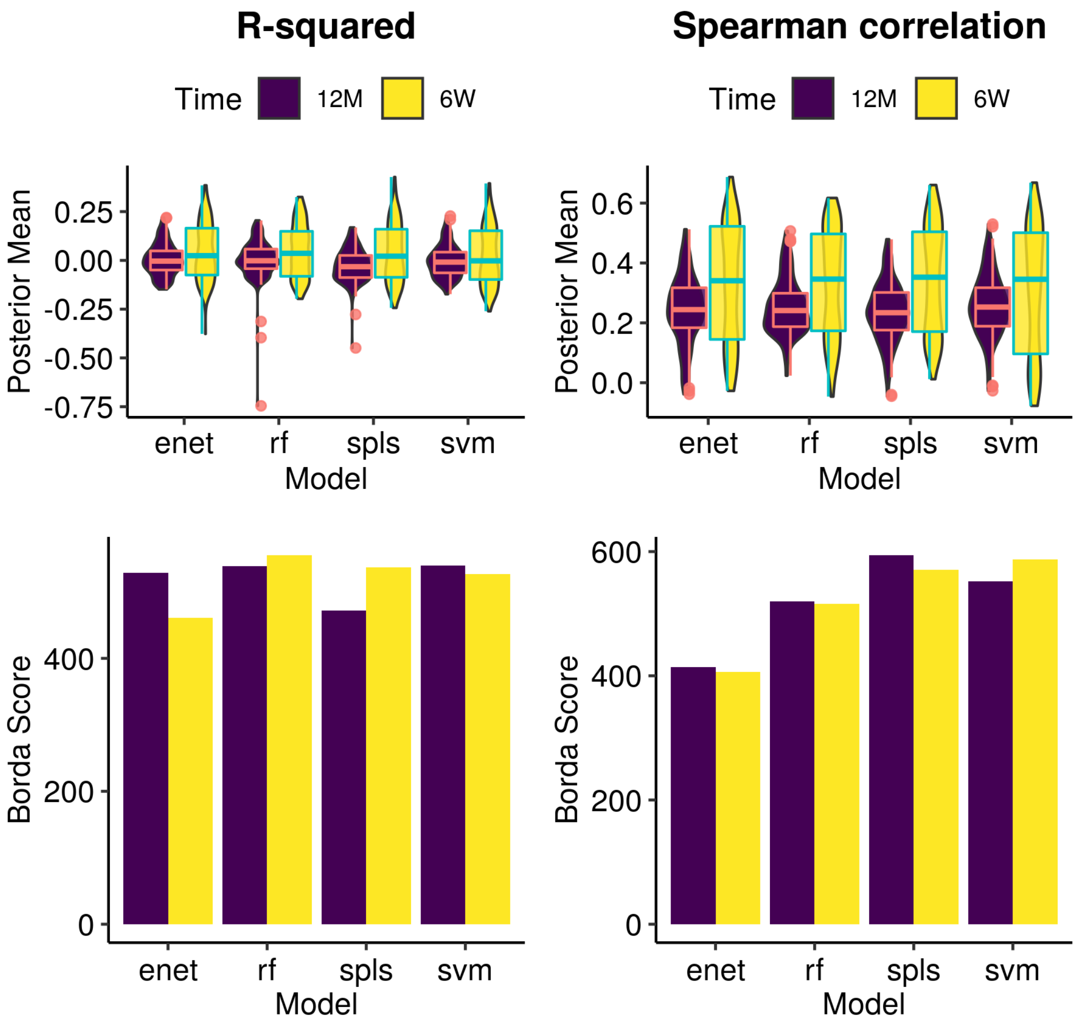

Figure B.3: Comparative analysis predictive model performance across all metabolites in the untargeted dataset for both 6-weeks (n = 158) and 12-months (n = 282) timepoints. Top panel shows superimposed boxplots and violin plots of the distribution of predictive posterior mean for each evaluation metric across all 208 spectral bins. Bottom panels show aggregated model rankings for all metabolites using R-squared (left) and spearman correlation (right) using Borda scores (Methods).

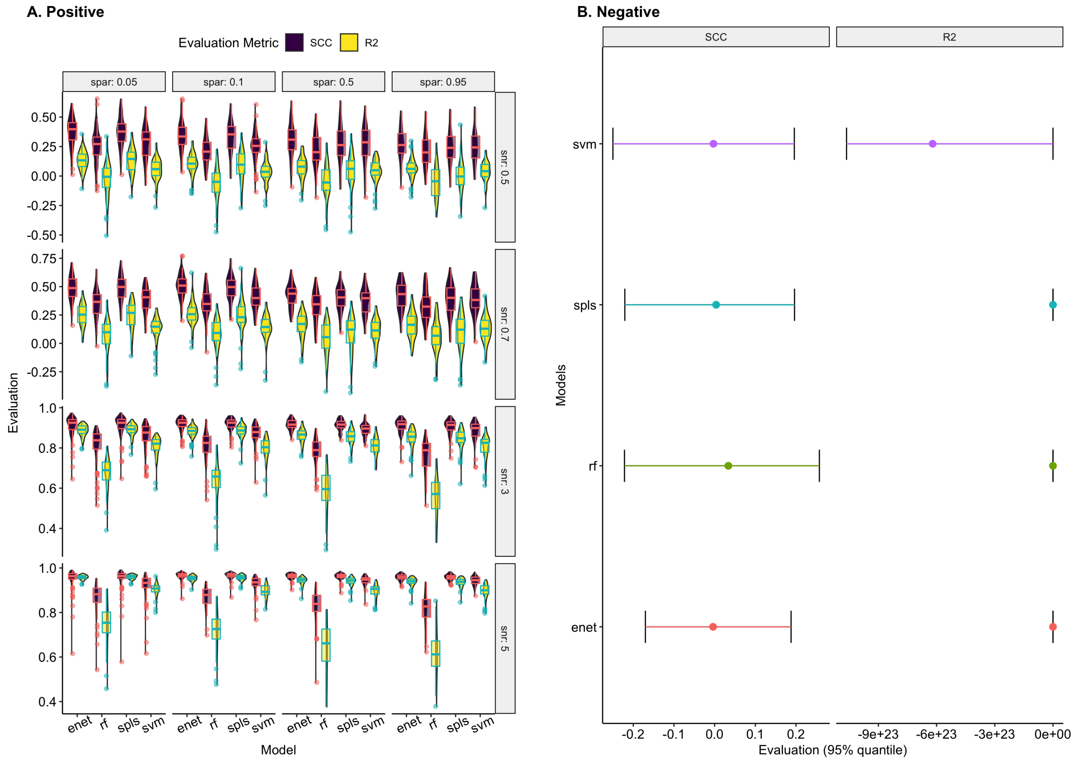

Figure B.4: Results for positive (Panel A) and negative simulations (Panel B). Positive simulations were conducted based on bootstrapped resamples of the original data (12-month time point) and a normally distributed outcome vector which represented a log-transformed metabolite profile. Different levels of model saturation (horizontal, model sparsity (spar) at 0.05, 0.1, 0.5, 0.95) and effect sizes (vertical, signal-to-noise ratio (snr) at 0.5, 0.7, 3, 5) were assessed, with 100 data sets generated for each setting combination. Negative simulations were conducted based on permutations of the original data (12-month time point), with a total of 1000 permutations. Highly negative outliers were removed for the purposes of visualization.

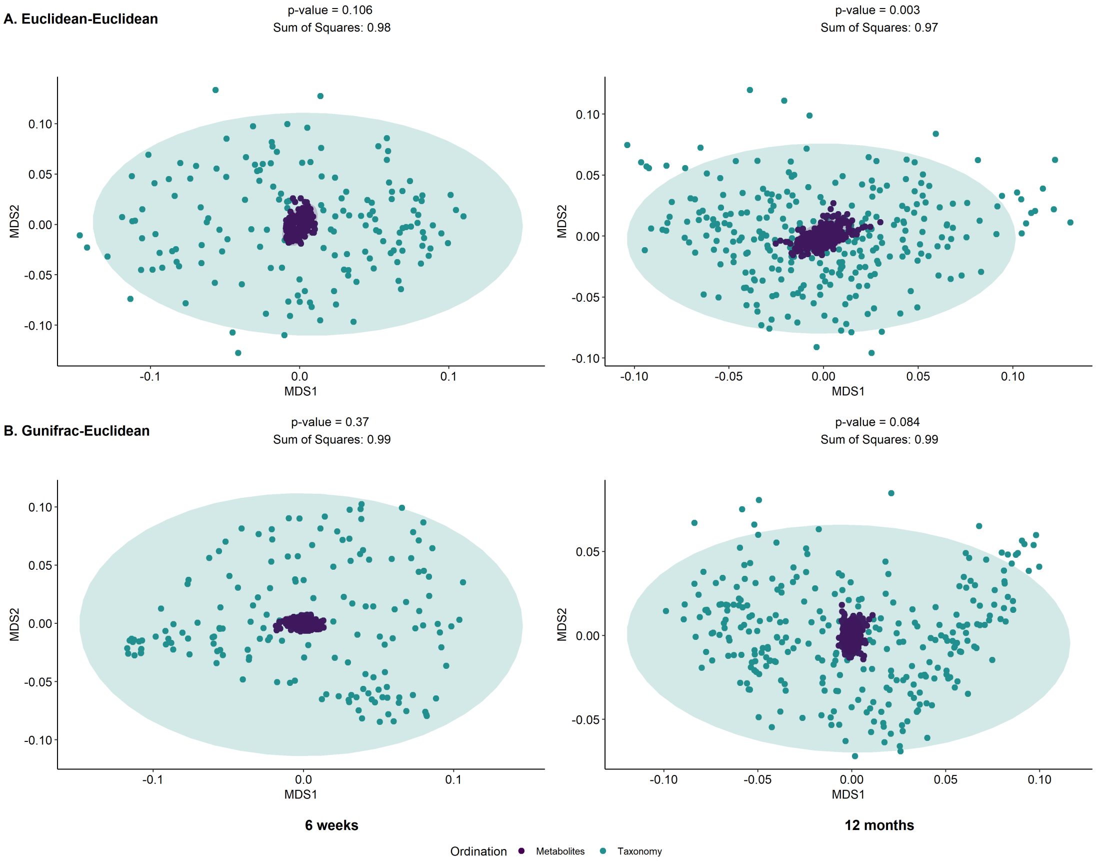

Figure B.5: Inter-omics Procrustes biplots comparing PCoA ordinations of targeted metabolite profiles and taxonomic relative abundances in the sensitivity analyses for 6 weeks (left panels) ($N = 65$) and 12 months (right panels) ($N = 65$). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of arcsine square root transformed metabolite relative abundances. Bottom panel presents analyses based on generalized Unifrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of arcsine square root transformed metabolite relative abundances.

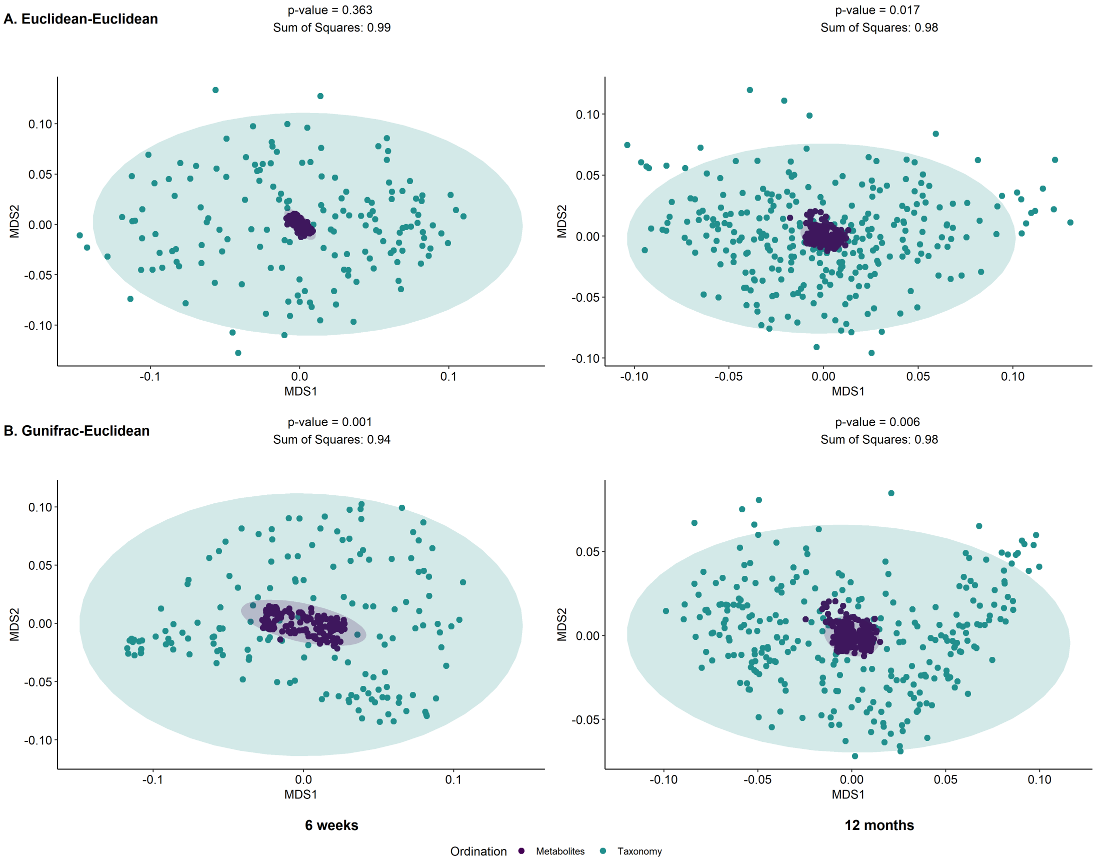

Figure B.6: Inter-omics Procrustes biplots comparing PCoA ordinations of untargeted metabolite bin relative concentrations and taxonomic relative abundances in the sensitivity analyses for 6 weeks (left panels) ($N = 65$) and 12 months (right panels) ($N = 65$). Top panels present analyses based on ordinations from Euclidean distances of genus level abundances after centered log ratio transformation and Euclidean distances of arcsine square root transformed metabolite relative abundances. Bottom panel presents analyses based on generalized Unifrac distance of amplicon sequence variant (ASV) relative abundances and Euclidean distances of arcsine square root transformed metabolite relative abundances.

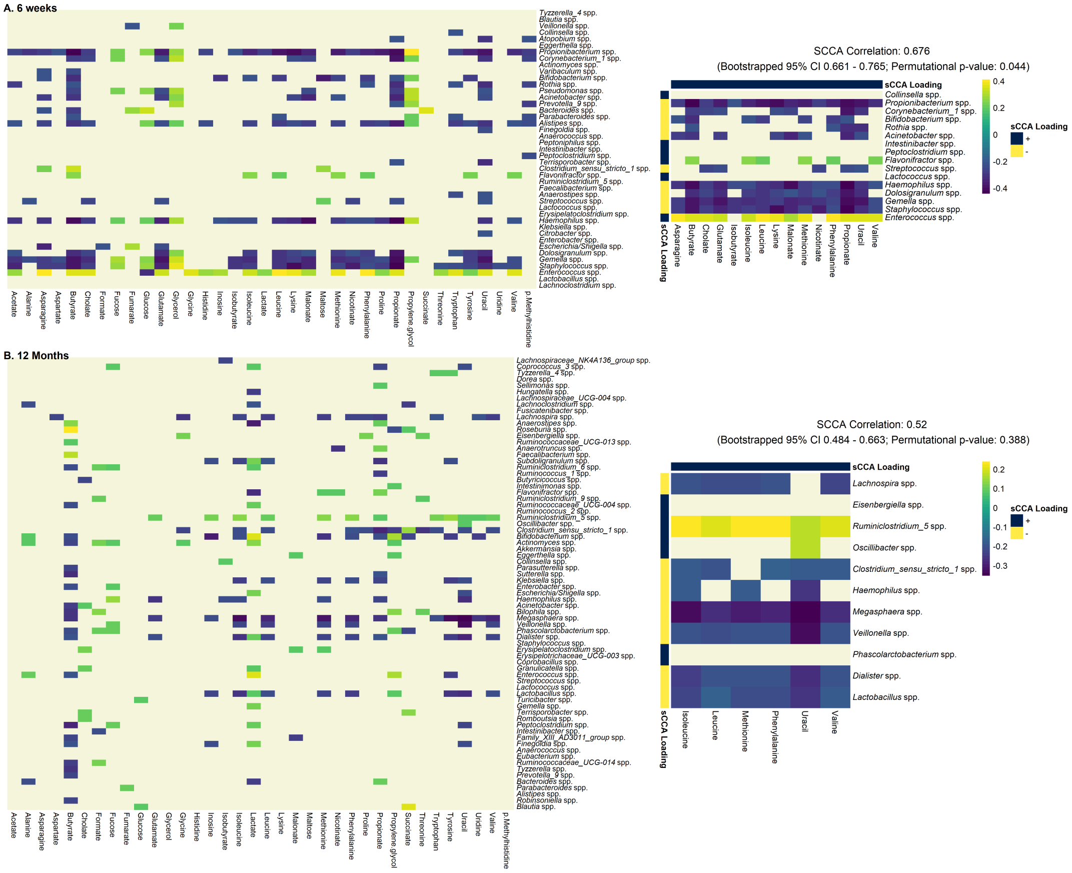

Figure B.7: Pairwise spearman correlation of concentration-fitted targeted metabolite concentrations and genus-level taxonomic abundances for 6-weeks (panel A, $N = 65$) and 12-months (panel B, $N = 65$) infants in sensitivity analyses. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (FDR controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sCCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval (nboot = 5000) and permutation p-value (nperm = 1000) are also reported.

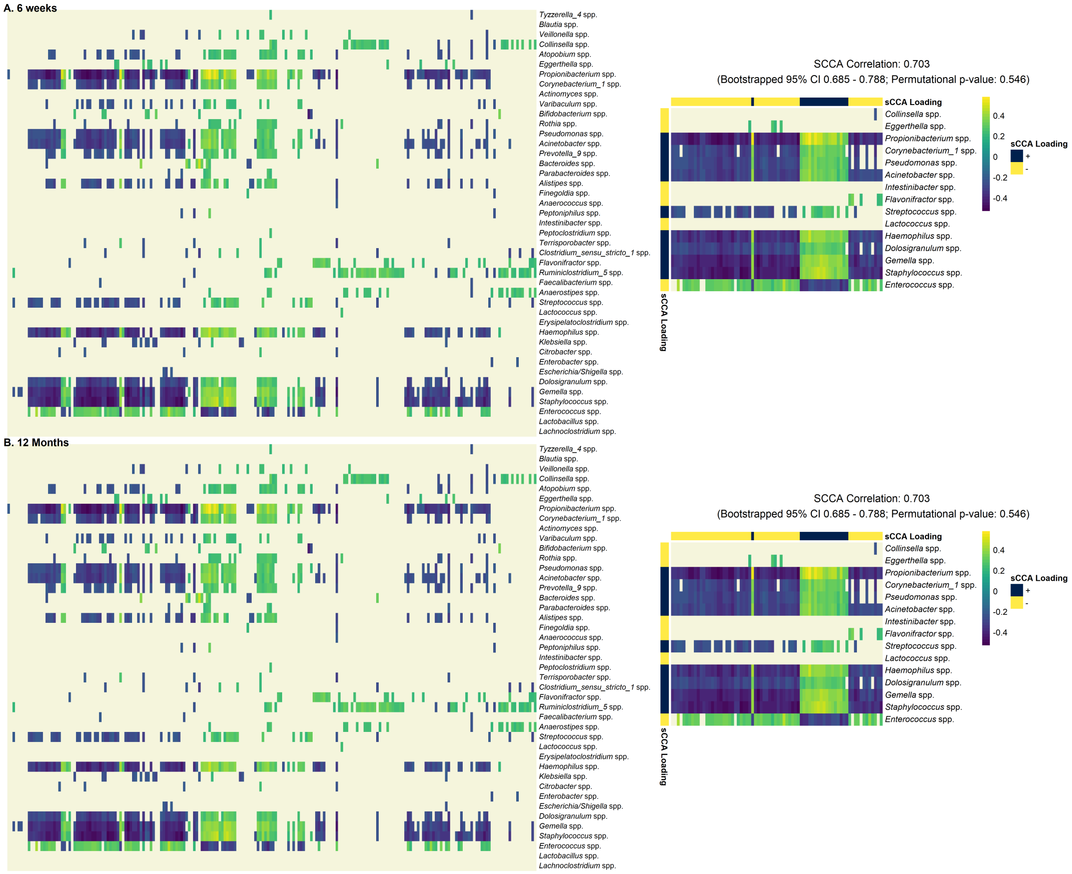

Figure B.8: Pairwise spearman correlation of untargeted metabolite bin relative concentrations and genus-level taxonomic abundances for 6-weeks (panel A, $N = 65$) and 12-months (panel B, $N = 65$) infants in sensitivity analyses. Left panel displays the overall correlation pattern, where non-significant correlations are not colored (FDR controlled q-value < 0.05). Right panel displays the same heatmap restricted to taxa and metabolites selected by the sCCA procedure. Additionally, correlation coefficient of the first sCCA variate pair, bootstrapped 95% confidence interval (`nboot` = 5000) and permutation p-value (`nperm` = 1000) are also reported.

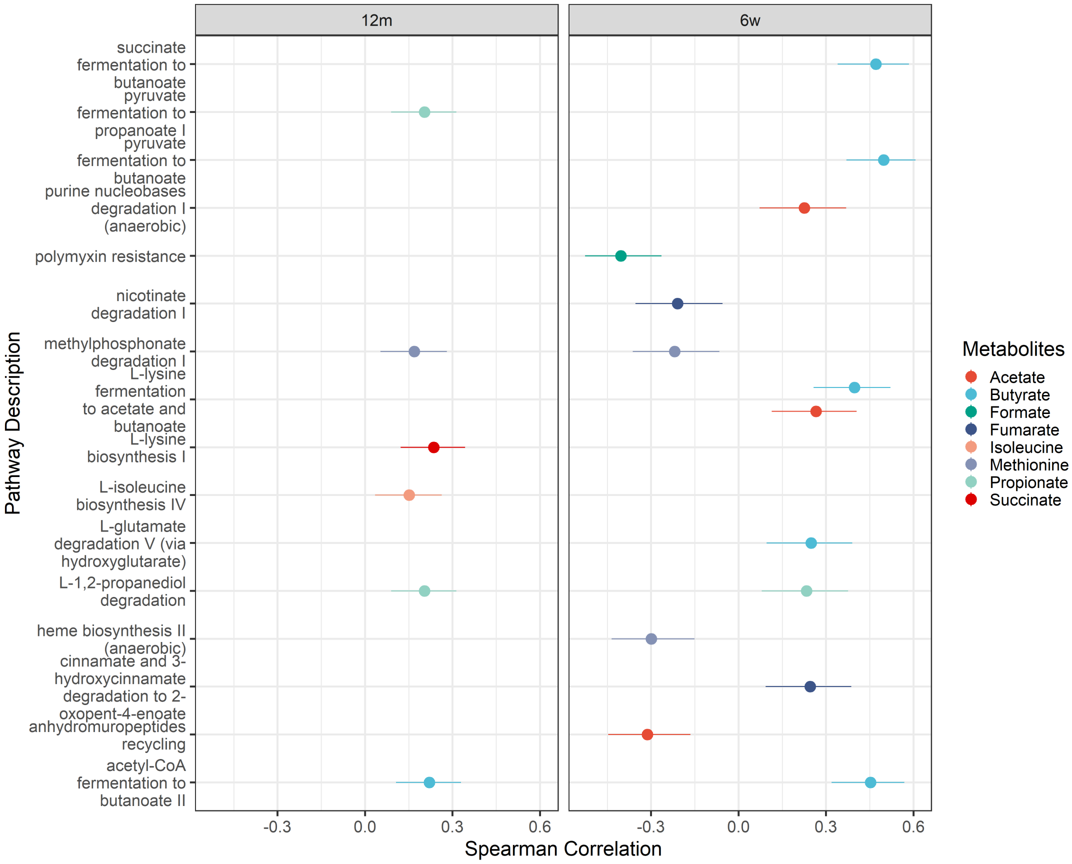

Figure B.9: Spearman correlation coefficients and 95% confidence intervals of significant correlations (q-value > 0.05) between metabolite concentrations in the targeted data set and the abundances of pathways that produce them. Pathway abundances were obtained via PICRUSt2 predictions, with pathway-metabolite relationship retrieved from MetaCyc database. Both 6-week ($N = 158$) and 12-month ($N = 282$) samples are represented.

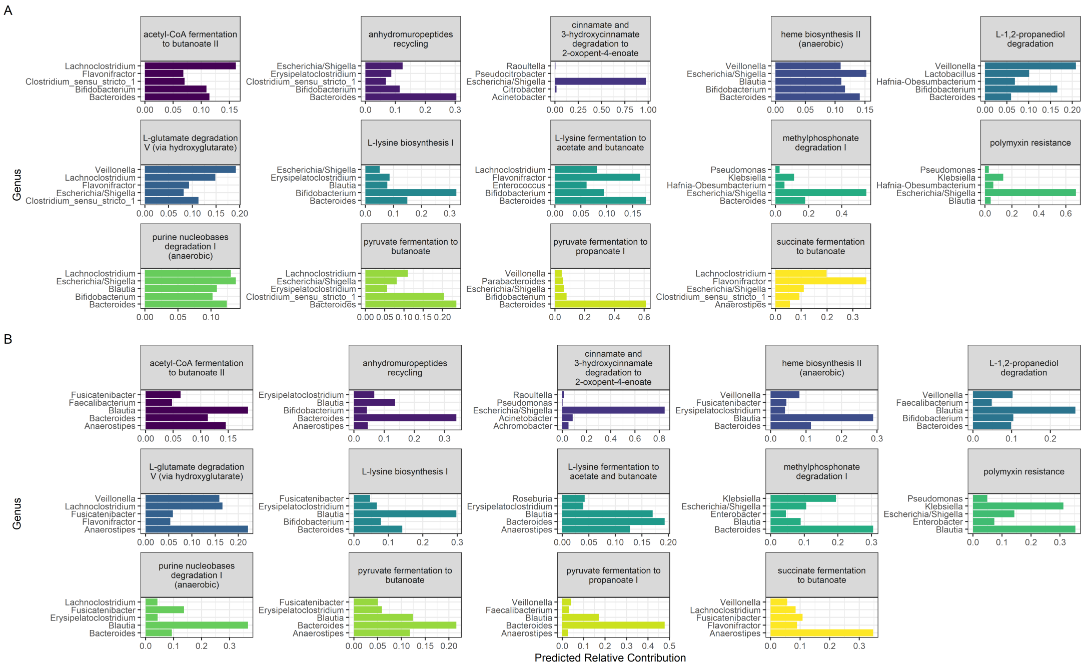

Figure B.10: Top five contributors at the Genus level for each significantly correlated pathway-metabolite pair obtained using observed metabolite concentrations and predicted pathway abundances (Spearman correlation with q-value ¡ 0.05). Panel A represents 6-week samples while panel B represents samples at 12-months. Relative contributions are calculated as the total number of copies of genes mapped to a pathway across all samples per Genus over the total number of gene copies assigned to that pathway.

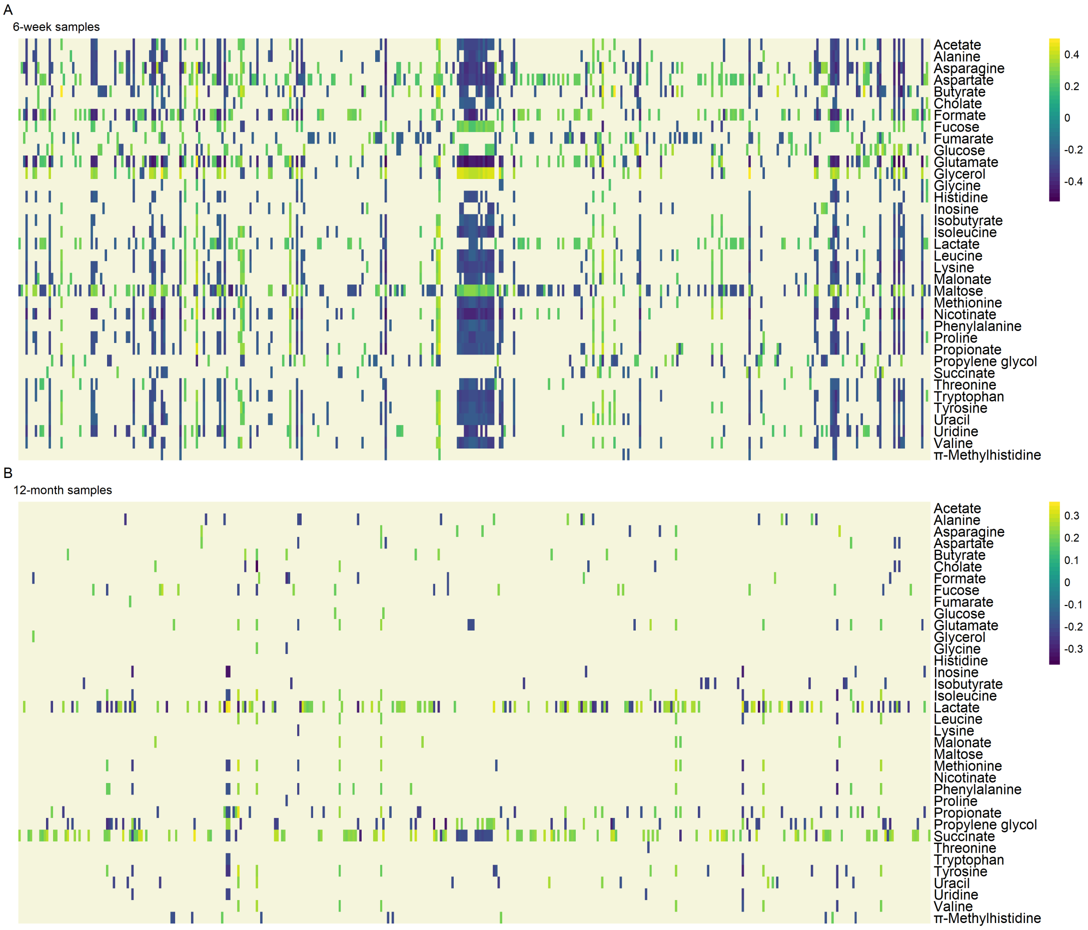

Figure B.11: Heatmap representing overall spearman correlations between predicted pathway abundances (obtained via PICRUSt2) and metabolite concentrations in the targeted data set regardless of pathway-metabolite annotations. Both 6-week ($N = 158$) (Panel A) and 12-month ($N = 282$) (Panel B) samples are presented. Non-significant correlations (q-value > 0.05) are not colored.

Appendix C

Supporting material for “CBEA: Competitive balances for taxonomic enrichment analysis”

Section C.1

Supplementary Note 1

In order to perform inference with CBEA, we estimated the null distribution empirically. This can be done either non-parametrically by constructing a null distribution through computing scores on multiple permutations of the data, or parametrically via estimating parameters of a distribution using the same permuted scores. This means that our null distribution for a given set is equivalent to scores computed for sets of similar sizes but containing randomly chosen taxa. We chose two distributional forms for the null: the normal distribution and a two-component mixture normal distribution. For the normal distribution, we estimated parameters using the maximum likelihood using the *fitdistrplus* package [66]. For the mixture normal distribution, we utilized the expectation-maximization procedure from the package *mixtools* [22].

An advantage of estimating the null using a parametric approach is the ability to correct for variance inflation due to inter-taxa correlation within the set compared to overall background correlation [324]. CBEA addresses this issue by combining the location (or mean) estimate from the column permuted raw score matrix with the spread (or variance) estimate from the original un-permuted scores. This still allows us to leverage the null generated via column permutation while using the proper variance estimate taken from scores where the correlation structure has not been disrupted. As such, this procedure assumes that the variance of the test statistic under the alternate hypothesis is the same as that of the null.

For the normal distribution, it is straightforward to combine mean and variance estimates from the respective raw score matrices. For the mixture normal distribution, however, due to the fact that the distribution is made of component-wise means, variances and mixing coefficients, we decided to take an optimization based approach to identifying the component wise variances σ_1 and σ_2 such the overall mean and variance estimates come from the respective raw cILR score matrices as detailed above. We can write the optimization problem as follows:

$$\begin{aligned} \min_{\sigma_1,\sigma_2} \quad & \sqrt{\left(\sqrt{(\sigma_1+\mu'_1-M')\lambda'_1+(\sigma_2+\mu'_2-M')\lambda'_2}-SD\right)^2} \\ \text{s.t.} \quad & \sigma_1 \geq 10^{-5}, \sigma_2 \geq 10^{-5} \\ \text{where} \quad & M', SD, \mu'_1, \mu'_2, \lambda'_1, \lambda'_2 \text{ are constants} \end{aligned} \tag{C.1}$$

λ'_1, λ'_2, μ'_1, μ'_2, and M' are component-wise mixing coefficients, component-wise means, and overall mean of the mixture distribution estimated from column-permuted scores while SD is the overall standard deviation of the mixture distribution estimated from unpermuted scores. We solve for this optimization problem using a quasi Newton method with box constraints (L-BFGS-B) with the default finite-difference approximation of the gradient, as implemented in the *optim* function in R.

There are also different variations to this approach. We can choose to vary both σ_1 and σ_2 or to keep either σ_1 or σ_2 constant and varying the remaining component. We can assume that null distribution is a two-component mixture distribution where there is one major component with smaller mean (representing the bulk of the distribution centered at the permuted mean) and one minor component with higher mean (representing the inflated right tail). Under this assumption, we can modify the optimization problem to only estimate the variance parameter of the smaller component (i.e. without loss of generalizability keeping σ_1 constant where $\lambda'_1 > \lambda'_2$). This allows for the optimization procedure to more properly capture the right tail distribution rather than increasing weight on the left tail of the distribution which impacts the computation of p-values for a one-sided test. Empirical experiments (data not shown) done on simulated data and random set analyses suggest that this adjustment improves performance of the adjusted CBEA under mixture-normal assumption. In the R implementation of CBEA, users can control this behaviour by specifying the *fix_comp* parameter as part of the *control* argument.

Section C.2

Supplementary Note 2

C.2.1. Design

Even though real data evaluations provide good estimates for performance of CBEA under typical analysis tasks, it does not allow for understanding of the behavior of the model under different effect sizes, correlations, and sparsity. As such, we also perform parametric numerical simulations by generating microbiome count data under the assumption that it follows a zero-inflated negative binomial distribution, which is a good fit for real microbiome relative abundance data [39]. Suppose X_{ij} are observed counts for a sample i and taxon j, then we have the following probability model

$$\mathbf{X}_{ij} = \begin{cases} 0 & \text{with probability } p_j \\ \mathbf{NB}(\mu_j, \phi_j) & \text{with probability } 1 - p_j \end{cases} \tag{C.2}$$

where μ_j and ϕ_j are mean and dispersion parameters, respectively. To incorporate a flexible correlation structure into our simulation model, we utilized the NorTA (Normal to Anything) method [43]. Given an n by p matrix of values $\mathbf{U}$ sampled from multivariate normal distribution with correlation matrix ρ, we can generate target microbiome count vector $\mathbf{X_{.j}}$ for taxa j following the marginal distribution $\mathbf{NB}$ characterized by the negative binomial cumulative distribution function $\mathbb{F}_{\mathbf{NB}}$:

$$\mathbf{X}_{.j} = {\mathbb{F}_{\mathbf{NB}}}^{-1}(\Phi_{U_i}) \tag{C.3}$$

In this instance, for each taxon j, we set elements in $\mathbf{U}_{.j}$ to be zero with probability p_j and applied $\mathbf{NB}^{-1}(\mu_j, \phi_j)$ on non-zero elements to generate our final count matrix $\mathbf{X}$. To ensure that our simulations match closely to real data, we fitted negative binomial distribution using a maximum likelihood approach (with the *fitdistrplus* package in R [66]) to non-zero counts for each taxon from 16S rRNA profiling of stool samples from the Human Microbiome Project (HMP). We take the median values of the estimated mean (μ_j) and dispersion parameters (ϕ_j) as the baseline of our simulations. For simplicity, we assumed that inter-taxa correlation follows an exchangeable structure with correlation equals to ρ.

C.2.2. Scenarios

Simulation scenarios for enrichment analysis at the sample level. To assess type I error rate and power for enrichment significance testing at the sample level, we simulated data based on the schema above, and assessed enrichment for one focal set. Type I error was obtained under the global null as the number of samples where the null hypothesis was rejected at $\alpha = 0.05$ over the total number of samples (which represents the total number of hypotheses tested). Power was obtained using the same formulation as type I error rate but under the global alternate. We treated type I error and power as estimates of binomial proportions and utilized the Agresti-Couli [3] formulation to calculate 95% confidence intervals. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$) and inter-taxa correlation within the set ($\rho = 0, 0.2, 0.5$). For type I error analysis, we also varied the size of the set (50, 100, 150). For power analyses, set size was kept constant at 100 but different effect sizes (fold change of 1.5, 2, and 3). All sample sizes were set at 10,000.

For classifiability, we evaluated the scores against the true labels per sample (indicating the sample has a set with inflated counts) using the area under the receiving operator curve (AUROC). This is a strategy used in Frost [95] which evaluates the informativeness of scores by assessing the relative ranking of samples (i.e. whether samples with inflated counts are highly ranked using estimated scores). DeLong 95% confidence intervals for AUROC [67] were obtained for each estimate. Simulation settings for classification performance were identical to power analyses as detailed in the previous paragraph.

Simulation scenarios for enrichment analysis at the population level. To assess type I error rate and power for inferece at the population level, we simulated data based on the schema above, and assessed the enrichment of 50 sets (with 100 taxa per set) across 10 replicates per simulation condition. Type I error is calculated as the number of enriched sets over the total number of sets for each simulation under the global null. Power is defined similarly, but instead under the global alternate hypothesis. Estimates and confidence intervals for type I error and power are calculated as cross-replicate mean and standard error. Across both analyses, we varied sparsity levels ($p = 0.2, 0.4, 0.6$), and inter-taxa correlation within the set ($\rho = 0, 0.2, 0.5$). For power analyse, we defined an enriched set as a set where all taxa within a set have inflated means of the same effect size. Half of the sets are defined as enriched across case/control status with varying effect sizes (fold change of 1.5, 2, and 3). Due to the compositional nature of microbiome taxonomic data, simple inflation of raw counts would cause an artificial decrease in the abundance of the remaining un-inflated sets. As such, we applied a compensation procedure as described in Hawinkel et al. [113] to ensure the validity of simulation results. All sample sizes were set at 500.

Simulation scenarios for downstream prediction. To assess predictive performance, we generated predictors based on the simulation schema presented above and evaluated prediction for both binary and continuous outcomes using a standard random forest model [33]. For binary outcomes, we use AUROC similar to the classification analyses above. For continuous outcomes, we used root mean squared error (RMSE). All predictive model fitting was performed using *tidymodels* [152] suite of packages. Across both learning tasks, we varied sparsity ($p = 0.2, 0.4, 0.6$), and inter-taxa correlation ($\rho = 0, 0.2, 0.5$). Continuous outcomes Y_{cont} were generated as linear combinations of taxa counts.

$$Y_{cont} = f(\mathbf{X}) + \epsilon \tag{C.4}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $f(\mathbf{X}) = \beta_0 + \mathbf{X}\beta$. For each simulation, we set β_0 to be $\frac{6}{\sqrt{10}}$ similar to [327]. The degree of model saturation (the number of non zero β values) were varied between 0.1 and 0.5, and signal to noise ratio (SNR $= \frac{\sigma(f(\mathbf{X}))}{\sigma_\epsilon}$) was varied between 1.5, 2, and 3.

For binary outcomes, we generate Y_{binary} as Bernoulli draws with probability

p_{binary}, where

$$p_{binary} = \frac{1}{1 + \exp(f(\mathbf{X}) + \epsilon)} \tag{C.5}$$

To ensure a balance of classes, we applied the strategy described in Dong et al. [71] where the associated β values are evenly split between positive and negative associations. All data sets generated from prediction tasks have 2,000 samples with 5,000 taxa over 50 sets with a size of 100 taxa per set.

C.2.3. Results

Statistical Inference. Fig C.1 demonstrate type I error evaluations for sample-level inference with CBEA compared to the Wilcoxon rank sum test, which uses a rank-based statistic to compare the mean count difference between taxa in the set its complement. All methods demonstrate good type I error control at $\alpha = 0.05$ under zero correlation across all simulation conditions. However, under both medium ($\rho = 0.2$) and high ($\rho = 0.5$) correlation settings, both the Wilcoxon test and unadjusted CBEA variants show high levels of inflated type I error, where Wilcoxon test performed the worst. On the other hand, adjusted CBEA methods (under both distributions) control for type I error at the appropriate α level even at high correlations. This is opposite from our real data evaluations, where adjusted CBEA demonstrated inflated type I error under random set evaluations.

We also assessed the ability to perform inference at the population level using CBEA similar to GSVA [111]. Here, we test for enrichment of sets across case/control status by generating CBEA scores and performing Welch's t-test as a difference in means test. We compared the performance of this approach with CBEA and two commonly used methods for differential abundance testing in the microbiome literature: DESeq2 [177] and corncob [191]. Fig. C.2 present results for simulation studies for both type I error evaluations. All methods were able to control for type I error

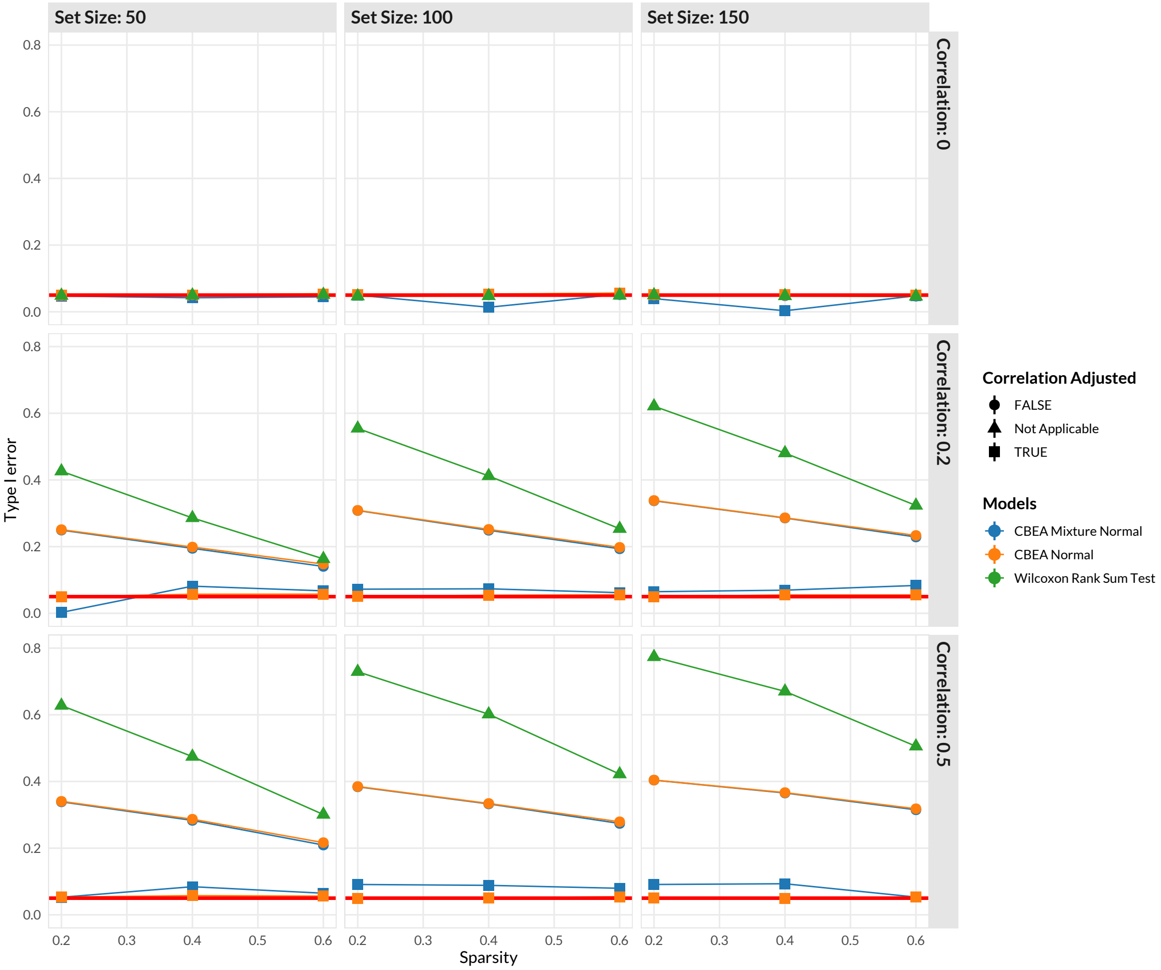

Figure C.1: Simulation results for type I error evaluation for CBEA sample-level inference. Type I error rate (y axis) was estimated for each approach across data sparsity levels (x axis) across different set sizes (horizontal) and inter-taxa correlation within the set (vertical). We compared variatns of CBEA against a Wilcoxon rank sum test at α of 0.05. For each scenario, a data set of 10,000 samples (equivalent to 10,000 hypotheses) was utilized. Confidence bounds were obtained using Agresti-Couli approach.

across both sparsity and correlation levels, where medium level sparsity ($p = 0.4$) and correlation ($\rho = 0.2$) showed the strongest performance. In these scenarios, the CDF values of CBEA generated under the adjusted mixture normal distribution performed the best. This is different than our real data evaluations, where corncob and DESeq2 showed increased type I error.

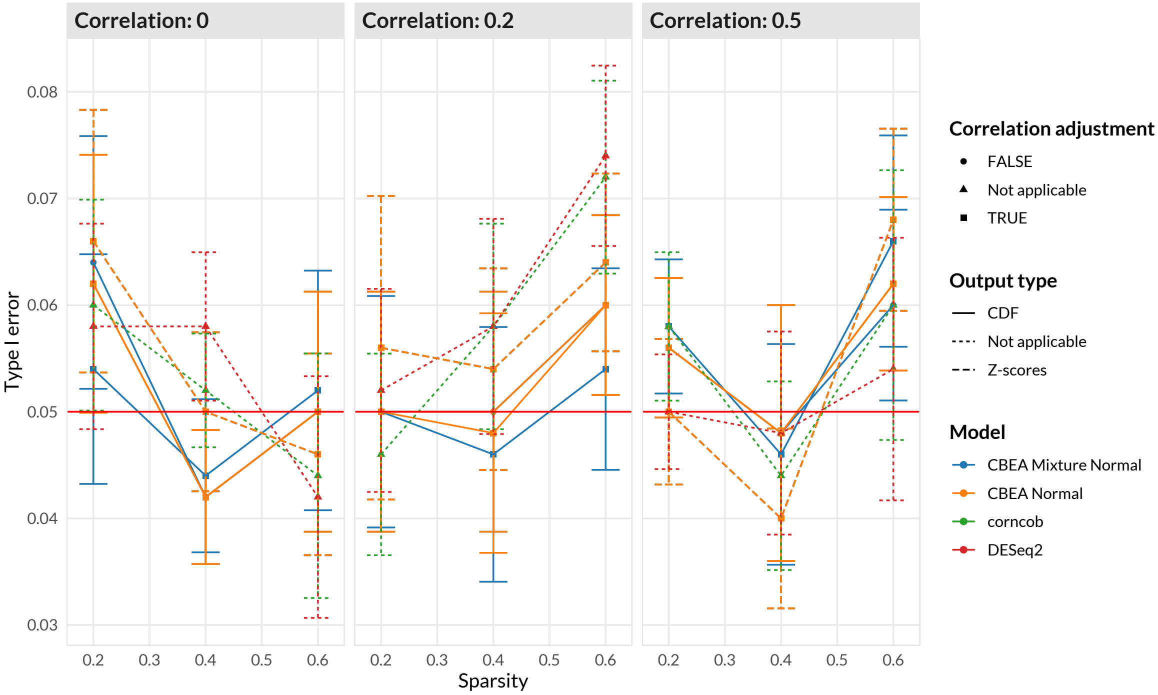

Figure C.2: Simulation results for type I error evaluation for CBEA population-level inference. Type I error (y-axis) was estimated as the average proportion of sets with significant enrichment at 0.05 across 10 replications per simulation condition under the global null. Error bars were estimated using standard errors computed across 10 replicated data sets. Performance was evaluated across different sparsity (x-axis) and inter-taxa correlation levels. For CBEA methods, enrichment analysis was performed using a Welch's t-test across case/control status with single sample scores representing set-based features generated by CBEA (across different output types and distributional assumptions). For corncob and DESeq2, set-based features were constructed using element-wise summations.

Phenotype relevance. We assessed phenotype relevance similar to the main manuscript by assessing statistical power and score rankings via AUROC. Results for this analysis is shown in Fig C.3. For statistical power (panel A), under low-correlation settings, all CBEA approaches demonstrate similar power, with the unadjusted methods being slightly more performant at low effect sizes. Notably, all CBEA variants outperformed the Wilcoxon rank sum test. However, as correlation increases, the adjusted CBEA variants showed much lower power, congruent with the perspective that the adjustment process is conservative and trades off power for type I error control. Even at the highest effect size (fold change of 3 in means) adjusted CBEA does not approach 0.8. For score rankings (panel B), all methods are close together in performance, with the Wilcoxon W statistic being the worst performer. These results are similar to that of our real data evaluations. Notably, performance values were not affected by sparsity and increases to near perfect prediction at higher effect sizes.

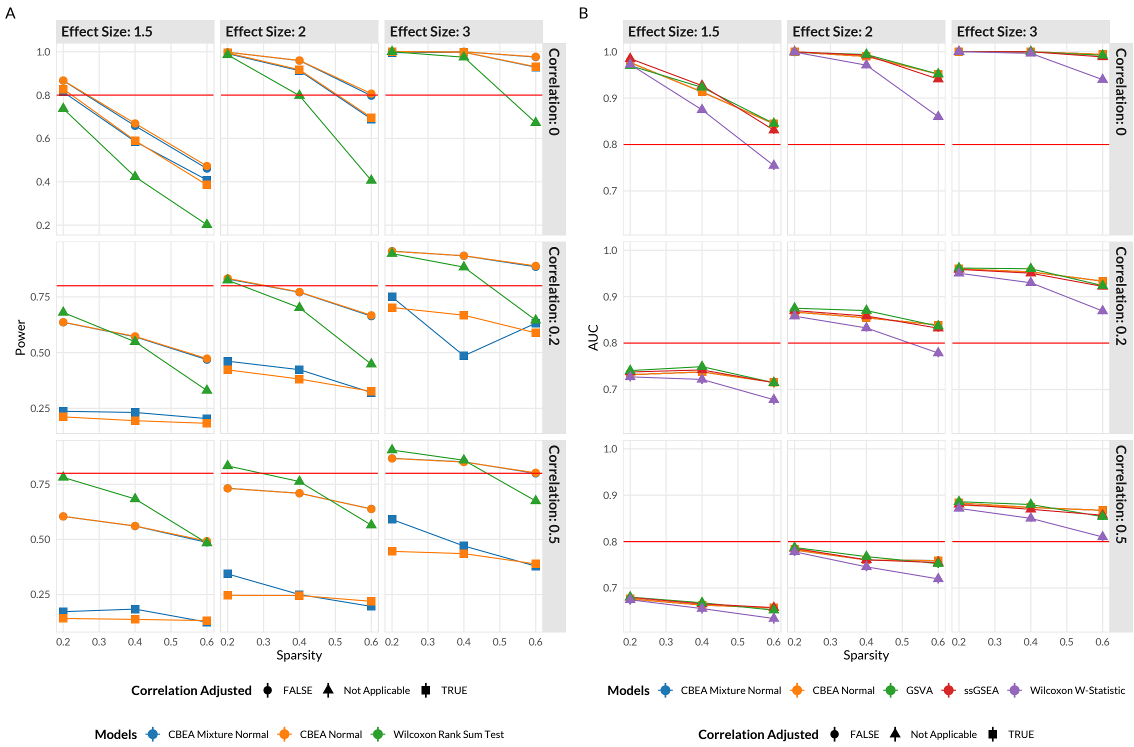

Figure C.3: Simulation results for phenotype relevance evaluation for CBEA sample-level inference. **(A)** demonstrate statistical power (y-axis) across different data sparsity levels (x-axis) and power **(B)** for differential abundance test across different parametric simulation scenarios. For CBEA methods, differential abundance analysis was performed using a difference in means test (either Wilcoxon rank-sum test or Welch's t-test) across case/control status using single sample scores generated by CBEA (across different output types and distributional assumptions). CBEA associated methods demonstrated similar type I error to conventional differential abundance analysis methods but with more power to detect differences even at small effect sizes.

For population-level analyses, we assessed phenotype relevance as statistical power to detect sets that were simulated to be significantly enriched. Fig C.4 showed these results. As expected, power decreases with increasing sparsity, where the effect was attenuated at lower effect sizes. Correlation did impact power, however the difference was not notable. Interestingly, at lower effect sizes both DESeq2 and corncob has comparable power with CBEA, however as effect size increases, the difference in performance values became more stark. This is different than our real data evaluations, where power was more comparable (with slight advantage to corncob and DESeq2).

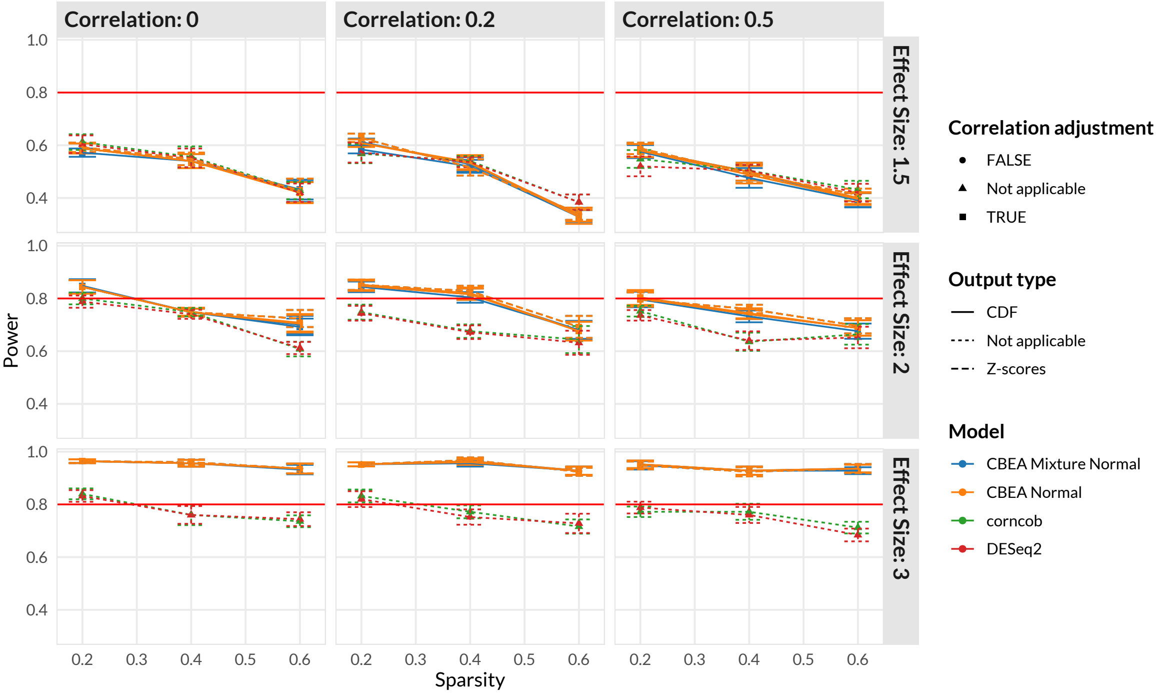

Figure C.4: Simulation results for phenotype relevance evaluation for CBEA population-level inference. Power (y-axis) was estimated as the average proportion of sets correctly identified as significantly enriched (at 0.05) across 10 replications per simulation condition under the global null. Error bars were estimated using standard errors computed across 10 replicated data sets. Performance was evaluated across different sparsity (x-axis) and inter-taxa correlation levels. For CBEA methods, enrichment analysis was performed using a Welch's t-test across case/control status with single sample scores representing set-based features generated by CBEA (across different output types and distributional assumptions). For corncob and DESeq2, set-based features were constructed using element-wise summations.

Predictive analysis. Fig C.5 shows results for simulation studies as detailed in the Methods section. Panel A presents results for a regression learning task with a continuous outcome. We observed no difference in performance values across different evaluation models. Overall, prediction error did not change across sparsity levels, and decreases with increasing signal-to-noise ratio (SNR). However, at higher correlation levels, the pattern was more erratic. For example, when ρ is set at 0.5, higher SNR decreases performance only at low sparsity $p = 0.2$, but had the expected pattern medium sparsity $p = 0.4$, where higher SNR correlates with improved performance. This effect seems to be specific to low effect saturation scenarios (only a limited number of taxa sets are associated with the outcome). Interestingly, higher sparsity levels produces better performance.

Panel B represent results for a classification task with a binary outcome with AUROC as the evaluation criteria. Here we observed similar results as that of our real data evaluations, where using CLR transfromed data produces more predictive models across all scenarios. Conversely, GSVA and ssGSEA were consistently under performing when compared to CBEA and CLR. Interestingly, the degree of difference varies across sparsity and inter-taxa correlations. We noticed that increasing sparsity and correlation decreases the gap in performance between CBEA and CLR, while increasing the gap in performance between CBEA/CLR and GSVA/ssGSEA. As such, we can hypothesize that both GSVA and ssGSEA are more sensitive to the degree of inter-taxa correlation and sparsity. Finally, effect saturation did not change model rankings, but did decrease overall performance.

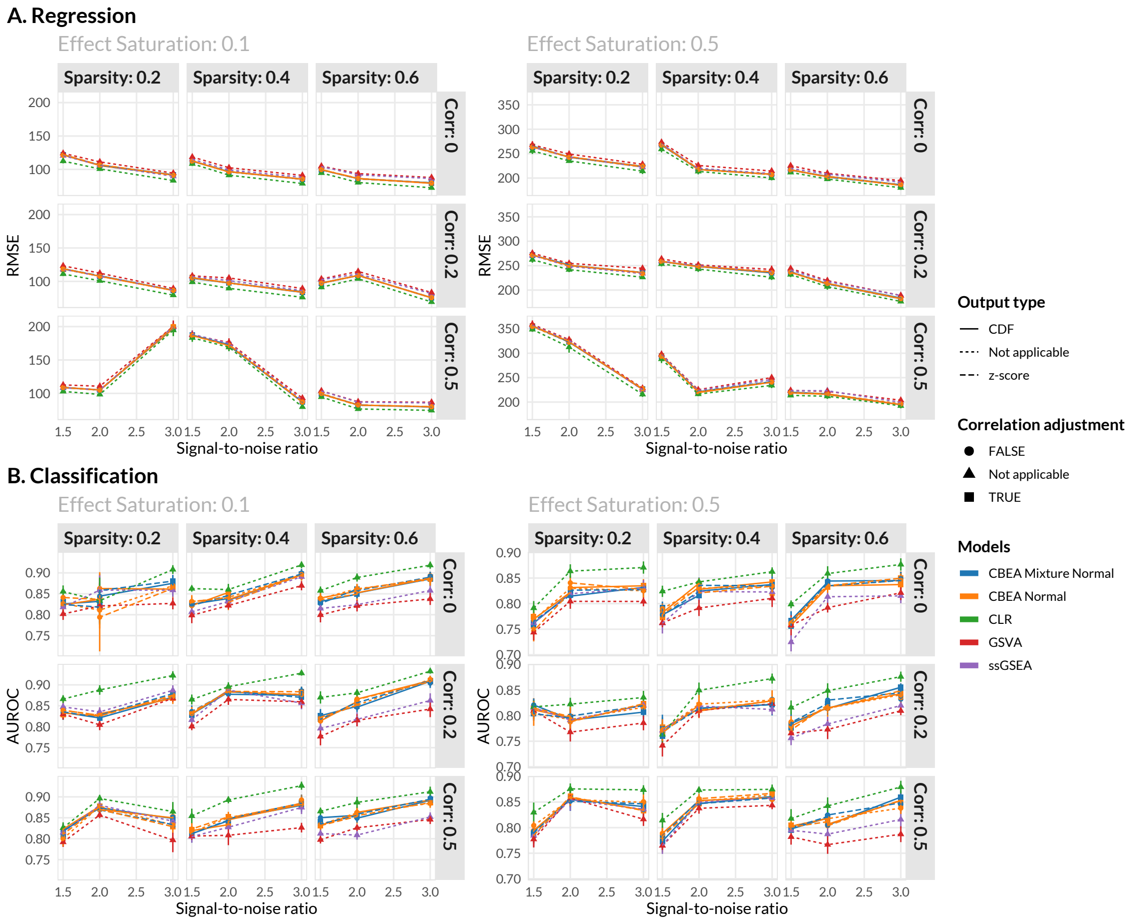

Figure C.5: Simulation results for predictive pefromance evaluation for CBEA. Predictive performance of a random forest model (with no hyperparameter tuning) trained on set-based features as inputs. Methods to generate these features include CBEA, ssGSEA, GSVA, and the CLR transformation applied on sum-aggregated sets. Simulation data was generated across different levels of data sparsity, inter-taxa correlation, effect saturation, and signal-to-noise ratio. Panel **(A)** presents performance on a regression task using RMSE (root mean squared error) as the evaluation measure. Panel **(B)** presents performance on a classification task with AUROC as the evaluation measure.

Section C.3

Supplementary Note 3

We implemented CBEA in the package *CBEA* on GitHub (`https://www.github.com/qpmnguyen/CBEA`). We evaluated computational time using the *bench* package in R. We applied CBEA to a standard data set generated using our simulation model consisting of 40 sets (of size 20 each) and 500 samples. Benchmark was performed on a single core using a node on the computing cluster (Specifications: Intel Xeon E5-2690 (2.6GHz) with 4GB of RAM).

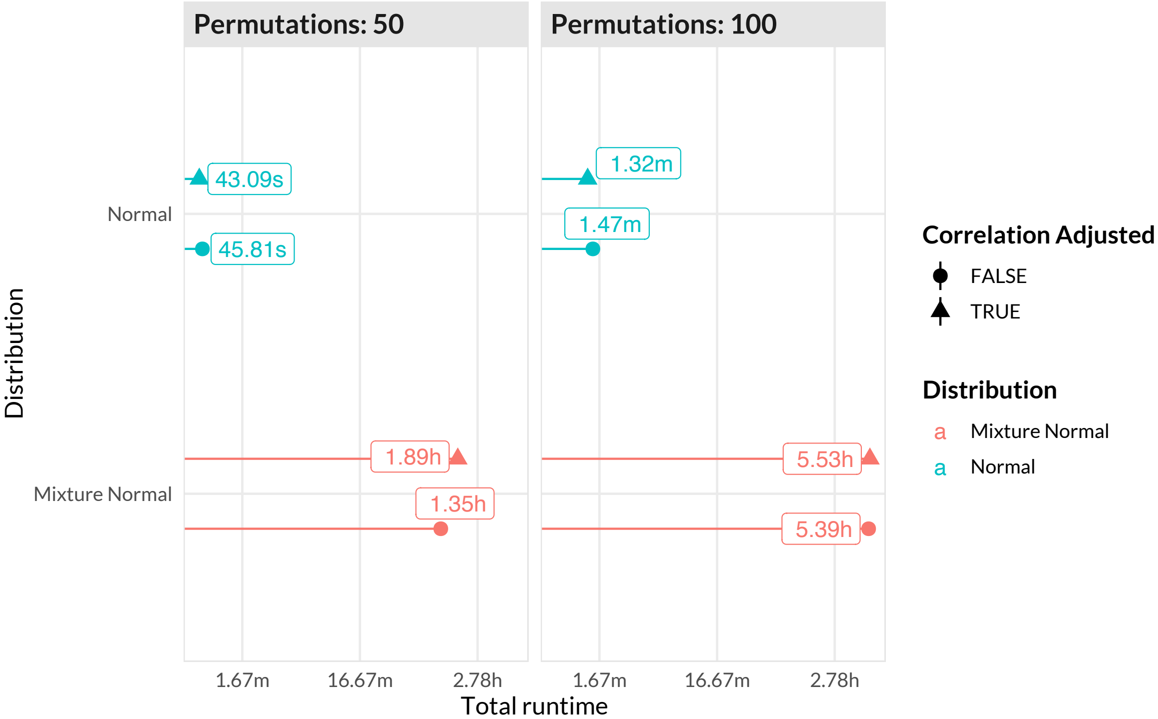

Figure C.6: Runtime performance. Overall runtime of CBEA under different parameters for a data set of 500 samples, 800 taxa (40 sets of size 20 each). This data set was generated via simulations.

In general, using the normal distribution is the fastest approach regardless of the total number of permutations performed (2 minutes). However, using the mixture normal distribution increased the runtime many folds, especially for the adjusted approach (highest was 5.53 hours). This time scales with the number of sets evaluated, as well as the number of samples. We believe this is due to the procedure used to estimate parameters of the mixture normal distrubiton as implemented in the *mixtools* package. The default parameters used in CBEA also increased the runtime in order to reduce convergence issues.

In order to reduce runtime, users can attempt the following: Since CBEA fits parametric distributions over permuted values of all samples within a data set (i.e. for $N = 100$ and 10 permutations, the fitting procedure will attempt to estimate parameters from a vector of size 1000 for each set), if the sample size is high users can reduce the number of permutations. Additionally, CBEA also implements a procedure to parallelize computation across sets, which might be applicable to situations where there are a lot of sets to evaluate. Finally, a lot of CBEA approaches work well without parametric fit, so users can use the non-parametric approaches like the permutation test or using raw CBEA scores.

Section C.4

Supplementary Note 4

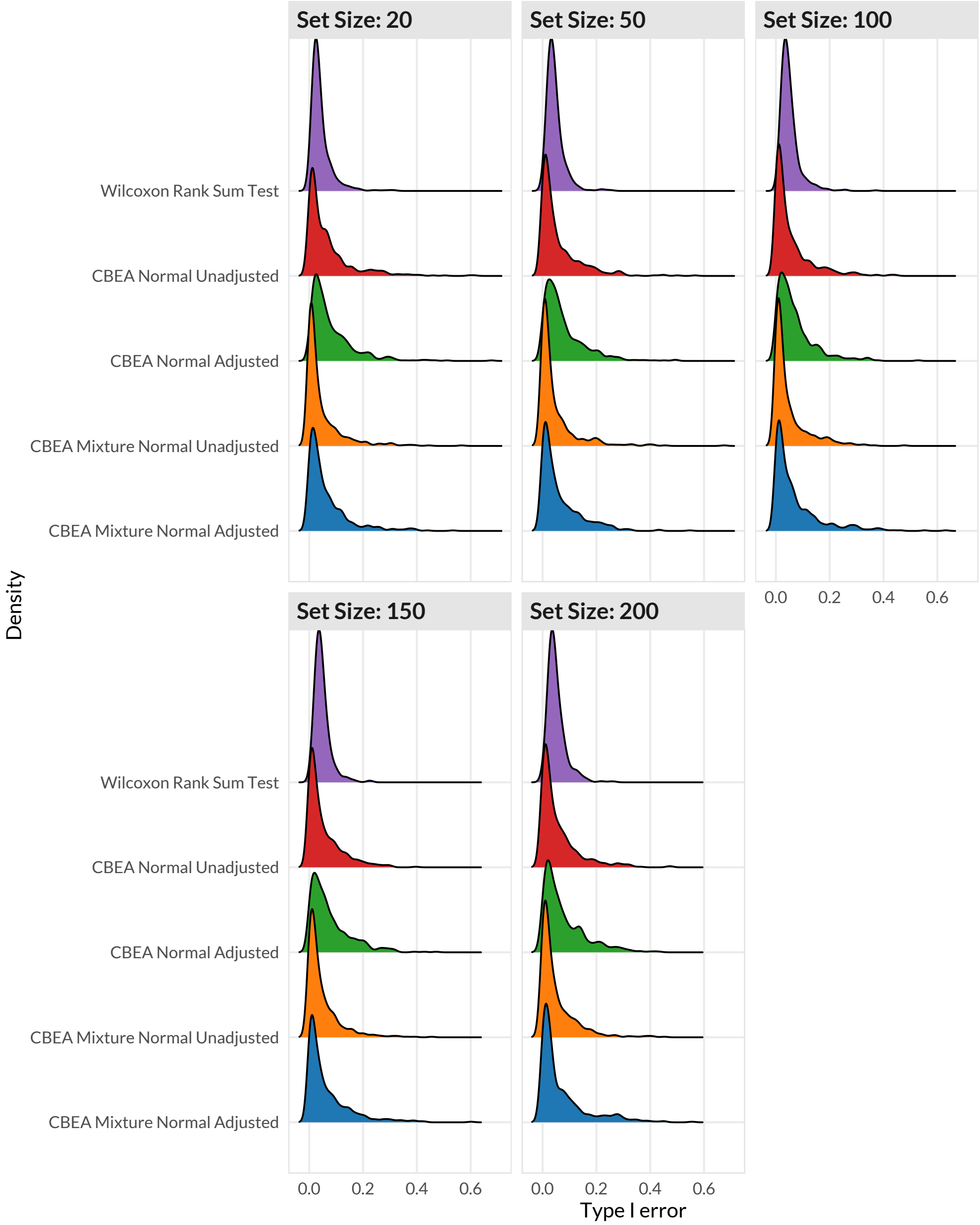

Figure C.7: Distribution of type I error values across all replications in real data random set evaluations for CBEA inference at the sample-level. Density (y-axis) for type I error values (x-axis) of each evaluated approach for sample-level inference using real data across 500 replications. Here, type I error was estimated as the proportion of samples where a randomly sampled set of different sizes where identified to be statistically significant at p-value threshold of 0.05.

Appendix D

Supporting material for “Evaluating trait databases for taxon set enrichment analysis”

Section D.1

Supplemental Figures

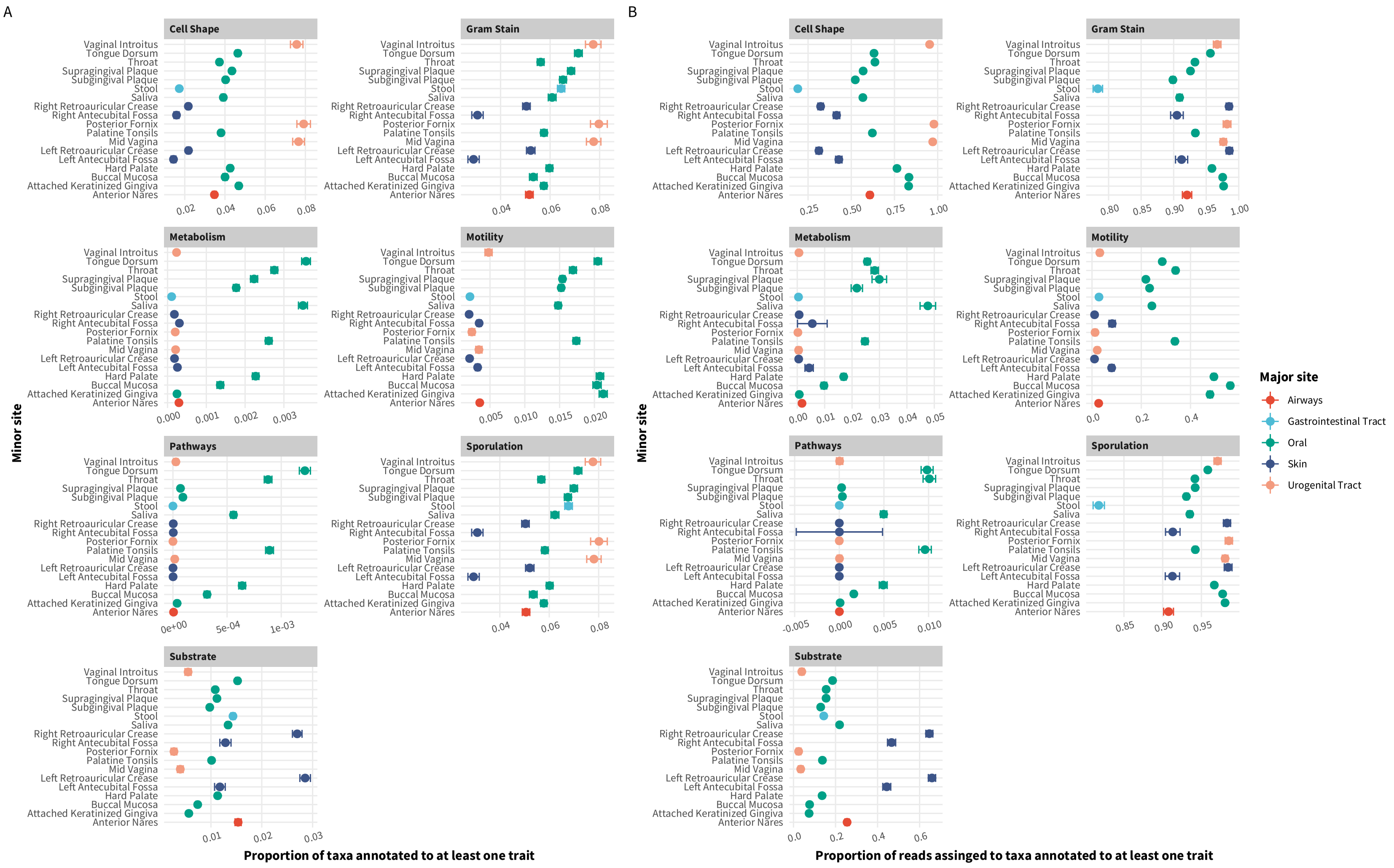

Figure D.1: Trait coverage statistics for samples profiled with 16S rRNA gene metabarcoding. Panel **(A)** illustrates the proportion of present taxa per sample annotated to at least one trait. Panel **(B)** illustrates the proportion of reads assigned to taxa annotated to at least one trait which accounts for taxa relative abundances. Each plot facet represents different trait categories that were evaluated. Error bar represents the standard error of the evaluation statistic of interest across the the total number of samples evaluated per body site.

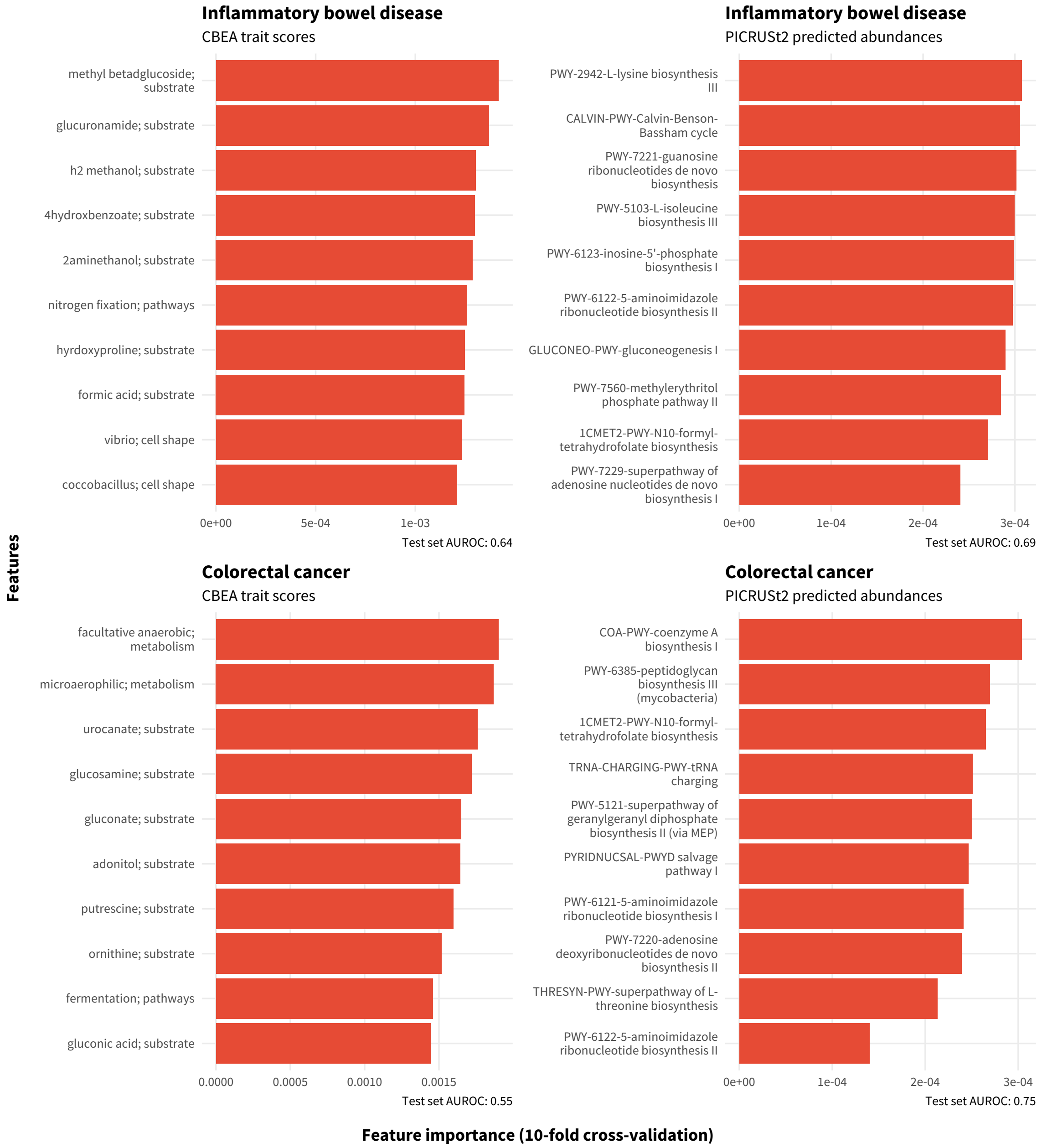

Figure D.2: Top 10 important features based on random forest model fitted different inputs from data sets profiled with 16S rRNA gene sequencing. Features were selected from mean decrease in Gini impurity averaged across 500 decision trees and 10-fold cross-validation (nested with the training set) as implemented in `scikit-learn`. AU-ROC scored on a held-out test set is also presented for each input type and disease condition.

Bibliography

[1] Jørn A. Aas, Bruce J. Paster, Lauren N. Stokes, Ingar Olsen, and Floyd E. Dewhirst, *Defining the normal bacterial flora of the oral cavity*, J Clin Microbiol **43** (2005), no. 11, 5721–5732.

[2] Marit Ackermann and Korbinian Strimmer, *A general modular framework for gene set enrichment analysis*, BMC bioinformatics **10** (2009), no. 1, 1–20.

[3] Alan Agresti and Brent A. Coull, *Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions*, The American Statistician **52** (1998), no. 2, 119–126.

[4] J. Aitchison, *The Statistical Analysis of Compositional Data*, Journal of the Royal Statistical Society: Series B (Methodological) **44** (1982), no. 2, 139–160.

[5] J. Aitchison and S.M. Shen, *Logistic-normal distributions:Some properties and uses*, Biometrika **67** (1980), no. 2, 261–272.

[6] John Aitchison, *Principles of compositional data analysis*, Lecture Notes-Monograph Series (1994), 73–81.

[7] Steven D. Allison and Jennifer B. H. Martiny, *Resistance, resilience, and redundancy in microbial communities*, PNAS **105** (2008), no. Supplement 1, 11512–11519.

[8] Dingding An, Sungwhan F. Oh, Torsten Olszak, Joana F. Neves, Fikri Y. Avci, Deniz Erturk-Hasdemir, Xi Lu, Sebastian Zeissig, Richard S. Blumberg, and Dennis L. Kasper, *Sphingolipids from a Symbiotic Microbe Regulate Homeostasis of Host Intestinal Natural Killer T Cells*, Cell **156** (2014), no. 1-2, 123–133.

[9] Antoine Aoun, Fatima Darwish, and Natacha Hamod, *The Influence of the Gut Microbiome on Obesity in Adults and the Role of Probiotics, Prebiotics, and Synbiotics for Weight Loss*, Prev Nutr Food Sci **25** (2020), no. 2, 113–123.

[10] Marie-Claire Arrieta, Leah T. Stiemsma, Nelly Amenyogbe, Eric M. Brown, and Brett Finlay, *The Intestinal Microbiome in Early Life: Health and Disease*, Front. Immunol. **5** (2014).

[11] Marie-Claire Arrieta, Leah T. Stiemsma, Pedro A. Dimitriu, Lisa Thorson, Shannon Russell, Sophie Yurist-Doutsch, Boris Kuzeljevic, Matthew J. Gold, Heidi M. Britton, Diana L. Lefebvre, Padmaja Subbarao, Piush Mandhane, Allan Becker, Kelly M. McNagny, Malcolm R. Sears, Tobias Kollmann, the CHILD Study Investigators, William W. Mohn, Stuart E. Turvey, and B. Brett Finlay, *Early infancy microbial and metabolic alterations affect risk of childhood asthma*, Sci. Transl. Med. **7** (2015), no. 307.

[12] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock, *Gene Ontology: Tool for the unification of biology*, Nat Genet **25** (2000), no. 1, 25–29.

[13] Funmilola A. Ayeni, Elena Biagi, Simone Rampelli, Jessica Fiori, Matteo Soverini, Haruna J. Audu, Sandra Cristino, Leonardo Caporali, Stephanie L.

Schnorr, Valerio Carelli, Patrizia Brigidi, Marco Candela, and Silvia Turroni, *Infant and Adult Gut Microbiome and Metabolome in Rural Bassa and Urban Settlers from Nigeria*, Cell Rep **23** (2018), no. 10, 3056–3067.

[14] Fredrik Bäckhed, Hao Ding, Ting Wang, Lora V. Hooper, Gou Young Koh, Andras Nagy, Clay F. Semenkovich, and Jeffrey I. Gordon, *The gut microbiota as an environmental factor that regulates fat storage*, Proc. Natl. Acad. Sci. U.S.A. **101** (2004), no. 44, 15718–15723.

[15] Fredrik Bäckhed, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, Yan Xia, Hailiang Xie, Huanzi Zhong, Muhammad Tanweer Khan, Jianfeng Zhang, Junhua Li, Liang Xiao, Jumana Al-Aama, Dongya Zhang, Ying Shiuan Lee, Dorota Kotowska, Camilla Colding, Valentina Tremaroli, Ye Yin, Stefan Bergman, Xun Xu, Lise Madsen, Karsten Kristiansen, Jovanna Dahlgren, and Jun Wang, *Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life*, Cell Host & Microbe **17** (2015), no. 5, 690–703.

[16] Monika Balvočiūtė and Daniel H. Huson, *SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?*, BMC Genomics **18** (2017), no. S2, 114.

[17] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch, Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov,

David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn, *Systematic RNA interference reveals that oncogenic* KRAS-*driven cancers require TBK1*, Nature **462** (2009), no. 7269, 108–112.

[18] Olaf Beckonert, Hector C. Keun, Timothy M. D. Ebbels, Jacob Bundy, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson, *Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts*, Nat Protoc **2** (2007), no. 11, 2692–2703.

[19] Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A Franzosa, and Nicola Segata, *Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3*, eLife **10** (2021), e65088.

[20] Francesco Beghini, Audrey Renson, Christine P. Zolnik, Ludwig Geistlinger, Mykhaylo Usyk, Thomas U. Moody, Lorna Thorpe, Jennifer B. Dowd, Robert Burk, Nicola Segata, Heidi E. Jones, and Levi Waldron, *Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study*, Ann Epidemiol **34** (2019), 18–25.e3.

[21] Yasmine Belkaid and Timothy W. Hand, *Role of the Microbiota in Immunity and Inflammation*, Cell **157** (2014), no. 1, 121–141.

[22] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young, *Mixtools: An R package for analyzing finite mixture models*, Journal of Statistical Software **32** (2009), no. 6, 1–29.

[23] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon, *Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis*, Journal of Machine Learning Research **18** (2017), no. 77, 1–36.

[24] Yoav Benjamini and Yosef Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological) **57** (1995), no. 1, 289–300.

[25] Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter, *Microbiome definition re-visited: Old concepts and new challenges*, Microbiome **8** (2020), no. 1, 103.

[26] Sharon Bewick, Eliezer Gurarie, Jake L. Weissman, Jess Beattie, Cyrus Davati, Rachel Flint, Peter Thielen, Florian Breitwieser, David Karig, and William F. Fagan, *Trait-based analysis of the human skin microbiome*, Microbiome **7** (2019), no. 1, 101.

[27] Luc Biedermann, Jonas Zeitz, Jessica Mwinyi, Eveline Sutter-Minder, Ateequr Rehman, Stephan J. Ott, Claudia Steurer-Stey, Anja Frei, Pascal Frei, Michael Scharl, Martin J. Loessner, Stephan R. Vavricka, Michael Fried, Stefan Schreiber, Markus Schuppler, and Gerhard Rogler, *Smoking Cessation Induces*

Profound Changes in the Composition of the Intestinal Microbiota in Humans, PLoS ONE **8** (2013), no. 3, e59260.

[28] F. Blachier, F. Mariotti, J. F. Huneau, and D. Tomé, *Effects of amino acid-derived luminal metabolites on the colonic epithelium and physiopathological consequences*, Amino Acids **33** (2007), no. 4, 547–562.

[29] Nicholas A. Bokulich, Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso, *Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin*, Microbiome **6** (2018), no. 1, 90.

[30] Nicholas A. Bokulich, Sathish Subramanian, Jeremiah J. Faith, Dirk Gevers, Jeffrey I. Gordon, Rob Knight, David A. Mills, and J. Gregory Caporaso, *Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing*, Nature methods **10** (2013), no. 1, 57.

[31] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights,

Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J. J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso, *Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2*, Nature Biotechnology **37** (2019), no. 8, 852–857.

[32] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik, *A Training Algorithm for Optimal Margin Classifiers*, Proceedings of the Fifth Annual Workshop on Computational Learning Theory (New York, NY, USA), COLT '92, ACM, 1992, pp. 144–152.

[33] Leo Breiman, *Random Forests*, Machine Learning **45** (2001), no. 1, 5–32.

[34] Glenn W. Brier, *VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY*, Mon. Wea. Rev. **78** (1950), no. 1, 1–3.

[35] SL Brilleman, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe, *Joint longitudinal and time-to-event models via Stan.*

[36] Lauren Brink, Sree Chintapalli, Kelly Mercer, Brian Piccolo, Sean Adams, Anne Bowlin, Katelin Matazel, Kartik Shankar, Thomas Badger, A. Andres, and Laxmi Yeruva, *Early Postnatal Diet Differentially Affects the Fecal Microbiome and Metabolome (FS04-02-19)*, Curr Dev Nutr **3** (2019), no. Supplement_1.

[37] David Broadhurst, Royston Goodacre, Stacey N. Reinke, Julia Kuligowski, Ian D. Wilson, Matthew R. Lewis, and Warwick B. Dunn, *Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies*, Metabolomics **14** (2018), no. 6, 72.

[38] Matteo Calgaro, *Mcalgaro93/sc2meta: Paper Release*, Zenodo, July 2020.

[39] Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo, *Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data*, Genome Biology **21** (2020), no. 1, 191.

[40] Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes, *DADA2: High-resolution sample inference from Illumina amplicon data*, Nature Methods **13** (2016), no. 7, 581–583.

[41] Patrice D. Cani, *Human gut microbiome: Hopes, threats and promises*, Gut **67** (2018), no. 9, 1716–1725.

[42] Dong-Sheng Cao, Shao Liu, Wen-Bin Zeng, and Yi-Zeng Liang, *Sparse canonical correlation analysis applied to -omics studies for integrative analysis and biomarker discovery*, Journal of Chemometrics **29** (2015), no. 6, 371–378.

[43] Marne C Cario, *Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix*, Tech. report, 1997.

[44] Hannah C. Carrow, Lakshmi E. Batachari, and Hiutung Chu, *Strain diversity in the microbiome: Lessons from Bacteroides fragilis*, PLoS Pathog **16** (2020), no. 12, e1009056.

[45] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp, *The MetaCyc database of metabolic pathways and enzymes - a 2019 update*, Nucleic Acids Research **48** (2020), no. D1, D445–D453.

[46] Scott Chamberlain and Zebulun Arendsee, *Taxizedb: Tools for working with 'taxonomic' databases*, 2021.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*, jair **16** (2002), 321–357.

[48] Jun Chen, Kyle Bittinger, Emily S. Charlson, Christian Hoffmann, James Lewis, Gary D. Wu, Ronald G. Collman, Frederic D. Bushman, and Hongzhe Li, *Associating microbiome composition with environmental covariates using generalized UniFrac distances*, Bioinformatics **28** (2012), no. 16, 2106–2113.

[49] Yiwen Cheng, Zongxin Ling, and Lanjuan Li, *The Intestinal Microbiota and Colorectal Cancer*, Front. Immunol. **11** (2020).

[50] Marlène Chiarello, Mark McCauley, Sébastien Villéger, and Colin R. Jackson, *Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold*, PLoS ONE **17** (2022), no. 2, e0264443.

[51] Marcus C Chibucos, Adrienne E Zweifel, Jonathan C Herrera, William Meza, Shabnam Eslamfam, Peter Uetz, Deborah A Siegele, James C Hu, and Michelle G Giglio, *An ontology for microbial phenotypes*, BMC Microbiol **14** (2014), no. 1, 294.

[52] Ilseung Cho and Martin J. Blaser, *The human microbiome: At the interface of health and disease*, Nature Reviews Genetics **13** (2012), no. 4, 260–270.

[53] Jasmine Chong, Peng Liu, Guangyan Zhou, and Jianguo Xia, *Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data*, Nature Protocols **15** (2020), no. 3, 799–821.

[54] Hyonho Chun and SÃ¼ndÃ¼z KeleÅ, *Sparse partial least squares regression for simultaneous dimension reduction and variable selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72** (2010), no. 1, 3–25.

[55] David S Clausen and Amy D Willis, *Evaluating replicability in microbiome data*, Biostatistics (2021), kxab048.

[56] Mette Rye Clausen and Per Br⊘bech Mortensen, *Fecal ammonia in patients with adenomatous polyps and cancer of the colon*, Nutrition and Cancer **18** (1992), no. 2, 175–180.

[57] M. O. Coker, A. G. Hoen, E. Dade, S. Lundgren, Z. Li, A. D. Wong, M. S. Zens, T. J. Palys, H. G. Morrison, M. L. Sogin, E. R. Baker, M. R. Karagas, and J. C. Madan, *Specific class of intrapartum antibiotics relates to maturation*

of the infant gut microbiota: A prospective cohort study, BJOG **127** (2020), no. 2, 217–227.

[58] James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje, *Ribosomal Database Project: Data and tools for high throughput rRNA analysis*, Nucl. Acids Res. **42** (2014), no. D1, D633–D642.

[59] The Human Microbiome Project Consortium, Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I.-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher,

Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H. Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, YuanQing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, and Owen White, *Structure, function and diversity of the healthy human microbiome*, Nature **486** (2012), no. 7402, 207–214.

[60] Microsoft Corporation and Steve Weston, *doParallel: Foreach parallel adaptor for the 'parallel' package*, 2019.

[61] Renan Corrêa-Oliveira, José Luís Fachi, Aline Vieira, Fabio Takeo Sato, and Marco Aurélio R Vinolo, *Regulation of immune cell function by short-chain fatty acids*, Clin Transl Immunology **5** (2016), no. 4, e73.

[62] Zhao-Lai Dai, *Amino acid metabolism in intestinal bacteria: Links between gut ecology and host health*, Frontiers in Bioscience **16** (2011), no. 1, 1768.

[63] Lawrence A. David, Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm, *Host lifestyle affects human microbiota on daily timescales*, Genome Biol **15** (2014), no. 7, R89.

[64] Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, Sudha B. Biddinger, Rachel J. Dutton, and Peter J. Turnbaugh, *Diet rapidly and reproducibly alters the human gut microbiome*, Nature **505** (2014), no. 7484, 559–563.

[65] Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman, and Benjamin J. Callahan, *Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data*, Microbiome **6** (2018), no. 1, 226.

[66] Marie Laure Delignette-Muller and Christophe Dutang, *Fitdistrplus: An R package for fitting distributions*, Journal of Statistical Software **64** (2015), no. 4, 1–34.

[67] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach*, Biometrics **44** (1988), no. 3, 837–845.

[68] Gijs den Besten, Katja Lange, Rick Havinga, Theo H. van Dijk, Albert Gerding, Karen van Eunen, Michael Müller, Albert K. Groen, Guido J. Hooiveld, Barbara M. Bakker, and Dirk-Jan Reijngoud, *Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids*, American Journal of Physiology-Gastrointestinal and Liver Physiology **305** (2013), no. 12, G900–G910.

[69] Anthony C. Dona, Beatriz Jiménez, Hartmut Schäfer, Eberhard Humpfer, Manfred Spraul, Matthew R. Lewis, Jake T. M. Pearce, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson, *Precision high-throughput proton NMR spec-*

troscopy of human urine, serum, and plasma for large-scale metabolic phenotyping, Anal. Chem. **86** (2014), no. 19, 9887–9894.

[70] Gregory P. Donaldson, S. Melanie Lee, and Sarkis K. Mazmanian, *Gut biogeography of the bacterial microbiota*, Nat. Rev. Microbiol. **14** (2016), no. 1, 20–32.

[71] Mei Dong, Longhai Li, Man Chen, Anthony Kusalik, and Wei Xu, *Predictive analysis methods for human microbiome data with application to Parkinson's disease*, PLOS ONE **15** (2020), no. 8, e0237779.

[72] Eun-Hee Doo, Christophe Chassard, Clarissa Schwab, and Christophe Lacroix, *Effect of dietary nucleosides and yeast extracts on composition and metabolic activity of infant gut microbiota in PolyFermS colonic fermentation models*, FEMS Microbiol. Ecol. **93** (2017), no. 8.

[73] Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille, *PICRUSt2 for prediction of metagenome functions*, Nature Biotechnology **38** (2020), no. 6, 685–688.

[74] Veronika B. Dubinkina, Alexander V. Tyakht, Vera Y. Odintsova, Konstantin S. Yarygin, Boris A. Kovarsky, Alexander V. Pavlenko, Dmitry S. Ischenko, Anna S. Popenko, Dmitry G. Alexeev, Anastasiya Y. Taraskina, Regina F. Nasyrova, Evgeny M. Krupitsky, Nino V. Shalikiani, Igor G. Bakulin, Petr L. Shcherbakov, Lyubov O. Skorodumova, Andrei K. Larin, Elena S. Kostryukova, Rustam A. Abdulkhakov, Sayar R. Abdulkhakov, Sergey Y. Malanin, Ruzilya K. Ismagilova, Tatiana V. Grigoryeva, Elena N. Ilina, and Vadim M. Govorun, *Links of gut microbiota composition with alcohol dependence syndrome and alcoholic liver disease*, Microbiome **5** (2017), no. 1, 141.

[75] Juliana Durack and Susan V. Lynch, *The gut microbiome: Relationships with disease and opportunities for therapy*, J Exp Med **216** (2019), no. 1, 20–40.

[76] Claire Duvallet, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm, *Meta-analysis of gut microbiome studies identifies disease-specific and shared responses*, Nat Commun **8** (2017), no. 1, 1784.

[77] Robert C. Edgar and Henrik Flyvbjerg, *Error filtering, pair assembly and error correction for next-generation sequencing reads*, Bioinformatics **31** (2015), no. 21, 3476–3482.

[78] Bradley Efron, *Large-Scale Simultaneous Hypothesis Testing*, Journal of the American Statistical Association **99** (2004), no. 465, 96–104.

[79] J. J. Egozcue and V. Pawlowsky-Glahn, *Groups of Parts and Their Balances in Compositional Data Analysis*, Mathematical Geology **37** (2005), no. 7, 795–828.

[80] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, *Isometric Logratio Transformations for Compositional Data Analysis*, Mathematical Geology **35** (2003), no. 3, 279–300.

[81] Raphael Eisenhofer, Jeremiah J. Minich, Clarisse Marotz, Alan Cooper, Rob Knight, and Laura S. Weyrich, *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations*, Trends in Microbiology **27** (2019), no. 2, 105–117.

[82] Sahar El Aidy, Muriel Derrien, Claire A. Merrifield, Florence Levenez, Joël Doré, Mark V. Boekschoten, Jan Dekker, Elaine Holmes, Erwin G. Zoetendal, Peter van Baarlen, Sandrine P. Claus, and Michiel Kleerebezem, *Gut bacteria–host metabolic interplay during conventionalisation of the mouse germfree colon*, The ISME Journal **7** (2013), no. 4, 743–755.

[83] Raphaël Enaud, Renaud Prevel, Eleonora Ciarlo, Fabien Beaufils, Gregoire Wieërs, Benoit Guery, and Laurence Delhaes, *The Gut-Lung Axis in Health and Respiratory Diseases: A Place for Inter-Organ and Inter-Kingdom Crosstalks*, Front. Cell. Infect. Microbiol. **10** (2020), 9.

[84] Alexander Eng and Elhanan Borenstein, *Taxa-function robustness in microbial communities*, Microbiome **6** (2018), 45.

[85] Felix G.M. Ernst, Sudarshan A. Shetty, Tuomas Borman, and Leo Lahti, *Mia: Microbiome analysis*, 2021.

[86] D Rose Ewald and Susan CJ Sumner, *Human Microbiota, Blood Group Antigens, and Disease*, Wiley Interdiscip Rev Syst Biol Med **10** (2018), no. 3, e1413.

[87] Karoline Faust and Jeroen Raes, *Microbial interactions: From networks to models*, Nat Rev Microbiol **10** (2012), no. 8, 538–550.

[88] Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, Lili Su, Xiaoping Li, Xin Li, Junhua Li, Liang Xiao, Ursula Huber-Schönauer, David Niederseer, Xun Xu, Jumana Yousuf Al-Aama, Huanming Yang, Jian Wang, Karsten Kristiansen, Manimozhiyan Arumugam, Herbert Tilg, Christian Datz, and Jun Wang, *Gut microbiome development along the colorectal adenoma-carcinoma sequence*, Nat Commun **6** (2015), 6528.

[89] Noah Fierer, Christian L. Lauber, Nick Zhou, Daniel McDonald, Elizabeth K. Costello, and Rob Knight, *Forensic identification using skin bacterial communities*, Proc Natl Acad Sci U S A **107** (2010), no. 14, 6477–6481.

[90] Betsy Foxman and Emily T. Martin, *Use of the Microbiome in the Practice of Epidemiology: A Primer on -Omic Technologies*, Am J Epidemiol **182** (2015), no. 1, 1–8.

[91] Eric A. Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C. Morgan, and Curtis Huttenhower, *Sequencing and beyond: Integrating molecular 'omics for microbial community profiling*, Nat Rev Microbiol **13** (2015), no. 6, 360–372.

[92] Eric A. Franzosa, Xochitl C. Morgan, Nicola Segata, Levi Waldron, Joshua Reyes, Ashlee M. Earl, Georgia Giannoukos, Matthew R. Boylan, Dawn Ciulla, Dirk Gevers, Jacques Izard, Wendy S. Garrett, Andrew T. Chan, and Curtis Huttenhower, *Relating the metatranscriptome and metagenome of the human gut*, Proc. Natl. Acad. Sci. U.S.A. **111** (2014), no. 22.

[93] Eric A. Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J. Haiser, Stefan Reinker, Tommi Vatanen, A. Brantley Hall, Himel Mallick, Lauren J. McIver, Jenny S. Sauk, Robin G. Wilson, Betsy W. Stevens, Justin M. Scott, Kerry Pierce, Amy A. Deik, Kevin Bullock, Floris Imhann, Jeffrey A. Porter, Alexandra Zhernakova, Jingyuan Fu, Rinse K. Weersma, Cisca Wijmenga, Clary B. Clish, Hera Vlamakis, Curtis Huttenhower, and Ramnik J. Xavier, *Gut microbiome structure and metabolic activity in inflammatory bowel disease*, Nature Microbiology **4** (2019), no. 2, 293–305.

[94] H Robert Frost, *Computation and application of tissue-specific gene set weights*, Bioinformatics **34** (2018), no. 17, 2957–2964.

[95] Hildreth Robert Frost, *Variance-adjusted Mahalanobis (VAM): A fast and accurate method for cell-specific gene set scoring*, Nucleic Acids Research **48** (2020), no. 16, e94–e94.

[96] Marcus Fulde, Felix Sommer, Benoit Chassaing, Kira van Vorst, Aline Dupont, Michael Hensel, Marijana Basic, Robert Klopfleisch, Philip Rosenstiel, André Bleich, Fredrik Bäckhed, Andrew T. Gewirtz, and Mathias W. Hornef, *Neonatal selection by Toll-like receptor 5 influences long-term gut microbiota composition*, Nature **560** (2018), no. 7719, 489–493.

[97] Jing Gao, Kang Xu, Hongnan Liu, Gang Liu, Miaomiao Bai, Can Peng, Tiejun Li, and Yulong Yin, *Impact of the Gut Microbiota on Intestinal Immunity Mediated by Tryptophan Metabolism*, Front Cell Infect Microbiol **8** (2018), 13.

[98] Simon Garnier, *Viridis: Default color maps from 'matplotlib'*, 2018.

[99] Graciela L. Garrote, Analía G. Abraham, and Martín Rumbo, *Is lactate an undervalued functional component of fermented food products?*, Front. Microbiol. **6** (2015).

[100] Ludwig Geistlinger, Gergely Csaba, Mara Santarelli, Marcel Ramos, Lucas Schiffer, Nitesh Turaga, Charity Law, Sean Davis, Vincent Carey, Martin Morgan, et al., *Toward a gold standard for benchmarking gene set enrichment analysis*, Briefings in bioinformatics **22** (2021), no. 1, 545–556.

[101] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier, *The Treatment-Naive Microbiome in New-Onset Crohn's Disease*, Cell Host & Microbe **15** (2014), no. 3, 382–392.

[102] R.J. Gibbons, *Bacterial Adhesion to Oral Tissues: A Model for Infectious Diseases*, J Dent Res **68** (1989), no. 5, 750–760.

[103] Diane Gilbert-Diamond, Kathryn L. Cottingham, Joann F. Gruber, Tracy Punshon, Vicki Sayarath, A. Jay Gandolfi, Emily R. Baker, Brian P. Jackson, Carol L. Folt, and Margaret R. Karagas, *Rice consumption contributes to arsenic exposure in US women*, Proc. Natl. Acad. Sci. U.S.A. **108** (2011), no. 51, 20656–20660.

[104] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue, *Microbiome Datasets Are Compositional: And This Is Not Optional*, Front. Microbiol. **8** (2017).

[105] Jelle J. Goeman and Peter Bühlmann, *Analyzing gene expression data in terms of gene sets: Methodological issues*, Bioinformatics **23** (2007), no. 8, 980–987.

[106] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein, and Rob Knight, *Qiita: Rapid, web-enabled microbiome meta-analysis*, Nature Methods **15** (2018), no. 10, 796–798.

[107] John Guittar, Ashley Shade, and Elena Litchman, *Trait-based community assembly and succession of the infant gut microbiome*, Nat Commun **10** (2019), no. 1, 512.

[108] Ankit Gupta, Darshan B. Dhakan, Abhijit Maji, Rituja Saxena, Vishnu Prasoodanan P K, Shruti Mahajan, Joby Pulikkan, Jacob Kurian, Andres M.

Gomez, Joy Scaria, Katherine R. Amato, Ashok K. Sharma, and Vineet K. Sharma, *Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India*, mSystems **4** (2019), no. 6, e00438–19.

[109] Andrew Brantley Hall, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K. Lagoudas, Tommi Vatanen, Nadine Fornelos, Robin Wilson, Madeline Bertha, Melissa Cohen, John Garber, Hamed Khalili, Dirk Gevers, Ashwin N. Ananthakrishnan, Subra Kugathasan, Eric S. Lander, Paul Blainey, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower, *A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients*, Genome Med **9** (2017), no. 1, 103.

[110] Geoffrey D. Hannigan, Melissa B. Duhaime, Mack T. Ruffin, Charlie C. Koumpouras, and Patrick D. Schloss, *Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome*, mBio **9** (2018), no. 6, e02248–18.

[111] Sonja Hänzelmann, Robert Castelo, and Justin Guinney, *GSVA: Gene set variation analysis for microarray and RNA-Seq data*, BMC Bioinformatics **14** (2013), no. 1, 7.

[112] Trevor Hastie, Robert Tibshirani, and J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed ed., Springer Series in Statistics, Springer, New York, NY, 2009.

[113] Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas, *A broken promise: Microbiome differential abundance methods do not control the false discovery rate*, Brief Bioinform **20** (2019), no. 1, 210–221.

[114] Almut Heinken and Ines Thiele, *Systems biology of host–microbe metabolomics*, Wiley Interdiscip Rev Syst Biol Med **7** (2015), no. 4, 195–219.

[115] Anna Heintz-Buschart, Patrick May, Cédric C. Laczny, Laura A. Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, Jochen G. Schneider, Angela Hogan, Carine de Beaufort, and Paul Wilmes, *Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes*, Nat Microbiol **2** (2017), no. 1, 16180.

[116] Anna Heintz-Buschart and Paul Wilmes, *Human Gut Microbiome: Function Matters*, Trends in Microbiology **26** (2018), no. 7, 563–574.

[117] Cian J. Hill, Denise B. Lynch, Kiera Murphy, Marynka Ulaszewska, Ian B. Jeffery, Carol Anne O'Shea, Claire Watkins, Eugene Dempsey, Fulvio Mattivi, Kieran Tuohy, R. Paul Ross, C. Anthony Ryan, Paul W. O' Toole, and Catherine Stanton, *Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort*, Microbiome **5** (2017), no. 1, 4.

[118] Katherine Hollywood, Daniel R. Brison, and Royston Goodacre, *Metabolomics: Current technologies and future trends*, Proteomics **6** (2006), no. 17, 4716–4723.

[119] Ruizhu Huang, Charlotte Soneson, Felix G.M. Ernst, Kevin C. Rue-Albrecht, Guangchuang Yu, Stephanie C. Hicks, and Mark D. Robinson, *TreeSummarizedExperiment: A S4 class for data with hierarchical structure*, F1000Research **9** (2021), 1246.

[120] Elizabeth R. Hughes, Maria G. Winter, Breck A. Duerkop, Luisella Spiga, Tatiane Furtado de Carvalho, Wenhan Zhu, Caroline C. Gillis, Lisa Büttner, Madeline P. Smoot, Cassie L. Behrendt, Sara Cherry, Renato L. Santos, Lora V. Hooper, and Sebastian E. Winter, *Microbial respiration and formate oxidation*

as metabolic signatures of inflammation-associated dysbiosis, Cell Host Microbe **21** (2017), no. 2, 208–219.

[121] Susan M Huse, Vincent B Young, Hilary G Morrison, Dionysios A Antonopoulos, John Kwon, Sushila Dalal, Rose Arrieta, Nathaniel A Hubert, Lici Shen, Joseph H Vineis, Jason C Koval, Mitchell L Sogin, Eugene B Chang, and Laura E Raffals, *Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects*, Microbiome **2** (2014), 5.

[122] Phillip B. Hylemon, Huiping Zhou, William M. Pandak, Shunlin Ren, Gregorio Gil, and Paul Dent, *Bile acids as regulatory molecules*, J. Lipid Res. **50** (2009), no. 8, 1509–1520.

[123] Umer Zeeshan Ijaz, Christopher Quince, Laura Hanske, Nick Loman, Szymon T. Calus, Martin Bertz, Christine A. Edwards, Daniel R. Gaya, Richard Hansen, Paraic McGrogan, Richard K. Russell, and Konstantinos Gerasimidis, *The distinct features of microbial 'dysbiosis' of Crohn's disease do not occur to the same extent in their unaffected, genetically-linked kindred*, PLoS One **12** (2017), no. 2, e0172605.

[124] S. Andrew Inkpen, Gavin M. Douglas, T. D. P. Brunet, Karl Leuschen, W. Ford Doolittle, and Morgan G. I. Langille, *The coupling of taxonomy and function in microbiomes*, Biol Philos **32** (2017), no. 6, 1225–1243.

[125] National Human Genome Research Institute, *Genetics vs. Genomics Fact Sheet*, https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics.

[126] Rafael A. Irizarry, Chi Wang, Yun Zhou, and Terence P. Speed, *Gene Set Enrichment Analysis Made Simple*, Stat Methods Med Res **18** (2009), no. 6, 565–575.

[127] K. B. M. Saiful Islam, Satoru Fukiya, Masahito Hagio, Nobuyuki Fujii, Satoshi Ishizuka, Tadasuke Ooka, Yoshitoshi Ogura, Tetsuya Hayashi, and Atsushi Yokota, *Bile acid is a host factor that regulates the composition of the cecal microbiota in rats*, Gastroenterology **141** (2011), no. 5, 1773–1781.

[128] Kieran James, Francesca Bottacini, Jose Ivan Serrano Contreras, Mariane Vigoureux, Muireann Egan, Mary O'connell Motherway, Elaine Holmes, and Douwe van Sinderen, *Metabolism of the predominant human milk oligosaccharide fucosyllactose by an infant gut commensal*, Scientific Reports **9** (2019), no. 1, 1–20.

[129] Yorick Janssens, Joachim Nielandt, Antoon Bronselaer, Nathan Debunne, Frederick Verbeke, Evelien Wynendaele, Filip Van Immerseel, Yves-Paul Vandewynckel, Guy De Tré, and Bart De Spiegeleer, *Disbiome database: Linking the microbiome to disease*, BMC Microbiol **18** (2018), no. 1, 50.

[130] Yue Jiang, Xuejian Xiong, Jayne Danska, and John Parkinson, *Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality*, Microbiome **4** (2016), no. 1, 2.

[131] Jian-Yu Jiao, Lan Liu, Zheng-Shuang Hua, Bao-Zhu Fang, En-Min Zhou, Nimaichand Salam, Brian P Hedlund, and Wen-Jun Li, *Microbial dark matter coming to light: Challenges and opportunities*, National Science Review **8** (2021), no. 3, nwaa280.

[132] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock, *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis*, Nature Communications **10** (2019), no. 1, 1–11.

[133] Benjamin D. Kaehler, Nicholas A. Bokulich, Daniel McDonald, Rob Knight, J. Gregory Caporaso, and Gavin A. Huttley, *Species abundance information improves sequence taxonomy classification accuracy*, Nat Commun **10** (2019), no. 1, 4643.

[134] Minoru Kanehisa, *Toward understanding the origin and evolution of cellular organisms*, Protein Sci. **28** (2019), no. 11, 1947–1951.

[135] Abhishek Kaul, Ori Davidov, and Shyamal D. Peddada, *Structural zeros in high-dimensional data with applications to microbiome studies*, Biostatistics **18** (2017), no. 3, 422–433.

[136] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada, *Analysis of Microbiome Data in the Presence of Excess Zeros*, Front. Microbiol. **8** (2017).

[137] Yuji Kawamata, Ryo Fujii, Masaki Hosoya, Masataka Harada, Hiromi Yoshida, Masanori Miwa, Shoji Fukusumi, Yugo Habata, Takashi Itoh, Yasushi Shintani, Shuji Hinuma, Yukio Fujisawa, and Masahiko Fujino, *A G protein-coupled receptor responsive to bile acids*, J. Biol. Chem. **278** (2003), no. 11, 9435–9440.

[138] S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb, *Picante: R tools for integrating phylogenies and ecology*, Bioinformatics **26** (2010), 1463–1464.

[139] Purvesh Khatri, Marina Sirota, and Atul J. Butte, *Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges*, PLOS Computational Biology **8** (2012), no. 2, e1002375.

[140] Evan Kiefl, Ozcan C. Esen, Samuel E. Miller, Kourtney L. Kroll, Amy D. Willis, Michael S. Rappé, Tao Pan, and A. Murat Eren, *Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution*, Preprint, Microbiology, March 2022.

[141] Gyungcheon Kim, Jaewoong Bae, Mi Jin Kim, Hyeji Kwon, Gwoncheol Park, Seok-Jin Kim, Yon Ho Choe, Jisook Kim, Sook-Hyun Park, Byung-Ho Choe, Hakdong Shin, and Ben Kang, *Delayed Establishment of Gut Microbiota in Infants Delivered by Cesarean Section*, Front. Microbiol. **11** (2020), 2099.

[142] Charles H. King, Hiral Desai, Allison C. Sylvetsky, Jonathan LoTempio, Shant Ayanyan, Jill Carrie, Keith A. Crandall, Brian C. Fochtman, Lusine Gasparyan, Naila Gulzar, Paul Howell, Najy Issa, Konstantinos Krampis, Lopa Mishra, Hiroki Morizono, Joseph R. Pisegna, Shuyun Rao, Yao Ren, Vahan Simonyan, Krista Smith, Sharanjit VedBrat, Michael D. Yao, and Raja Mazumder, *Baseline human gut microbiota profile in healthy people and standard reporting template*, PLoS ONE **14** (2019), no. 9, e0206484.

[143] Juma Kisuse, Orawan La-ongkham, Massalin Nakphaichit, Phatthanaphong Therdtatha, Rie Momoda, Masaru Tanaka, Shinji Fukuda, Siam Popluechai, Kongkiat Kespechara, Kenji Sonomoto, Yuan-Kun Lee, Sunee Nitisinprasert, and Jiro Nakayama, *Urban Diets Linked to Gut Microbiome and Metabolome Alterations in Children: A Comparative Cross-Sectional Study in Thailand*, Front. Microbiol. **9** (2018).

[144] Jonathan L. Klassen, *Defining microbiome function*, Nat Microbiol **3** (2018), no. 8, 864–869.

[145] Jeremy E. Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D. Fricker, Jesse Stombaugh, Rob Knight, Largus T. Angenent, and Ruth E. Ley, *Succession of microbial consortia in the developing infant gut microbiome*, PNAS **108** (2011), no. Supplement 1, 4578–4585.

[146] Raivo Kolde, *Pheatmap: Pretty heatmaps*, 2019.

[147] Aleksandar D. Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöhö, Ismo Mattila, Harri Lähdesmäki, Eric A. Franzosa, Outi Vaarala, Marcus de Goffau, Hermie Harmsen, Jorma Ilonen, Suvi M. Virtanen, Clary B. Clish, Matej Orešič, Curtis Huttenhower, Mikael Knip, and Ramnik J. Xavier, *The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes*, Cell Host & Microbe **17** (2015), no. 2, 260–273.

[148] Yan Kou, Xiaomin Xu, Zhengnong Zhu, Lei Dai, and Yan Tan, *Microbe-set enrichment analysis facilitates functional interpretation of microbiome profiling data*, Sci Rep **10** (2020), no. 1, 21466.

[149] Sascha Krause, Xavier Le Roux, Pascal A. Niklaus, Van Bodegom, Peter M, Jay T. Lennon, Stefan Bertilsson, Hans-Peter Grossart, Laurent Philippot, and Paul L. E. Bodelier, *Trait-based approaches for understanding microbial biodiversity and ecosystem functioning*, Front. Microbiol. **5** (2014).

[150] A. Kröger, V. Geisler, E. Lemma, F. Theis, and R. Lenger, *Bacterial fumarate respiration*, Arch. Microbiol. **158** (1992), no. 5, 311–314.

[151] Max Kuhn, *Tidyposterior: Bayesian analysis to compare models using resampling statistics*, 2018.

[152] Max Kuhn and Hadley Wickham, *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.*, 2020.

[153] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau, *Sparse and Compositionally Robust Inference of Microbial Ecological Networks*, PLOS Computational Biology **11** (2015), no. 5, e1004226.

[154] Jean-Christophe Lagier, Saber Khelaifia, Maryam Tidjani Alou, Sokhna Ndongo, Niokhor Dione, Perrine Hugon, Aurelia Caputo, Frédéric Cadoret, Sory Ibrahima Traore, El Hadji Seck, Gregory Dubourg, Guillaume Durand, Gaël Mourembou, Elodie Guilhot, Amadou Togo, Sara Bellali, Dipankar Bachar, Nadim Cassir, Fadi Bittar, Jérémy Delerce, Morgane Mailhe, Davide Ricaboni, Melhem Bilen, Nicole Prisca Makaya Dangui Nieko, Ndeye Mery Dia Badiane, Camille Valles, Donia Mouelhi, Khoudia Diop, Matthieu Million, Didier Musso, Jônatas Abrahão, Esam Ibraheem Azhar, Fehmida Bibi, Muhammad Yasir, Aldiouma Diallo, Cheikh Sokhna, Felix Djossou, Véronique Vitton, Catherine Robert, Jean Marc Rolain, Bernard La Scola, Pierre-Edouard Fournier, Anthony Levasseur, and Didier Raoult, *Culture of previously uncultured members of the human gut microbiota by culturomics*, Nat Microbiol **1** (2016), no. 12, 16203.

[155] William Michael Landau, *The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing*, Journal of Open Source Software **6** (2021), no. 57, 2959.

[156] Morgan G. I. Langille, *Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome*, mSystems **3** (2018), no. 2, e00163–17.

[157] Martin F. Laursen, Martin I. Bahl, Kim F. Michaelsen, and Tine R. Licht, *First Foods and Gut Microbes*, Front. Microbiol. **8** (2017).

[158] Melissa A. E. Lawson, Ian J. O'Neill, Magdalena Kujawska, Sree Gowrinadh Javvadi, Anisha Wijeyesekera, Zak Flegg, Lisa Chalklen, and Lindsay J. Hall, *Breast milk-derived human milk oligosaccharides promote Bifidobacterium interactions within a single ecosystem*, The ISME Journal **14** (2020), no. 2, 635–648.

[159] Gwénaëlle Le Gall, Samah O. Noor, Karyn Ridgway, Louise Scovell, Crawford Jamieson, Ian T. Johnson, Ian J. Colquhoun, E. Kate Kemsley, and Arjan Narbad, *Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome*, J. Proteome Res. **10** (2011), no. 9, 4208–4218.

[160] Dagmar Hajkova Leary, W. Judson Hervey, Jeffrey R. Deschamps, Anne W. Kusterbeck, and Gary J. Vora, *Which metaproteome? The impact of protein extraction bias on metaproteomic analyses*, Molecular and Cellular Probes **27** (2013), no. 5-6, 193–199.

[161] Jean Guy LeBlanc, Florian Chain, Rebeca Martín, Luis G. Bermúdez-Humarán, Stéphanie Courau, and Philippe Langella, *Beneficial effects on host energy metabolism of short-chain fatty acids and vitamins produced by commensal and probiotic bacteria*, Microb Cell Fact **16** (2017).

[162] Jean Guy LeBlanc, Christian Milani, Graciela Savoy de Giori, Fernando Sesma, Douwe van Sinderen, and Marco Ventura, *Bacteria as vitamin suppliers to their*

host: A gut microbiota perspective, Current Opinion in Biotechnology **24** (2013), no. 2, 160–168.

[163] Katherine P. Lemon, Gary C. Armitage, David A. Relman, and Michael A. Fischbach, *Microbiota-Targeted Therapies: An Ecological Perspective*, Sci Transl Med **4** (2012), no. 137, 137rv5.

[164] Ruth E. Ley, Daniel A. Peterson, and Jeffrey I. Gordon, *Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine*, Cell **124** (2006), no. 4, 837–848.

[165] Hongzhe Li, *Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis*, Annual Review of Statistics and Its Application **2** (2015), no. 1, 73–94.

[166] ———, *Statistical and Computational Methods in Microbiome and Metagenomics*, Handbook of Statistical Genomics (David Balding, Ida Moltke, and John Marioni, eds.), Wiley, first ed., July 2019, pp. 977–550.

[167] Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Al-Aama, Sherif Edris, Huanming Yang, Jian Wang, Torben Hansen, Henrik Bjørn Nielsen, Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S. Dusko Ehrlich, MetaHIT Consortium, Peer Bork, Jun Wang, and MetaHIT Consortium, *An integrated catalog of reference genes in the human gut microbiome*, Nat Biotechnol **32** (2014), no. 8, 834–841.

[168] M. Li, B. Wang, M. Zhang, M. Rantalainen, S. Wang, H. Zhou, Y. Zhang, J. Shen, X. Pang, M. Zhang, H. Wei, Y. Chen, H. Lu, J. Zuo, M. Su, Y. Qiu, W. Jia, C. Xiao, L. M. Smith, S. Yang, E. Holmes, H. Tang, G. Zhao, J. K. Nicholson, L. Li, and L. Zhao, *Symbiotic gut microbes modulate human metabolic phenotypes*, Proceedings of the National Academy of Sciences **105** (2008), no. 6, 2117–2122.

[169] Xue Li, Xinlei Wang, and Guanghua Xiao, *A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications*, Briefings in Bioinformatics, 2019.

[170] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo, *The Molecular Signatures Database Hallmark Gene Set Collection*, Cell Systems **1** (2015), no. 6, 417–425.

[171] Hsi-Chiang Lin and Willard J. Visek, *Colon Mucosal Cell Damage by Ammonia in Rats*, The Journal of Nutrition **121** (1991), no. 6, 887–893.

[172] Huang Lin and Shyamal Das Peddada, *Analysis of compositions of microbiomes with bias correction*, Nat Commun **11** (2020), no. 1, 3514.

[173] Shili Lin, *Rank aggregation methods*, Wiley Interdisciplinary Reviews: Computational Statistics **2** (2010), no. 5, 555–570.

[174] Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G. Graeber, A. Brantley Hall, Kathleen Lake, Carol J. Landers, Himel Mallick, Damian R. Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A.

White, Jonathan Braun, Lee A. Denson, Janet K. Jansson, Rob Knight, Subra Kugathasan, Dermot P. B. McGovern, Joseph F. Petrosino, Thaddeus S. Stappenbeck, Harland S. Winter, Clary B. Clish, Eric A. Franzosa, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower, *Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases*, Nature **569** (2019), no. 7758, 655–662.

[175] Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, Heather H. Creasy, Carrie McCracken, Michelle G. Giglio, Daniel McDonald, Eric A. Franzosa, Rob Knight, Owen White, and Curtis Huttenhower, *Strains, functions and dynamics in the expanded Human Microbiome Project*, Nature (2017).

[176] Stilianos Louca, Martin F. Polz, Florent Mazel, Michaeline B. N. Albright, Julie A. Huber, Mary I. O'Connor, Martin Ackermann, Aria S. Hahn, Diane S. Srivastava, Sean A. Crowe, Michael Doebeli, and Laura Wegener Parfrey, *Function and functional redundancy in microbial systems*, Nature Ecology & Evolution (2018), 1.

[177] Michael I. Love, Wolfgang Huber, and Simon Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*, Genome Biology **15** (2014), no. 12, 550.

[178] Catherine A. Lozupone, Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight, *Diversity, stability and resilience of the human gut microbiota*, Nature **489** (2012), no. 7415, 220–230.

[179] Ruth Ann Luna, Numan Oezguen, Miriam Balderas, Alamelu Venkatachalam, Jessica K. Runge, James Versalovic, Jeremy Veenstra-VanderWeele,

George M. Anderson, Tor Savidge, and Kent C. Williams, *Distinct Microbiome-Neuroimmune Signatures Correlate With Functional Abdominal Pain in Children With Autism Spectrum Disorder*, Cellular and Molecular Gastroenterology and Hepatology **3** (2017), no. 2, 218–230.

[180] Sara N. Lundgren, Juliette C. Madan, Jennifer A. Emond, Hilary G. Morrison, Brock C. Christensen, Margaret R. Karagas, and Anne G. Hoen, *Maternal diet during pregnancy is related with the infant stool microbiome in a delivery mode-dependent manner*, Microbiome **6** (2018), no. 1, 109.

[181] Siyuan Ma, Boyu Ren, Himel Mallick, Yo Sup Moon, Emma Schwager, Sagun Maharjan, Timothy L. Tickle, Yiren Lu, Rachel N. Carmody, Eric A. Franzosa, Lucas Janson, and Curtis Huttenhower, *A statistical model for describing and simulating microbial community profiles*, PLoS Comput Biol **17** (2021), no. 9, e1008913.

[182] George T. Macfarlane and Sandra Macfarlane, *Bacteria, Colonic Fermentation, and Gastrointestinal Health*, J AOAC Int **95** (2012), no. 1, 50–60.

[183] Juliette C. Madan, Anne G. Hoen, Sara N. Lundgren, Shohreh F. Farzan, Kathryn L. Cottingham, Hilary G. Morrison, Mitchell L. Sogin, Hongzhe Li, Jason H. Moore, and Margaret R. Karagas, *Effects of Cesarean delivery and formula supplementation on the intestinal microbiome of six-week old infants*, JAMA Pediatr **170** (2016), no. 3, 212–219.

[184] Joshua S. Madin, Daniel A. Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin K. M. Engqvist, Noah Fierer, Jemma L. Geoghegan, Michael Gillings, Nikos C. Kyrpides, Elena Litchman, Christopher E. Mason, Lisa Moore, Søren L. Nielsen, Ian T. Paulsen, Nathan D. Price, T. B. K. Reddy,

Matthew A. Richards, Eduardo P. C. Rocha, Thomas M. Schmidt, Heba Shaaban, Maulik Shukla, Fran Supek, Sasha G. Tetu, Sara Vieira-Silva, Alice R. Wattam, David A. Westfall, and Mark Westoby, *A synthesis of bacterial and archaeal phenotypic trait data*, Sci Data **7** (2020), no. 1, 170.

[185] Morgane Mailhe, Davide Ricaboni, Véronique Vitton, Jean-Michel Gonzalez, Dipankar Bachar, Grégory Dubourg, Frédéric Cadoret, Catherine Robert, Jérémy Delerce, Anthony Levasseur, Pierre-Edouard Fournier, Emmanouil Angelakis, Jean-Christophe Lagier, and Didier Raoult, *Repertoire of the gut microbiota from stomach to colon using culturomics and next-generation sequencing*, BMC Microbiol **18** (2018), no. 1, 157.

[186] Farhad Maleki, Katie Ovens, Daniel J. Hogan, and Anthony J. Kusalik, *Gene Set Analysis: Challenges, Opportunities, and Future Research*, Front. Genet. **11** (2020), 654.

[187] Himel Mallick, Eric A. Franzosa, Lauren J. Mclver, Soumya Banerjee, Alexandra Sirota-Madi, Aleksandar D. Kostic, Clary B. Clish, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower, *Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences*, Nat Commun **10** (2019), no. 1, 1–11.

[188] Ohad Manor and Elhanan Borenstein, *Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome*, Cell Host & Microbe **21** (2017), no. 2, 254–267.

[189] A. Marcobal, M. Barboza, E.D. Sonnenburg, N. Pudlo, E.C. Martens, P. Desai, C.B. Lebrilla, B.C. Weimer, D.A. Mills, J.B. German, and J.L. Sonnenburg, *Bacteroides in the Infant Gut Consume Milk Oligosaccharides via Mucus-Utilization Pathways*, Cell Host Microbe **10** (2011), no. 5, 507–514.

[190] Perrine Marquet, Sylvia H. Duncan, Christophe Chassard, Annick Bernalier-Donadille, and Harry J. Flint, *Lactate has the potential to promote hydrogen sulphide formation in the human colon*, FEMS Microbiology Letters **299** (2009), no. 2, 128–134.

[191] Bryan D. Martin, Daniela Witten, and Amy D. Willis, *Modeling microbial abundances and dysbiosis with beta-binomial regression*, The Annals of Applied Statistics **14** (2020), no. 1, 94–115.

[192] J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, *Model-based replacement of rounded zeros in compositional data: Classical and robust approaches*, Computational Statistics & Data Analysis **56** (2012), no. 9, 2688–2704.

[193] Ian H. McHardy, Maryam Goudarzi, Maomeng Tong, Paul M. Ruegger, Emma Schwager, John R. Weger, Thomas G. Graeber, Justin L. Sonnenburg, Steve Horvath, Curtis Huttenhower, Dermot PB McGovern, Albert J. Fornace, James Borneman, and Jonathan Braun, *Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships*, Microbiome **1** (2013), no. 1, 17.

[194] Donald T. McKnight, Roger Huerlimann, Deborah S. Bower, Lin Schwarzkopf, Ross A. Alford, and Kyall R. Zenger, *Methods for normalizing microbiome data: An ecological perspective*, Methods in Ecology and Evolution **10** (2019), no. 3, 389–400.

[195] Michael R McLaren, Amy D Willis, and Benjamin J Callahan, *Consistent and correctable bias in metagenomic sequencing experiments*, eLife **8** (2019), e46923.

[196] Paul J. McMurdie and Susan Holmes, *Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data*, PLoS ONE **8** (2013), no. 4, e61217.

[197] ______, *Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible*, PLOS Computational Biology **10** (2014), no. 4, e1003531.

[198] Ryan McNulty, Duluxan Sritharan, Shichen Liu, Sahand Hormoz, and Adam Z. Rosenthal, *Droplet-based single cell RNA sequencing of bacteria identifies known and previously unseen cellular states*, Preprint, Microbiology, March 2021.

[199] Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium, Damian Rafal Plichta, Agnieszka Sierakowska Juncker, Marcelo Bertalan, Elizabeth Rettedal, Laurent Gautier, Encarna Varela, Chaysavanh Manichanh, Charlène Fouqueray, Florence Levenez, Trine Nielsen, Joël Doré, Ana Manuel Dantas Machado, Mari Cristina Rodriguez de Evgrafov, Torben Hansen, Torben Jørgensen, Peer Bork, Francisco Guarner, Oluf Pedersen, Morten O. A. Sommer, S. Dusko Ehrlich, Thomas Sicheritz-Pontén, Søren Brunak, and H. Bjørn Nielsen, *Transcriptional interactions suggest niche segregation among microorganisms in the human gut*, Nat Microbiol **1** (2016), no. 11, 16152.

[200] MetaHIT Consortium (additional members), Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin,

Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork, *Enterotypes of the human gut microbiome*, Nature **473** (2011), no. 7346, 174–180.

[201] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, arXiv (2013).

[202] Christian Milani, Sabrina Duranti, Francesca Bottacini, Eoghan Casey, Francesca Turroni, Jennifer Mahony, Clara Belzer, Susana Delgado Palacio, Silvia Arboleya Montes, Leonardo Mancabelli, Gabriele Andrea Lugli, Juan Miguel Rodriguez, Lars Bode, Willem de Vos, Miguel Gueimonde, Abelardo Margolles, Douwe van Sinderen, and Marco Ventura, *The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota*, Microbiol. Mol. Biol. Rev. **81** (2017), no. 4, e00036–17.

[203] Frédéric Moens, Stefan Weckx, and Luc De Vuyst, *Bifidobacterial inulin-type fructan degradation capacity determines cross-feeding interactions between bifidobacteria and Faecalibacterium prausnitzii*, Int J Food Microbiol **231** (2016), 76–85.

[204] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster, *Sustainable data analysis with Snakemake*, F1000Res **10** (2021), 33.

[205] Rebecca E. Moore and Steven D. Townsend, *Temporal development of the infant gut microbiome*, Open Biol. **9** (2019), no. 9, 190128.

[206] Shirin Moossavi, Faisal Atakora, Kelsey Fehr, and Ehsan Khafipour, *Biological observations in microbiota analysis are robust to the choice of 16S rRNA gene sequencing processing algorithm: Case study on human milk microbiota*, BMC Microbiol **20** (2020), no. 1, 290.

[207] Livia H. Morais, Henry L. Schreiber, and Sarkis K. Mazmanian, *The gut microbiota–brain axis in behaviour and brain disorders*, Nat Rev Microbiol **19** (2021), no. 4, 241–255.

[208] Michael J. Morowitz, Erica M. Carlisle, and John C. Alverdy, *Contributions of Intestinal Bacteria to Nutrition and Metabolism in the Critically Ill*, Surgical Clinics of North America **91** (2011), no. 4, 771–785.

[209] Kayla Morrell and Martin Morgan, *BiocSet: Representing different biological sets*, 2021.

[210] Douglas J. Morrison and Tom Preston, *Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism*, Gut Microbes **7** (2016), no. 3, 189–200.

[211] James T. Morton, Alexander A. Aksenov, Louis Felix Nothias, James R. Foulds, Robert A. Quinn, Michelle H. Badri, Tami L. Swenson, Marc W. Van Goethem, Trent R. Northen, Yoshiki Vazquez-Baeza, Mingxun Wang, Nicholas A. Bokulich, Aaron Watters, Se Jin Song, Richard Bonneau, Pieter C. Dorrestein, and Rob Knight, *Learning representations of microbe–metabolite interactions*, Nat Methods **16** (2019), no. 12, 1306–1314.

[212] James T. Morton, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S. Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight, *Establishing microbial composition measurement standards with reference frames*, Nature Communications **10** (2019), no. 1.

[213] James T. Morton, Jon Sanders, Robert A. Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A. Navas-Molina, Se Jin Song, Jessica L. Metcalf, Embriette R. Hyde, Manuel Lladser, Pieter C. Dorrestein, and Rob Knight, *Balance Trees Reveal Microbial Niche Differentiation*, mSystems **2** (2017), no. 1, e00162–16.

[214] James T. Morton, Justin Silverman, Gleb Tikhonov, Harri Lähdesmäki, and Rich Bonneau, *Scalable estimation of microbial co-occurrence networks with Variational Autoencoders*, Preprint, Bioinformatics, November 2021.

[215] Andrés Moya and Manuel Ferrer, *Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance*, Trends in Microbiology **24** (2016), no. 5, 402–413.

[216] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy, *Genomes OnLine Database (GOLD) v.8: Overview and updates*, Nucleic Acids Research **49** (2021), no. D1, D723–D733.

[217] Efrat Muller, Yadid M. Algavi, and Elhanan Borenstein, *A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations*, Microbiome **9** (2021), no. 1, 203.

[218] Iftekhar Naim and Daniel Gildea, *Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients*, ICML (2012), 8.

[219] Jacob T. Nearing, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau, and Morgan G. I. Langille, *Microbiome differential abundance methods produce different results across 38 datasets*, Nat Commun **13** (2022), no. 1, 342.

[220] Evelien P. J. G. Neis, Cornelis H. C. Dejong, and Sander S. Rensen, *The Role of Microbial Amino Acid Metabolism in Host Metabolism*, Nutrients **7** (2015), no. 4, 2930–2946.

[221] Josef Neu and W. Allan Walker, *Necrotizing Enterocolitis*, N Engl J Med **364** (2011), no. 3, 255–264.

[222] Ryan J. Newton, Sandra L. McLellan, Deborah K. Dila, Joseph H. Vineis, Hilary G. Morrison, A. Murat Eren, and Mitchell L. Sogin, *Sewage Reflects the Microbiomes of Human Populations*, mBio **6** (2015), no. 2.

[223] Quang P. Nguyen, Anne G. Hoen, and H. Robert Frost, *CBEA: Competitive balances for taxonomic enrichment analysis*, Preprint, Bioinformatics, September 2021.

[224] Jeremy K. Nicholson, Elaine Holmes, James Kinross, Remy Burcelin, Glenn Gibson, Wei Jia, and Sven Pettersson, *Host-Gut Microbiota Metabolic Interactions*, Science **336** (2012), no. 6086, 1262–1267.

[225] H. Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, Laurent Gautier, Anders G. Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde,

Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B. Quintanilha Dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S. Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S. Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W. Ussery, Takuji Yamada, MetaHIT Consortium, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, S. Dusko Ehrlich, and MetaHIT Consortium, *Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes*, Nat Biotechnol **32** (2014), no. 8, 822–828.

[226] Cecilia Noecker, Hsuan-Chao Chiu, Colin P. McNally, and Elhanan Borenstein, *Defining and Evaluating Microbial Contributions to Metabolite Variation in Microbiome-Metabolome Association Studies*, mSystems **4** (2019), no. 6.

[227] Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagner, *Vegan: Community ecology package*, 2019.

[228] Kaitlyn Oliphant and Emma Allen-Vercoe, *Macronutrient metabolism by the human gut microbiome: Major fermentation by-products and their impact on host health*, Microbiome **7** (2019), no. 1, 91.

[229] Chana Palmer, Elisabeth M. Bik, Daniel B. DiGiulio, David A. Relman, and Patrick O. Brown, *Development of the Human Infant Intestinal Microbiota*, PLOS Biology **5** (2007), no. 7, e177.

[230] E. Paradis and K. Schliep, *Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R*, Bioinformatics **35** (2018), 526–528.

[231] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B. Dowd, Curtis Huttenhower, Martin Morgan, Nicola Segata, and Levi Waldron, *Accessible, curated metagenomic data through ExperimentHub*, Nature Methods **14** (2017), no. 11, 1023–1024.

[232] Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata, *Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights*, PLOS Computational Biology **12** (2016), no. 7, e1004977.

[233] Helle Krogh Pedersen, Valborg Gudmundsdottir, Henrik Bjørn Nielsen, Tuulia Hyotylainen, Trine Nielsen, Benjamin A. H. Jensen, Kristoffer Forslund, Falk Hildebrand, Edi Prifti, Gwen Falony, Emmanuelle Le Chatelier, Florence Levenez, Joel Doré, Ismo Mattila, Damian R. Plichta, Päivi Pöhö, Lars I. Hellgren, Manimozhiyan Arumugam, Shinichi Sunagawa, Sara Vieira-Silva, Torben Jørgensen, Jacob Bak Holm, Kajetan Trošt, MetaHIT Consortium, Karsten Kristiansen, Susanne Brix, Jeroen Raes, Jun Wang, Torben Hansen, Peer Bork, Søren Brunak, Matej Oresic, S. Dusko Ehrlich, and Oluf Pedersen, *Human gut microbes impact host serum metabolome and insulin sensitivity*, Nature **535** (2016), no. 7612, 376–381.

[234] Thomas Lin Pedersen, *Patchwork: The composer of plots*, 2020.

[235] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research **12** (2011), 2825–2830.

[236] Luying Peng, Zhong-Rong Li, Robert S. Green, Ian R. Holzman, and Jing Lin, *Butyrate Enhances the Intestinal Barrier by Facilitating Tight Junction Assembly via Activation of AMP-Activated Protein Kinase in Caco-2 Cell Monolayers*, J Nutr **139** (2009), no. 9, 1619–1625.

[237] Fátima C. Pereira and David Berry, *Microbial nutrient niches in the gut*, Environ Microbiol **19** (2017), no. 4, 1366–1378.

[238] Pedro R. Peres-Neto and Donald A. Jackson, *How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test*, Oecologia **129** (2001), no. 2, 169–178.

[239] Ana Elena Pérez-Cobas, Laura Gomez-Valero, and Carmen Buchrieser, *Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses*, Microb Genom **6** (2020), no. 8.

[240] Enrica Pessione, *Lactic acid bacteria contribution to gut microbiota complexity: Lights and shadows*, Front. Cell. Infect. Microbiol. **2** (2012).

[241] Joseph M. Pickard and Alexander V. Chervonsky, *Intestinal fucose as a mediator of host-microbe symbiosis*, J Immunol **194** (2015), no. 12, 5588–5593.

[242] John C. Platt, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, ADVANCES IN LARGE MARGIN CLASSIFIERS, MIT Press, 1999, pp. 61–74.

[243] Kevin J. Portune, Martin Beaumont, Anne-Marie Davila, Daniel Tomé, François Blachier, and Yolanda Sanz, *Gut microbiota role in dietary protein*

metabolism and health-related outcomes: The two sides of the coin, Trends in Food Science & Technology **57** (2016), 213–232.

[244] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin, *FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments*, PLOS ONE **5** (2010), no. 3, e9490.

[245] Lita M. Proctor, Heather H. Creasy, Jennifer M. Fettweis, Jason Lloyd-Price, Anup Mahurkar, Wenyu Zhou, Gregory A. Buck, Michael P. Snyder, Jerome F. Strauss, George M. Weinstock, Owen White, Curtis Huttenhower, and The Integrative HMP (iHMP) Research Network Consortium, *The Integrative Human Microbiome Project*, Nature **569** (2019), no. 7758, 641–648.

[246] Benoit Pugin, Weronika Barcik, Patrick Westermann, Anja Heider, Marcin Wawrzyniak, Peter Hellings, Cezmi A. Akdis, and Liam O'Mahony, *A wide diversity of bacteria from the human gut produces and degrades biogenic amines*, Microb. Ecol. Health Dis. **28** (2017), no. 1, 1353881.

[247] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner, *The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools*, Nucleic Acids Res **41** (2013), no. D1, D590–D596.

[248] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata, *Shotgun metagenomics, from sampling to analysis*, Nat Biotechnol **35** (2017), no. 9, 833–844.

[249] Thomas P Quinn, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley, *A field guide for the compositional analysis of any-omics data*, GigaScience **8** (2019), no. giz107.

[250] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley, *Understanding sequencing data as compositions: An outlook and review*, Bioinformatics **34** (2018), no. 16, 2870–2878.

[251] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[252] Maitreyi Raman, Iftikhar Ahmed, Patrick M. Gillevet, Chris S. Probert, Norman M. Ratcliffe, Steve Smith, Rosemary Greenwood, Masoumeh Sikaroodi, Victor Lam, Pam Crotty, Jennifer Bailey, Robert P. Myers, and Kevin P. Rioux, *Fecal Microbiome and Volatile Organic Compound Metabolome in Obese Humans With Nonalcoholic Fatty Liver Disease*, Clinical Gastroenterology and Hepatology **11** (2013), no. 7, 868–875.e3.

[253] Jacques Ravel, Pawel Gajer, Zaid Abdo, G. Maria Schneider, Sara S. K. Koenig, Stacey L. McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol O. Tacket, Rebecca M. Brotman, Catherine C. Davis, Kevin Ault, Ligia Peralta, and Larry J. Forney, *Vaginal microbiome of reproductive-age women*, Proc. Natl. Acad. Sci. U.S.A. **108** (2011), no. supplement_1, 4680–4687.

[254] Aspen T Reese, Eugenia H Cho, Bruce Klitzman, Scott P Nichols, Natalie A Wisniewski, Max M Villa, Heather K Durand, Sharon Jiang, Firas S Midani, Sai N Nimmagadda, Thomas M O'Connell, Justin P Wright, Marc A Deshusses, and Lawrence A David, *Antibiotic-induced changes in the microbiota disrupt redox dynamics in the gut*, eLife **7** (2018), e35987.

[255] Liam J. Revell, *Phytools: An R package for phylogenetic comparative biology (and other things).*, Methods in Ecology and Evolution **3** (2012), 217–223.

[256] Jason M. Ridlon, Dae Joong Kang, Phillip B. Hylemon, and Jasmohan S. Bajaj, *Bile Acids and the Gut Microbiome*, Curr Opin Gastroenterol **30** (2014), no. 3, 332–338.

[257] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle, *Balances: A New Perspective for Microbiome Analysis*, mSystems **3** (2018), no. 4, e00053–18.

[258] Audrey Rivière, Marija Selak, David Lantin, Frédéric Leroy, and Luc De Vuyst, *Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut*, Front Microbiol **7** (2016).

[259] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller, *pROC: An open-source package for R and S+ to analyze and compare ROC curves*, BMC Bioinformatics **12** (2011), 77.

[260] N. Sato, T. Nakano, H. Kawakami, and T. Idota, *In vitro and in vivo effects of exogenous nucleotides on the proliferation and maturation of intestinal epithelial cells*, J. Nutr. Sci. Vitaminol. **45** (1999), no. 1, 107–118.

[261] Henning Schiebenhoefer, Tim Van Den Bossche, Stephan Fuchs, Bernhard Y. Renard, Thilo Muth, and Lennart Martens, *Challenges and promise at the interface of metaproteomics and genomics: An overview of recent progress in metaproteogenomic data analysis*, Expert Review of Proteomics **16** (2019), no. 5, 375–390.

[262] Lucas Schiffer, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B Dowd, Nicola Segata, and Levi

Waldron, *HMP16SData: Efficient access to the human microbiome project through bioconductor*, American Journal of Epidemiology (2019).

[263] Thomas S.B. Schmidt, Jeroen Raes, and Peer Bork, *The Human Gut Microbiome: From Association to Modulation*, Cell **172** (2018), no. 6, 1198–1215.

[264] Conrad L. Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi, *NCBI Taxonomy: A comprehensive update on curation, resources and tools*, Database (Oxford) **2020** (2020), baaa062.

[265] Julie Schulthess, Sumeet Pandey, Melania Capitani, Kevin C. Rue-Albrecht, Isabelle Arnold, Fanny Franchini, Agnieszka Chomka, Nicholas E. Ilott, Daniel G. W. Johnston, Elisabete Pires, James McCullagh, Stephen N. Sansom, Carolina V. Arancibia-Cárcamo, Holm H. Uhlig, and Fiona Powrie, *The Short Chain Fatty Acid Butyrate Imprints an Antimicrobial Program in Macrophages*, Immunity **50** (2019), no. 2, 432–445.e7.

[266] Yan Shao, Samuel C. Forster, Evdokia Tsaliki, Kevin Vervier, Angela Strang, Nandi Simpson, Nitin Kumar, Mark D. Stares, Alison Rodger, Peter Brocklehurst, Nigel Field, and Trevor D. Lawley, *Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth*, Nature **574** (2019), no. 7776, 117–121.

[267] Michael Shapira, *Gut Microbiotas and Host Evolution: Scaling Up Symbiosis*, Trends in Ecology & Evolution **31** (2016), no. 7, 539–549.

[268] Sapna Sharma and Prabhanshu Tripathi, *Gut microbiome and type 2 diabetes: Where we are and where to go?*, The Journal of Nutritional Biochemistry **63** (2019), 101–108.

[269] Pixu Shi, Anru Zhang, and Hongzhe Li, *Regression analysis for microbiome compositional data*, Ann. Appl. Stat. **10** (2016), no. 2, 1019–1040. MR MR3528370

[270] Andrew B. Shreiner, John Y. Kao, and Vincent B. Young, *The gut microbiome in health and in disease*, Curr Opin Gastroenterol **31** (2015), no. 1, 69–75.

[271] Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David, *Naught all zeros in sequence count data are the same*, Comput Struct Biotechnol J **18** (2020), 2789–2798.

[272] Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David, *A phylogenetic transform enhances analysis of compositional microbiota data*, eLife **6** (2017), e21887.

[273] Maggie A. Stanislawski, Dana Dabelea, Brandie D. Wagner, Nina Iszatt, Cecilie Dahl, Marci K. Sontag, Rob Knight, Catherine A. Lozupone, and Merete Eggesbø, *Gut Microbiota in the First 2 Years of Life and the Association with Body Mass Index at Age 12 in a Norwegian Birth Cohort*, mBio **9** (2018), no. 5.

[274] Christopher J. Stewart, Nadim J. Ajami, Jacqueline L. O'Brien, Diane S. Hutchinson, Daniel P. Smith, Matthew C. Wong, Matthew C. Ross, Richard E. Lloyd, HarshaVardhan Doddapaneni, Ginger A. Metcalf, Donna Muzny, Richard A. Gibbs, Tommi Vatanen, Curtis Huttenhower, Ramnik J. Xavier, Marian Rewers, William Hagopian, Jorma Toppari, Anette-G. Ziegler, Jin-Xiong She, Beena Akolkar, Ake Lernmark, Heikki Hyoty, Kendra Vehik, Jef-

frey P. Krischer, and Joseph F. Petrosino, *Temporal development of the gut microbiome in early childhood from the TEDDY study*, Nature **562** (2018), no. 7728, 583–588.

[275] Christopher J. Stewart, Nicholas D. Embleton, Emma C. L. Marrs, Daniel P. Smith, Tatiana Fofanova, Andrew Nelson, Tom Skeath, John D. Perry, Joseph F. Petrosino, Janet E. Berrington, and Stephen P. Cummings, *Longitudinal development of the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls*, Microbiome **5** (2017), no. 1, 75.

[276] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov, *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS **102** (2005), no. 43, 15545–15550.

[277] Jane Tang, *Microbial Metabolomics*, CG **12** (2011), no. 6, 391–403.

[278] Qiang Tang, Ge Jin, Gang Wang, Tianyu Liu, Xiang Liu, Bangmao Wang, and Hailong Cao, *Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices*, Front. Cell. Infect. Microbiol. **10** (2020), 151.

[279] Andrew Maltez Thomas, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, Serena Manara, Nicolai Karcher, Chiara Pozzi, Sara Gandini, Davide Serrano, Sonia Tarallo, Antonio Francavilla, Gaetano Gallo, Mario Trompetto, Giulio Ferrero, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Jakob Wirbel, Petra Schrotz-King, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, Georg Zeller, Francesca Cordero, Emmanuel Dias-Neto, João Carlos Setubal, Adrian Tett, Barbara

Pardini, Maria Rescigno, Levi Waldron, Alessio Naccarati, and Nicola Segata, *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation*, Nat Med **25** (2019), no. 4, 667–678.

[280] Carissa M. Thomas, Teresa Hong, Jan Peter van Pijkeren, Peera Hemarajata, Dan V. Trinh, Weidong Hu, Robert A. Britton, Markus Kalkum, and James Versalovic, *Histamine derived from probiotic Lactobacillus reuteri suppresses TNF via modulation of PKA and ERK signaling*, PLoS ONE **7** (2012), no. 2, e31951.

[281] T. Thurnheer, N. Bostanci, and G. N. Belibasakis, *Microbial dynamics during conversion from supragingival to subgingival biofilms in an in vitro model*, Mol Oral Microbiol **31** (2016), no. 2, 125–135.

[282] Liang Tian, Xu-Wen Wang, Ang-Kun Wu, Yuhang Fan, Jonathan Friedman, Amber Dahlin, Matthew K. Waldor, George M. Weinstock, Scott T. Weiss, and Yang-Yu Liu, *Deciphering functional redundancy in the human microbiome*, Nat Commun **11** (2020), no. 1, 6217.

[283] Lu Tian, Steven A. Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S. Kohane, and Peter J. Park, *Discovering statistically significant pathways in expression profiling studies*, PNAS **102** (2005), no. 38, 13544–13549.

[284] David Tilman, Forest Isbell, and Jane M. Cowles, *Biodiversity and Ecosystem Functioning*, Annual Review of Ecology, Evolution, and Systematics **45** (2014), no. 1, 471–493.

[285] Martha E Trujillo, Svetlana Dedysh, Paul DeVos, Brian Hedlund, Peter Kämpfer, Fred A Rainey, and William B Whitman (eds.), *Bergey's Manual of Systematics of Archaea and Bacteria*, first ed., Wiley, April 2015.

[286] Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata, *MetaPhlAn2 for enhanced metagenomic taxonomic profiling*, Nature Methods **12** (2015), no. 10, 902–903.

[287] Peter J. Turnbaugh and Jeffrey I. Gordon, *An invitation to the marriage of metagenomics and metabolomics*, Cell **134** (2008), no. 5, 708–713.

[288] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon, *A core gut microbiome in obese and lean twins*, Nature **457** (2009), no. 7228, 480–484.

[289] Luke K Ursell, Jessica L Metcalf, Laura Wegener Parfrey, and Rob Knight, *Defining the human microbiome*, Nutrition Reviews **70** (2012), S38–S44.

[290] K. Gerald van den Boogaart, Raimon Tolosana-Delgado, and Matevz Bren, *Compositions: Compositional data analysis*, 2019.

[291] Tommi Vatanen, Eric A. Franzosa, Randall Schwager, Surya Tripathi, Timothy D. Arthur, Kendra Vehik, Åke Lernmark, William A. Hagopian, Marian J. Rewers, Jin-Xiong She, Jorma Toppari, Anette-G. Ziegler, Beena Akolkar, Jeffrey P. Krischer, Christopher J. Stewart, Nadim J. Ajami, Joseph F. Petrosino, Dirk Gevers, Harri Lähdesmäki, Hera Vlamakis, Curtis Huttenhower, and Ram-

nik J. Xavier, *The human gut microbiome in early-onset type 1 diabetes from the TEDDY study*, Nature **562** (2018), no. 7728, 589–594.

[292] Eric M. Velazquez, Henry Nguyen, Keaton T. Heasley, Cheng H. Saechao, Lindsey M. Gil, Andrew W. L. Rogers, Brittany M. Miller, Matthew R. Rolston, Christopher A. Lopez, Yael Litvak, Megan J. Liou, Franziska Faber, Denise N. Bronner, Connor R. Tiffany, Mariana X. Byndloss, Austin J. Byndloss, and Andreas J. Bäumler, *Endogenous Enterobacteriaceae underlie variation in susceptibility to Salmonella infection*, Nat Microbiol **4** (2019), no. 6, 1057–1064.

[293] Nathan C Verberkmoes, Alison L Russell, Manesh Shah, Adam Godzik, Magnus Rosenquist, Jonas Halfvarson, Mark G Lefsrud, Juha Apajalahti, Curt Tysk, Robert L Hettich, and Janet K Jansson, *Shotgun metaproteomics of the human distal gut microbiota*, ISME J **3** (2009), no. 2, 179–189.

[294] Arnau Vich Vila, Floris Imhann, Valerie Collij, Soesma A. Jankipersadsing, Thomas Gurry, Zlatan Mujagic, Alexander Kurilshikov, Marc Jan Bonder, Xiaofang Jiang, Ettje F. Tigchelaar, Jackie Dekens, Vera Peters, Michiel D. Voskuil, Marijn C. Visschedijk, Hendrik M. van Dullemen, Daniel Keszthelyi, Morris A. Swertz, Lude Franke, Rudi Alberts, Eleonora A. M. Festen, Gerard Dijkstra, Ad A. M. Masclee, Marten H. Hofker, Ramnik J. Xavier, Eric J. Alm, Jingyuan Fu, Cisca Wijmenga, Daisy M. A. E. Jonkers, Alexandra Zhernakova, and Rinse K. Weersma, *Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome*, Sci Transl Med **10** (2018), no. 472, eaap8914.

[295] Sara Vieira-Silva, Gwen Falony, Youssef Darzi, Gipsi Lima-Mendez, Roberto Garcia Yunta, Shujiro Okuda, Doris Vandeputte, Mireia Valles-Colomer, Falk Hildebrand, Samuel Chaffron, and Jeroen Raes, *Species–function*

relationships shape ecological properties of the human gut microbiome, Nature Microbiology **1** (2016), no. 8, 16088.

[296] Emily Vogtmann, Xing Hua, Georg Zeller, Shinichi Sunagawa, Anita Y. Voigt, Rajna Hercog, James J. Goedert, Jianxin Shi, Peer Bork, and Rashmi Sinha, *Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing*, PLoS One **11** (2016), no. 5, e0155362.

[297] S. M. Waititu, F. Yin, R. Patterson, A. Yitbarek, J. C. Rodriguez-Lecompte, and C. M. Nyachoti, *Dietary supplementation with a nucleotide-rich yeast extract modulates gut immune response and microflora in weaned pigs in response to a sanitary challenge*, Animal **11** (2017), no. 12, 2156–2164.

[298] Alan W. Walker, Sylvia H. Duncan, Petra Louis, and Harry J. Flint, *Phylogeny, culturing, and metagenomics of the human gut microbiota*, Trends in Microbiology **22** (2014), no. 5, 267–274.

[299] Brian H. Walker, *Biodiversity and Ecological Redundancy*, Conservation Biology **6** (1992), no. 1, 18–23.

[300] William A. Walters, Zech Xu, and Rob Knight, *Meta-analyses of human gut microbes associated with obesity and IBD*, FEBS Letters **588** (2014), no. 22, 4223–4233.

[301] Stephen Wandro, Stephanie Osborne, Claudia Enriquez, Christine Bixby, Antonio Arrieta, and Katrine Whiteson, *The microbiome and metabolome of preterm infant stool is personalized, and not driven by health outcomes including necrotizing enterocolitis and late-onset sepsis*, (2018), –.

[302] Ji Wang, Wei-Dong Chen, and Yan-Dong Wang, *The Relationship Between Gut Microbiota and Inflammatory Diseases: The Role of Macrophages*, Front. Microbiol. **11** (2020), 1065.

[303] Zeneng Wang, Elizabeth Klipfell, Brian J. Bennett, Robert Koeth, Bruce S. Levison, Brandon Dugar, Ariel E. Feldstein, Earl B. Britt, Xiaoming Fu, Yoon-Mi Chung, Yuping Wu, Phil Schauer, Jonathan D. Smith, Hooman Allayee, W. H. Wilson Tang, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen, *Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease*, Nature **472** (2011), no. 7341, 57–63.

[304] Alex D Washburne, Justin D Silverman, Jonathan W Leff, Dominic J Bennett, and John L Darcy, *Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets.*, PeerJ (2017), 26.

[305] Aaron Weimann, Kyra Mooren, Jeremy Frank, Phillip B. Pope, Andreas Bremges, and Alice C. McHardy, *From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer*, mSystems **1** (2016), no. 6, e00101–16.

[306] Alexa R. Weingarden, Chi Chen, Aleh Bobr, Dan Yao, Yuwei Lu, Valerie M. Nelson, Michael J. Sadowsky, and Alexander Khoruts, *Microbiota transplantation restores normal fecal bile acid composition in recurrent* Clostridium difficile *infection*, American Journal of Physiology-Gastrointestinal and Liver Physiology **306** (2014), no. 4, G310–G319.

[307] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight, *Normalization and microbial differential abundance strategies depend upon data characteristics*, Microbiome **5** (2017), no. 1.

[308] Jake L. Weissman, Sonia Dogra, Keyan Javadi, Samantha Bolten, Rachel Flint, Cyrus Davati, Jess Beattie, Keshav Dixit, Tejasvi Peesay, Shehar Awan, Peter Thielen, Florian Breitwieser, Philip L. F. Johnson, David Karig, William F. Fagan, and Sharon Bewick, *Exploring the functional composition of the human microbiome using a hand-curated microbial trait database*, BMC Bioinformatics **22** (2021), no. 1, 306.

[309] Aalim M. Weljie, Jack Newton, Pascal Mercier, Erin Carlson, and Carolyn M. Slupsky, *Targeted Profiling: Quantitative Analysis of ^{1} H NMR Metabolomics Data*, Analytical Chemistry **78** (2006), no. 13, 4430–4442.

[310] Hadley Wickham, *The split-apply-combine strategy for data analysis*, Journal of Statistical Software **40** (2011), no. 1, 1–29.

[311] ______, *Ggplot2: Elegant graphics for data analysis*, Springer-Verlag New York, 2016.

[312] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani, *Welcome to the tidyverse*, Journal of Open Source Software **4** (2019), no. 43, 1686.

[313] Claus O. Wilke, *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*, 2019.

[314] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel,

the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt., *Caret: Classification and regression training*, 2019.

[315] Jakob Wirbel, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, Anita Y. Voigt, Albert Palleja, Ruby Ponnudurai, Shinichi Sunagawa, Luis Pedro Coelho, Petra Schrotz-King, Emily Vogtmann, Nina Habermann, Emma Niméus, Andrew M. Thomas, Paolo Manghi, Sara Gandini, Davide Serrano, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Tatsuhiro Shibata, Shinichi Yachida, Takuji Yamada, Levi Waldron, Alessio Naccarati, Nicola Segata, Rashmi Sinha, Cornelia M. Ulrich, Hermann Brenner, Manimozhiyan Arumugam, Peer Bork, and Georg Zeller, *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer*, Nat Med **25** (2019), no. 4, 679–689.

[316] Daniela Witten, Rob Tibshirani, Sam Gross, and Balasubramanian Narasimhan, *PMA: Penalized multivariate analysis*, 2019.

[317] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*, Biostatistics **10** (2009), no. 3, 515–534.

[318] Andrea C. Wong and Maayan Levy, *New Approaches to Microbiome-Based Therapies*, mSystems **4** (2019), no. 3, e00122–19.

[319] Julia M. W. Wong, Russell de Souza, Cyril W. C. Kendall, Azadeh Emam, and David J. A. Jenkins, *Colonic Health: Fermentation and Short Chain Fatty Acids:*, Journal of Clinical Gastroenterology **40** (2006), no. 3, 235–243.

[320] Erik S. Wright, *Using DECIPHER v2.0 to analyze big biological sequence data in r*, The R Journal **8** (2016), no. 1, 352–359.

[321] Marvin N. Wright, Andreas Ziegler, and Inke R. König, *Do little interactions get lost in dark random forests?*, BMC Bioinformatics **17** (2016), no. 1, 145.

[322] Chong Wu, Jun Chen, Junghi Kim, and Wei Pan, *An adaptive association test for microbiome data*, Genome Med **8** (2016), no. 1, 56.

[323] Chong Wu and Wei Pan, *MiSPU: Microbiome based sum of powered score (MiSPU) tests*, 2016.

[324] Di Wu and Gordon K. Smyth, *Camera: A competitive gene set test accounting for inter-gene correlation*, Nucleic Acids Res **40** (2012), no. 17, e133.

[325] Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, Rohini Sinha, Erin Gilroy, Kernika Gupta, Robert Baldassano, Lisa Nessel, Hongzhe Li, Frederic D. Bushman, and James D. Lewis, *Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes*, Science **334** (2011), no. 6052, 105–108.

[326] Guojun Wu, Naisi Zhao, Chenhong Zhang, Yan Y. Lam, and Liping Zhao, *Guild-based analysis for understanding gut microbiome in human health and diseases*, Genome Med **13** (2021), no. 1, 22.

[327] Jian Xiao, Li Chen, Yue Yu, Xianyang Zhang, and Jun Chen, *A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data*, Frontiers in Microbiology **9** (2018).

[328] Nan Xiao, *Ggsci: Scientific journal and sci-fi themed color palettes for 'ggplot2'*, 2018.

[329] Jian Xu, Magnus K. Bjursell, Jason Himrod, Su Deng, Lynn K. Carmichael, Herbert C. Chiang, Lora V. Hooper, and Jeffrey I. Gordon, *A genomic view of the human-Bacteroides thetaiotaomicron symbiosis*, Science **299** (2003), no. 5615, 2074–2076.

[330] Zhenjiang Xu, Daniel Malmer, Morgan G I Langille, Samuel F Way, and Rob Knight, *Which is more important for classifying microbial communities: Who's there or what they can do?*, ISME J **8** (2014), no. 12, 2357–2359.

[331] Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, Fumie Hosoda, Hirofumi Rokutan, Minori Matsumoto, Hiroyuki Takamaru, Masayoshi Yamada, Takahisa Matsuda, Motoki Iwasaki, Taiki Yamaji, Tatsuo Yachida, Tomoyoshi Soga, Ken Kurokawa, Atsushi Toyoda, Yoshitoshi Ogura, Tetsuya Hayashi, Masanori Hatakeyama, Hitoshi Nakagama, Yutaka Saito, Shinji Fukuda, Tatsuhiro Shibata, and Takuji Yamada, *Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer*, Nat Med **25** (2019), no. 6, 968–976.

[332] Jing Yang, Ji Pu, Shan Lu, Xiangning Bai, Yangfeng Wu, Dong Jin, Yanpeng Cheng, Gui Zhang, Wentao Zhu, Xuelian Luo, Ramon Rosselló-Móra, and Jianguo Xu, *Species-Level Analysis of Human Gut Microbiota With Metataxonomics*, Front. Microbiol. **11** (2020), 2029.

[333] Moran Yassour, Tommi Vatanen, Heli Siljander, Anu-Maaria Hämäläinen, Taina Härkönen, Samppa J. Ryhänen, Eric A. Franzosa, Hera Vlamakis, Curtis Huttenhower, Dirk Gevers, Eric S. Lander, Mikael Knip, on behalf of the DIABIMMUNE Study Group, and Ramnik J. Xavier, *Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain di-*

versity and stability, Science Translational Medicine **8** (2016), no. 343, 343ra81–343ra81.

[334] Shin Yoshimoto, Tze Mun Loo, Koji Atarashi, Hiroaki Kanda, Seidai Sato, Seiichi Oyadomari, Yoichiro Iwakura, Kenshiro Oshima, Hidetoshi Morita, Masahira Hattori, Masahisa Hattori, Kenya Honda, Yuichi Ishikawa, Eiji Hara, and Naoko Ohtani, *Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome*, Nature **499** (2013), no. 7456, 97–101.

[335] Noelle E. Younge, Christopher B. Newgard, C. Michael Cotten, Ronald N. Goldberg, Michael J. Muehlbauer, James R. Bain, Robert D. Stevens, Thomas M. O'Connell, John F. Rawls, Patrick C. Seed, and Patricia L. Ashley, *Disrupted Maturation of the Microbiota and Metabolome among Extremely Preterm Infants with Postnatal Growth Failure*, Scientific Reports **9** (2019), no. 1, 1–12.

[336] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao Yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, Xiaokai Wang, Xiaoqiang Xu, Ning Chen, William Ka Kei Wu, Jumana Al-Aama, Hans Jørgen Nielsen, Pia Kiilerich, Benjamin Anderschou Holbech Jensen, Tung On Yau, Zhou Lan, Huijue Jia, Junhua Li, Liang Xiao, Thomas Yuen Tung Lam, Siew Chien Ng, Alfred Sze-Lok Cheng, Vincent Wai-Sun Wong, Francis Ka Leung Chan, Xun Xu, Huanming Yang, Lise Madsen, Christian Datz, Herbert Tilg, Jian Wang, Nils Brünner, Karsten Kristiansen, Manimozhiyan Arumugam, Joseph Jao-Yiu Sung, and Jun Wang, *Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer*, Gut **66** (2017), no. 1, 70–78.

[337] Georg Zeller, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina

Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R Mende, Martin A Schneider, Petra Schrotz-King, Christophe Tournigand, Jeanne Tran Van Nhieu, Takuji Yamada, Jürgen Zimmermann, Vladimir Benes, Matthias Kloor, Cornelia M Ulrich, Magnus Knebel Doeberitz, Iradj Sobhani, and Peer Bork, *Potential of fecal microbiota for early-stage detection of colorectal cancer*, Mol Syst Biol **10** (2014), no. 11, 766.

[338] Chenhong Zhang and Liping Zhao, *Strain-level dissection of the contribution of the gut microbiome to human metabolic disease*, Genome Med **8** (2016).

[339] Jianbo Zhang, Shana Sturla, Christophe Lacroix, and Clarissa Schwab, *Gut Microbial Glycerol Metabolism as an Endogenous Acrolein Source*, mBio **9** (2018), no. 1.

[340] Xu Zhang, Leyuan Li, James Butcher, Alain Stintzi, and Daniel Figeys, *Advancing functional and translational microbiome research using meta-omics approaches*, Microbiome **7** (2019), no. 1, 154.

[341] Danping Zheng, Timur Liwinski, and Eran Elinav, *Interaction between microbiota and immunity in health and disease*, Cell Res **30** (2020), no. 6, 492–506.

[342] Yi-Hui Zhou and Paul Gallins, *A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction*, Front. Genet. **10** (2019).

[343] Jonas Zierer, Matthew A. Jackson, Gabi Kastenmüller, Massimo Mangino, Tao Long, Amalio Telenti, Robert P. Mohney, Kerrin S. Small, Jordana T. Bell, Claire J. Steves, Ana M. Valdes, Tim D. Spector, and Cristina Menni, *The fecal metabolome as a functional readout of the gut microbiome*, Nat Genet **50** (2018), no. 6, 790–795.

[344] Hui Zou and Trevor Hastie, *Regularization and Variable Selection via the Elastic Net*, Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67** (2005), no. 2, 301–320.

[345] Cristal Zuñiga, Livia Zaramela, and Karsten Zengler, *Elucidation of complexity and prediction of interactions in microbial communities*, Microb. Biotechnol. **10** (2017), no. 6, 1500–1522.