INTEGRATING FUNCTIONAL AND TAXONOMIC DATA TYPES FOR MICROBIOME DATA ANALYSIS

by

Gavin M. Douglas

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2020

The idea that all bacteria are bad is crazy. Some are just artifacts.

Table of Contents

List of Tables

List of Figures

Abstract

Communities of microbes in natural environments, referred to as microbiomes, are commonly profiled with DNA sequencing approaches. Sequencing results are typically partitioned so that the relative abundances of microbes (taxonomic data) and genes (functional data) are analyzed separately. It is challenging to biologically interpret these data, partially due to the lack of computational frameworks for joint analysis of taxonomic and functional data. Herein, I present my work to address this issue from three perspectives.

First, I did so in the context of an investigation into the microbiome of pediatric Crohn's disease patients. Our main goal with this work was to compare the performance of microbiome data types for classifying samples in both independent and combined models. We found that genera identified through marker-gene sequencing performed best in these models, but that in combined models functions performed best for classifying treatment response. Although these and other insights were valuable, it became clear that improved methods for generating and analyzing taxa-function links were needed.

One method for generating these links is through metagenome prediction methods. Although these approaches are widely used, they suffer from several major caveats and have been inconsistently validated. Accordingly, I developed a new bioinformatic method, PICRUSt2, for generating predicted taxon-function links based on several hypothesized improvements. Although I confirmed that this new approach performed moderately better than alternative methods, I also identified issues with analyzing metagenome predictions in general.

My final project focused on partially addressing these and related problems in functional data analysis. I did this by developing a novel method to better integrate taxonomic and functional data types to identify functional biomarkers. This tool, POMS, accurately identified genes under selection in simulated data and performed well when applied to actual case-control metagenomics datasets.

Taken together, this thesis represents several valuable developments in joint taxa-function analysis that enabled improved interpretation of microbiome data. In several instances, particularly with the application POMS, this joint analysis approach yielded novel insights that would be overlooked by analyzing each data type individually.

List of Abbreviations Used

16S/18S	16/18 Svedberg
ABC transporter	ATP-binding cassette transporter
ALDEx2	Analysis of Variance (ANOVA)-like Differential Expression
APE	Analyses of Phylogenetics and Evolution
ASV	Amplicon sequence variant
ATP	Adenosine triphosphate
BY	Benjamini-Yekutieli-corrected P-values
bp	Base-pair(s)
CD	Crohn's Disease
CLR	Centred log-ratio
COG	Clusters of Orthologous Genes
CN	Healthy colon controls
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
EEN	Exclusive enteral nutrition
EMPANADA	Evidence-based Metagenomic Pathway Assignment using geNe Abundance DAta
ENA	European Nucleotide Archive
EPA-ng	Evolutionary Placement Algorithm (next-generation)
FAPROTAX	Functional Annotation of Prokaryotic Taxa
FDR	False discovery rate
FishTaco	Functional Shifts' Taxonomic Contributors
GI	Gastrointestinal
GRS	Genetic risk score
HMM	Hidden Markov model
HMP	Human Microbiome Project
HSP	Hidden state prediction
HUMAnN	HMP Unified Metabolic Analysis Network
IBD	Inflammatory bowel disease
ILR	Isometric log-ratio

IMG	Integrated Microbial Genomes
ITS	Internal transcribed spacer
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG ortholog
LOOCV	Leave-one-out cross-validation
LPS	Lipopolysaccharide
MAG	Metagenome assembled genome
MinPath	Minimal set of Pathways
MGS	Shotgun metagenomics sequencing
MLI	Mucosal-luminal interface
MP	Maximum Parsimony
MSA	Multiple sequence alignment
mRNA	Messenger ribonucleic acid
NADH	Nicotinamide adenine dinucleotide (reduced)
NF-κB	Nuclear Factor kappa-light-chain-enhancer of activated B cells
NOD2	Nucleotide-oligmerization-domain-2
NR	Non-responders
NSTI	Nearest sequenced taxon index
OOB	Out-of-bag
OTU	Operational taxonomic unit
PanFP	Pangenome-based Functional Profiles
PAPRICA	PAthway PRedction by phylogenetIC plAcement
Pfam	Protein families
PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States
PCR	Polymerase chain reaction
PE	Paired-end
PIC	Phylogenetic Independent Contrasts
POMS	Phylogenetic Organization of Metagenome Signals
PRR	Pattern Recognition Receptor
PTW	Paired sample, two-tailed Wilcoxon tests

RF	Random Forest
RISK	Risk Stratification and Identification of Immunogenetic and Microbial Markers of Rapid Disease Progression in Children with Crohn’s Disease
rRNA	Ribosomal ribonucleic acid
RS	Responders
SD	Standard deviation
SNP	Single nucleotide polymorphism
SRA	Short read archive
SSU	Small subunit
UniRef	UniProt Reference Clusters
UniProt	Universal Protein Resource
USCG	Universal single-copy gene
QIIME	Quantitative Insights Into Microbial Ecology
v	Version

Acknowledgements

I would first like to thank all of my collaborators from throughout my PhD. In particular, Dr. Richard Hansen and Dr. Johan Van Limbergen were integral to the pediatric Crohn's disease study described in Chapter 3. Similarly, Dr. Elhanan Borenstein played a crucial role for my work on POMS, which is described in Chapter 5.

I would also like to thank all of my committee members for their feedback throughout my PhD: Dr. Robert Beiko, Dr. Joseph Bielawski, Dr. Zhenyu Cheng, Dr. Andrew Stadnyk, and Dr. Johan Van Limbergen. A special thanks to those also on my exam committee, and particularly to Dr. Stadnyk for his kind guidance as my co-supervisor.

I would like to thank Dr. Laura Parfrey from the University of British Columbia for taking the time to be an external reviewer of this work.

Many thanks to Marsha Scott-Meldrum for helping to keep my program requirements organized and on schedule during my PhD.

For the last few years I have been financially supported by an Alexander Graham Bell doctoral scholarship from NSERC and also by the President's Award from Dalhousie University. I am grateful for this support and would like to thank both organizations.

I would also like to thank my supervisor Dr. Morgan Langille for his guidance. Thanks to you I think I am both a better researcher and person than when I first joined the lab.

To all members of the Langille lab: thank you for your feedback, friendship, and for pushing me to be a more thoughtful scientist.

To my parents, Andrew and Shirley: thank you for everything you have done for me throughout my life. I trust this enticing read will be adequate compensation.

To Heather: thank you for always being there for me. Your love helps me trust my ideas.

Chapter 1 - Introduction

Microbial communities encompass most of the genetic and species-level diversity on Earth. These communities are commonly characterized through DNA sequencing, which can be used to identify the presence and relative abundance of microbes in a community. These communities, including both the microbes, their constituent genes, and metabolites, are referred to as microbiomes. Due to technological improvements and the reduced cost of sequencing, the number of sequenced microbiomes has substantially grown in recent years. For instance, in 2017 the Earth Microbiome Project published a meta-analysis of 23,828 sequencing samples from all seven continents (Thompson et al. 2017). This data represented 109 environmental groupings and 21 major biomes, such as animal secretions, saline water, and soil. A key goal of microbial ecology research is to robustly analyze and correctly interpret these and other such microbial profiles.

But is DNA sequencing the best method for characterizing microbial communities? It is commonly observed that microbiome research would benefit from more emphasis on culturing, which enables individual microbes to be isolated and precisely studied in the lab. Traditionally, microbial communities were difficult to study by culturing alone because the vast majority of environmental microbes, particularly bacteria, could not be grown under standard culturing conditions (Staley & Konopka 1985). This issue remains unresolved even after gradual improvements to standard culturing conditions; a recent evaluation of six major environments only identified 34.9% of bacteria as culturable under standard conditions (Martiny 2019). However, modified culturing conditions can largely resolve this problem. By systematically applying 66 different conditions it was demonstrated that 95% of bacterial species in human stool samples could be grown in the lab (Lau et al. 2016). Therefore, it is no longer true for human stool samples, and likely other environments as well, that the majority of constituent bacteria cannot be cultured.

Despite these advances, a clear remaining advantage of DNA sequencing is that it enables microbial communities to be characterized in place, which theoretically enables the exact community relative abundances to be profiled. In practice, biases during sample collection and sequencing library preparation can perturb microbial relative abundances (Jones et al. 2015; Bukin et al. 2019; Watson et al. 2019). But nonetheless, DNA sequencing provides a more accurate view of the relative abundances of the community members than would be possible from culturing alone. For this reason, DNA sequencing remains the predominant method for characterizing microbial communities, although it is well-complemented by culturing (Lau et al. 2016).

DNA sequencing is typically analyzed to identify specific associations between individual features (e.g. individual microbes) and sample groupings of interest. Most commonly, researchers are interested in identifying associations between disease states and the relative abundances of features. A similar goal is often to investigate whether different measures of diversity in the studied dataset are associated with the sample groupings. These measures of diversity are divided into alpha and beta diversity (Goodrich et al. 2014). Alpha diversity metrics refer to within-sample measures, such as richness, the number of taxa, and the Shannon diversity index (or entropy), which incorporates both the abundance and evenness of taxa within a sample (Jost 2006). In contrast, beta diversity refers to metrics that summarize variation between samples, which is most often performed by metrics that take the presence and abundance of features into account, such as the Bray-Curtis dissimilarity metric (Goodrich et al. 2014). Other microbiome-specific metrics have also been developed, such as the weighted UniFrac distance, which also takes the phylogenetic distance between taxa into account (Lozupone & Knight 2005). There is often more statistical power to detect overall differences based on alpha and beta diversity metrics than to detect associations with individual features, but diversity-level insights are also less actionable (Shade 2017).

There are many sub-categories of DNA sequencing approaches for characterizing microbial communities. One key distinction is between approaches that aim to characterize taxa (i.e. a group of organisms) and those that characterize genes and pathways, referred to as functions, that could be active in the community. These data types are referred to as taxonomic and functional microbiome data, respectively. Biologically this dichotomy is counter-intuitive; clearly genes are encoded in the genomes of taxa. So why does this distinction exist?

The reason is entirely related to methodological challenges. The most common and cost-effective sequencing approach focuses on sequencing marker genes. This method provides no direct information on the genomes of sequenced microbes, and instead is used to profile taxa. In contrast, shotgun metagenomics sequencing (MGS) provides information on all DNA present in a sample. MGS data can be used for analyzing both taxonomic and functional profiles. However, it is difficult to integrate the two data types, largely due to the complexity of microbial communities: it is relatively straight-forward to identify genes in MGS data but challenging to determine from which genomes they originated. In this thesis, I present several projects that leverage both data types with the common theme of integrating functional and taxonomic microbiome data (Langille 2018) to yield more robust and novel insights.

This thesis is in publication-format, which means that each main chapter is a published or draft manuscript. The exception is that the methods of each manuscript are all presented in Chapter 2. In Chapter 3 I present a reproduction of my work exploring the microbiome pediatric Crohn's disease patients (Douglas et al. 2018). Our main goal with this work was to compare the performance of microbiome data types for classifying samples in both separate and combined models. Although our insights based on this approach were useful, it became clear that improved focus on taxa-function links is needed. One method for generating these links is through metagenome prediction methods. In Chapter 4 I present my work developing an improved prediction method and also the novel validations I performed (Douglas et al. 2020). As described below, improved statistical approaches for analyzing taxa-function links are sorely needed in the microbiome field. To help address this issue, I developed a novel bioinformatics method which I present in Chapter 5. Last, I briefly present my overall conclusions and key discussion points arising from these projects in Chapter 6.

A consequence of presenting this thesis in publication format is that the depth of introductory information provided is constrained by journal formatting requirements. Accordingly, this chapter will provide pertinent introductory material to complement the subsequent chapters. I will first provide a detailed introduction to microbiome data types, followed by a discussion of the analysis challenges particular to these data. Next, I will explicitly discuss areas where microbiome data types have or could be integrated. One focus of this discussion is on generating links between data types with predicted metagenome data. I also describe several approaches for integrating taxonomic and functional data types in statistical analyses. In closing I will focus on a case example of a human disease highly associated with the microbiome. This section will not only aid the reader better appreciate my work on Crohn’s disease pediatric patient microbiomes, but also provide useful examples of representative associations between microbes and traits of interest.

The paragraphs below that discuss metagenomics assembly are reproduced from a review paper I previously published (Douglas & Langille 2019). This was done with permission from the publisher (see Appendices) and is also indicated again directly above that sub-section.

1.1 - Marker-gene Sequencing

The earliest developed and most common form of microbiome sequencing is marker-gene sequencing, also known as amplicon sequencing. Under this approach specific genes are PCR-amplified and then sequenced. There are two key requirements for robust marker genes. First, they must be encoded by all taxa of interest. Second, the observed sequence divergence between orthologs should be approximately equal to the neutral mutation fixation rate multiplied by double the divergence time between orthologs (Woese 1987). Note that the divergence time should be doubled because mutations could accumulate in either lineage since the organisms diverged. Genes displaying this second requirement have been referred to as molecular chronometers. This term highlights the close link between these marker genes and the concept of the molecular clock (Zuckerkandl & Pauling 1965): given equal mutation rates and equal fixation rates for neutral mutations, the number of neutral substitutions between organisms is directly proportional to the evolutionary divergence between them.

However, there are many reasons why a gene might be an unreliable molecular chronometer (Janda & Abbott 2007). One reason is that if a gene varies in function across taxa then contrasting selection pressures could result in different non-synonymous substitution rates (Wheeler et al. 2016). For instance, as previously observed (Woese 1987), the cytochrome complex gene is a useful molecular chronometer in eukaryotes, but suffers from drawbacks. This gene was valuable for building early phylogenetic trees representing both long evolutionary distances across eukaryotes and between human populations (Fitch & Margoliash 1967). However, within prokaryotes the cytochrome complex systematically varies in size, which is believed to be due to positive selection (Ambler et al. 1979). Because positive selection is likely driving divergence between orthologous cytochrome complexes, in at least some cases it would be an invalid molecular chronometer to study in prokaryotes. Similarly, if a gene is sufficiently divergent between organisms then it can be difficult to accurately align residues. Misalignments lead to inaccurate estimates of evolutionary divergence, which is particularly true if the gene accumulates insertions and deletions. Such highly divergent regions, particularly in areas under no selective constraint, have been referred to as "evolutionary stopwatches" (Woese 1987), because they are useful only at short evolutionary distances. Therefore, to select a robust marker gene one should adhere in some ways to the Goldilocks principle: some nucleotide conservation is needed, but not too much.

The 16 Svedberg (16S) ribosomal RNA (rRNA) gene fits well with this principle. This gene features highly conserved regions surrounding nine less conserved regions (referred to as variable regions). It is also encoded by all prokaryotes and represents 50 helical RNA regions encoded by approximately 1,500 base-pairs (Woese et al. 1980). This high number of independent functional domains is valuable in a marker gene (Woese 1987). This is because if there are non-random substitutions within a single domain, but substitutions in the majority of other domains are driven by random processes, there likely would be little effect on estimates of evolutionary divergence. This gene also encodes a highly conserved function across both prokaryotes and eukaryotes (where it is called the 18S rRNA gene). The 16S rRNA molecule is part of the 30S small subunit (SSU) of the ribosome, which helps initiate protein synthesis by binding the Shine-Dalgarno sequence in messenger RNA (mRNA) to align the ribosome with the encoded start codon. Many changes in the highly conserved region of the 16S rRNA gene affect its binding affinity to the ribosome and mRNA. The strong negative selection acting against such substitutions makes these regions valuable for detecting rare substitutions between distant relatives, anchoring alignments of 16S rRNA genes, and for primer design (Wang et al. 2013b).

Since the 16S rRNA gene was identified as a useful molecular chronometer, it has been the prime marker gene used to develop phylogenetic models of the tree of life. Most famously, an alignment of 16S (and 18S) rRNA gene sequences from across life lead to distinguishing archaea, bacteria, and eukaryotes into distinct domains (Woese & Fox 1977). In these early days, research focused on analyzing the rRNA sequences of isolated microbes. This was painstaking work, as illustrated by the prediction in 1987 that future research groups could plausibly sequence on the order of one hundred 16S rRNAs a year (Woese 1987).

Thirty-three years later, through next-generation sequencing technology, insufficient availability of sequenced rRNA genes is no longer a common complaint. Databases such as SILVA contain enormous collections of sequenced SSU fragments; as of August 2020 SILVA contained 9,469,124 non-clustered, independent sequences (Quast et al. 2013). Software such as redbiom also enables unique 16S rRNA gene variants to be compiled from the growing number of 16S rRNA gene sequencing (hereafter referred to as 16S sequencing) studies (McDonald et al. 2019). These 16S datasets are produced to characterize and compare the relative abundances of prokaryotes across communities. However, despite the ubiquity of such datasets, they are non-trivial to analyze and interpret. There are numerous methodological reasons for this difficulty.

First, due to sequencing length constraints, only certain 16S rRNA gene variable regions are typically amplified and sequenced. Each variable region has particular strengths and limitations (Chen et al. 2019; Johnson et al. 2019). My colleagues and I have previously compared the biases between the amplified fragments from variable regions four and five and from regions six to eight (written as V4-V5 and V6-V8, respectively) on a mock community from the Human Microbiome Project (HMP) (Comeau et al. 2017). We found the V4-V5 region overrepresented Firmicutes and Bacteroides while drastically underestimating Actinobacteria. In contrast, the V6-V8 region overrepresented Proteobacteria and underrepresented Bacteroides. These biases highlight that choice of variable region can depend on which taxa are of interest. For example, region V4-V5 was recently shown to be superior to region V6-V8 for identifying archaea in the North Atlantic Ocean (Willis et al. 2019). In this case the authors were particularly interested in archaeal diversity so the V4-V5 region was more appropriate.

Typically, however, the taxonomic scope of interest and region biases in a particular environment are not clear and little or no rationale is given for the variable region selection. This is a problem, because analyses of the same communities with different variable regions can result in not only systematic biases in the raw data, but also in strikingly different biological interpretations. For example, key species that modulate human vaginal health are underrepresented or missing in V1-V2 sequencing datasets, such as *Gardnerella vaginalis*, *Bifidobacterium bifidum*, and *Chlamydia trachomatis* (Graspeuntner et al. 2018). Application of this region for profiling vaginal samples, instead of the more appropriate choice of the V3-V4 region, can result in entirely missing associations between vaginal health and the microbiome. Similarly, a comparison of the tick microbiome based on six sequenced 16S rRNA gene regions found a wide range of the number of prokaryotic families and in the Shannon diversity index for each individual tick (Sperling et al. 2017). The problem of such biases in variable region selection is beginning to recede as long-read technology enables full-length 16S sequencing (Callahan et al. 2019; Johnson et al. 2019). However, it will remain an important issue for the foreseeable future as long as the microbiome is largely studied by short-read sequencing.

Regardless of the sequenced region, most reads originating from the same biological molecule will differ due to sequencing errors. Raw reads are either clustered based on sequence identity into operational taxonomic units (OTUs) or alternatively errors are corrected to produce amplicon sequence variants (ASVs). OTUs are typically clustered at 97% identity (Goodrich et al. 2014), which often results in merging different species into a single OTU (Mysara et al. 2017). This issue has long plagued 16S rRNA gene-based analyses. For instance, *Bacillus globisporus* and *Bacillus psychrophilus* are problematic cases because their 16S genes share 99.5% sequence identity, but are highly distinct at the genome level (Fox et al. 1992).

In contrast to clustering approaches, error-correcting approaches, referred to as denoising methods, theoretically can correct raw reads sufficiently well to produce exact biological molecules. Several different denoising approaches have emerged recently. DADA2 is the most sophisticated approach, which generates a different parametric error model for every input sequencing dataset (Callahan et al. 2016a). The raw sequencing reads are then corrected to generate ASVs based on this error model. Deblur and UNOISE3 are two other denoising tools that are based on rapidly clustering raw reads and using predetermined hard cut-offs related to the expected error rates to generate ASVs. My colleagues and I evaluated the performance of these three tools and open-reference OTU clustering (which combines both *de novo* and reference-based clustering) and found that all three denoising methods result in similar overall microbial communities (Nearing et al. 2018). In contrast, we found that open-reference OTU clustering resulted in a high rate of spurious OTUs compared to these methods. Nonetheless, there were important differences between the three methods, particularly in terms of richness and when profiling rare taxa (Nearing et al. 2018). A more recent independent validation based on a higher number of test datasets reached similar conclusions (Prodan et al. 2020).

In addition to the choice of clustering or denoising method and selected variable regions, there are numerous other steps of sample preparation and analysis that are contentious (Pollock et al. 2018). My co-authors and I have been concerned about these and related issues throughout (and before) my doctoral program. Accordingly, throughout my thesis analyses of 16S rRNA gene sequencing data were based on the recommended best-practices at the time, but these best-practices have rapidly changed over the years. For instance, in Chapter 3 we clustered raw sequencing data into OTUs, while in Chapter 4 the focus is on ASVs. Although we acknowledge that these and other similar differences in analysis can affect the interpretation of analyses, in key places we used multiple approaches to ensure our interpretations were robust to the bioinformatics pipeline employed.

In addition to 16S rRNA gene sequencing data, one section of this thesis focuses on eukaryotic marker genes specifically. As mentioned above, the 18S rRNA gene is the homolog of the 16S rRNA gene in eukaryotes and is widely used to profile that domain. However, fungi are more difficult to distinguish based on the 18S rRNA gene, because fungi lack several variable regions for this gene (Schoch et al. 2012). Instead, the internal transcribed spacer (ITS) region, although not strictly a marker gene, is more often amplified to study fungal communities, because it typically has more resolution to distinguish fungi than the 18S rRNA gene (Liu et al. 2015a). This region is within the nuclear rRNA cistron of fungi genomes, which contains the 18S, 5.8S, and the 28S rRNA genes. The ITS regions encompasses the two intergenic regions, which have relatively high rates of insertions and deletions, and the 5.8S rRNA gene (Schoch et al. 2012). Only a single intergenic region is typically amplified, referred to as ITS1 or ITS2, which have better discriminatory resolution for the major phyla Basidiomycota and Ascomycota, respectively (Saroj et al. 2015).

Although the above described marker genes are the most commonly profiled loci, in many cases there are marker genes more appropriate for specific lineages. For example, several halophilic species of *Haloarcula* encode multiple 16S copies that can differ by more than 5% sequence identity within the same genome (Sun et al. 2013). Consequently, different marker genes are often used when building phylogenetic trees representing a single species or genera. For instance, six housekeeping genes have been shown to be valuable for discriminating strains of *Helicobacter pylori* (Palau et al. 2016), which could be separately amplified and sequenced in gastric biopsy samples to produce complementary OTUs (Palau et al. 2020). More generally, marker genes for specialized comparisons are often chosen to match the defining function of a given lineage. For example, the methyl coenzyme M redundance A (*mrcA*) gene and a nitrate reductase gene have been previously profiled to explore the diversity of methanogens (Hallam et al. 2003) and nitrogen-fixing microbes (Comeau et al. 2019), respectively.

1.2 - Shotgun Metagenomics Sequencing

Shotgun metagenomics (MGS) is a qualitatively different method from marker-gene sequencing, because it involves sequencing all DNA in a community. This advantage means that MGS data can profile any taxa, including viruses and microbial eukaryotes. MGS approaches were first applied to study ocean water communities through a Fosmid cloning approach (Stein et al. 1996). Building upon such early studies, the potential for leveraging MGS was widely publicized by an investigation into the microbial diversity of the Sargasso Sea (Venter et al. 2004). This study identified 1.2 million previously unknown genes and many other microbial features that would be impossible to study with 16S rRNA gene sequencing. These and other related observations sparked an explosion of interest in profiling microbial communities with MGS approaches. This interest has culminated in the generation of enormous MGS datasets such as the ongoing work on the Earth Microbiome Project (Thompson et al. 2017) and the Human Microbiome Project (Lloyd-Price et al. 2017).

There are two main approaches for analyzing MGS data: read-based workflows and metagenomics assembly. Each of these approaches has strengths and weaknesses, but in both cases the generated profiles will likely only imprecisely reflect biological reality. For instance, the number of species identified by read-based methods can differ by three orders of magnitude (McIntyre et al. 2017). The exact species relative abundances can also drastically differ across tools as well, as recently shown in a comparison of read-based methods applied to simulated datasets (Ye et al. 2019). As discussed below, different approaches for metagenomic assembly will produce different assembled contigs and microbial profiles as well (Olson et al. 2019). Unsurprisingly, given this wide variation, there is also low concordance between 16S sequencing and MGS data taken from the same samples. For example, one comparison found that fewer than 50% of phyla identified in water samples based on 16S sequencing were also identified in the corresponding MGS profiles (Tessler et al. 2017). This wide variation in results highlights that any interpretation of MGS profiles should be done cautiously. It is crucial to appreciate that any approach will have important weaknesses and that the generated profile will only partially represent the actual microbial diversity.

With those important caveats in mind, an understanding of the different approaches is nonetheless important to give context to MGS data analysis. Read-based workflows involve little or no assembly of the reads and instead each read (or pair of reads) is treated independently. This is the most common approach for analyzing MGS data, particularly because it can be performed with low sequencing depth (Hillmann et al. 2018) and in complex communities (Zhou et al. 2015). However, an important disadvantage of this approach is that taxonomic and functional annotations are typically generated and treated as entirely independent data types. When generated independently, the most common approach for generating taxonomic profiles is either based on a marker-gene or k-mer method.

Marker-gene approaches are based on the insight that specific genes can be used to identify the presence and relative abundance of certain taxa. An extreme example is to use solely the 16S rRNA gene for taxonomic classification (Hao & Chen 2012). More commonly, marker-gene approaches base classifications on many genes. For instance, PhyloSift (Darling et al. 2014) leverages 37 nearly universal prokaryotic marker-genes (Wu et al. 2013) in addition to eukaryotic and viral gene sets to make a combined set of approximately 800 (mainly viral) gene families for classification. Aligned reads are placed into a phylogenetic tree of reference sequences and taxonomic classification is performed based on summing the likelihood of each taxa based on each read placement (Darling et al. 2014). MetaPhlAn2 is a contrasting approach that instead bases taxonomic predictions on the presence of clade-specific marker genes, which are genes only found in that given lineage, and found in all members (Truong et al. 2015). This method has rapidly become the most popular marker-gene MGS approach and is what my co-authors and I used when analyzing the data presented in Chapter 3. Given that this approach is limited by the existence of robust clade-specific genes, it is not surprising that it tends to have low sensitivity (Tessler et al. 2017; Miossec et al. 2020), meaning that it misses taxa that are actually present.

In contrast, k-mer-based approaches are much more sensitive but have slightly lower specificity than marker-gene methods (Miossec et al. 2020). These approaches search for exact matches of short DNA sequences (k-mers) within reference genomes. An algorithm such as lowest-common ancestor is then performed to determine the likely taxonomic classification based on all matching genomes. Two common kmer-based approaches are kraken2 (Wood et al. 2019) and centrifuge (Kim et al. 2016), both of which match k-mers against a compressed database of reference genomes. In contrast to the marker-gene results, the main challenge of analyzing taxonomic profiles output by these methods is the high number of rare taxa of different ranks identified, some of which may be false positives. Summarizing the output profiles with an additional approach, such as the Bayesian abundance re-estimation tool Bracken (Lu et al. 2017) in the case kraken2 data, can help partially mitigate this problem.

In contrast to taxonomic analyses, there are fewer options for functional annotation of MGS reads. Most functional read-based methods are based on a similarity search of reads against a database of known gene families. This is primarily done in protein space, because protein similarity matches are more informative and the database requirements are lower (Koonin & Galperin 2003). The common similarity searching tool BLASTX is prohibitively slow when scanning millions of reads, which has driven the development of faster alternatives like DIAMOND (Buchfink et al. 2015) and MMseqs2 (Steinegger & Söding 2017). These faster alternatives are leveraged by workflows implemented in software such as MEGAN (Huson et al. 2007) and HUMAnN2 (Franzosa et al. 2018) to identify gene family matches and output overall metagenome profiles. HUMAnN2 is a unique approach in that it first screens reads that map to reference genomes of taxa identified as present with MetaPhlAn2. This step enables a small subset of gene families to be linked directly to particular taxa. However, the vast majority of gene families typically have no taxonomic links and are only part of the community-wide metagenome. There are clear issues with the general approach implemented by these gene profiling approaches, as has been previously observed: "genes are expressed in cells, not in a homogenized cytoplasmic soup" (McMahon 2015).

Linking functional annotations to specific taxa by assembling raw reads is the ideal approach to resolve this problem, but this too comes with caveats. Most importantly, insufficiently high read depth, which depends on the complexity of a sample, can result in too few assembled contigs to sensibly analyze. Nonetheless, with sufficiently high read depth metagenome assembly can be a valuable way to leverage information about microbial communities. Note that the below description of metagenome assembly workflows (which runs until the end of this section) is a lightly edited reproduction of a review paper section that I previously published (Douglas & Langille 2019).

There has recently been enormous growth in the number of metagenome assembled genomes (MAGs) generated from MGS data (Pasolli et al. 2019; Almeida et al. 2019). With this growth of available genomes there have been renewed discussions about the need for genome quality control, particularly for MAGs (Shaiber & Eren 2019). Either composite assemblies of multiple taxa or incomplete genomes missing genes of interest could result in false inferences. One extreme example is of the tardigrade genome, which was falsely identified as having 17% of genes acquired through horizontal gene transfer due to contaminant sequences within the assembly (Koutsovoulos et al. 2016). Such false inferences are more likely in metagenome assemblies compared with genome assemblies due largely to the challenge of distinguishing many organisms at different abundances (Ayling et al. 2019). Mis-assemblies can also affect the detection of other genic events as well. For instance, repetitive regions of assemblies are difficult to resolve with current short-read sequencing (Chin et al. 2013), which can make duplication events difficult to identify. Due to these challenges an understanding of the workflows for generating MAGs is needed.

There are many metagenome assembly tools currently available, which are predominately based on De Bruijn graphs of overlapping k-mers (Ayling et al. 2019; Vollmers et al. 2017). The outputs of these tools are assembled contigs, which typically vary in length from ~500 bp to near-complete genomes. Some of the most popular freely available assembly tools are MetaSPAdes (Nurk et al. 2017), Ray Meta (Boisvert et al. 2012), Omega (Haider et al. 2014), IDBA-UD (Peng et al. 2012), and Megahit (Li et al. 2015). Choice of assembly tool can have a major influence on the resulting assembled contigs, and so careful consideration needs to be taken at this stage. An independent evaluation of these and other methods found that MetaSPAdes performed best overall with the caveat that it may not be appropriate for distinguishing highly similar genomes (Vollmers et al. 2017). However, no assembly tool performed best across all environments and it was suggested that the best choice of assembly tool depends on the study environment and research question.

Contig binning, where contigs from the same species or strain are grouped, is another key step when generating MAGs. Binning approaches typically group contigs based on sequence composition (e.g. GC or tetranucleotide content) and similar coverage of mapped reads (Ayling et al. 2019). The most popular freely available binning tools are CONCOCT (Alneberg et al. 2014), MaxBin2 (Wu et al. 2016), and MetaBAT (Kang et al. 2015). As above, the choice of binning software can have drastic effects on the resulting MAGs (Meyer et al. 2018). One partial solution to this issue is to run multiple binning tools and use the software Das Tool (Sieber et al. 2018) to identify the consensus output, which has been shown to produce high quality bins (Meyer et al. 2018).

Evaluating the quality of MAGs is a crucial step once the final contig bins have been generated and guidelines on how to categorize MAGs based upon quality metrics have recently been established (Bowers et al. 2017). The two key metrics are completeness and contamination, which are based on the counts of universal single-copy genes (USCGs) identified in an assembly. Completeness is measured based on the proportion of USCGs identified in an assembly and contamination is defined as the proportion of USCGs found more than once in an assembly. Hard cut-offs for these metrics have been suggested for categorizing the overall quality of a MAG, for instance high-quality draft MAGs are defined as being >90% complete with <5% contamination (Bowers et al. 2017). CheckM (Parks et al. 2015) and BUSCO (Simão et al. 2015) are two tools that will estimate the completeness and contamination of prokaryotic assemblies and BUSCO can also be used to evaluate eukaryotic assemblies. Determining strain heterogeneity, the degree of contamination due to different strains, within an assembly is also important, which can be measured using CheckM or alternatively custom methods to identify polymorphisms in an assembly (Pasolli et al. 2019). An assembly with high strain heterogeneity can still be useful but should be considered differently than an assembly of a single strain.

1.3 - Characteristics of Microbiome Count Data

Regardless of the sequencing technology and workflow used for profiling a microbial community, the final product is typically a count table. This is true for many sequencing approaches, such as RNA sequencing, but there are several differences. First, unlike in the case of RNA sequencing where there are a known number of genomic loci, novel taxa and functions are frequently identified in microbiome data. For instance, novel OTUs, ASVs, and contigs are frequently identified in taxonomic analyses. Similarly, 25-85% of proteins in MGS are novel microbial genes of unknown function (Prakash & Taylor 2012). Second, no statistical distribution fits microbiome data in all contexts. For instance, many statistical distributions, including the negative binomial (Love et al. 2014), beta binomial (Martin et al. 2020), and Poisson (Faust et al. 2012) distributions have been proposed as appropriate fits to microbiome data. However, upon analysis with real data these and other distributions fit with inconsistent accuracy (Weiss et al. 2017; Calgaro et al. 2020). Last, microbiome count tables typically have high sparsity, meaning that there is a high proportion of features not found across many samples (Thorsen et al.

2016). Both of these characteristics make microbiome data analysis challenging for all taxonomic analyses and most functional analyses (see Microbial Functions section).

These challenges are exacerbated by the inherent compositionality of sequencing data. Compositional data refers to data that is constrained to an arbitrary constant sum (Aitchison 1982), such as the arbitrary number of raw sequencing reads output per sample. This characteristic means that the observed abundance of any given feature is dependent on the observed abundance of all other features. A simple example can help illustrate the implications of this characteristic. Imagine a microbe, microbe X, at low relative abundance in sample A and at high relative abundance in sample B. An observer might naively infer that there is more of microbe X in sample B than in sample A. However, there are many reasons this could be false. For instance, the absolute abundance of microbe X could be the same in each sample but the abundance of other microbes in general might be higher in sample A. This higher total microbial load would push the relative abundance of microbe X in sample A down. Depending on the total microbial cell count it is even possible that the absolute abundance of microbe X could be higher in sample A than in sample B, but that the relative abundance is simply lower. This example highlights a necessary consideration regarding microbiome sequencing data analysis: it only provides information on the relative abundances, or percentages, of features and does not provide insight on feature absolute abundances.

This important characteristic was not widely appreciated in the field until relatively recently, when researchers identified fatal issues with common approaches for analyzing microbiome data (Gloor et al. 2016, 2017). Standard differential abundance approaches, such as the t-test and Wilcoxon test, when applied to relative abundances, and microbiome-specific tools such as LEfSe (Segata et al. 2011) do not account for this compositionality. Common summary metrics for microbiome data, such as the UniFrac distance, also suffer from this problem (Gloor et al. 2017). This is a major issue, because ignoring this characteristic is known to lead to spurious discoveries with compositional data (Aitchison 1982; Jackson 1997; Fernandes et al. 2014).

Fortunately, there is active work in the field to resolve this issue and numerous compositional approaches have been developed. The focus has primarily been on developing novel correlation (Friedman & Alm 2012; Kurtz et al. 2015; Schwager et al.

2017) and differential abundance approaches, such as ALDEx2 (Fernandes et al. 2013, 2014) and ANCOM (Mandal et al. 2015). A common theme of these compositional approaches is that the data is transformed based on the ratio of feature relative abundances to some reference frame (Aitchison 1982; Morton et al. 2019). This choice of reference frame varies substantially between approaches. For instance, ALDEx2 transforms relative abundances by the centred log-ratio (CLR) transformation (Fernandes et al. 2013), which essentially normalizes feature relative abundances by the mean relative abundance per sample. This approach transforms the original data but maintains the interpretation of individual features. In contrast, it has been suggested that analyses could instead be based on ratios between features (Morton et al. 2019), which converts the data type into comparisons of features rather than individual features.

There are no best-practices regarding approaches that compositionally transform individual features. More generally, differential abundance tests commonly produce widely different sets of significant taxa from each other (Thorsen et al. 2016; Weiss et al. 2017; Hawinkel et al. 2019). This wide variation is largely due to specific characteristics of microbiome count data. A large proportion of the variation in results is driven by high false discovery rates. Although many methods advertise that only approximately 5% of significant taxa are likely false positives, it has been estimated that for some methods the actual false discovery rate is substantially higher (Hawinkel et al. 2019). This particular validation observed this trend for several methods, including ANCOM (Mandal et al. 2015) and metagenomeSeq (Paulson et al. 2013), two microbiome-oriented methods that are otherwise considered conservative (Paulson et al. 2013; Weiss et al. 2017). In addition, a recent evaluation of differential abundance tools found that compositional methods are actually less robust than several non-compositional alternatives (Calgaro et al. 2020).

Given this wide variation in differential abundance tool performance and unclear best-practices, how is a microbiome researcher to proceed? One possible answer is that a change in expectations regarding the interpretability of microbiome data analysis is needed. In particular, analyses using ratios between the relative abundances of taxa has been shown to be robust, although it comes at the cost of interpretability (Morton et al. 2019). However, an important issue is how to determine which taxa should be the numerator and denominator of each ratio. One solution is to leverage the bifurcating structure of a clustered tree (Egozcue & Pawlowsky-Glahn 2011; Morton et al. 2017) or phylogenetic tree (Silverman et al. 2017) of features. Analyses can be focused on the ratios in relative abundances between features on the left-hand and right-hand of each node in the tree. Despite the potential of this approach, it is rarely used for standard microbiome analyses because it is unclear how to biologically interpret any differences in the values of these ratios across samples. In Chapter 5 I discuss a novel bioinformatics approach that leverages these tree-based compositional approaches and integrates microbial functional information to provide improved interpretability.

The above discussion focused on taxonomic features, which were either based on 16S sequencing or read-based MGS data analysis. However, it is important to emphasize that count tables produced from MAGs do not resolve this issue. In fact, attempting to account for these challenging characteristics of microbiome count data and the links between taxa and function makes the analysis more difficult.

1.4 - Microbial Functions

To this point I have only discussed functional microbiome data in vague terms as referring to microbial gene abundances. When based on DNA sequencing data this information summarizes the functional potential, meaning the functions that are present, but not necessarily active in a community. However, rather than individual gene sequences, research is typically focused on gene families, which are gene clusters. Alternatively, the focus is sometimes on higher-order functional categories like pathways. To complicate matters further, there are several different functional ontologies, which are different frameworks for studying functions at different resolutions. Depending on which of these functional ontologies and sub-categories are analyzed, the characteristics of the data can drastically differ.

The Universal Protein Resource (UniProt) Reference Clusters (UniRef) database contains all protein sequences from the Swiss-Prot (manually curated) and TrEMBL (automated) databases clustered at either 50%, 90%, or 100% identity (Apweiler et al. 2004). The most recent versions of these clusters have been generated with the MMseqs2 algorithm (Steinegger & Söding 2018). As of June 30th, 2020, the 100% identity clusters (called UniRef100), corresponded to 235,561,514 unique protein sequences, which provides a detailed summary of almost all known protein sequences. Despite being clustered at lower identity thresholds, UniRef50 and UniRef90 nonetheless contain enormous numbers of protein clusters: 41,883,832 and 115,885,342, respectively.

The UniRef database contrasts with another common functional ontology, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa & Goto 2000; Kanehisa et al. 2016). KEGG is based on 23,530 individual gene families (as of September 10th, 2020), which are called KEGG orthologs (KOs). The advantage of KOs is that the majority have well-described molecular functions that can be linked to higher-order KEGG pathways and modules. Accordingly, any analysis of KEGG data will likely result in less sparse count tables than the corresponding UniRef-based database, simply because KOs are shared across more taxa than UniRef clusters. To illustrate this point, my colleagues and I previously compared the taxonomic coverage of each function within these two functional ontologies and each sub-category (Inkpen et al. 2017). We found that all UniRef functions, including those in UniRef50 clusters, are on average found in a single domain and encoded by fewer than four species. In contrast, we found that KOs were encoded in 1.3 domains and 184.3 species on average. Similarly, the high-level KEGG modules and pathways were predicted to be potentially active in a mean of 1.7 and 2.5 domains and 671 and 1267.6 species, respectively (Inkpen et al. 2017). Based on these statistics, clearly a shift in the abundance of a UniRef cluster should not be treated the same as a KEGG function: the former corresponds to the activity of a small number of species while the latter could correspond to a large assemblage. This example highlights that the choice of how function is defined in a given analysis can have profound effects on the biological interpretation.

In addition to UniRef and KEGG, several other functional ontologies were leveraged for analyses that are presented in this thesis. These additional function types include: Clusters of Orthologous Genes (COGs) (Tatusov et al. 2000; Makarova et al. 2015), Enzyme Commission (EC) numbers, Protein families (Pfam) (Punta et al. 2012; Finn et al. 2014), and TIGRFAMs (Haft et al. 2003). These categories represent a range of approaches for defining gene families and functional categories.

The COG strategy for functional annotation was originally intended to phylogenetically classify proteins into groups of orthologs (Tatusov et al. 2000). This one-to-one approach of matching individual orthologs has now been expanded to allow for more complex relationships between genes, such as paralogs and horizontally transferred homologs (Makarova et al. 2015; Galperin et al. 2019). As of 2015, there were 4,631 independent COGs (Galperin et al. 2015). The COG framework is similar to that of the eggNOG database (Jensen et al. 2008), which is a more high-throughput, automated approach. However, the key advantage of the COG database is that orthologous genes are clustered into 26 interpretable functional categories, which are expanded from categories originally defined to functionally bin *Escherichia coli* genes (Riley 1993).

The EC number framework, which was developed in 1992 by the "International Union of Biochemistry and Molecular Biology", is a contrasting approach for functional annotation. Instead of focusing on orthologous genes, EC numbers specify particular enzyme-catalyzed reactions. An interesting characteristic of this database is that these reactions can be performed by non-homologous isofunctional enzymes (Omelchenko et al. 2010). As of August 12th, 2020, there were 6,520 EC numbers, which correspond to one of four levels of granularity. For example, the accession EC 3.5.1.2 corresponds to glutaminases, while the higher-level categories correspond to hydrolases (3.-.-.-), that act on carbon-nitrogen bonds other than peptide bonds (3.5.-.-), and that are in linear amides (3.5.1.-). One major advantage of EC numbers is that because they specify exact enzymatic reactions they are straight-forward to link into pathway ontologies based on reactions, such as MetaCyc pathways (Caspi et al. 2013).

The Pfam database categorizes protein families, which are protein regions that share sequence homology (Punta et al. 2012). Individual proteins with multiple domains can thus belong to multiple Pfam families. Each Pfam family is represented by a hidden Markov model (HMM), which models the likely amino acids at each residue and the likely adjacent amino acids based on curated alignments of representative protein sequence. This approach identified homologous protein regions, which are often hypothesized to have a shared evolutionary history, but not necessarily. As of May 2020, there were 18,259 Pfam families.

Lastly, TIGRFAMs are manually curated protein families, which are also identified based on HMMs, but also additional pertinent information (Haft et al. 2003). As of September 16th, 2014, there were 4,488 TIGRFAMs. The distinguishing feature for this database is that different information supplements each HMM. For instance, certain TIGRFAM are annotated based on species metabolic context and neighbouring genes, while others are based on validated functions from the scientific literature. This database has been less commonly analyzed in recent years and is best known as the annotation system for early large-scale metagenomics projects (Venter et al. 2004). Alternative approaches, such as the FIGfam protein database are now more commonly used than TIGRFAMs. FIGfams are based on a similar approach, but instead of being manually curated they are aggregated into isofunctional groups based on shared roles in specific subsystems (Meyer et al. 2009).

A recurrent question thus far has been that given a range of comparable, or contrasting, bioinformatics options, how is one to proceed? Fortunately, in the case of selecting functional ontologies, the choice is much clearer than other bioinformatics areas. Each functional database typically excels for different purposes. For instance, UniRef is useful for identifying uncharacterized genes that may be of interest in an environment, but quickly becomes challenging to interpret and analyze in diverse communities.

In contrast, KEGG is useful for looking for shifts in well-described functions at a high level, which means this database is more robust to granular functional diversity. Due to also being more robust to granular functional diversity and because they are more interpretable, pathway-level functions are often of particular interest. For instance, obesity is associated with an enrichment of phosphotransferase systems involved in carbohydrate processing in human and mouse gut microbiomes (Turnbaugh et al. 2008, 2009a). This straight-forward explanation quickly communicates the pertinent biological details, which might be lost by focusing on more granular levels. However, it is worth noting that pathways identified based on DNA sequencing are merely predictions based on the identified gene families. Although there are several pathway reconstruction approaches, they all require some mapping from gene families or reactions to pathways.

This mapping can be structured, meaning that optional and required contributors can be specified, or non-structured, meaning that all genes and/or reactions are treated equally.

The naïve approach for pathway reconstruction is to assume that a pathway is present if any gene or reaction involved is present in the community. This was the predominant approach used for pathway inference in early functional analyses (Moriya et al. 2007; Meyer et al. 2008) and in several pathway inference tools such as PICRUSt (Langille et al. 2013). Pathway abundance under this framework is calculated by summing the abundance of each contributing gene family. This approach errs towards avoiding missing the presence of a pathway, which is a concern in metagenomes as key genes may be missing due to mis-annotations. However, this approach comes at the cost of spurious annotations. Based on the naïve mapping approach the human genome was previously annotated as including the KEGG pathway equivalent of the reductive carboxylate cycle (Ye & Doak 2011). This pathway is restricted to autotrophic microbes and is similar to reversing the Krebs cycle. Consequently, several gene families are shared in both processes. Under the naïve mapping approach, the presence of genes involved in the Krebs cycle are also evidence for the predicted presence of this atypical microbial pathway in humans. Similarly, vitamin C biosynthesis would also be predicted in humans based on the naïve approach (Ye & Doak 2011). However, the *GLO* gene, which encodes the protein involved in the key last step of vitamin C biosynthesis in mammals, is pseudogenized in humans (Drouin et al. 2011), which makes vitamin C biosynthesis impossible.

The Minimal set of Pathways (MinPath) approach is an approach developed to address this issue (Ye & Doak 2011). This tool identifies the smallest set of pathways, based on maximum parsimony, that are required to explain the presence of a set of proteins. In this way, the approach is more conservative than naïve mapping and also accounts for incomplete protein sets. This method has been applied in numerous contexts, including for the “HMP Unified Metabolic Analysis Network 2” (HUMAnN2) (Abubucker et al. 2012; Franzosa et al. 2018) MGS gene family profiling and pathway reconstruction framework. This popular framework reconstructs pathways based on MinPath and infers pathway abundance based on different approaches, depending if the pathway mapping is structured. For unstructured mappings, the arithmetic mean of the upper half of individual gene family abundances is taken to be the pathway abundance (Abubucker et al. 2012). For structured mappings, the harmonic mean of the key (i.e. required) genes families is computed for pathway abundance (Franzosa et al. 2018). Both these approaches are motivated by the need to be robust to variable abundance in alternative gene families.

Although this approach for MGS pathway reconstruction is commonly performed, it is important to emphasize that it has not been universally accepted and there remains disagreement about best-practices. For example, Evidence-based Metagenomic Pathway Assignment using geNe Abundance DAta (EMPANADA) is a method that addresses the same issue as MinPath and HUMAnN2 in a different way (Manor & Borenstein 2017a). This method focuses pathway reconstruction on distinguishing genes that are shared with multiple pathways from those that are unique to a single pathway. Pathway support weightings are first given by the average abundance of gene families unique to each given pathway. The abundance of all shared gene families is then partitioned between all pathways according to their relative support values. Pathway abundances are then taken as the sum of the unique gene family relative abundances and the partitioned relative abundances of the shared gene families (Manor & Borenstein 2017a).

The exact reconstructed pathways and their respective abundances differ depending on whether naïve mapping, MinPath/HUMAnN2, or EMPANADA are used (see Chapter 4). Validating pathway reconstructions is challenging without a gold-standard comparison, particularly in metagenomes. Even in isolated genomes, as demonstrated by the above examples of the human pathway reconstructions, pathway reconstruction is non-trivial. However, the advantage in these cases is that experimental validation of pathway reconstructions is possible (Francke et al. 2005; Oberhardt et al. 2008). Such validations would be possible if predictions are based on individual members of a microbiome, but it is less clear what experiments could validate pathways predicted for an overall community. In MGS data pathways are typically inferred as though all gene families were free to interact with each other. In other words, they are inferred as though there was universal cross-feeding. All three approaches described above are intended to be used for such community-wide gene family profiles. However, as mentioned above, this assumption is invalid because clearly not all proteins and metabolites in the microbiome can freely interact (McMahon 2015). The implications of this assumption being invalid remain unclear, but nonetheless it is an important caveat when interpreting pathway reconstruction data based on community-wide MGS data.

This section would be incomplete without addressing the most common discussion regarding microbiome functional data: its ostensible high stability. Functional pathways are commonly at similar relative abundances across the same sample-types whereas taxonomic features, such as phyla, can vary substantially (Turnbaugh et al. 2009a; Burke et al. 2011; HMP-consortium 2013; Louca et al. 2016). This functional consistency is often taken to be evidence of environmental selection for particular microbial functions (Turnbaugh et al. 2009a; Louca & Doebeli 2017). However, the validity of comparing variation between these two data types is rarely discussed. My colleagues and I investigated this question from a philosophical perspective and concluded that any meaningful comparison of the relative variation between taxonomic and functional profiles is likely impossible (Inkpen et al. 2017). This difficulty is largely because it is unclear which levels of granularity would be meaningful to compare between each data type. In other words, each data type is qualitatively different from the other and the choice of how to compare the two is based on arbitrary decisions.

For instance, as discussed above, the sparsity and number of possible functional categories differs drastically across ontologies and sub-categories. My colleagues and I demonstrated how observations of functional and taxonomic stability are entirely dependent on how function and taxa are defined (Inkpen et al. 2017). We did this by comparing human stool sample profiles at each possible taxonomic rank and also each functional level for both the KEGG and UniRef functional ontologies. As expected, phyla were less stable across the samples than KEGG pathways, but more stable than UniRef50 protein clusters. However, this area remains an area of active debate. Others have also argued that taxonomic variability never unambiguously reflects functional variation, which they believe is strong evidence for functional conservation (Louca et al. 2018a). Nonetheless, this example demonstrates once again the common theme throughout this section: “function” has many meanings.

1.5 - Metagenome Prediction Methods

Ideally, analyses of microbial functions are based on MGS data. However, including in the discussion on Crohn's disease and the microbiome presented below (Morgan et al. 2012; Davenport et al. 2014), predicted functions based on 16S rRNA gene (hereafter 16S) sequencing are often analysed instead. Metagenome prediction, predicting complete genomes for each individual ASV or taxon weighted by their relative abundance, when based on 16S data is much cheaper than performing MGS.

There are additional advantages of predicted metagenomes over actual MGS data. Namely, MGS is often prohibitively expensive for samples where host DNA overwhelms microbial DNA. The high read depths required to yield sufficient microbial read depths is infeasible in many cases (Gevers et al. 2014). Similarly, low-biomass samples are difficult to accurately quantify with MGS, but they can be profiled with PCR-based 16S sequencing. For example, applying MGS to profile human tumours is currently infeasible, but it is straight-forward to apply 16S sequencing (Nejman et al. 2020). In both cases, for host DNA contaminated and low-biomass samples, metagenome prediction based on 16S profiles is a useful alternative to MGS.

However, metagenome prediction suffers from important drawbacks. The key problematic assumption is that the marker gene used for predictions, typically the 16S, is strongly associated with genome content. This broad assumption is correct: genera such as *Lactobacillus* and *Desulfobacter* can be easily distinguished based on the 16S and they are enriched for extremely different functions. Namely, *Lactobacillus* can often perform lactic acid fermentation whereas *Desulfobacter* can typically oxidize acetate to CO_2. Such comparisons of characteristic functions between distantly related taxa are uncontroversial. The difficulty arises when approaches attempt to predict entire genome contents for an entire community, including for closely related taxa.

This issue is highlighted by classic DNA hybridization experiments (Mandel 1966; Brenner 1973). These experiments are based upon mixing single-stranded DNA from two organisms and recording the melting temperature required to separate the strands. Higher melting temperatures are required to break apart DNA that shares more complementary bases connected by hydrogen bonds. This approach provides a rough estimate of the genetic distance between different strains or species. An early comparison of these genetic distances with 16S dissimilarity across 34 bacteria computed a linear correlation of 0.728 (Devereux et al. 1990). However, the relationship between these two metrics is not linear: many bacteria with highly similar 16S genes have hybridization rates much lower than 70% (Stackebrandt & Goebel 1994), which is the traditional cut-off for delineating species. This trend has been corroborated across diverse prokaryotes (Hauben et al. 1997, 1999; Kang et al. 2007). In addition, a meta-analysis of 16S gene sequencing and DNA hybridization data from 45 bacterial genera further clarified these observations (Keswani & Whitman 2001). This analysis established that 78% of the variability in hybridization rates could be accounted for by 16S similarity, based on a non-linear model. However, they also identified that a minority of hybridization rates were extremely poorly predicted by 16S similarity (Keswani & Whitman 2001).

These observations agree well with genomic comparisons of strains, which can drastically differ in genome content. For example, across 17 *E. coli* genomes there are ~13,000 genes that are variably distributed and only ~2,200 core genes (Rasko et al. 2008). This enormous range of genomic variation is not reflected at the 16S level, where *E. coli* strains are typically >99% identical (Suardana 2014). These genomic differences can translate to enormous variation at higher taxonomic levels as well. For instance, a comparison of the genomes from 11 *Yersinia* species found a range of genome sizes from 3.7 – 4.8 megabases (Chen et al. 2010). A closer comparison of three pathogenic species of *Yersinia* determined that they shared 2,558 protein clusters while 2,603 were variably distributed. These species-level differences are also not proportionally reflected by divergence in *Yersinia* species 16S genes, which are typically > 97% identical (Ibrahim et al. 1993). These examples highlight that 16S similarity can be a poor predictor of genomic similarity. This issue is compounded when there are divergent 16S copies within the same genome, although typically these are >99.5% identical (Větrovský & Baldrian 2013).

Variation in gene content within a taxonomic lineage is a recurrent observation across microbial communities. Variably present genes are often linked to putative niche-specific adaptations (Wilson et al. 2005), such as genes affecting antibiotic resistance (Kallonen et al. 2017), carbohydrate catabolism (Arboleya et al. 2018), and wound healing (Kalan et al. 2019). Based on these and other observations, the understanding of bacterial genomic content has shifted from that of a static genome to a pan-genome, consisting of core and variable genes (Tettelin et al. 2005). Variably present genes are transmitted between genomes through horizontal gene transfer, which typically occurs between closely related organisms (Popa & Dagan 2011). However, horizontal gene transfer can also occur between distantly related organisms, such as between different bacterial phyla (Beiko et al. 2005; Kloesges et al. 2011; Martiny et al. 2013).

The high variability between bacterial genomes and extensive horizontal gene transfer highlights the major challenges facing metagenome prediction. Despite these challenges, interest in performing metagenome predictions has continued, supported by several observations. First, although there are important outliers, 16S sequence identity does logarithmically correlate well with the average nucleotide identity between genomes, with an R^2 of 0.79 (Konstantinidis & Tiedje 2005). Second, 16S sequence similarity does provide some information on the ecological similarity of bacteria (Chaffron et al. 2010). This was demonstrated by the fact that co-occurring environmental bacteria are more likely to have similar 16S sequences. In addition, overall differences in inferred KEGG pathway potential are strongly associated with 16S divergence (Chaffron et al. 2010). Last, within a given environment, such as the human gut, 16S divergence was shown to be particularly predictive of divergence in average gene content (Zaneveld et al. 2010).

Originally, metagenome prediction workflows were based on matching 16S sequences to reference genomes. In addition to predicting microbial functions linked to Crohn's disease (Morgan et al. 2012), this approach has also been used to profile diet-related microbial functions across mammals (Muegge et al. 2011) and the functions of invasive bacteria within corals (Barott et al. 2012). Although bioinformatics tools for metagenome prediction are now typically used for performing this task, this 16S-matching approach is still used for custom analyses (Verster & Borenstein 2018; Bradley & Pollard 2020).

The first metagenome prediction tool to expand beyond this approach, and specifically intended for 16S sequencing data, is 'Phylogenetic Investigation of Communities by Reconstruction of Unobserved States' (PICRUSt) (Langille et al. 2013). To distinguish this software from the tool I present in Chapter 4, I will hereafter refer to this version as PICRUSt1. This tool is based on leveraging classical ancestral-state reconstruction methods, which have been widely used in phylogenetics (Zaneveld & Thurber 2014). The key extension of this framework is to extend trait predictions from internal, or ancestral, nodes in a phylogenetic tree to tips with unknown trait values. This approach has been termed hidden-state prediction (HSP) (Zaneveld & Thurber 2014). PICRUSt1 bases genome predictions on genomes from the Integrated Microbial Genomes (IMG) database (Markowitz et al. 2012). Generating predictions for individual 16S sequences is time and memory-consuming with this approach, so for typical usage predictions for all OTUs from the Greengenes database (DeSantis et al. 2006) are pre-computed in advance and then provided to users. However, it is possible for users to conduct ancestral-state reconstruction with custom reference genome databases, although this is a laborious process. In fact, the construction of one alternative database focused on cow rumen-associated bacteria, called CowPI, was deemed sufficient work to be published as a standalone paper (Wilkinson et al. 2018).

This approach requires a phylogenetic tree based on an alignment of 16S sequences from reference genomes and 16S sequences without known genome content. The known genome annotations typically correspond to KEGG or COG gene families by default. A range of phylogenetic ancestral state reconstruction methods can be used to reconstruct trait values at ancestral nodes. Many of these methods are implemented by the Analyses of Phylogenetics and Evolution (APE) R package (Paradis et al. 2004), which is used by PICRUSt1. Faster and extended versions of these methods were recently implemented in the castor R package as well (Louca & Doebeli 2018). The default PICRUSt1 predictions are based on the Phylogenetic Independent Contrasts (PIC) reconstruction method (Felsenstein 1985). PIC is a fast method for reconstructing continuous traits, which can also be applied to discrete traits like gene counts if they are treated continuously. Under the PIC model evolution occurs through Brownian motion (i.e. a random walk), where differences between organisms are normally distributed around a mean of 0 with a variance proportional to their phylogenetic distance. A least squares approach is used to predict ancestral trait values, based on this model and the input data (Langille et al. 2013). The key hidden-state prediction step implemented in PICRUSt1 is performed for each OTU (with unknown genome content) by taking the average of its ancestral and close relatives' value weighted by the branch-length distance to each value. Metagenome prediction can then be performed by multiplying the relative abundance of each OTU by the abundance of each gene family within each OTU's predicted genome. PICRUSt1 predicts pathway levels through naïve mapping of genes to pathways (see Microbial Functions section).

PICRUSt1 introduced the step of normalizing relative abundances by the predicted number of 16S copies within each genome, which is intended to control biases in 16S sequencing due to copy number (Farrelly et al. 1995). Importantly, although 16S copy number correction has become a common step for metagenome prediction (Angly et al. 2014), accurately predicting 16S copy number is particularly challenging. An independent validation of several 16S copy number prediction methods, including PICRUSt1, identified poor agreement of predicted copy numbers against existing reference genomes (Louca et al. 2018b). In some cases, less than 10% of the variance in actual 16S copy number was explained by these predictions. In addition, these predictions were often only slightly correlated between prediction methods.

The original validation of PICRUSt1 predictions focused on comparing predicted KEGG orthologs (KOs), pathways, and modules with the corresponding functions identified with MGS data on matched samples (Langille et al. 2013). This approach uncovered a high Spearman correlation between these data types on the same samples. In particular, predicted KOs from Human Microbiome Project (HMP) samples across numerous body sites were highly correlated with the matching MGS-identified data (Spearman $r = 0.82$). Principal component analysis based on 16S-predicted and MGS-derived KEGG modules also identified close matching by sample-type, regardless of whether the data was predicted or not. Lastly, PICRUSt1 predictions also matched with particular functional shifts in MGS datasets. For example, uronic acid metabolism was found to be at higher levels across all compared HMP body sites in both data types, with the exception of urogenital samples where this signal was largely absent (Langille et al. 2013).

Although PICRUSt1 was shown to perform well in these validations, it nonetheless suffers from certain limitations. First, as mentioned above, by default PICRUSt1 requires input sequences to be OTUs that can be linked with the Greengenes database. This means that some form of closed-reference OTU picking is required to identify these links, which eliminates the possibility of generating predictions for novel 16S sequences. This is a major disadvantage, because in certain communities, such as in soil and ocean samples, *de novo* 16S sequences can account for a large proportion of the data. In addition, due to this limitation, PICRUSt1's default implementation is incompatible with sequence denoising methods (Callahan et al. 2016a; Edgar 2016; Amir et al. 2017), which are rapidly becoming the predominant approach as they enable sequence resolution down to the single-nucleotide level. This improved resolution allows closely related organisms to be better distinguished and thus more precise gene annotations are associated with a given 16S sequence. Another major drawback is that PICRUSt1 cannot be used with 18S rRNA gene and internal transcribed spacer sequencing data as its database is limited to prokaryotic community predictions. The feasibility of making such eukaryotic predictions with PICRUSt1 is unclear. Lastly, the prokaryotic reference databases used by PICRUSt1 have not been updated since 2013 and lack many recently added gene families and pathway mappings. I hypothesized that making these and other improvements to the PICRUSt workflow, which I implemented in PICRUSt2, would improve the accuracy and flexibility of the tool. I also identified an important issue related to how to interpret concordance between 16S-predicted and MGS-derived functional data. As described in the manuscript (see Chapter 4), high correlations can occur in these comparisons by chance, which has major implications for how metagenome prediction approaches in general are evaluated.

Since PICRUSt1 was published a number of similar metagenome prediction tools have been developed. All of these approaches aim to capture the shared phylogenetic signal in the distribution of functions across taxa; however, they differ in many fundamental ways. Tax4Fun (Aßhauer et al. 2015) is a similar approach that pre-computes predictions in advance and restricts input data to OTUs corresponding to the SILVA database (Quast et al. 2013). Predictions are based on converting SILVA OTUs to prokaryotic KEGG organisms, which have pre-computed functional profiles. An updated version of this tool, Tax4Fun2 (Wemheuer et al. 2020), extended this approach to allow for ASVs and *de novo* OTUs to be used. Input 16S sequences are mapped against reference sequences in Tax4Fun2 to identify the closest match, which allows novel sequences to be used.

Piphillin is another metagenome prediction approach, which is based on a nearest-neighbour search between input and reference 16S sequences within a global alignment (Iwai et al. 2016). The genome content of the matching nearest neighbour is taken to be the genome for each input 16S sequence. This procedure is fast and is re-run for each input dataset, meaning that either reference OTUs or novel ASVs can be input. A major advantage of this tool is that it has an online web-portal, which allows straight-forward usage. This tool was shown to have minor performance improvements over PICRUSt1, which at the time was hypothesized to be due to the higher number of genomes included in the reference database (Iwai et al. 2016). The Piphillin authors also introduced the notion of using the concordance of differential abundance results between actual and predicted metagenomics profiles as a novel validation. Although this is a valuable idea, thus far it has only been applied with a limited number of tool and differential abundance test comparisons (Iwai et al. 2016; Narayan et al. 2020; Sun et al. 2020).

Pangenome-based Functional Profiles (PanFP) is a contrasting method that bases metagenome prediction on the taxonomic classifications of 16S sequences, rather than on exact sequences (Jun et al. 2015). The taxonomic lineage for each OTU is mapped to RefSeq lineages and then genome predictions are performed by averaging the genome content of all organisms within a given lineage. PICRUSt1 differs from this approach, in addition to Tax4Fun and Piphillin, because PICRUSt1 explicitly accounts for each unknown organism's position in a phylogenetic tree. Accounting for this phylogenetic information enables more sophisticated methods for inferring hidden states to be employed.

PAthway PRedction by phylogenetIC plAcement (PAPRICA) is another approach that leverages phylogenetic information (Bowman & Ducklow 2015). This method uses a phylogenetic placement approach, originally pplacer (Matsen et al. 2010) and now the next-generation of the Evolutionary Placement Algorithm (EPA-ng) (Barbera et al. 2019), to place input 16S sequences into a reference phylogenetic tree. The advantage of placing sequences into a reference tree is that it more accurately represents the phylogenetic relationship between reference and query sequences than if an entirely new alignment and tree are created (Janssen et al. 2018). Genome predictions are acquired differently depending on how query sequences are placed (Bowman & Ducklow 2015). If a query's most likely placement is on an edge leading to a tree tip, then the genome content is taken to be that tip. Alternatively, if query sequences are placed at internal nodes then the predicted genome content is taken to be the core genome of all descendent tips. The major drawback of PAPRICA is that it has not been evaluated using standard MGS-16S comparisons. Instead, it has largely been evaluated by its utility in practice for explaining variation across aquatic microbial communities compared with PICRUSt (Bowman & Ducklow 2015). The original database for this tool also was focused on marine microbes and originally the tool was challenging to use because it required raw reads, rather than OTUs or ASVs, to be input. Another important difference of this tool compared with the others is that it focuses on EC number predictions rather than KOs or COGs.

In addition to these prediction methods that predict gene families across metagenomes, alternative methods have focused instead on predicting more interpretable phenotypes. The Functional Annotation of Prokaryotic Taxa (FAPROTAX) is a database that enables prokaryotic taxonomic labels to be linked with specific high-level functions (Louca et al. 2016). This database focuses on over 80 interpretable functions, such as nitrate respiration and methanogenesis. These taxa-function links were collated through a systematic literature review. BugBase is a similar approach that instead focuses on categorizing six main phenotypes: Gram stain, oxygen tolerance, biofilm forming, mobile element content, pathogenicity, and oxidative stress tolerance (Ward et al. 2017). Rather than literature searches, these phenotypes were inferred based on the genome content of reference genomes. The key advantage of both FAPROTAX and BugBase is that the high-level traits they predict are typically more robust within taxonomic lineages than lower level functions, such as individual gene families (Martiny et al. 2013).

1.6 - Current State of the Integration of Taxonomic and Functional Data Types

The above discussion has described the many faces of microbiome data types. Taxonomic and functional microbiome data are typically generated independently, but in some cases can be directly linked. Regardless of the exact processing workflow for these data types, I have yet to address one question: how are they integrated?

For independent taxonomic and functional data types this is largely done anecdotally. For example, this is commonly done in regards to the nine genera that are the primary producers of short-chain fatty acids (SCFAs) in the human gut (Moya & Ferrer 2016). SCFA levels have long had an ambiguous link with Crohn's disease (CD) (Treem et al. 1994), although they are typically negatively associated with disease activity (Venegas et al. 2019). Due to this association, there has been long-standing interest in identifying microbial taxa that are associated with altered SCFA levels. Accordingly, CD microbiome studies commonly hypothesize that shifts in the relative abundance of any known SCFA-producing taxa likely cause altered SCFA levels. For example, *Faecalibacterium prausnitzii* is a well-known commensal SCFA-producer in the human gut and is consistently found at lower levels in the CD patient microbiomes (Wright et al. 2015). Although potential links between lower levels of this species, in addition to other taxa such as *Roseburia* (Laserna-Mendieta et al. 2018), and SCFA levels are often discussed, this is rarely formally investigated.

More often, anecdotal links between function and taxa are based on observed associations between significant features. Several such cases have previously been noted as representative examples (Manor & Borenstein 2017b). For instance, *Propionibacterium acnes* has been identified as strongly correlated with NADH dehydrogenase levels in the skin microbiome (Oh et al. 2014). Consequently, this species was implicated as the likely cause for changes in NADH dehydrogenase levels. Similarly, *Bacteroides thetaiotaomicron* relative abundance has been identified as positively correlated with microbial genes involved with the degradation of complex sugars and starch in the infant gut (Bäckhed et al. 2015). Based on this observation, this species was hypothesized to be the key contributor to increased levels of these degradation genes. Such insights are valuable, but as previously discussed (Manor & Borenstein 2017b), these anecdotal links alone are not convincing evidence that particular taxa are the primary contributors to functional shifts.

Linked taxonomic and functional data alone is not sufficient to resolve this issue. There are substantial challenges facing the integration of these data types besides simply generating a combined format. For example, two massive datasets have recently been published as part of the next iteration of the Human Microbiome Project. Both datasets include numerous sequencing and profiling technologies, including 16S and MGS, from the stool and various body-sites of IBD (Lloyd-Price et al. 2019) and individuals with pre-diabetes (Zhou et al. 2019). However, in each case there was little integration of microbiome functional and taxonomic data types. Instead, these features were largely tested independently, despite the availability of links between the data types, and associations between top features were discussed (Lloyd-Price et al. 2019; Zhou et al. 2019).

In contrast to these examples, there have been calls for improved integration of these microbiome data types, which is rooted in a systems-level biology outlook (Greenblum et al. 2013). ‘Functional Shifts’ Taxonomic Contributors’ (FishTaco) is one bioinformatics method developed for this purpose, which quantifies taxonomic contributions to functional shifts (Manor & Borenstein 2017b). Significant shifts in functional abundances are first identified using a standard differential abundance test, typically a Wilcoxon test. Subsequently, a permutation analysis is undertaken, which consists of randomly shifting the relative abundance of a subset of taxa, while maintaining the rest. A large collection of such permutations is performed, which include permutations of single and multiple taxa in different replicates. Based on this approach an estimate of the relative contribution of each taxon to a functional shift can be estimated (Manor & Borenstein 2017b). These relative contributions are then presented as stacked bar charts breaking down the direction and magnitude of each functional contribution. These visualizations help distinguish when a functional shift is due to the enrichment or depletion of taxa and also which sample grouping the shift occurred within. This approach was motivated by Shapley values, which were introduced in game-theory to summarize the contribution of each player in a multiplayer game (Shapley 1953). Specifically, FishTaco leverages a modified version of this approach that enables the contribution of individual features to be estimated in large datasets without exhaustively testing every possible permutation (Keinan et al. 2004).

FishTaco represents an important advancement in integration and improved interpretability of taxonomic and functional microbiome data. However, it nonetheless suffers from major limitations. First, although the taxonomic breakdown of contributors to a function is valuable, the FishTaco approach requires significant functions to be identified based on the relative abundance of individual gene families and pathways. This is done by systematically testing all functions across the entire metagenome, which is problematic when performed with a non-compositional approach like a Wilcoxon test. This approach also treats gene families under the bag-of-genes model, which is inappropriate, as discussed above. An improved method would conduct a compositionally sound analysis and integrate taxonomic information when identifying significant functions.

An alternative method is phylogenize, which does address each of these issues (Bradley et al. 2018; Bradley & Pollard 2020). This approach tests for significant associations between the presence of a taxa within a given sample grouping and the probability that a taxon encodes a given gene family. This is performed through phylogenetic linear regression, which accounts for the genetic similarity of co-occurring taxa that might trivially be due to a shared evolutionary history. A separate phylogenetic linear model is fitted for each gene family. The key distinction of this approach from a normal linear model is that instead of the residuals being independent and normally distributed, they covary so that phylogenetically similar microbes have higher covariance (Bradley et al. 2018). This overall approach was partially motivated by an attempt to address a similar problem by comparing the species and gene trees of gut and non-gut microbes (Lozupone et al. 2008). Based on simulated random data (i.e. data with no real functional shifts) the phylogenize authors demonstrated that performing standard linear models without controlling for phylogenetic structure results in false positive rates ranging from 20% - 68%. In contrast, controlling for phylogenetic structure with phylogenize resulted in a uniform P-value distribution and an appropriate false positive rate of 5%. One interesting feature is that phylogenize does not directly analyze relative abundances. Instead, the tool converts taxa relative abundance into one of three formats: (1) binary presence/absence across all samples, (2) overall prevalence within each sample grouping, (3) or the specificity within each sample grouping (Bradley et al. 2018).

Although phylogenize is undeniably an invaluable contribution to microbiome data analysis, it also has several limitations. First, information on taxa abundance is discarded entirely in favour of presence/absence data. From one perspective this is an advantage; eliminating taxa relative abundances enables phylogenize to circumvent compositionality issues. However, relative abundance data is often more important to investigate, because key taxonomic shifts might not be detected by presence/absence alone. In addition, phylogenize reports significant gene families for each phylum in a dataset. This is performed to reduce the memory usage and to enable phylum-specific rates of evolution for each function (Bradley et al. 2018). This focus on the phylum level makes the results difficult to interpret for two reasons. First, it is insufficiently broad, because it limits the potential to identify functions distributed across multiple phyla that might be linked with a condition of interest. From another perspective, this focus on the phylum level is also not specific enough; although phylum-function associations are valuable they do not provide information on the relative contributions of lower-level taxa, such as species, to the association. Accordingly, there is room for improvement in both the statistical analysis and interpretation of the phylogenize approach.

Despite the availability of approaches for integrating functional and taxonomic data, they have yet to become a mainstay of microbiome analyses. However, it is becoming common to visualize stacked bar-charts of taxonomic contributors to functions of interest. This is typically performed on predicted metagenome output by PICRUSt or alternatively on HUMAnN2 output, although this could be performed with any linked taxa-function data. As discussed above, the HUMAnN2 pipeline includes a step for identifying particular strains in MGS dataset, which allows gene families to be linked to those strains (Franzosa et al. 2018). In some cases this approach enables complete links between taxa and function to be identified. For instance, *F. prausnitzii* was shown to be the obvious principal contributor to glutaryl-CoA biosynthesis in the HMP gut MGS samples (Franzosa et al. 2018). However, more commonly there are numerous taxonomic contributors to a single given function, and it is difficult to interpret which taxa are the key contributors by looking at visualizations alone. Nonetheless, even in the presence of many taxonomic contributors, the HUMAnN2 authors demonstrated that these visualizations can provide information about the diversity of taxa contributing to a function, termed the contributional diversity (Franzosa et al. 2018). This is most often quantified with the Gini-Simpson index, which is the complement of Simpson's evenness (Jost 2006).

Contributional diversity has been shown to be a useful approach for delineating housekeeping pathways encoded by many taxa, intermediate pathways, and those rarely encoded, which can correspond to opportunists or keystone species. For instance, *F. prausnitzii* has previously been linked with several human microbiome pathways identified through MGS that have intermediate contributional diversities (Abu-Ali et al. 2018). When present, this species tended to contribute the majority of all pathways it encoded.

This approach has also been valuable for profiling shifts in the contributions to microbial pathways over time, such as in the infant gut profiled with MGS (Vatanen et al. 2018). In this case, several microbial pathways, such as siderophore biosynthesis, were found to display decreasing contributional diversity with age. This is an interesting observation because siderophores are costly to produce but are highly beneficial in the human gut. In particular, siderophores can confer a strong benefit to multiple community members, including those that do not produce siderophores, by providing access to iron. Siderophores have previously been presented as microbial functions whose distribution is consistent with the Black Queen Hypothesis (Morris et al. 2012). This hypothesis states that adaptive gene loss may occur for functions that are costly to produce, provided that the function is provided by other community members. This hypothesis was discussed in the context of the infant microbiome as an explanation for why siderophore contributional diversity decreases over time (Vatanen et al. 2018): perhaps gene loss confers an adaptive benefit by avoiding the production of a costly metabolite. Although this is an interesting hypothesis, a less controversial interpretation of this result is that siderophores became less stably encoded over time.

Related to this point, two additional metrics have also been developed to summarize the stability of taxonomic contributions to microbial functions (Eng & Borenstein 2018). More specifically, these metrics are intended to summarize functional robustness across samples, which is the stability in the relative abundance for a given function in response to taxonomic perturbation. This is performed by generating a taxa-response curve that describes the average change in functional relative abundances in response to taxonomic perturbations of different magnitudes. Two metrics are then computed based upon these curves: attenuation and buffering. Attenuation captures how rapidly a function shifts with increasing taxonomic perturbation magnitudes. In contrast, buffering represents how well functional shifts are suppressed at smaller taxonomic perturbation magnitudes.

Applying these metrics to PICRUSt-predicted metagenomes from 16S sequencing of human body sites, validated by a subset of MGS samples, yielded several novel perspectives. First, attenuation and buffering were conserved across body sites for microbial house-keeping pathways but varied for several others. For instance, robustness in the biosynthesis of unsaturated fatty acids varied substantially across body sites. In addition, human gut samples were found to have higher values of both attenuation and buffering than compared to vaginal samples. These trends were shown to be driven by more than simply lower richness in vaginal samples by subsampling to comparable diversity levels across each body-site (Eng & Borenstein 2018). These observations are consistent with the controversial hypothesis that microbial communities may be under varying selection strengths for functional robustness, depending on the environment (Naeem et al. 1998; Ley et al. 2006).

The development of these metrics for summarizing functional contributions represent an important goal of microbiome research, which is to leverage sequencing data to yield novel biological insights. In contrast, another major goal is to answer a more practical question: how useful is microbiome data for classification and prediction tasks?

There is great interest in applying machine learning approaches to microbiome sequencing data (Knights et al. 2011). Most commonly this is performed with either Support Vector Machine or Random Forest (Breiman 2001) models. Applications of these and other machine learning approaches to microbiome data are primarily aimed at classifying healthy and diseased samples (Zhou & Gallins 2019). On rare occasions this is performed on functional data types, which was the focus of one MGS meta-analysis that identified informative functional biomarkers across several human diseases (Armour et al. 2019). However, more commonly taxonomic features are the focus of these machine learning approaches, which is true for both 16S (Duvallet et al. 2017) and MGS (Pasolli et al. 2016) data.

Regardless of the data type, one major motivation of these projects is to leverage the gut microbiome to improve disease diagnoses. The microbiome of IBD patients has been analysed through machine learning for this purpose on numerous occasions (Gevers et al. 2014; Tedjo et al. 2016; Sprockett et al. 2019). In addition to classifying patients by diseases, this work has also focused on leveraging 16S-based taxonomic profiles to classify patients by remission status (Tedjo et al. 2016) and disease sub-type (Gevers et al. 2014). This latter approach has particularly been valuable as a proof-of-concept that intestinal biopsies from multiple regions of the GI tract perform comparably in a machine learning context. In general, the observed within-study performances typically are relatively high. For instance, one influential study computed accuracy scores (the area under the curve [AUC] in this case) ranging from 0.66 – 0.85 (Gevers et al. 2014). However, the generalizability of these models is rarely assessed across different study cohorts. This issue applies not only to CD, but also to most human diseases associated with the microbiome.

One major exception is microbiome-based models for colorectal cancer, which in one investigation were shown to be generalizable across five independent datasets (Wirbel et al. 2019). This landmark study also systematically compared the utility of functional and taxonomic data types in these models and found them to be comparable overall. This finding is consistent with a past comparison of the classification performance of 16S-based taxa and predicted metagenome data (Ning & Beiko 2015). In the case of predicted metagenomes, which are based on 16S profiles, it is perhaps less surprising that they yield comparable classification performance. However, with MGS data in particular it might be possible to detect robust, informative functions that might be undetectable with taxonomy alone due to taxonomic variability (Doolittle & Booth 2017).

Despite this great interest in applying machine learning to different microbiome data types, there has been little focus on integrating across them. The aforementioned comparison of 16S-based taxa and predicted functions is one exception where a hybrid classification model of both data types was created (Ning & Beiko 2015). In this case, there was a small increase in classification performance for distinguishing nine human oral sub-locations. The original OTU and KO-based models yielded accuracies of 76.2% and 76.1%, respectively, while the hybrid model resulted in an accuracy of 77.7% (Ning

& Beiko 2015). This result indicates that predicted functions may provide some additional information in combination with taxonomic data, but the consistency and biological significance of this small effect remains unclear. Further investigation into the integration of these data types within a machine learning context is needed to ensure that the highest-quality models possible are constructed.

1.7 - Microbial Associations with Disease: Crohn’s Disease as a Case Example

The above sections have described microbiome data types in detail. I have discussed the different sequencing technologies, data types, and introduced challenges facing microbiome data analysis. However, I have neglected to describe the motivation driving most human microbiome research: to identify links between the microbiome and disease.

Myriad associations have been identified between microbiome features and disease, such as asthma (Arrieta et al. 2015; Hufnagl et al. 2020), obesity (Ley et al. 2005; Turnbaugh et al. 2009a; Castaner et al. 2018), and colorectal cancer (Zeller et al. 2014; Flemer et al. 2018). A review of microbial associations with disease in general is beyond the scope of this work, and detailed reviews on this topic are available elsewhere (Durack & Lynch 2019; Nørreslet et al. 2020; Willis & Gabaldón 2020). Instead, I will focus on a single disease, Crohn’s disease (CD), which has a complicated etiology associated with the microbiome. Understanding this disease is important to appreciate Chapter 3, which focuses on the microbiomes of intestinal biopsy samples from pediatric CD patients.

CD is an inflammatory bowel disease (IBD) characterized by chronic mucosal inflammation. Any part of the gastrointestinal (GI) tract can be discontinuously affected, but the ileum and/or colon are most often inflamed (Thia et al. 2010). This disease has many symptoms, including diarrhea, abdominal pain, GI bleeding, and fatigue. In addition to physical issues, these symptoms often have negative social and psychological impacts on CD patients, which drastically decrease patients’ quality-of-life (Casati & Toner 2000; Sewitch et al. 2001). Common treatments for CD include prescribing corticosteroids, 5-aminosalicylates, purine antimetabolites, and anti-tumour necrosis factor suppressive strategies (Ho & Khalil 2015). In addition, 71% of CD patients eventually undergo invasive intestinal surgery (a surgical resection), which unfortunately leads to relapses (up to 10 years later) in 44% of cases (Bernell et al. 2000).

CD diagnosis has become more common in the Western world over the last 60 years. As of 2011, CD prevalence varied from 44 – 201 per 100,000 persons across North American cohorts (Cosnes et al. 2011). Although historically CD has typically developed in young adults (Ekbom et al. 1991), there is increasing incidence in pediatric patients (Jabandziev et al. 2020). Canada has one of the highest incidence rates for pediatric CD: 9.68 per 100,000 children under 16 from 1999-2010 (Benchimol et al. 2017). Although all surveyed Canadian provinces have similar incidence rates, Nova Scotia has the highest incidence rate for CD. In addition, younger children seem to be a growing risk group for CD in Canada as disease incidence was found to be increasing only for children under five (Benchimol et al. 2017). In general, CD in younger individuals can be more complex and aggressive (Goodhand et al. 2010). Current treatments can be less effective for treating these cases, or alternatively can have severe side effects on pediatric patients, such as corticosteroids that may interfere with growth during puberty (Alemzadeh et al. 2002). Immunosuppressive treatments are also associated with increased infection risks, likely caused by impaired immune responses (Bonovas et al. 2016). These issues, in addition to the lack of long-term solutions for CD, have motivated renewed efforts to develop novel treatments.

Exclusive enteral nutrition (EEN) is one such alternative first-line therapy to standard approaches such as corticosteroids. This treatment involves providing all nutrition via a liquid formula, which can be done orally or with a feeding tube for 6-8 weeks (Whitten et al. 2012). EEN has been investigated with variable success for over 30 years (Morain et al. 1984). With the exception of Japan, EEN is not an effective treatment for adults, partially because it requires strict dietary adherence (Wall et al. 2013). In contrast, EEN is highly effective for treating active disease and inducing remission in pediatric patients (Day et al. 2006; Critch et al. 2012). Although the mechanism of action remains unknown, EEN provides additional benefits than corticosteroids, such as higher rates of mucosal healing, greater weight gain, and enhanced bone turnover (Wall et al. 2013).

There is a high genetic predisposition for CD as indicated by twin-studies (Orholm et al. 2000; Halfvarson et al. 2003; Halme et al. 2006). A meta-analysis of six CD twin-studies determined that the overall concordance rate is 30.3% and 3.6% for identical and fraternal twins, respectively (Brant 2011). The higher concordance rate in identical twins indicates that CD risk is highly heritable. Accordingly, there have been many investigations into genetic variants underlying CD (Liu & Anderson 2014). Many risk loci have been identified and one interesting discovery has been that the genetic variants associated with CD susceptibility are largely independent from those related to disease prognosis (Lee et al. 2017).

A major focus has been to develop genetic risk score models based on combining the odds-ratios of hundreds of single-nucleotide polymorphisms (SNPs). This approach has partially accounted for CD heritability estimated from twin-studies and to develop CD screening methods based on relatively few genetic variants (Wang et al. 2013a; Zupančič et al. 2016). Genetic risk score estimates for 34,819 IBD patients based on approximately 200 SNPs demonstrated that this approach could also be used for distinguishing disease sub-types (Cleynen et al. 2016). In particular, CD could not only be distinguished from another IBD, ulcerative colitis, but ileal and colonic CD sub-types could also be distinguished. This result highlights that genetic risk score can in principle be leveraged to rapidly stratify patients into different disease sub-types, which could have treatment implications.

The primary disease risk gene for CD was identified in 2001 as the nucleotide-oligmerization-domain-2 (*NOD2*) locus (Hugot et al. 2001; Ogura et al. 2001). This gene is an apoptosis regulator expressed in monocytes and variants of the encoded protein activate the pro-inflammatory transcription factor NF-κB. NOD2 is also a pattern recognition receptor that can detect cell wall components of both Gram-positive and Gram-negative bacteria as well as other microbes (Moreira & Zamboni 2012). *NOD2* variants have also been associated with extreme microbial shifts following antibiotic usage and with failure to control microbial pathogens (Al Nabhani et al. 2017).

Many environmental factors are also thought to contribute to the etiology of CD, which likely explains the rising rates of pediatric CD (Feeney et al. 2002). Smoking is strongly associated with CD, with an odds ratio of 3 to 4-fold increased risk of developing disease (Halfvarson et al. 2006; Ng et al. 2012). Early-life GI infections and antibiotic usage are also associated with developing CD (Nguyen et al. 2020). In addition, differences in dietary intake have been linked to CD, for instance diets with high fibre and fruit content are protective (Hou et al. 2011). Lastly, reduced exposure to sunlight may be linked with CD risk as vitamin D insufficiency is common in CD patients (Raftery & O’Sullivan 2015). Many of these etiological factors, such as dietary differences, GI infections, and antibiotic usage are intuitively potentially linked with the microbiome. Consequently, there has been a long-standing interest in investigating links between the gut microbiome and CD etiology.

The most consistent microbial signal in the gut microbiome of CD patients has been reduced alpha diversity (Manichanh et al. 2006; Hansen et al. 2012; Pascal et al. 2017). However, different alpha diversity metrics are often significant (or tested) across different studies. For instance, it was previously typical to assess alpha diversity based on richness alone (Manichanh et al. 2006). In some cases richness has not significantly differed between CD and control samples, but other measures of alpha diversity such as Shannon’s entropy and evenness, have significantly differed (Hansen et al. 2012). Therefore, lower alpha diversity metric values in general are characteristic of the gut microbiome of CD patients, but there can be subtleties regarding how to interpret that difference depending on the cohort and analysis. Importantly, CD patients also have significantly lower microbial load compared with healthy control individuals, which likely accounts for the observed differences in alpha diversity metrics (Vandeputte et al. 2017).

However, higher alpha diversity should not be equated with healthiness in general (Shade 2017). EEN typically causes a rapid reduction in alpha diversity, but nonetheless is able to induce remission in 80% of patients (MacLellan et al. 2017). In fact, patients who exhibit sustained remission after EEN have significantly lower alpha diversity following treatment, whereas an increase in alpha diversity has been observed in patients who relapse after treatment (Dunn et al. 2016a). After several weeks the microbiome of patients in sustained remission eventually becomes more similar to healthy samples (Lewis et al. 2015), which suggests that EEN does not cause a permanent reduction in alpha diversity.

Specific bacterial taxa are also at differential relative abundances in the gut microbiome of CD patients. In general, many bacterial commensals are at lower levels or absent in CD patients (Duvallet et al. 2017), which is consistent with the lower alpha diversity in these samples. The most consistent signal is that of higher relative abundances of Proteobacteria and lower relative abundances of Firmicutes (Frank et al. 2007; Sartor 2008; Gevers et al. 2014). In particular, members of the Proteobacteria family Enterobacteriaceae, such as *E. coli*, are often highly associated with CD (Frank et al. 2007; Willing et al. 2009). In addition, certain *NOD2* risk alleles are associated with increased levels of enterobactin synthesis, which enables Enterobacteriaceae to inhibit the bactericidal host enzyme myeloperoxidase (Bonder et al. 2016). Inhibiting this enzyme likely confers a survival advantage to *E. coli* and other Enterobacteriaceae in the gut (Singh et al. 2015).

Although most CD microbiome research has focused on bacteria, fungi and viruses are also associated with patient disease state. Interestingly, fungal alpha diversity is largely unaffected in CD samples, which indicates that fungi may have an advantage over bacteria under inflammatory conditions (Sokol et al. 2017). In addition, there are reproducibly higher levels of Basidiomycota and lower levels of Ascomycota in CD patient microbiomes (Mukhopadhya et al. 2015; Sokol et al. 2017). The pathogenic Ascomycota *Candida albicans* is an important exception, which frequently infects CD patients (Sokol et al. 2017; Stamatiades et al. 2018).

Bacteriophages are highly associated with the specific bacteria they infect, so perhaps unsurprisingly many bacteriophages have also been linked with CD (Norman et al. 2015; Ungara et al. 2019b). However, these associations do not necessarily merely reflect trivial associations with CD-linked bacteria: bacteriophages can also directly modulate host immunity. This was demonstrated by inoculating germ-free mice with bacteriophages specific to *Lactobacillus*, *Escherichia*, and *Bacteroides* (Gogokhia et al. 2019). Direct stimulation of toll-like-receptor nine was shown to result in the production of the pro-inflammatory cytokine interferon gamma. The key bacteriophage signature of IBD is an increase in *Caudovirales* taxonomic richness (Norman et al. 2015). In contrast to bacteriophages, any connections between CD and eukaryotic viruses are less clear (Ungara et al. 2019b), but nonetheless several associations have been identified. In particular, *Hepeviridae* and *Virgaviridae* have been observed at higher and lower relative abundances, respectively, in the gut of CD patients compared with control subjects (Ungara et al. 2019a). Importantly however, analyzing and preparing viral MGS data is particularly challenging (Rose et al. 2016): these results should be interpreted with the understanding that the field of viral metagenomics is rapidly changing and improving.

In humans, it remains unclear whether these microbial differences are a cause or an effect of disease. In mice, there are clearer links between the relative abundances of certain bacteria and intestinal anti-inflammatory and pro-inflammatory responses (Khan et al. 2019). In particular, *Clostridia* and *Bacteroides* are known to induce regulatory T cells that initiate anti-inflammatory cytokine responses (Geuking et al. 2011; Atarashi et al. 2013). In contrast, segmented filamentous bacteria are sufficient to induce pro-inflammatory cytokine activation by T helper 17 cells (Ivanov et al. 2009).

Although identifying similar causal links is challenging in humans, it is possible to make more robust inferences by controlling for confounding factors in observational studies. One important confounding factor is the effect of treatment on the microbiome. This was the motivation driving the collection of samples at or near the time of diagnosis in the large BISCUIT (Hansen et al. 2013) and RISK (Gevers et al. 2014) IBD cohorts. Similarly, the microbial profile of stool samples represents an aggregate of microbial communities from across the GI tract. Accordingly, stool sample profiles often substantially differ from biopsy microbial profiles. Even in the case of colon biopsies, which correspond to the region of the GI tract with the highest microbial load (Hillman et al. 2017), stool samples nonetheless are highly distinct (Stearns et al. 2011). Because CD can affect the GI tract at different locations it is important to differentiate microbial signals identified in particular GI tract locations compared with in the stool overall. Sequencing biopsy samples of inflamed regions of the GI tract can address this issue. Based on this idea, a comparison of the 16S sequencing profiles of the stool, ileal biopsies, and rectal biopsies taken from CD patients was undertaken as part of the RISK project (Gevers et al. 2014). The stool profiles qualitatively differed from the biopsy sample-types, but interestingly there were only minor differences overall between the microbial composition between the two biopsy locations. In fact, microbial profiles at either biopsy site could accurately diagnose CD regardless of the location of inflammation (Gevers et al. 2014). Although the authors performed MGS on a subset of stool samples, this was deemed infeasible at the time for the biopsy samples due to the high proportion of unintended human DNA that would be sequenced.

MGS profiling of the stool of CD patients has nonetheless yielded valuable insights regarding the disease (Morgan et al. 2012; Gevers et al. 2014). One key observation is that the shift in the gut microbiota from obligate anaerobes, such as many Firmicutes, to facultative anaerobes, such as Enterobacteriaceae, is consistent with the gut environment shifting to a more oxygenated environment (Byndloss et al. 2018). IBD is known to result in higher levels of free oxygen species and oxidant levels in general, particularly near inflamed tissue (Keshavarzian et al. 2003). More generally, a shift from obligate anaerobic to facultatively aerobic commensals has been observed in several contexts where the microbiome is associated with diseases or treatments. For instance, similar microbial shifts are observed following antibiotic therapy, colorectal cancer, and *Salmonella* infection (Litvak et al. 2018).

An oxygen-dependent shift in the metabolism of colonocytes could account for these observations. Under homeostatic conditions the colon is typically an anaerobic environment, that is partially maintained by the microbial production of butyrate, a short-chain fatty acid, through the digestion of dietary fibres primarily by Clostridia (Rivera-Chávez et al. 2016). To perform one of the colon's primary functions, water absorption, an osmotic gradient is created by first absorbing Na^+ into colonocytes. The majority of ATP required for Na^+ transport by colonocytes under these conditions is generated by oxidizing butyrate, which produces CO_2 (Velázquez et al. 1997). This process maintains the hypoxic conditions of the colonocytes, and because oxygen can diffuse freely across biological membranes, it is also thought to greatly contribute to depleted oxygen levels in the lumen. A key observation was that if there is a perturbation to this process at one of several key steps, this can result in a shift of colonocyte ATP production to require less oxygen (Rivera-Chávez et al. 2017). This could occur through direct inflammatory damage to colonocytes or the perturbation of butyrate-producing commensal bacteria (Litvak et al. 2018). In either case, facultative aerobes will bloom at the expense of obligate anaerobes, which agrees well with the common microbial signatures in CD patient microbiomes.

In the context of this larger model, several more specific microbial functions have also been linked with disease state. In particular, bacteria that produce the short-chain fatty acids acetate and propionate, in addition to butyrate, are commonly at lower levels in CD patient microbiomes (Wang et al. 2014; Takahashi et al. 2016). This observation has also been corroborated with metabolite data from CD patient samples (Lloyd-Price et al. 2019). Bacteria involved with nucleotide biosynthesis are also often found at lower levels in CD patient microbiomes (Morgan et al. 2012). One hypothesis for this recurrent observation is that there are fewer carbohydrates available for bacterial metabolism in inflamed tissue, which could select for bacteria that are able to perform amino acid and lipid metabolism (Davenport et al. 2014). In addition, microbes that encode genes related to glutathione production and sulfate transport have been found at higher levels in CD microbiomes, which could indicate a response to inflammation-related oxidative stress (Morgan et al. 2012). Lastly, microbial functions related to xenobiotic degradation, such as nitrotoluene degradation, have also been linked with CD (Dunn et al. 2016b). The mechanism underlying this observation is unclear, but it could have implications for CD treatment efficacies (Clarke et al. 2019).

1.8 - Outlook

As discussed above, an increased focus on the integration of functional and taxonomic microbiome data types is needed. In many cases, linked taxa and functions are available, but integration between them is only performed anecdotally. In addition, although there are methods available for analyzing integrated datasets, there are major issues with the implementation and interpretation of the results. There are several areas where methodological improvements could resolve these issues and yield improved insights regarding microbiomes. My goal was to address these issues, which I present from three perspectives in this thesis.

My first results chapter focuses on my work that compared the utility of different microbiome data types for classifying pediatric CD (Douglas et al. 2018). This is presented in Chapter 3 and the methods are presented in Chapter 2, where I describe all methods used in this thesis. These analyses focused on functional and taxonomic data types from both 16S and MGS data at different levels of granularity. One motivation of this work was to both corroborate known and identify novel microbial links with CD based on these data types. Another major goal was to compare the performance of machine learning models based on each individual data type and on combined data types.

Although this machine learning approach yielded valuable insights, it is difficult to make more than anecdotal biological insights regarding taxa and function without explicit links between them. These linkages are commonly generated with 16S data by performing metagenome prediction. Although this approach has been widely applied, particularly with the tool PICRUSt, there are limitations of this approach. In addition, there have been suggestions that past validations have not fully captured the variability in performance across metagenome prediction tools. I developed PICRUSt2 to address these problems (Douglas et al. 2020), which is presented in Chapter 4. I show that improved metagenome predictions, and the resulting taxa-function links, are produced by this method compared with alternative metagenome prediction approaches. In addition, my evaluations highlighted major issues with the reproducibility and interpretability of both predicted and actual functional data.

Such issues are commonly encountered when analyzing microbiome data, which could be partially addressed by better integrating data types in statistical analyses. To complement existing integration approaches like FishTaco and phylogenize, I developed a novel bioinformatics tool called POMS. This method is described in Chapter 5. The main goal of POMS is to incorporate functional information into a pre-existing framework for identifying taxonomic differences. This framework provides a method for organizing functional signals by how consistently they are found in independent taxonomic groups at relatively higher or lower abundance in sample groupings. The intuition underlying this approach is that functions that show consistent signals of enrichment are more likely to represent actual cases of enrichment due to the effects of the function rather than due to indirect effects. Based on applying POMS on simulated and real data we showed that the top functional features are more reliable than current differential abundance methods. This body of work represents several useful advances in microbiome data analysis, particularly from the perspective of improved integration of functional and taxonomic data types. In some cases, as discussed below, my work resulted in more questions than answers, and identified alarming issues with current analysis approaches. However, although not always the answers we might want, this gradual progress is needed to provide substantive improvements to the microbiome field.

Chapter 2 – Materials & Methods

This section contains the methods sections corresponding to the manuscripts presented in Chapters 3, 4, and 5. This separate presentation of the methods is due to a Microbiology & Immunology department thesis formatting requirement.

2.1 – Methods for Chapter 3

2.1.1 - Sequenced Samples

Intestinal biopsies were previously taken from 20 Crohn's disease (CD) and 20 normal colon control patients as part of the "Bacteria in Inflammatory bowel disease in Scottish Children Undergoing Investigation before Treatment" (BISCUIT) cohort (Hansen et al. 2012, 2013). We did not perform a power test to predict what effect sizes could be detected with this sample size, but instead chose this sample size due to sequencing cost constraints. These patients were all under 17 years old with a mean age of 12.7 years. CD biopsies were obtained at the diagnostic endoscopy prior to commencing any therapy. We based CD diagnosis on the Paris Classification (Levine et al. 2011). None of these patients used systemic antibiotics or steroids in the 3 months prior to their colonoscopy or immunosuppression at any point. Treatment response was classified as sustained remission following induction treatment response and was defined by physician global assessment and the requirement for treatment escalation (repeat induction therapy) before 24 weeks.

2.1.2 - Metagenomics Sequencing and Bioinformatics Pipeline

Shotgun MGS preparation and sequencing was conducted by Génome-Québec (McGill University, Montréal, Québec) on an Illumina HiSeq. A mean of 110 million PE 100 base-pair (bp) MGS reads were produced with a range of 72.7-135 million reads over all samples. We first concatenated FASTQ files containing forward and reverse reads into a single FASTQ per sample. We then screened out contaminant sequences by mapping all reads against the human (hg19) and PhiX (RTA) genomes using bowtie2 (Langmead & Salzberg 2012) (v2.1.6), which resulted in a mean of 90% of reads being excluded. This high percentage of contaminant reads is mainly due to the high proportion of human cells in biopsy samples, which is less of an issue for microbiome studies that focus on stool samples. After screening out these non-microbial reads we classified the remaining reads taxonomically using MetaPhlAn2 (Segata et al. 2012) (v2.1.0) with the "–very-sensitive" global alignment option and into KEGG orthologs (KOs) using HUMAnN2 (Abubucker et al. 2012) (v0.11.1; http://huttenhower.sph.harvard.edu/humann2). Importantly, we found that running bowtie2 in local alignment mode with MetaPhlAn2 resulted in many spurious hits, which were mainly represented by viruses. These taxa were not identified when global alignment was performed. We ran MUSiCC (Manor & Borenstein 2015) (v1.0.2) to normalize the KO abundances within each sample by the median universal single-copy gene abundance, which controls for inter-sample variation in microbial genome sizes. We then ran HUMAnN2 on these normalized values to reconstruct KEGG module and pathway abundances within each sample. No taxa or functions were identified in the MGS of two samples, S34 and S38 (16S sequencing also failed for these samples, see below), which were excluded from downstream microbiome analyses.

2.1.3 - Calling Human Variants

Due to the large percentage of human DNA in our MGS (see above), we were also able to call human variants from the same dataset. Although we used 133 loci for calculating the genetic risk score (see below), we called genome-wide variants to improve imputation accuracy in cases where samples were missing data at these sites. We began by mapping all MGS reads to the human genome (hg19) using the Burrows-Wheeler Alignment Tool's (Li & Durbin 2009) (v0.7.12) mem algorithm, which resulted in a 98% mapping rate. This mapping rate is higher than the rate for the metagenomic microbial pipeline due to the different algorithms used for each workflow. We then followed the Genome Analysis ToolKit's (McKenna et al. 2010) (GATK) Best Practices workflow (DePristo et al. 2011; Van der Auwera et al. 2013) for variant calling. Pre-processing steps included marking duplicate reads, recalibrating base quality scores based on a model trained on known variants, and re-alignment of reads around known insertions and deletions. We then ran the GATK (v3.5) program HaplotypeCaller to call variants using default parameters and variant quality score recalibration per the Best Practices workflow. These steps resulted in 16,333,869 raw variants based on a genome-wide mean coverage of 7.5

reads across all 40 individuals. Due to the low genome-wide coverage, we also discarded variants based on several hard filters implemented by VCFtools (Danecek et al. 2011) (v0.1.13): any variant not in Hardy-Weinberg equilibrium (cut-off significance of P < 1×10^{-4}), any variant called by < 6 reads, or any variant with > 50% missing data. We retained 7,604,626 variants following these hard cut-offs. The 133 known risk loci were not required to pass these hard cut-offs.

2.1.4 - Imputing Missing Genotypes

After calling variants genome-wide, we next imputed the missing genotypes for the 133 known CD risk loci. Three variants (rs9264942, rs11209026, rs6927022) were missing genotype calls in all samples and were excluded. Haplotype phasing and the first pass of imputation was performed with SHAPEIT (Delaneau et al. 2012) (v2.r837). IMPUTE2 (Howie et al. 2009, 2011) (v2.3.2) was then run on SHAPEIT's phased output to impute the final genotypes. The HapMap phase II b37 genetic map was used for both imputation steps and the 1000 Genomes Phase 3 (1000 Genome Project Consortium 2015) phased haplotypes were used as reference haplotypes. Default parameters were used for running both SHAPEIT and IMPUTE2.

2.1.5 - Genetic Risk Scores

A custom Perl script was used to parse the IMPUTE2 output into variant call format and then PLINK (Purcell et al. 2007) (v1.90b3.29) was used to convert this table into PED and MAP files. Per-sample genetic risk scores (GRS) were calculated using the Mangrove R package (Jostins et al. 2013). To calculate the GRS, we used the genotypes at these imputed risk loci, odds-ratio information for risk alleles, and minor allele frequencies from previously published genome-wide association studies (Jostins et al. 2012; Liu et al. 2015b). We assumed a CD prevalence of 1% when calculating GRS (K value = 0.01).

2.1.6 - 16S rRNA Gene Sequencing

The intestinal biopsy samples were prepared for 16S sequencing using our Microbiome Amplicon Sequencing Workflow (Comeau et al. 2017). Briefly, the pre-extracted DNA

(Hansen et al. 2013) was first amplified in duplicate using dual-indexing Illumina primers (forward: ACGCGHNRAACCTTACC; reverse: ACGGGCRGTGWGTRCAA) that targeted the V6-V8 region (438 bp) of the bacterial 16S rRNA gene. The pooled duplicate PCR products were verified using high-throughput E-gels (Invitrogen), then purified and normalized using the SequalPrep 96-well Plate Kit (Invitrogen). Following quantification, the pooled samples were run on an Illumina MiSeq using PE 300+300 bp v3 chemistry at the Integrated Microbiome Resource (Dalhousie University, Halifax, Nova Scotia).

2.1.7 - 16S rRNA Gene Bioinformatic Pipeline

We followed the Microbiome Helper standard operating procedure (Comeau et al. 2017) to process the 16S rRNA gene data. Two CD samples (S34 and S38) were excluded from this pipeline due to low DNA quality and repeated sequencing failures, which left a total of 38 samples remaining (20 CN and 18 CD). A mean of 21,793 raw PE read pairs were produced over these remaining samples (min=9,503; max=40,392). Forward and reverse reads were then stitched together using PEAR (Zhang et al. 2014) (v0.9.6) with an assembly rate >80% for all samples except for sample S22 (68.7% of reads assembled). We then filtered out stitched reads with a quality score < 30 over 90% of bases using the FASTX toolkit (v0.0.14; http://hannonlab.cshl.edu/fastx_toolkit/). We also filtered out reads < 400 bp or that did not have exact matches to the forward and reverse primers using BBMap (v35.82; https://sourceforge.net/projects/bbmap/). An average of 18.7% of the assembled reads per sample were discarded by these filters. Next, we removed chimeric sequences using UCHIME (Edgar et al. 2011) (v6.1) with the parameters mindiv=1.5 and minh=0.2, which resulted in an average of 16.3% of the assembled reads being discarded. Following these filters a mean of 13,815 reads were remaining per sample (min=4,427; max=27,472). We ran open-reference 97% OTU picking using QIIME (v1.9.0) wrapper scripts with these filtered reads. Reference OTU picking was run against the Greengenes (DeSantis et al. 2006) (v13_8) database using SortMeRNA (Kopylova et al. 2012) (v2.0-dev, 29/11/2014) with a minimum query coverage of 80% and *de novo* OTU picking using SUMACLUST (v1.0.00; https://git.metabarcoding.org/obitools/sumatra/wikis/home/). We filtered out OTUs that were called by < 0.1% of reads and then rarefied read counts to 4,000 reads per sample, which resulted in a final set of 984 OTUs. PICRUSt (Langille et al. 2013) (v1.0.0) was used to predict KEGG ortholog and pathway abundances based on reference OTU abundances. We compared the rarified OTU abundances to non-rarified abundances after performing a centered log-ratio transformation (Gloor et al. 2016). Read counts were imputed with the count zero multiplicative method in the zCompositions R package (Palarea-Albaladejo & Martín-Fernández 2015) (v1.1.1) before performing the centered log-ratio transformation. We compared these workflows by evaluating how well models performed using abundance tables produced by each workflow. To evaluate concordance between MGS and 16S-identified genera we calculated the Spearman's correlation (ρ) of the relative abundances of 16S genera at greater than 10% frequency and identified in both datasets.

2.1.8 - RISK Validation Cohort

We downloaded single-end sequencing of the V4 region of the 16S gene produced for the "Risk Stratification and Identification of Immunogenetic and Microbial Markers of Rapid Disease Progression in Children with Crohn's Disease" (RISK) cohort (Gevers et al. 2014) from the National Center for Biotechnology Information under study accession PRJEB13679. We reduced this data to 773 biopsy samples that were either controls or CD patients and <= 18 years old. To process this data, we first merged together sequencing replicates for the same samples. We then trimmed all reads to 130 nucleotides using Trimmomatic (Bolger et al. 2014) (v0.36). The remaining steps were the same as the 16S processing pipeline described above. The OTU table was rarefied to 4000 reads (42 samples with depth below this cut-off were discarded), which resulted in 2564 OTUs being called over 731 samples.

2.1.9 - Random Forest Classification

For each dataset, we ran random forest (RF) models to classify disease state and treatment response separately. Each dataset was pre-processed so only features with > 10% non-zero values were retained. Each table was then standardized by sample (subtracted the sample's mean and then divided by the sample's standard deviation). We ran RF models using the randomForest (Liaw & Wiener 2002) (v4.6.12) R package with default *mtry* values and used 712 as the random seed. All models were run with 10,001 trees except for the KO models which were run with 501 trees to reduce running time. RF model significance was determined by the permutation test implemented in the rfUtilities (Murphy et al. 2016) (v2.0.0) R package. This test involves building a null distribution of out-of-bag (OOB) errors from RF models with randomized classes (e.g. the disease state column of the input table was randomized). Model significance is then determined by calculating whether < 5% of random permutation models have an OOB error less than or equal to the observed OOB error. Significance of RF models as tested by the above permutation procedure was treated as an omnibus test for any association between the signal derived from genetic data and the feature labels of each sample. This allowed us to identify at what level (e.g., family, genus and species) further investigation was warranted, and supported our investigation of variable importance in some "datasets" and not others. Note that RF models make no assumptions about how the input features are distributed. Leave-one-out cross-validation was also run on each dataset to output an accuracy for each model with the R package caret (Kuhn 2008) (v6.0.77).

2.2 – Primary Methods for Chapter 4

This section describes the methods primarily related to the main-text results. In contrast, the Supplementary Materials section which is later in this file describes the methods underlying the Supplementary Results.

2.2.1 - Data Availability

The raw sequencing reads analyzed in this study are available from the following online repositories. The Human Microbiome Project (HMP) raw data is available from https://www.hmpdacc.org/HMIWGS/healthy/. The mammalian stool sequencing data is available from the Short Read Archive (SRA) under accessions SRP115632 (shotgun metagenomics [MGS]) and SRP115643 (16S rRNA gene). The ocean sample sequencing data is available from SRA project SRP056891. The blueberry soil samples are available at SRA project accessions PRJNA484230 (MGS) and PRJNA389786 (16S rRNA gene). The Cameroonian MGS data is available under European Nucleotide Archive (ENA) project PRJEB27005 and the 16S rRNA gene sequencing data is available at MG-RAST under accession mgp15238. All of the Indian sequencing data is available under ENA project PRJNA397112. The primate metagenomics data and 16S rRNA gene sequencing data (with processed outputs) are available in the QIITA repository under accession 11212.

The blueberry 18S rRNA gene sequencing data are available under SRA project accessions PRJNA391782 (soil) and PRJNA434067 (root). The matching MGS data for these blueberry root and soil samples are available under accession PRJNA484230. The processed ITS output files for the wine fermentation dataset are available as part of a GigaDB dataset (http://gigadb.org/dataset/100309), and the raw MGS data are available as part of SRA project accession PRJNA305659.

2.2.2 - PICRUSt Pipeline Updates

The analyses in this paper are based on PICRUSt2 version 2.1.0-b. In addition to the improvements reported in the main text, several other updates have also been made to the PICRUSt pipeline. Since the hidden-state prediction (HSP) step is now run using the castor R package, other inference approaches like maximum parsimony (MP) may be performed in realistic time-frames besides phylogenetic independent contrasts (Felsenstein 1985), which was the default approach in PICRUSt1. The default HSP method is now MP with a parameter weighting the contribution of branch lengths set to 0.5 (*edge_exponent* option in castor package). This parameter value was chosen because setting this parameter to a non-zero value resulted in more reproducible predictions.

In addition, now that any study sequences can be input to PICRUSt, and not just Greengenes closed-reference OTUs, a nearest-sequence taxon index (NSTI) screening step is recommended to eliminate sequences above a certain cut-off. The default NSTI cut-off in PICRUSt2 is two, which was chosen as an extremely lenient cut-off intended to eliminate problematic sequences. Only one ASV in the test inflammatory bowel disease dataset (see below) was above this cut-off, which corresponded to a mitochondrial sequence. The only sequences above this cut-off in the HMP validation dataset corresponded to two 18S rRNA gene ASVs that were clustered within the 16S rRNA gene dataset. Similarly, although 13/1148 of ASVs in the ocean dataset were above the NSTI cut-off of two, these ASVs corresponded to candidate taxonomic groups that have no representative reference genomes in the default PICRUSt2 database. Based on these observations, we believe this cut-off should be suitable for most scenarios; however, users can select a NSTI value that best fits their study design and environment (i.e. whether to maximize precision or recall).

Transforming gene family predictions to pathway abundances in PICRUSt1 was done by assuming that the abundance of each gene family contributed equal abundance to all pathways containing the gene family (i.e. if a gene family can be involved in 10 pathways the gene family abundance would be added equally to the abundance of all 10 pathways). Although this approach is easy to understand, it results in a high false-positive rate of identifying pathways present. To improve on this approach, we adapted the approach taken by HUMAnN2 (Franzosa et al. 2018) v0.11.1 into the PICRUSt2 pipeline. MinPath (Ye & Doak 2011) (v1.2 as modified for the HMP workflow (Abubucker et al. 2012)) is first run to identify the minimum pathways present given the gene families present. By default, these predictions are made based on the EC number predictions after regrouping them to MetaCyc reactions to predict MetaCyc pathway abundances. The mappings files and code for regrouping to MetaCyc reactions and mapping from reactions to structured pathways were taken from HUMAnN2. We further split the pathway mapping files into prokaryotic and fungal sets based on the taxa where these pathways have been identified (as reported in the MetaCyc online database).

2.2.3 - 16S rRNA Gene Database Processing

The 16S rRNA gene sequences and gene family counts from a total of 52,217 genomes were acquired from IMG on 8 Nov. 2017. These data were based on IMG annotations and we did not work with the raw genome sequences. Genomes lacking a 16S rRNA gene length of at least 1,250 bp or that were identified as eukaryotic marker genes were removed. The gene families in these annotations corresponded to these databases: Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2012) (KEGG; v77.1), Protein Families(Finn et al. 2014) (Pfam; v30), The Institute for Genomic Research's database of protein FAMilies (Haft et al. 2003) (TIGRFAM; v15), Clusters of Orthologous Genes (Tatusov et al. 2000) (COG; v2014), and Enyzme Commission (EC) numbers (as of 21 Jan 2016).

We also created an alternative trait database containing phenotypes defined by IMG (Chen et al. 2013). Use of this database was motivated by the predictions made by the tools FAPROTAX (Louca et al. 2016) and Bugbase (Ward et al. 2017). These phenotypes are more directly interpretable than the gene family databases described above. Files listing IMG genome ids positive for one of 65 phenotypes were downloaded on 8 Jan 2019. The presence and absence of phenotypes was re-coded as 1 and 0. Prototrophic and auxotrophic phenotypes for the same compound were combined into a single prototrophic phenotype (auxotrophs are coded as 0 and unknown phenotypes are coded as NA). After this merging step, and removing two extremely rare phenotypes, there was a final set of 41 phenotypes remaining.

To identify low-quality, incomplete genomes, and possible misassembled assemblies we calculated the median number of single-copy KEGG orthologs (KO) as previously identified for the tool MUSiCC (Manor & Borenstein 2015). Since these genes are expected to be found in single copies within each genome, we reasoned that incomplete or contaminated genomes could be identified by a median copy number less or greater than one, respectively. Accordingly, we discarded all genomes with a median number of single-copy genes that differed from one. In addition, we discarded genomes lacking a sufficient number of genes within any gene family. The minimum number of gene families per genome was chosen based on visualizing the distribution of gene family numbers over all genomes and choosing a cut-off that eliminated outliers (minimum cut-offs of unique gene families were 500, 250, 500, 750, and 350 for the COG, EC, KO, Pfam, and TIGRFAM databases). Importantly, this filtering means that endosymbionts and other organisms with reduced genome sizes will be underrepresented in the PICRUSt2 reference database. After these filtering steps, a total of 10,291 genomes were discarded, producing a final set of 41,926 genomes. Gene family copy numbers higher than 10 in this final set were re-coded to be 10 to decrease the number of possible prediction states.

Since prokaryotes often have multiple 16S rRNA gene copies, the centroid 16S rRNA gene per genome was identified using the VSEARCH (Rognes et al. 2016) (v2.4.4) *cluster-fast* command with an identity cut-off of 90%. In cases where multiple centroid sequences were found, a single centroid was chosen randomly. Identical centroid 16S rRNA gene sequences across genomes were then identified to produce genome clusters. There were 3,002 such clusters with more than one genome and the sequences contained in these clusters made up 59.5% of the original 41,926 genomes. Among the total 20,000 sequence clusters (including clusters with one genome), there was a mean of 2.1 sequences per cluster (standard deviation = 562.5). The cluster with the highest sequence count of 1,379 corresponded to strains of *Staphylococcus aureus*. We observed a mean clustered-sequence length of 1489.6 base-pairs (bp; standard deviation = 65.8) overall.

The final 16S rRNA gene sequences were used to build a multiple sequence alignment (MSA) using ssu-align (v0.1.1; http://eddylab.org/software/ssu-align/) against the bacteria alignment model. Note that although the majority of the reference 16S rRNA gene sequences correspond to bacteria, there are archaeal sequences as well, for which the bacteria ssu-align alignment model is likely less appropriate. Only weak masking of this output MSA was performed (*ssu-mask* options: *--pf 0.001 --pt 0*). The custom Python script *derep_fasta.py* was used to identify sequences in this alignment after this masking step. A phylogenetic tree was built from this MSA with RAxML-ng (Kozlov et al. 2019) (v0.6.0) using the *GTR+G* model. The custom Python script *mean_16S_function_counts.py* was used to calculate the mean gene family abundances for all sequences within a cluster. These values were then rounded to the nearest integer. The full NCBI taxonomic lineage of all 16S rRNA gene clusters was called by the taxizedb R package (https://github.com/ropensci/taxizedb) using the species name provided by the IMG FASTA metadata. Gene family trait depths were calculated using the castor R package function *get_trait_depth* with default settings, which is based on the consenTRAIT metric (Martiny et al. 2013).

2.2.4 - Amplicon Dataset Processing

As described in the main text, we analyzed the following datasets:

1. 57 stool samples from Cameroonian individuals (Morton et al. 2015; Lokmer et al. 2019)
2. 91 stool samples from Indian individuals (Dhakan et al. 2019)
3. 137 samples from different body sites, but primarily stool (part of the Human Microbiome Project (Huttenhower et al. 2012) [HMP])

4. 77 non-human primate stool samples (Amato et al. 2019)
5. Eight mammalian stool samples (Finlayson-Trick et al. 2017)
6. Six ocean samples (Gillies et al. 2015)
7. 22 bulk soil and blueberry rhizosphere samples (Yurgel et al. 2019)

An in-depth comparison of the technologies and sequencing depths for each of the seven 16S rRNA gene validation datasets is shown in Table 4.2. The processing pipelines and filtering criteria differed for each dataset due to technical differences between them. The key difference was that DADA2 (Callahan et al. 2016a) was run for the HMP dataset because this was Roche 454 sequence data. Deblur (Amir et al. 2017) was run on all other validation datasets since they were Illumina sequence data (Table 4.2).

The HMP reads were filtered using DADA2 (v1.6.0) options. Denoising these sequences with DADA2 resulted in 1,865 ASVs after discarding ASVs with a minimum frequency of 10 and discarding 17 samples with fewer than 2,000 reads. The reverse complement of these sequences was taken before running the functional prediction pipelines. The mammalian stool dataset was run using deblur with the default options in QIIME 2 (Bolyen et al. 2019) (v2017.12), which resulted in 323 ASVs after discarding two samples with final read counts less than 3,220. The Cameroonian, Indian, ocean, and blueberry soil datasets were also processed with deblur and no post-processing was done (besides discarding samples that did not overlap with the MGS data) because all samples had high depth, which resulted in 4,077, 2,237, 1,148, and 3,333 ASVs for each dataset respectively. The primate dataset previously processed by deblur was acquired through the QIITA database (Gonzalez et al. 2018), which contained 7,452 ASVs after excluding samples with no MGS data available.

Before running PICRUSt1 on each dataset, ASVs were matched against the Greengenes v13_8 OTUs using the VSEARCH command *–usearch_global* with an identity cut-off of 97%. The ASVs were then regrouped to be the best matching Greengenes OTU to be compatible with the default PICRUSt1 pipeline.

The alternative functional prediction tools discussed in this paper were the following versions (with default reference databases): PICRUSt version 1.1.3, Tax4Fun2 version 1.1.3, PanFP from a specific GitHub commit (1f49bd1b7341b47d46fa7eaa45d7771044d0efde), Piphillin online interface as of 7th Nov.

2019 at http://piphillin.secondgenome.com (KEGG v88.1) and PAPRICA (Bowman & Ducklow 2015) version 0.5.2.

We also created several shuffled prediction datasets to help evaluate the PICRUSt2 predictions. These shuffled datasets were based on shuffling the predicted genome content and not the relative abundance of the ASVs across samples. This shuffling was performed on entire predicted genomes, i.e. each ASV was assigned the genomic content of a randomly sampled ASV (see below). This approach allowed us to assess how randomizing predicted genomic content for ASVs across samples in a dataset affects the concordance with MGS data. These shuffled datasets also provide a baseline of the performance expected by chance for the differential abundance validations. In other words, they give a baseline of the expected concordance (e.g. precision and recall) based on differential abundance testing compared to MGS data given the same ASV relative abundance across samples but shuffled predicted genomes.

More specifically, these datasets were produced by shuffling the ASV ids in each PICRUSt2 prediction table (i.e. the first column of the PICRUSt2 output prediction tables). This was performed 10 times for each dataset and then averaged to produce a single shuffled table per dataset. Shuffled MetaCyc pathway abundances were produced by running the PICRUSt2 pathway pipeline on the shuffled EC metagenome tables. These shuffled datasets are referred to as the "Shuffled ASVs" category in the main-text and figures.

2.2.5 - Shotgun Metagenomics Sequencing Validation Dataset Processing

All shotgun metagenomic sequencing (MGS) datasets were processed using the same pipeline, which is described below. Each dataset was filtered using kneaddata (v0.6.1; https://bitbucket.org/biobakery/kneaddata/wiki/Home) to run (1) Trimmomatic (Bolger et al. 2014) (v0.36) to exclude low-quality reads with the options *SLIDINGWINDOW:4:20* and *MINLEN:50* and (2) bowtie2 (Langmead & Salzberg 2012) (v2.3.2) to exclude reads that mapped to the human and PhiX genomes with the options *–very-sensitive* and --*dovetail*. For the blueberry-associated samples we also mapped reads against the northern highbush blueberry (*Vaccinium corymbosum)* genome (Gupta et al. 2015) version W8520 (downloaded from https://www.vaccinium.org on October 29, 2018) to exclude additional contaminant reads. For samples with paired-end reads the forward and reverse reads were concatenated into the same file. HUMAnN2 was then run to identify the abundances of annotated UniRef50 gene families in each sample. The abundances of gene families were regrouped into other gene family databases as indicated in the text using *humann2_regroup_table*. Pathway abundances and coverages were produced by the default HUMAnN2 mapping files. These steps were parallelized when possible with GNU Parallel (Tange 2011) (v20170722).

To compare how different metagenomics processing pipelines can affect the resulting functional abundance tables, we also ran HUMAnN2 directly against the KEGG v56 database using default options. This pipeline only involved mapping translated reads against this database (i.e. it skipped the nucleotide alignment step) and results in KO abundances directly for each sample. This differs from the other pipeline described above, because in the above case UniRef50 ids were regrouped to KOs based on a mapping file rather than based on direct read mappings. These KO abundances are referred to as the alternative MGS pipeline ("Alt. MGS") in the main-text and figures.

2.2.6 - 16S rRNA Gene-MGS Validation Analyses

The simulation approach to illustrate the issue of high null Spearman correlation coefficients was based on the following procedure. First, for each *N* in the set of integers that span 1 to 100, two subsets of *N* genomes were sampled randomly from the reference database. The abundance of each sampled genome was taken from a negative binomial distribution family implemented in the R function, *rnbinom*, with parameters *size=10* and *prob=0.7*, which aimed to simulate the over-dispersed count data frequently encountered in sequenced data sets. Gene family abundance tables were then computed for each of the two subsets of genomes based on the abundance of each genome and the abundances of gene families within each genome. Spearman correlation coefficients were then computed between the two gene family tables. This procedure was replicated 10 times for each *N*.

This simulation inspired the use of null distributions in our validation analyses. We calculated the correlations between the MGS metagenome or gene table and a synthetic gene table comprised of the mean gene count number across all reference genomes in the database, which is referred to as the "null expectation" through-out the main-text. The null Spearman correlation coefficient distributions of pathway abundances and coverages were similarly based on the reference genome pathways inferred from the EC reference database.

For the purposes of comparing functional prediction tools, gene family tables were filtered to only those gene families present in the databases of all tested functional prediction tools. Gene families absent in all samples were retained as zeros and were not removed. We converted the predictions to binary presence (positive) and absence (negative) format before calculating precision and recall (any abundance greater than zero was considered as evidence for presence). Precision and recall are defined as $TP/(TP + FP)$ and $TP/(TP + FN)$, respectively, where TP=number of true positives (i.e. functions correctly called as present), FP=number of false positives, TN=number of true negatives, and FN=number of false negatives. The F1 score is the harmonic mean of precision and recall: 2 * ((precision * recall) / (precision + recall)).

We also conducted differential abundance tests to compare how results differ between predicted metagenomes and actual MGS data. To conduct these analyses, we subset four of the validation datasets into sample groupings appropriate for pairwise testing. The other datasets were excluded because they resulted in no significant results based on the MGS data after restricting the data to appropriate sample groupings. The comparisons were:

1. Stool samples from 19 *Entamoeba*-positive vs 36 *Entamoeba*-negative Cameroonian individuals
2. 22 supragingival plaque vs 36 tongue dorsum samples from the Human Microbiome Project (i.e. samples from two different oral body sites).
3. Stool samples of 51 individuals from Bhopal, India vs 38 individuals from Kerala, India
4. Stool samples of 29 old world monkeys vs 29 new world monkeys

In the main-text the differential abundance results are focused on Wilcoxon tests on relative abundance values for the KOs (after normalizing function abundances by the median number of universal single-copy genes per sample (Manor & Borenstein 2015)). The pathway differential abundance results are also based on Wilcoxon tests, but on the relative abundance of pathways instead.

As a comparison point we also ran ALDEx2 (Fernandes et al. 2014) (Wilcoxon test with 128 Monte Carlo samples) and DESeq2 (Love et al. 2014) to see how these choices affect the resulting significant functions. DESeq2 was run twice on each dataset: once with default options and separately with options specifically recommended for microbiome data. In the latter case, these options included first calculating the geometric means of the data to estimate size factors and then using the option fitType="local". This second method is referred to as "DESeq2 GMeans" in Additional File 1. We also tested for differential prevalence of functions based on the presence and absence of functions with Fisher's exact tests.

ALDEx2 is a general method for compositional data analysis that uses a Dirichlet-multinomial model to infer sampling and biological variation. The significance reported by this tool is based on Wilcoxon tests after abundances are estimated from the count data. DESeq2 is also a compositional data analysis method for read counts, based on the negative binomial distribution. Lastly, the Fisher's exact tests tested for differential prevalence rather than abundance. In this case, the tests were focused on the counts of how many samples were positive (i.e. had the specific function) compared to those that were negative in each group based on the binary presence/absence of functions.

ALDEx2 and DESeq2 require count tables as input, which required the prediction tables to be rounded. Tax4Fun2 and PanFP were excluded from these analyses, because they do not output predictions in a format that corresponds to the original ASV relative abundances (i.e. the output tables cannot be meaningfully rounded). In all cases significant functions were identified based on Benjamini-Hochberg corrected P-values < 0.05.

2.2.7 - Additional Notes on Statistical Analyses

No tests for statistical power were conducted to determine sample sizes for this study. In the boxplots throughout this paper the centre line corresponds to the median and the lower and upper hinges (i.e. the edges of the "box") represent the 25th and 75th percentiles, respectively. The boxplot whiskers extend to the most extreme values no further than 1.5 multiplied by the inter-quartile range in each direction. The points overlaid on the boxplots correspond to each individual sample unless otherwise stated.

2.3 –Methods for Chapter 4 Supplementary Results

This section describes the methods relevant to the supplementary results published as additional online available with the letter presented in Chapter 4. In addition, the methods for the additional, unpublished analysis presented in Chapter 4 are also described below in Section 2.3.4.

2.3.1 - Inflammatory Bowel Disease Data Processing and Analyses

We ran PICRUSt2 on a dataset of ileal biopsies from an inflammatory bowel disease (IBD) cohort to highlight how metagenome inferences can be generated for datasets where MGS is infeasible. Raw 16S rRNA gene reads and processed human host transcriptome and metabolome tables were downloaded from https://ibdmdb.org. The 16S rRNA gene data was processed using deblur and QIIME 2 as described above for the validation datasets. Only ASVs found in at least two samples and called by at least 10 reads were retained, which resulted in 1,419 final ASVs. The MGS raw reads were processed using the same workflow as the validation datasets described above. PICRUSt2 was run with default options except for the option *--per_sequence_contrib*, which was set to get pathway abundances within each predicted genome for ASV-specific analyses. The unstratified pathway abundances were then calculated by summing over the pathways contributed by each ASV within each sample. Only features present in at least 33% of samples were retained for all analyses. In addition, any pathways described as “superpathways” or “engineered” were excluded from these analyses. ALDEx2 (Fernandes et al. 2014) (v1.12.0) was run with default options to identify features at differential relative abundance between Crohn’s disease and control subjects for taxa and pathways independently.

Partial Spearman correlations between predicted pathway abundances and both the metabolomic and transcriptome data was conducted with the R package ppcor (v1.1). Subject consent age was controlled for when calculating the partial correlations. Before calculating these correlations, pathway abundance data was first transformed by the arcsine square-root transformation, and the metabolomic and transcriptomic datasets were transformed by $\log_{10}$ after adding a pseudocount of one. Metabolites were limited to those with non-empty compound names and the gene expression data was limited to 11 genes known to be biomarkers of CD-associated ileal inflammation (Haberman et al. 2014): DUOXA2, MMP3, AQP9, IL8, DUOX2, APOA1, NAT8, AGXT2, CUBN, FAM151A, and NOD2. Because several of these genes are highly correlated, we removed redundant genes based on hierarchical clustering of the complement of Spearman correlation coefficients between all genes. Six clear clusters of genes were then identified and the following six representative genes for each cluster were retained for further analyses (the other genes in each cluster are indicated in parentheses): DUOX2 (DUOXA2), MMP3, AQP9 (IL8), APOA1, NAT8 (AGXT2, CUBN, FAM151A), and NOD2.

2.3.2 - 18S rRNA Gene and ITS Database Processing

A total of 574 publicly available fungi genomes were downloaded from the 1000 Fungal Genomes Database (http://1000.fungalgenomes.org) on November 16, 2018. The 18S rRNA genes were annotated using barrnap (v0.9-dev; https://github.com/tseemann/barrnap), and 18S rRNA genes were parsed from the genomes using the custom Python script *rRNA_from_gff3.py*. ITS sequences were identified and parsed from all genomes using ITSx (Bengtsson-Palme et al. 2013) (v1.0.11) using the *--only_full T* and *–heuristics* options. Sequence length cut-offs for the 18S rRNA genes and ITS sequences were 605-3,076 bp and 146-2,570 bp, respectively. BUSCO (Simão et al. 2015) (3.0.2) was run to identify incomplete and contaminated genomes with the *fungi_odb9* database. Only the genomes with a completeness of at least 70%, and a duplicated metric (which is based on the copy number of single-copy genes) of at maximum 10% were retained. After restricting genomes to those that passed these quality cut-offs that also had at least one passing amplicon region, there were 229 genomes in the 18S rRNA gene database and 201 genomes in the ITS database.

The 18S rRNA gene and ITS sequences were then dereplicated using the same approach as used for the 16S rRNA genes. The 18S rRNA gene MSA was built using the ssu-align pipeline as for the prokaryotic database (using the *eukarya* alignment model) whereas the ITS MSA was built using MAFFT (Katoh et al. 2002) (v7.407) with the *–genafpair* and *–maxiterate 1000* options. Phylogenetic trees for both MSAs were built with RAxML-ng (v0.8.0) as for the prokaryotic database except a guide tree enforcing a taxonomic topology was also used. EC number copy numbers per genome were downloaded for these genomes also from the 1000 Fungal Genomes Database. Mean EC number abundances were calculated for dereplicated amplicon sequences using the same approach implemented for the prokaryotic databases.

2.3.3 - 18S rRNA Gene and ITS Amplicon Data Processing

The blueberry soil 18S rRNA gene data was processed the same as the blueberry soil 16S rRNA gene data except the output ASVs were restricted to those classified as fungi. This resulted in a total of 1,048 ASVs and a minimum sample depth of 1,981 reads. A total of 3,691 ASVs and a minimum depth of 2,091 ASVs was produced when re-running this pipeline with all blueberry-associated 18S rRNA gene samples (i.e. including blueberry root and soil samples with no matching MGS data). The R package rfPermute (v2.1.6) was run with 501 trees, 1,000 replicates, and the default *mtry* setting to identify significantly different predicted pathways between the three environments based on PICRUSt2 predictions run on this full blueberry-associated dataset. Previously clustered ITS sequences and a processed abundance table were acquired for the wine fermentation dataset (Sternes et al. 2017). These files were used because no raw ITS reads could be located for this dataset. When comparing these amplicon datasets with the corresponding shotgun metagenomics data, the percent of eukaryotic DNA in the MGS data was identified with Metaxa2 (Bengtsson-Palme et al. 2015) (v2.2), which parses rRNA genes from the raw reads.

2.3.4 – Evaluating the Contribution of Individual Updates to PICRUSt2 Performance

The validation 16S rRNA gene datasets were re-run with PICRUSt2 v1.1.0-b with a range of input files and parameter settings. This was done to better compare the relative contributions of the individual updates to the increased performance observed for PICRUSt2. To perform these analyses, the original PICRUSt1 reference files (the abundance tables of KOs and 16S copy numbers across IMG genomes) were formatted for PICRUSt2. In addition, the Greengenes 13_5 phylogenetic tree and multiple-sequence alignment for all corresponding Greengenes OTUs to these IMG genomes were prepared for input. Last, we created trees with FastTree (Price et al. 2010) for all validation datasets that included both query and reference sequences as tips. These trees were computed based on MAFFT (Katoh et al. 2002) alignments of all 16S sequences per dataset. This workflow was performed with QIIME 2 v2020.2 commands.

2.4 - Methods for Chapter 5

2.4.1 - POMS Framework

The standard Phylogenetic Organization of Metagenome Signals (POMS) workflow starts by pre-processing three input files: (1) a table of taxon abundances, (2) a tree of the taxa in this table, and (3) a table of the gene copy numbers in the genome of each of these taxa. Two separate groupings of samples must also be specified. Currently only two-group comparisons can be performed by default. Rare gene families are first excluded, which by default includes functions that occur across fewer than 10 taxa and/or less than 0.1% of taxa. If an unrooted tree is input, it is rooted using midpoint rooting. Last, nodes in the input tree with sufficient underlying taxa are then identified. By default this includes nodes with at least ten underlying taxa on the left and right-hand sides.

Balances at each passing node are then calculated for all samples (see Isometric Log-Ratio section below). By default, the POMS pipeline tests for nodes with significantly different balances between sample groupings based on Wilcoxon tests. This testing is also conservative by default: only nodes with Benjamini-Yekutieli corrected P-values (BY) < 0.05 are identified as significantly different. This multiple correction approach was chosen because it better controls for false positives in the presence of dependencies between variables compared with standard methods (Benjamini & Yekutieli 2001). However, users can specify the significance cut-off and multiple-test correction approach to use for specific applications. The taxonomic breakdown of taxa on each side of a significant node can be parsed out based on the lowest possible taxonomic grouping shared by at least 75% of taxa.

Although this overall testing approach is effective, a major advantage of converting relative abundances to phylogenetic balances is that the resulting balances are orthonormal and can be used with a wide range of statistical approaches. Accordingly, the user can also use an alternative statistical approach to the Wilcoxon test and simply identify the significant nodes for POMS to use instead. This capability is especially useful for cases where controlling for confounding variables is required when testing for differences in balances.

Enriched gene families are then identified within taxonomic lineages on one side of each tested node compared with the other. In the main text we predominately focus on KOs, but POMS is agnostic to the functional ontology used. In addition, this is performed for all tested nodes and not only for those that are significant. This is crucial, because it enables a pseudo-null distribution to be generated based on different subsets of significant nodes (see below). In addition, the number of enrichments for each gene family at non-significant nodes is also provided to the user to provide context on how commonly this function varies across lineages. These enrichment tests are performed with Fisher's exact tests and by default raw P-values < 0.05 are taken as a cut-off for enrichment (although this can be changed by the user).

This functional enrichment information is then combined with the significant nodes to determine the direction of functional enrichment. In other words, POMS will determine which group has relatively higher levels of the taxa (based on the balances at that node) that are enriched for the function. In the main text we refer to this as being positive or negative when the enrichment is in the direction of case and control patients, respectively.

Finally, the key output by POMS is a table that summarizes the number of nodes that are positively or negatively enriched for each gene family from the perspective of the first sample group specified. Optionally additional information can also be provided to the user, such as the mean internode distances of all nodes that are either positively or negatively enriched for each function. This information is provided to enable further exploratory analyses.

Significantly enriched gene families can be identified based on a pseudo-null distribution approach (see below). Whether based on this approach or simply hard cut-offs to identify outliers, enriched pathways can be identified based on a set of gene families of interest with Fisher's exact test. This test compares the numbers (and proportion) of gene families within the set of significant genes to all genes in the background.

2.4.2 - POMS Dependencies

POMS is written in R (R Core Team 2019) and is dependent on the following R packages (versions used in this manuscript are indicated, but these exact versions are not required): ape v5.3 (Paradis et al. 2004), parallel v3.6.0, phangorn v2.5.5 (Schliep 2011), and stringr v1.4.0 (Wickham 2019). R v3.6.0 and RStudio v1.2.5033 were used for testing and developing this tool, which was on a server running Ubuntu v16.04.5.

Several additional R packages are required to follow the current analysis workflow after running POMS (again the versions indicated were used for this paper, but are not required versions): ggtree v1.16.1 (Yu et al. 2017; Yu 2020), ggplot2 v3.3.0 (Wickham 2016), plyr v1.8.4 (Wickham 2011), and reshape2 v1.4.3 (Wickham 2007). All multi-panel plots displayed were created with the cowplot (v1.0.0) R package (Wilke 2019).

2.4.3 - Isometric Log-Ratio

Phylogenetic balances in POMS are calculated based on the isometric log-ratio of taxa on one side of the node compared to the other (Morton et al. 2017; Silverman et al. 2017). The balance for a sample at node i is calculated based on this equation:

$$b_i = \sqrt{\frac{n_{Li} n_{Ri}}{n_{Li} + n_{Ri}}} \log \frac{g(y_{Li})}{g(y_{Ri})}$$

Where n_{Li} and n_{Ri}correspond to the number of taxa on the right and left-hand side of the node. Similarly, $g(y_{Li})$ and $g(y_{Ri})$ correspond to the geometric mean of the relative abundances of taxa on the left and right-hand side of the node. Note that the choice of which lineage is considered the left or right-hand side of a given node is arbitrary. The ratio of geometric means is the key component of this approach that converts microbiome relative abundance data to ratios. The fraction including the numbers of taxa on each of a node is included simply to scale the balance to give it unit length (i.e. to make the balances comparable despite varying numbers of taxa at each node).

The geometric mean of the relative abundance of a set of taxa (e.g. on the left-hand side) is calculated based on the below equation:

$$g(y_{Li}) = \left(\prod_{j=1}^{n_{Li}} y_j\right)^{\frac{1}{n_{Li}}} = \sqrt[\frac{1}{n_{Li}}]{y_1 y_2 \dots y_{n_{Li}}}$$

2.4.4 - Pseudo-Null Distribution

To identify significantly enriched nodes, we employed a pseudo-null distribution approach. This method is based on re-sampling random nodes as the "significant" set from the set of all tested nodes. The POMS pipeline then proceeds with the directionality given for sampled nodes based on whichever sample grouping has higher mean balances. This approach yields the number of nodes that are positively and negatively enriched for each function for each of 1000 replicates. A P-value can then be calculated for each gene family based on the proportion of the permutation replicates with absolute enrichment values (the absolute difference between the number of positively and negatively enriched nodes) greater or equal to the observed value. Based on this approach we identified gene families as significant with a BY < 0.05 in the main text.

Importantly, this approach can result in different sets of significant genes compared with simply setting a hard cut-off for a given absolute enrichment value. This is because the pseudo-null distribution is different for each gene family as it depends on the dispersion and variation in encoding of each function. In other words, a gene family might be significant based on the pseudo-null distribution based on an intermediate absolute enrichment that would otherwise be difficult to notice by eye.

We refer to this approach as generating a "pseudo-null" distribution, because we think it would be misleading to suggest that it represents a true null distribution; we do not believe it is guaranteed to represent what the pattern of gene enrichments would look like in the absence of signal between sample groupings. As discussed in the main text, there are many factors that could result in co-occurrences between significant nodes and increased likelihoods of certain nodes being significant over others. Nonetheless, this approach is a convenient method for identifying major outliers in the POMS output.

2.4.5 - Metagenome-Assembled Genome-Based Simulations

The MAG-based simulations were based on 704 control samples from a large human meta-analysis dataset (Almeida et al. 2019). These simulations proceeded as described in the Results section. First, the samples were randomly split into two groupings for each of the 500 replicates. Then, for each replicate a random focal gene family was randomly chosen and then within one group all MAGs encoding this gene family underwent a ten-fold increase in relative abundance. These simulated profiles were then input to POMS and the results are referred to as the "focal gene" profiles in the main text. We also performed parallel simulations where the relative abundance of random taxa were randomly inflated by ten-fold in one group only. Importantly, the same number of taxa were perturbed as were affected in each matching focal gene simulation replicate. The resulting output based on these simulated profiles are referred to as the "random taxa" profiles. Significant gene families were identified conservatively based on a cut-off of BY < 0.0001 with Wilcoxon tests. Default input parameters were used with POMS. In addition, all comparisons of summary distributions between the two tests were themselves compared with Wilcoxon tests. The summary metrics for these tests are reported in the main text.

2.4.6 - Reference Genome-Based Simulations

The reference genome-based simulations were based on genomes from the Integrated Microbial Genomes database (Markowitz et al. 2012) that were previously parsed for use with PICRUSt2 (Douglas et al. 2020). Per-genome KO annotations were taken from the default PICRUSt2 database. We created a de novo phylogenetic based on a set of universal single-copy genes (USCGs) with GToTree v1.4.16 (Lee & Ponty 2019). This approach parses out USCGs from genome sequences and wraps several tools to build a phylogenetic tree. The tool was run with the bacterial hidden-Markov model setting and with FastTree v2.1.10 (Price et al. 2010). GToTree also returns estimates of the percent completeness and redundancy for each genome. We excluded all genomes with completeness below 95% and/or redundancy above 5%. We then randomly sampled 3000 of the remaining high-quality genomes for the subsequent analyses.

We next simulated random abundances of these genomes across 1000 samples based on the zero-inflated beta-distribution implemented in the rBEZI function of the gamlss.dist v5.1.7 R package (Stasinopoulos & Rigby 2020). Simulations under this model can be modified with three key metrics: mu (the mean), nu (the probability of zero abundance), and sigma (the standard deviation). The selection of these parameters has drastic effects on the resulting abundance profiles. Accordingly, we chose parameter values for mu and nu to represent a range of genome abundances across samples. The sigma parameter was set to 1 in all cases. Specifically, we generated abundance tables with four parameter settings: mu=0.1 and nu=0.5 (Setting 1); mu=0.1 and nu=0.9 (Setting 2); mu=0.1 and nu=0.99 (Setting 3); and mu=0.01 and nu=0.99 (Setting 4). These tables are referred to based on each respective setting number in the Results section.

2.4.7 – Case-Control Shotgun Metagenomics Dataset Validations

We focused our validation analyses on three datasets that were part of the large meta-analysis of human shotgun metagenomics datasets (Almeida et al. 2019). These datasets are defined based on dataset accession identifiers in the European Nucleotide Archive. The largest dataset focused on obese and control individuals (which we refer to as the primary obesity dataset) that corresponded to data accession ERP002061. The secondary obesity dataset corresponds to accession ERP003612 and the colorectal cancer data is under accession ERP012177. We used the previously generated MAGs, sample MAG abundance profiles, and MAG phylogenetic tree as input to POMS after performing pre-processing. Importantly, we excluded MAGs from each sample with mapped read coverage lower than 25%.

The alternative differential abundance tools compared in this study were: Wilcoxon tests based on relative abundances, ALDEx2 v1.16.0 (Fernandes et al. 2014), and Limma-Voom v3.40.6 (Law et al. 2014). Venn diagrams comparing the number of overlapping significant gene families identified by these approaches were created with ggVennDiagram v0.3 (Gao 2019).

2.4.8 - Code Availability

The code for all analyses presented in this manuscript is available at: https://github.com/gavinmdouglas/POMS_manuscript/. The source code for POMS is available at: https://github.com/gavinmdouglas/POMS

Chapter 3 - Multi-omics Differentially Classify Disease State and Treatment Outcome in Pediatric Crohn's Disease

This chapter is a close reproduction of the paper of the same name published in the journal Microbiome (Douglas et al. 2018). I was co-first author on this work with Dr. Richard Hansen, a clinician in the department of Paediatric Gastroenterology at the Royal Hospital for Children in Glasgow, United Kingdom. Dr. Hansen designed the project, recruited all participants, collected raw data from patients, contributed to reviewing treatment response, and performed all DNA extraction. I conducted all analyses and wrote the paper.

The additional authors on this paper were: Casey M. A. Jones, Katherine A. Dunn, André M. Comeau, Joseph P. Bielawski, Rachel Tayler, Emad M. El-Omar, Richard K. Russell, Georgina L. Hold, Morgan G. I. Langille, and Johan Van Limbergen.

This paper was published under a Creative Commons CC BY license, which allows unrestricted use with attribution (see Appendix – Copyright Permissions). All additional files referred to in this chapter are freely available as part of the publication on the Microbiome journal website. The original author funding statements, author contributions breakdown, and acknowledgements are also available in the original publication.

3.1 – Abstract

Crohn's disease (CD) has an unclear etiology, but there is growing evidence of a direct link with a dysbiotic microbiome. Many gut microbes have previously been associated with CD, but these have mainly been confounded with patients' ongoing treatments. Additionally, most analyses of CD patients' microbiomes have focused on microbes in stool samples, which yield different insights than profiling biopsy samples.

We sequenced the 16S rRNA gene (16S) and carried out shotgun metagenomics (MGS) from the intestinal biopsies of 20 treatment-naïve CD and 20 control pediatric patients. We identified the abundances of microbial taxa and inferred functional categories within each dataset. We also identified known human genetic variants from the MGS data. We then used a machine learning approach to determine the classification

accuracy when these datasets, collapsed to different hierarchical groupings, were used independently to classify patients by disease state and by CD patients' response to treatment. We found that 16S-identified microbes could classify patients with higher accuracy in both cases. Based on follow-ups with these patients, we identified which microbes and functions were best for predicting disease state and response to treatment, including several previously identified markers. By combining the top features from all significant models into a single model, we could compare the relative importance of these predictive features. We found that 16S-identified microbes are the best predictors of CD state whereas MGS-identified markers perform best for classifying treatment response.

We demonstrate for the first time that useful predictors of CD treatment response can be produced from shotgun MGS sequencing of biopsy samples despite the complications related to large proportions of host DNA. The top predictive features that we identified in this study could be useful for building an improved classifier for CD and treatment response based on sufferers' microbiome in the future. The BISCUIT project is funded by a Clinical Academic Fellowship from the Chief Scientist Office (Scotland)- CAF/08/01.

3.2 - Background

Crohn’s disease (CD) is an inflammatory bowel disease (IBD) classically characterized by abdominal pain, rectal bleeding and weight loss. Recurring flares of IBD cause lifelong, far-reaching consequences for patients that can affect lifestyle and overall health (Cleynen et al. 2016; Neovius et al. 2013). CD differs from the other form of IBD - ulcerative colitis - in that CD can affect any part of the gastrointestinal tract, can be discontinuous, and can involve granulomatous inflammation (Cho 2008). There is a growing need to understand the etiology of CD due to the worldwide increase in annual incidence (Molodecky et al. 2012), particularly in children (Henderson et al. 2012).

Although the etiology of CD is unclear (Ananthakrishnan 2015), there is growing evidence for the dysbiosis hypothesis. This model postulates that a shift in the balance between commensal and pathogenic intestinal microbes interacting with the host’s immune system contributes to CD onset. In support of this model, large-scale differences in bacterial abundances have long been associated with CD (Seksik et al. 2003). The most reproducible finding has been a decrease in alpha diversity in CD patients compared to controls (Gevers et al. 2014; Hansen et al. 2012; Pascal et al. 2017). Several particular changes in taxonomic abundances have been linked to this dysbiotic state, for instance Firmicutes tend to be at lower proportion and Gammaproteobacteria at higher proportion in CD patients (Sokol & Seksik 2010). Most taxonomic profiles of CD patients have been based on stool samples, which yield drastically different insights into CD pathogenesis when compared with mucosal washing of the mucosal-luminal interface (MLI) and intestinal biopsy samples (Gevers et al. 2014). Irrespective of body site, it is unclear whether these shifts in microbiota are a cause or a symptom of the disease. However, there is reason to believe that the microbiome contributes to CD etiology due to several observations. Firstly, children that are exposed to antibiotics in the first year of life are more likely to develop IBD (Shaw et al. 2010), which could be related to acquiring a dysbiotic state. Also, many CD risk loci are linked to pattern recognition receptors (PRRs) and cytokines that regulate the host immune system (McGovern et al. 2015).

PRRs generate responses against pathogenic bacteria while identifying commensal bacteria within the human microbiome. The best-known example of a PRR linked to CD is the nucleotide-binding oligomerization domain containing 2 (*NOD2*) gene that codes for an intracellular PRR. Loss of function mutations in the gene lead to increased inflammation due to impaired clearance of intestinal bacteria that are harmful to the gut (De Souza & Fiocchi 2016). Despite these reproducible links to CD, risk mutations account for <14% of disease variance across patients (Jostins et al. 2012). However, the concordance rate of CD between monozygotic twins ranges from 20-50%, which is higher than several other complex diseases (Halme et al. 2006). Nonetheless, risk loci alone do not explain CD onset and the relative importance of the microbiome in the onset of this disease is not well understood.

Here, we compare the relative importance of genetic risk loci and microbiota identified from intestinal biopsy samples for classifying treatment-naïve pediatric patients by disease state. We also demonstrate that CD patients' treatment response status can be classified by microbial features with high accuracy. Taxonomic and functional profiles discussed in this study are based on both 16S sequencing and MGS sequencing of the same intestinal biopsy samples. To our knowledge, this is the first report of shotgun MGS of CD intestinal biopsy samples.

3.3 - Results

3.3.1 - Identifying Crohn's Disease Related SNPs, Microbial Taxa, and Functions from Intestinal Biopsy Samples

To investigate which microbial and genetic features best classify pediatric CD patients by disease state and treatment response, we sequenced the intestinal microbiomes of 20 CD and 20 normal colon controls prior to any treatments. Both MGS and 16S sequencing were performed on the same biopsy samples. Much of the MGS data was comprised of human DNA (90%), which was separated from the microbial DNA and used to call human genotypes. We combined the human genotypes at 133 known CD risk loci with known odds-ratios and allele frequencies to calculate a genetic risk score (Jostins et al. 2013) (GRS) per sample. We then used the remaining microbial MGS reads, a mean of 10.7 million paired-end (PE) reads per sample, to call 115 independent taxa (summarized at the class level in Figure 3.1).

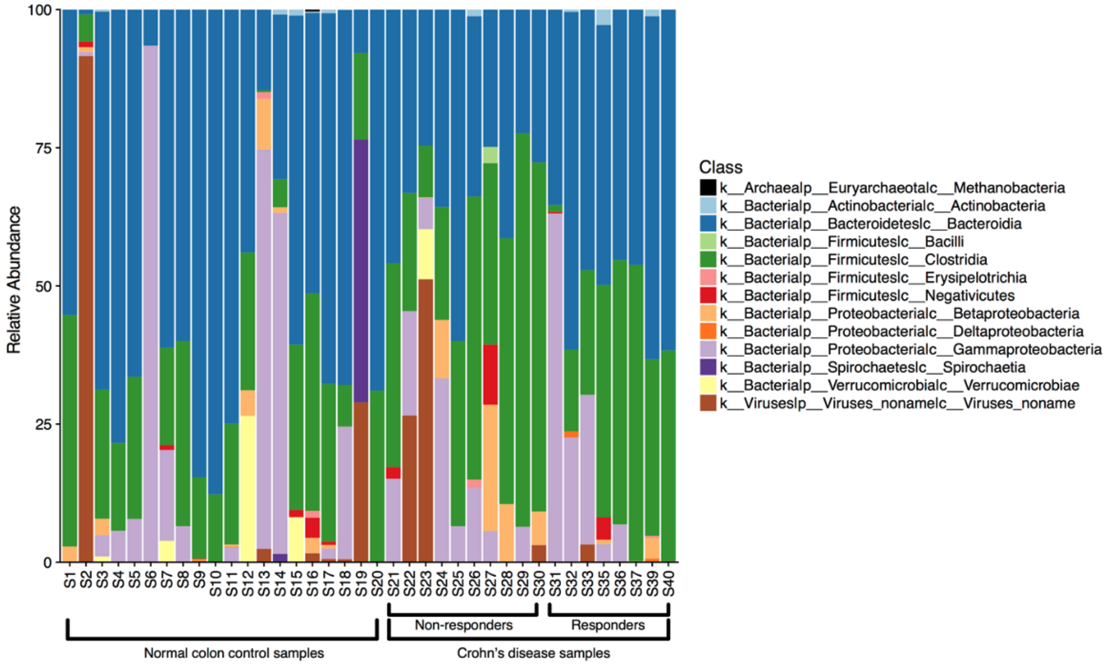


***Figure 3.1: Stacked bar-chart showing percentages of classes across metagenomic samples**. Note the presence of archaea and viruses, which are absent in the 16S data.*

Also, note the high prevalence of viral DNA in several samples. The metadata groupings of these samples are indicated at the bottom.

All microbial MGS reads were also used to identify the relative abundances of Kyoto Encyclopedia of Genes and Genomes (Kanehisa & Goto 2000) (KEGG) orthologs, pathways, and modules within each sample. Similarly, after filtering the 16S amplicon reads, we retained an average of 13,815 stitched reads per sample. We performed open-reference clustering to call 984 operational taxonomic units (OTUs; summarized at the class level in Figure 3.2).

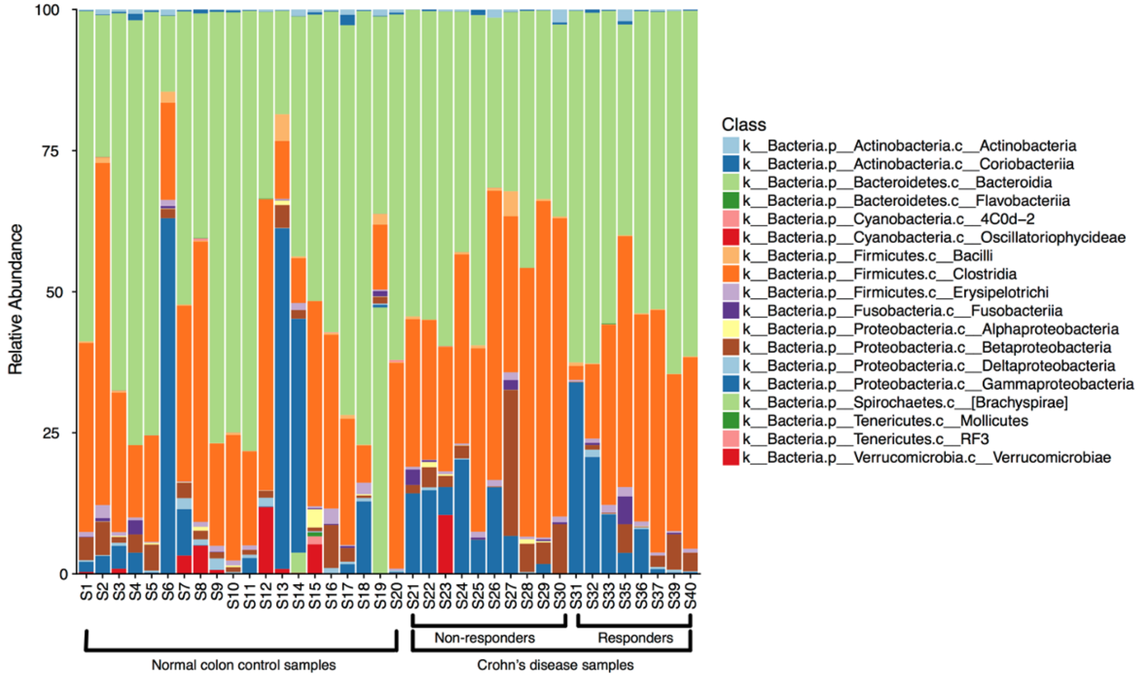


Figure 3.2: Stacked bar-chart showing percentages of classes across 16S rRNA gene sequencing samples*. Colours were chosen to help with discerning different taxa; however, several taxa have the same the colour so the taxa ordering should be considered when interpreting this figure. The metadata groupings of these samples are indicated at the bottom.*

Overall the relative abundances of MGS and 16S-identified genera were similar within the same biopsy samples (mean Spearman's ρ=0.51, standard deviation=0.18). Since sequencing read counts are a form of compositional data, we tested whether a centered log-ratio transform of the non-rarified read counts (Gloor et al. 2016) would result in improved model performance compared to rarefaction of all samples. Although the compositional-based methods performed slightly better for some feature tables, in the majority of cases this transformation resulted in less accurate classification of patients (Figure 3.3) and so we focused on the rarified datasets for our analyses. We used these OTUs to infer the relative abundances of KEGG orthologs and pathways within each sample (see Additional File 2 for sample sequencing coverage and metadata). Two of the CD patients' microbial profiles were discarded due to low 16S and MGS sequencing depth. These different datasets are outlined in Figure 3.4 (see Table 3.1 for sample details).

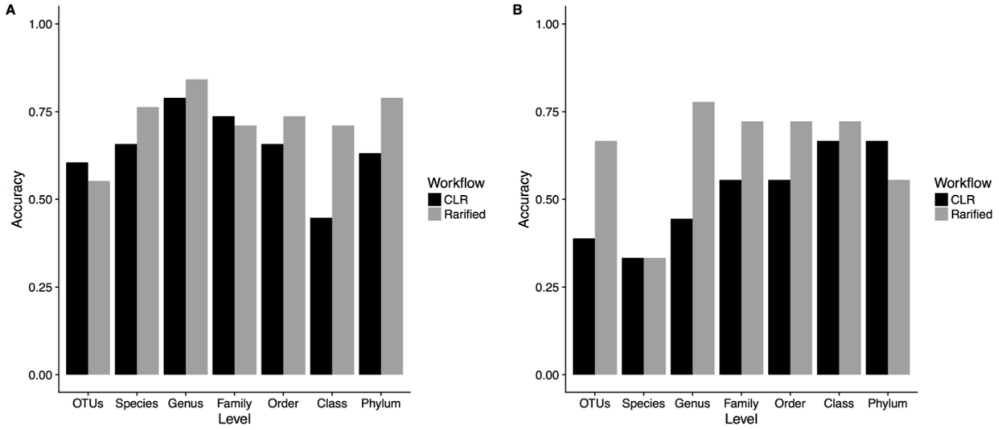


Figure 3.3: Comparison of taxonomic dataset accuracies either transformed by centered log-ratio or rarified*. The accuracies of random forest models trained on taxonomic datasets for each taxonomic level to classify patients by (A) disease and (B) treatment response are shown.*

We then replicated two well-known predictors of CD: increased GRS (Cleynen et al. 2016) and a reduction in microbial alpha diversity as proof of principle. We chose the simplest measure of alpha diversity: the observed number of OTUs per sample (# OTUs). Both GRS (one-tailed Mann-Whitney-Wilcoxon [M-W-W] test, W=288, P=0.00837) and # OTUs (one-tailed M-W-W test, W=261.5, P=0.00894) significantly differed between patients based on disease state in the expected directions (Additional file 1; Figure 3.5).

To make these known predictors comparable to classification accuracies using datasets containing multiple features, we used an analogous method to calculate accuracy. Importantly, these metrics produced only marginal accuracies when used to classify patients by disease state (GRS: 62.5%; # OTUs: 71.1%).

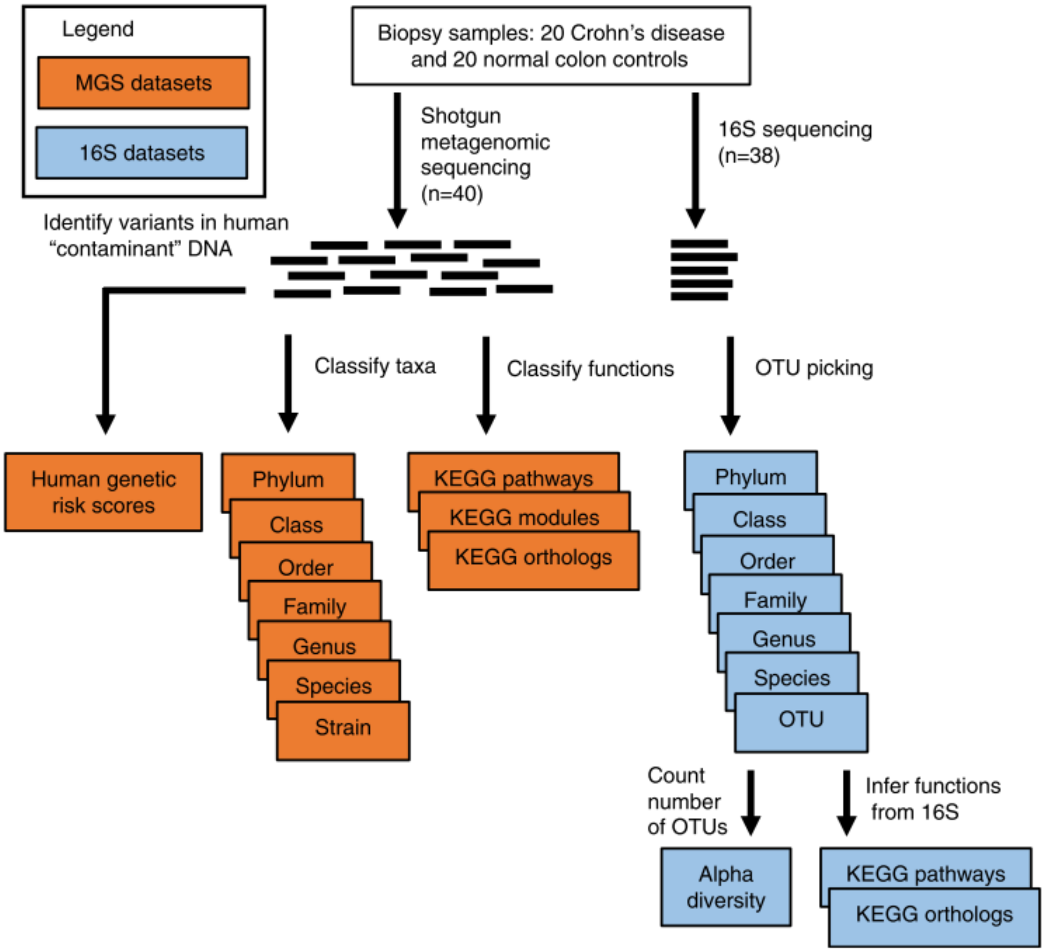
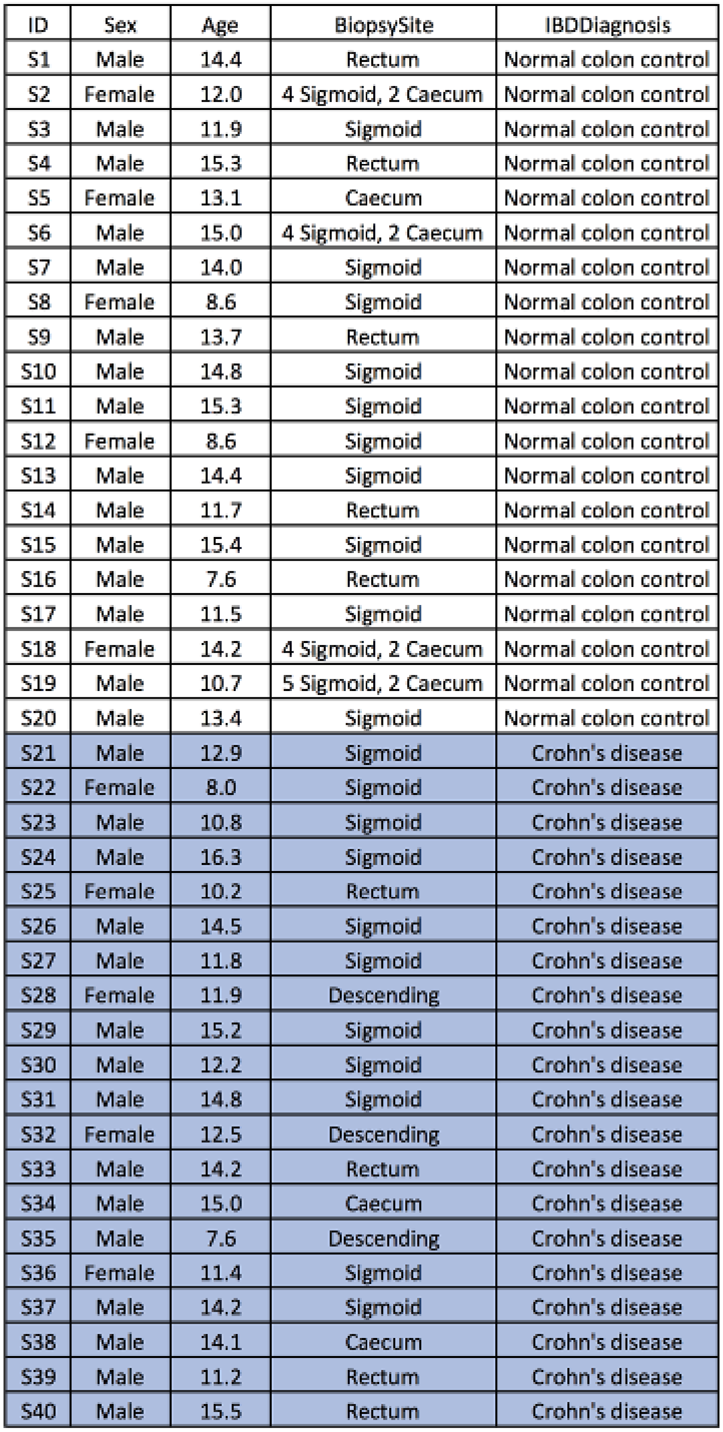


Figure 3.4: Diagram of the different datasets used for classification in this study. *Datasets in orange were derived from the shotgun metagenomic sequencing (MGS) data (n=40) and the datasets in blue were derived from the 16S rRNA gene (16S) sequencing data (n=38*). These datasets were used to classify both disease state and treatment response as input to random forest machine learning models. *Note two Crohn's disease samples were removed from both the 16S sequencing and MGS datasets due to low sequencing coverage, but their genetic profile was inferred from the MGS.*

Table 3.1: Demographic and phenotypic characteristics of pediatric patients. with Crohn's disease (highlighted in light blue) and normal colon controls from the BISCUIT study.



ID	Sex	Age	BiopsySite	IBDDiagnosis
S1	Male	14.4	Rectum	Normal colon control
S2	Female	12.0	4 Sigmoid, 2 Caecum	Normal colon control
S3	Male	11.9	Sigmoid	Normal colon control
S4	Male	15.3	Rectum	Normal colon control
S5	Female	13.1	Caecum	Normal colon control
S6	Male	15.0	4 Sigmoid, 2 Caecum	Normal colon control
S7	Male	14.0	Sigmoid	Normal colon control
S8	Female	8.6	Sigmoid	Normal colon control
S9	Male	13.7	Rectum	Normal colon control
S10	Male	14.8	Sigmoid	Normal colon control
S11	Male	15.3	Sigmoid	Normal colon control
S12	Female	8.6	Sigmoid	Normal colon control
S13	Male	14.4	Sigmoid	Normal colon control
S14	Male	11.7	Rectum	Normal colon control
S15	Male	15.4	Sigmoid	Normal colon control
S16	Male	7.6	Rectum	Normal colon control
S17	Male	11.5	Sigmoid	Normal colon control
S18	Female	14.2	4 Sigmoid, 2 Caecum	Normal colon control
S19	Male	10.7	5 Sigmoid, 2 Caecum	Normal colon control
S20	Male	13.4	Sigmoid	Normal colon control
S21	Male	12.9	Sigmoid	Crohn's disease
S22	Female	8.0	Sigmoid	Crohn's disease
S23	Male	10.8	Sigmoid	Crohn's disease
S24	Male	16.3	Sigmoid	Crohn's disease
S25	Female	10.2	Rectum	Crohn's disease
S26	Male	14.5	Sigmoid	Crohn's disease
S27	Male	11.8	Sigmoid	Crohn's disease
S28	Female	11.9	Descending	Crohn's disease
S29	Male	15.2	Sigmoid	Crohn's disease
S30	Male	12.2	Sigmoid	Crohn's disease
S31	Male	14.8	Sigmoid	Crohn's disease
S32	Female	12.5	Descending	Crohn's disease
S33	Male	14.2	Rectum	Crohn's disease
S34	Male	15.0	Caecum	Crohn's disease
S35	Male	7.6	Descending	Crohn's disease
S36	Female	11.4	Sigmoid	Crohn's disease
S37	Male	14.2	Sigmoid	Crohn's disease
S38	Male	14.1	Caecum	Crohn's disease
S39	Male	11.2	Rectum	Crohn's disease
S40	Male	15.5	Rectum	Crohn's disease

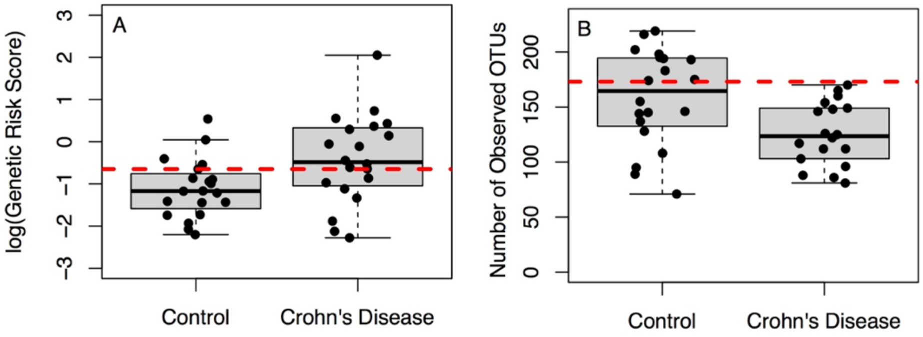


***Figure 3.5: Boxplots of (A) genetic risk scores on a natural log scale and (B) the number of observed OTUs**, which is a measure of alpha diversity. Samples are shown as black points. Red dotted lines correspond to the best cut-offs to distinguish the classes. There are 20 control and 18 Crohn's disease samples shown in each panel. The Mann-Whitney-Wilcoxon test were used to compare these groups since it is a non-parametric test whose main assumption is only that the data-points be independently distributed.*

3.3.2 - Classifying Samples by Disease State

We next investigated how well microbial datasets classify CD disease state. MGS and 16S taxonomic datasets included strain and OTU-level relative abundances respectively and were also collapsed at each level from species to phylum (Figure 3.4). Functional datasets included KEGG ortholog and pathway counts for both sequencing technologies, as well as KEGG modules for MGS samples. In total, 19 datasets were entered as classifiers for disease state after standardization (each mean-centered and scaled by the standard deviation for each sample). We ran independent random forest (RF) models to determine each dataset's classification accuracy (Figure 3.6A; see Additional File 3). Each of the 16S taxonomic datasets, except for the OTU level, could classify patients by disease state with high accuracy (maximum accuracy of 84.2% and P < 0.001 based on genus level). The MGS strain, genus, family, and phylum taxonomic datasets also classified patients, but with lower accuracy than the 16S datasets (maximum accuracy of 68.4% and P=0.016 based on strain level). The predicted KO abundances based on the

16S data and the MGS-identified KEGG modules both significantly classified patients as well (accuracies of 68.4% and 65.8% respectively).

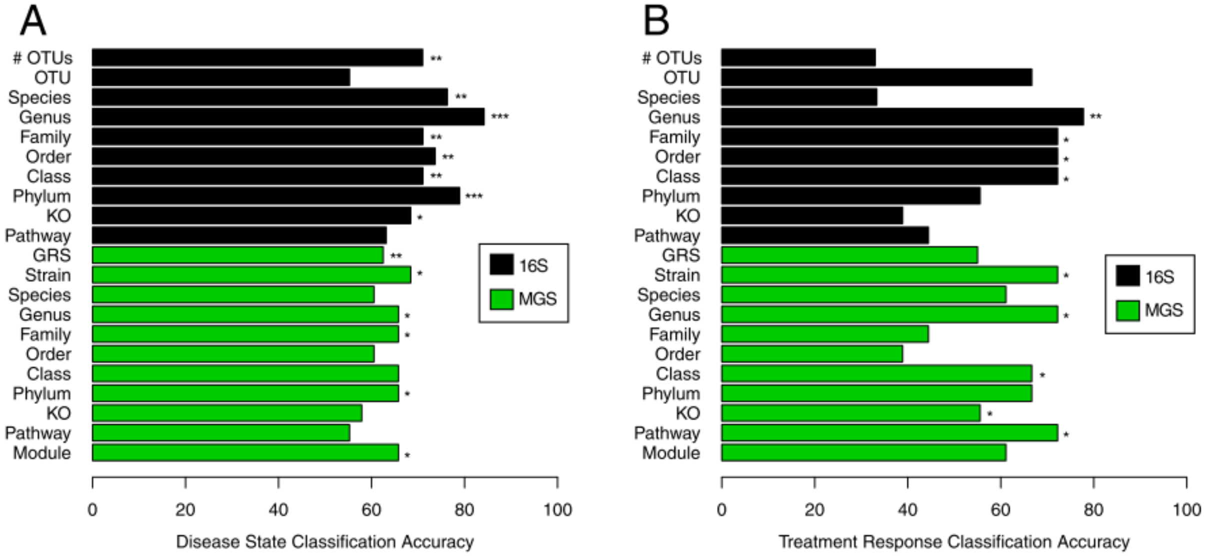


Figure 3.6: Classification accuracies for all datasets classifying (A) disease state and (B) treatment response*. Each bar corresponds to a different model. Accuracies are based on random forest (RF) leave-one-out cross-validation (LOOCV) in all cases, except for number of observed OTUs (# OTUs) and genetic risk scores (GRS) which are based on LOOCV of simple linear cut-off models. The symbols *, **, and *** indicate significance at P < 0.05, P < 0.01, and P < 0.001, respectively. RF model significances were based on a permutation test. P-values for # OTUs and GRS are based on one-tailed Mann-Whitney-Wilcoxon Tests.*

One advantage of RF models is that they output variable importance metrics for each feature used in a model. We considered each RF model to be an omnibus test for each dataset, which enabled us to look at the ranking of variable importance in significant models to identify important features (see Additional File 4). Based on these metrics, the three most informative 16S genera were *Desulfovibrio, Akkermansia*, and *Butyricimonas* (Figure 3.7), whereas the top MGS genera were *Alistipes, Oscillibacter*, and *Dorea*. These top genera could differ since both top 16S genera were close to the detection limit threshold of the MGS data: they were only identified in a small number of samples (Figure 3.8). Nonetheless, *Akkermansia* was ranked 4th in the MGS genus model despite being missed in several samples. The top features in the MGS strain model were strains of *Alistipes putredinis*, *Clostridium symbiosum,* and *Faecalibacterium prausnitzii*. The 16S-inferred KOs and the MGS modules were the only functional datasets that significantly classified samples by disease state (Accuracy=68.4%, P=0.043 and Accuracy=65.8%, P=0.03, respectively). The three top 16S KOs were (1) K03785, which is involved in amino acid biosynthesis, (2) K09013, an Fe-S cluster assembly ATP-binding protein, and (3) K03809, a tryptophan repressor binding protein. The three top MGS modules were (1) M00144, NADH: quinone oxidoreductase, (2) M00362, nucleotide sugar biosynthesis, and (3) M00239, peptides/nickel transport system. Importantly, the datasets collapsed to different taxonomic and functional levels were not independent from each other, which is reflected by the fact that the top features in each taxonomic dataset tended to be part of the same lineage (e.g. the ranks above *Desulfovibrio* and *Akkermansia* were also top hits).

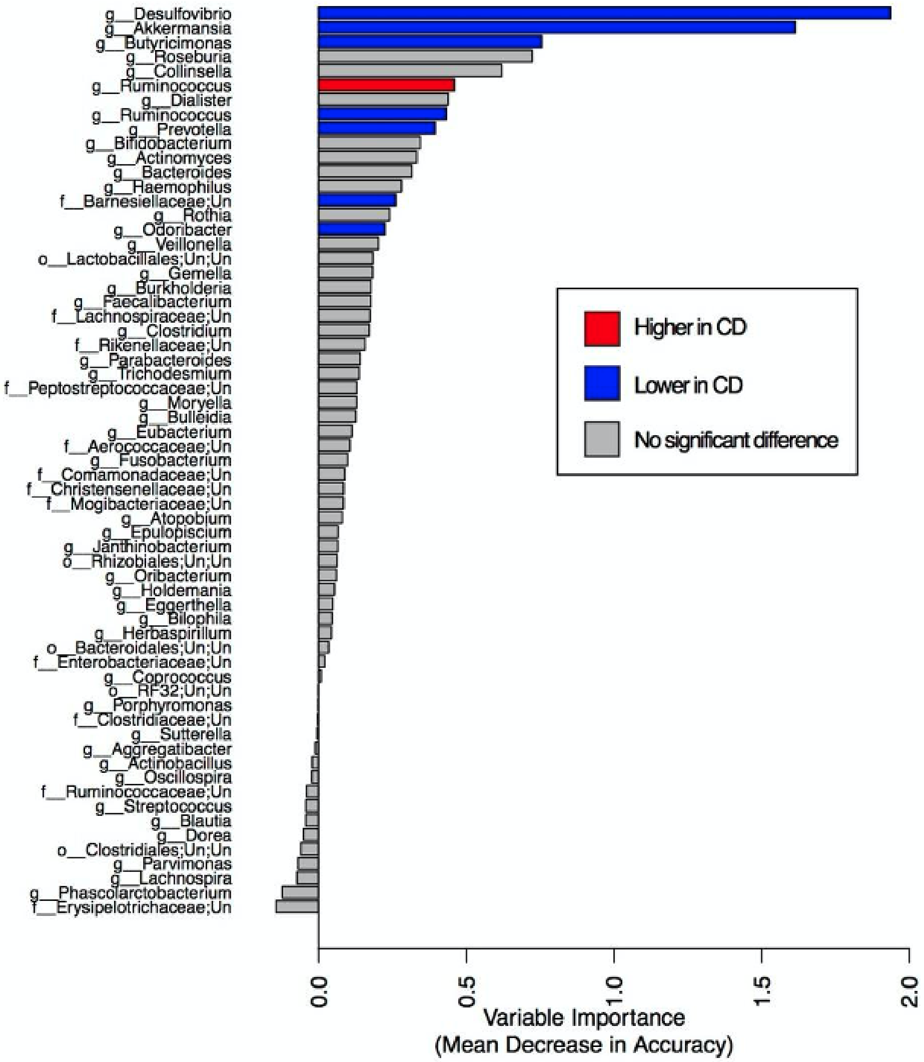


Figure 3.7: Genera identified through 16S rRNA gene sequencing ranked by their importance for classifying disease state*. Features that significantly differed (raw P < 0.05) between Crohn's disease (CD) and healthy colon control patients based on a two-tailed Mann-Whitney-Wilcoxon are indicated in red (if more abundant in CD patients) or blue (if lower in CD patients). Features that did not differ between the two classes are shown in grey.*

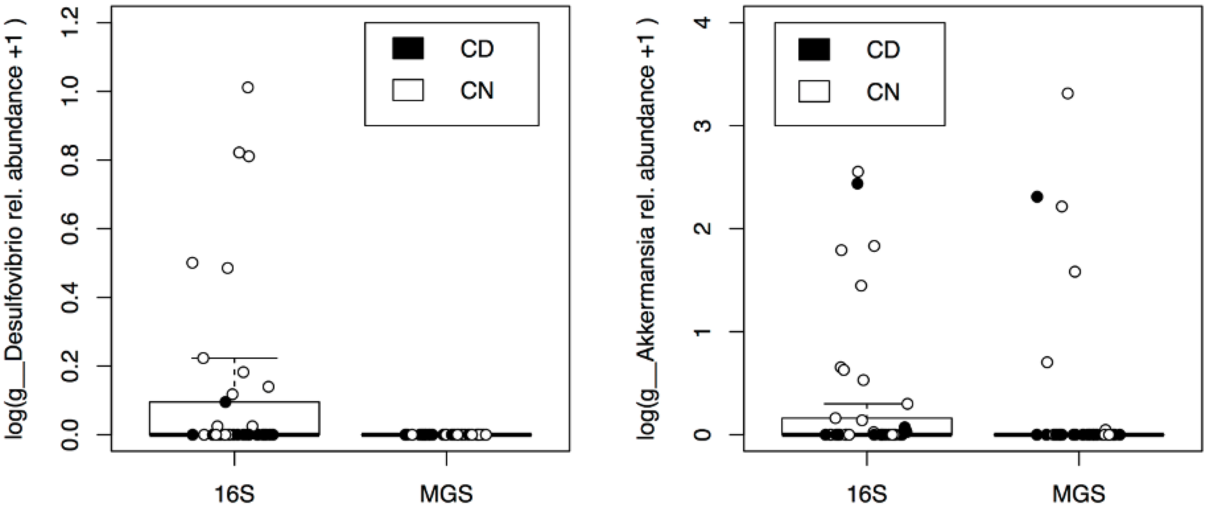


Figure 3.8 Boxplots of the natural log relative abundance of the genera (A) **Desulfovibrio** ***and (B)*** **Akkermansia** ***for both sequencing technologies.*** *These two genera had the highest variable importance in the most accurate disease classification random forest based on 16S rRNA gene sequencing data. Crohn's disease (CD) patients are indicated by black points and healthy colon controls (CN) are indicated as white points. A pseudocount of 1 was added to each sample's relative abundance since the log of 0 cannot be taken. Note that* Desulfovibrio *was absent in all metagenomic samples.*

3.3.3 - Classifying Samples by Treatment Response

Next, we used these same 19 microbial datasets, after excluding normal colon control patients, to classify the CD patients as responders (RS) and non-responders (NR) to induction of remission treatments, started at the time of diagnosis (Figure 3.6B; see Additional File 3). Clinical CD phenotypes were heterogeneous, but all included active colonic disease at the sampled location. Treatments were similarly not consistent across all patients, reflecting heterogeneity of phenotype, but instead were different combinations of exclusive enteral nutrition (EEN) therapy and immunosuppressive medications, as such representing 'real-world' CD treatment: 11 patients were on EEN, 3 were on Prednisolone and EEN therapy, 4 were on Mesalazine alone, and 2 were on Prednisolone alone, as decided by their gastroenterologist at the time of diagnosis. Sustained response or non-response was defined as need for a second induction within 150 days of diagnosis or not (Table 3.2). After classifying CD patients based on their response to induction treatment, 16S genera were again the top dataset (accuracy=77.8%; P=0.008). However, the MGS strain (P=0.029), genus (P=0.013), and KEGG pathway (P=0.018) datasets could also classify patients with only slightly lower accuracy (Accuracy=72.2% for all three). We also found that alpha diversity and GRS did not significantly differ between RS and NR patients (Figure 3.9).

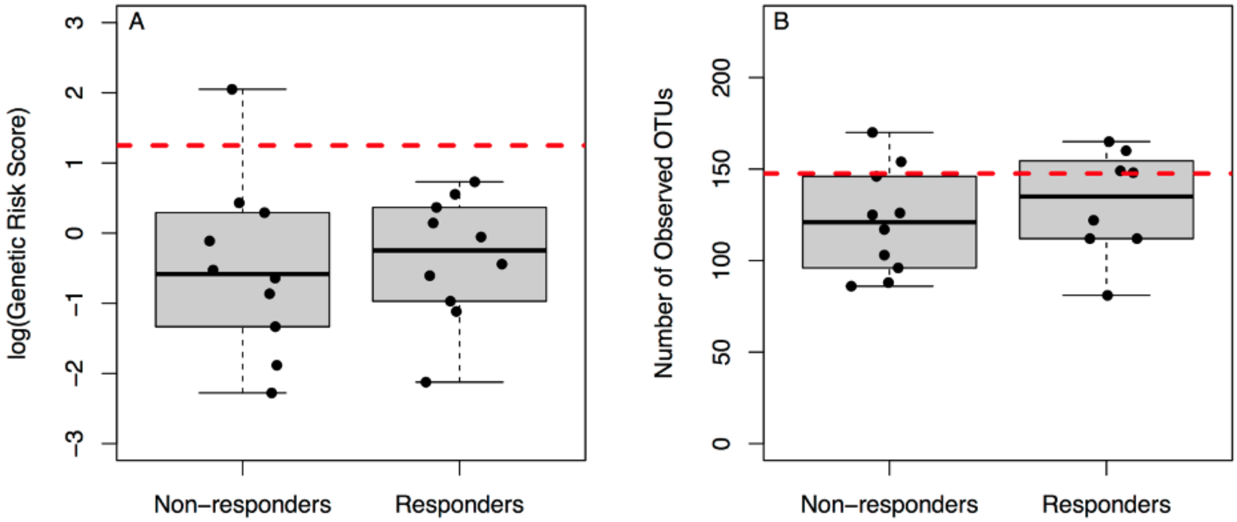
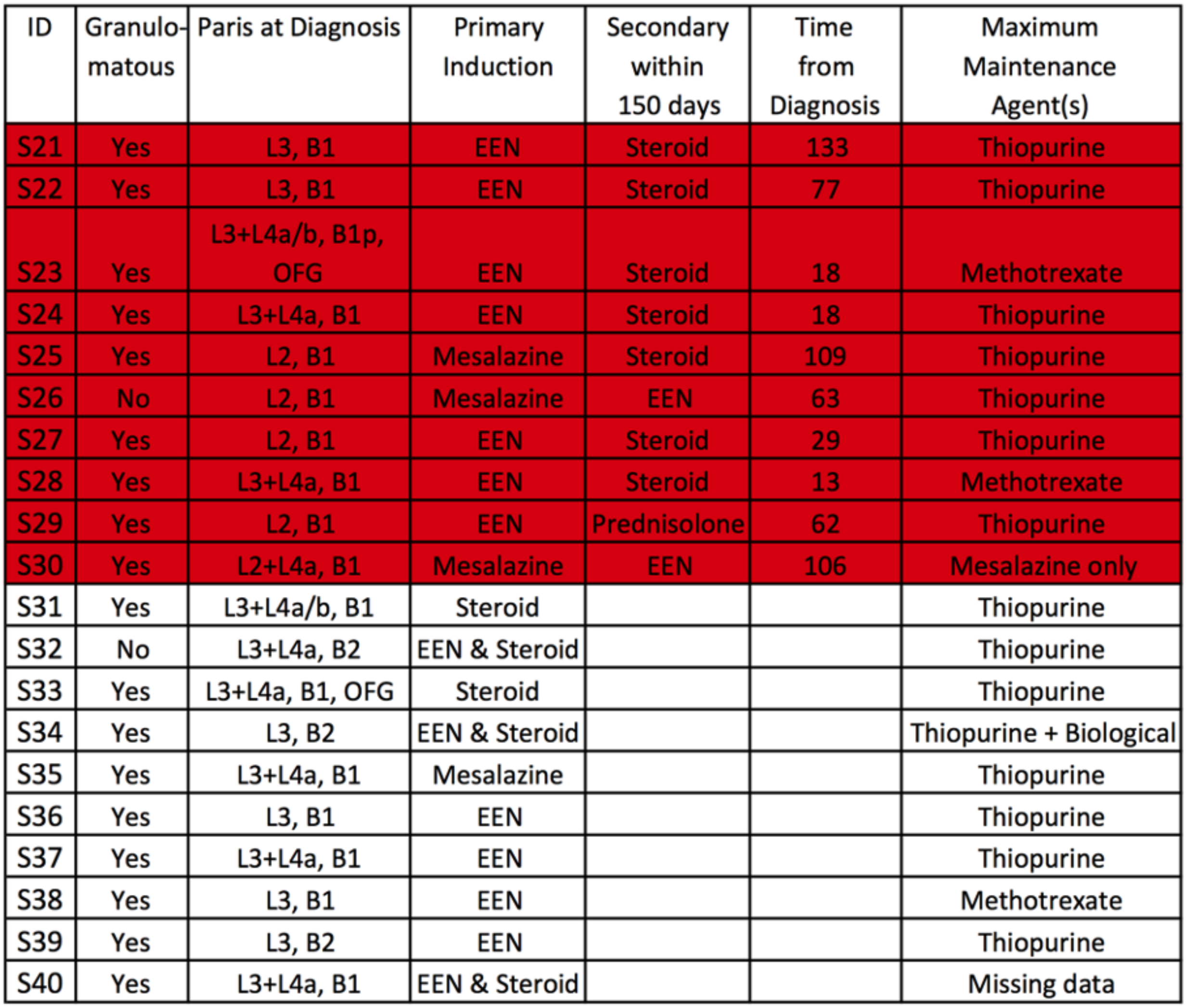


Figure 3.9: Boxplots of (A) natural log genetic risk scores (GRS) and (B) number of observed OTUs (# OTUs) based on Crohn's disease patients' response to treatment. *Both metrics did not significantly differ between non-responders and responders based on one-tailed Mann-Whitney-Wilcoxon tests (GRS: W=42, P=0.736; # OTUs: W=47, P=0.282). One-tailed tests were conducted based on the hypothesis that responders would have lower GRS and increased # OTUs.*

Table 3.2: Phenotypic characteristics and treatments of children with Crohn's disease from the BISCUIT study. *Non-responders to treatment are the red rows while responders to treatment are the white rows. The EEN treatment was Modulen.*



ID	Granulo-matous	Paris at Diagnosis	Primary Induction	Secondary within 150 days	Time from Diagnosis	Maximum Maintenance Agent(s)
S21	Yes	L3, B1	EEN	Steroid	133	Thiopurine
S22	Yes	L3, B1	EEN	Steroid	77	Thiopurine
S23	Yes	L3+L4a/b, B1p, OFG	EEN	Steroid	18	Methotrexate
S24	Yes	L3+L4a, B1	EEN	Steroid	18	Thiopurine
S25	Yes	L2, B1	Mesalazine	Steroid	109	Thiopurine
S26	No	L2, B1	Mesalazine	EEN	63	Thiopurine
S27	Yes	L2, B1	EEN	Steroid	29	Thiopurine
S28	Yes	L3+L4a, B1	EEN	Steroid	13	Methotrexate
S29	Yes	L2, B1	EEN	Prednisolone	62	Thiopurine
S30	Yes	L2+L4a, B1	Mesalazine	EEN	106	Mesalazine only
S31	Yes	L3+L4a/b, B1	Steroid			Thiopurine
S32	No	L3+L4a, B2	EEN & Steroid			Thiopurine
S33	Yes	L3+L4a, B1, OFG	Steroid			Thiopurine
S34	Yes	L3, B2	EEN & Steroid			Thiopurine + Biological
S35	Yes	L3+L4a, B1	Mesalazine			Thiopurine
S36	Yes	L3, B1	EEN			Thiopurine
S37	Yes	L3+L4a, B1	EEN			Thiopurine
S38	Yes	L3, B1	EEN			Methotrexate
S39	Yes	L3, B2	EEN			Thiopurine
S40	Yes	L3+L4a, B1	EEN & Steroid			Missing data

Using the same omnibus test approach as above, we were again able to identify the most informative features in each significant dataset (see Additional File 5). The top 16S genera were *Dialister, Bilophila*, and *Aggregatibacter* in this analysis. The top MGS strains were subtypes of *Parabacteroides merdae*, *Sutterella wadsworthensis*, and an unclassified strain within the Lachnospiraceae family. The top MGS genera included *Parabacteroides, Bacteroides*, and an unclassified genus of Lachnospiraceae. The top MGS KEGG pathways included (1) ko00633, nitrotoluene degradation, (2) ko00250, alanine, aspartate and glutamate metabolism, and (3) ko00230, purine metabolism. The top KOs were (1) K02954, a ribosomal protein, (2) K07259, which is involved in peptidoglycan biosynthesis, and (3) K07793, a putative tricarboxylic transport membrane protein.

3.3.4 - Comparing the Relative Importance of Top Features

Although comparing RF model accuracies allows individual datasets to be evaluated, it does not allow the relative importance of features across datasets to be evaluated. To this end, we next compared the relative importance of the overall top features by running RF models using the top three features from the significant datasets for both CD state (Figure 3.10A) and treatment response (Figure 3.10B). The combined model for disease state classification performed with high accuracy (Accuracy=78.9*%,* P<0.001), but notably this was lower than the 16S genera alone. In contrast, the combined model for treatment response classification performed better than the independent datasets (Accuracy=94.4%, P<0.001). As expected, many of these features in both models are highly correlated (Figure 3.11 and Figure 3.12), nonetheless this approach yielded several useful results. Firstly, *Akkermansia muciniphila* was ranked as the most important feature for classifying disease state, followed by Verrucomicrobia and Verrucomicrobiales, which represent the phylum and order of *A. muciniphila* respectively. Number of OTUs was ranked 4th amongst these features, whereas GRS and other MGS-derived features were ranked lower. Notably, 29/37 (78%) of the microbial features in this model were at lower relative abundances in CD patients compared to controls. The top three features for classifying treatment response in the combined model were ko00633, the nitrotoluene degradation pathway, K07793, the putative tricarboxylic transport membrane protein, and Erysipelotrichi (the class containing the family Erysipelotrichaceae). Unlike for the combined disease model, MGS-derived functions were among the most highly ranked features (all 6 MGS functions are within the first 8 top features).

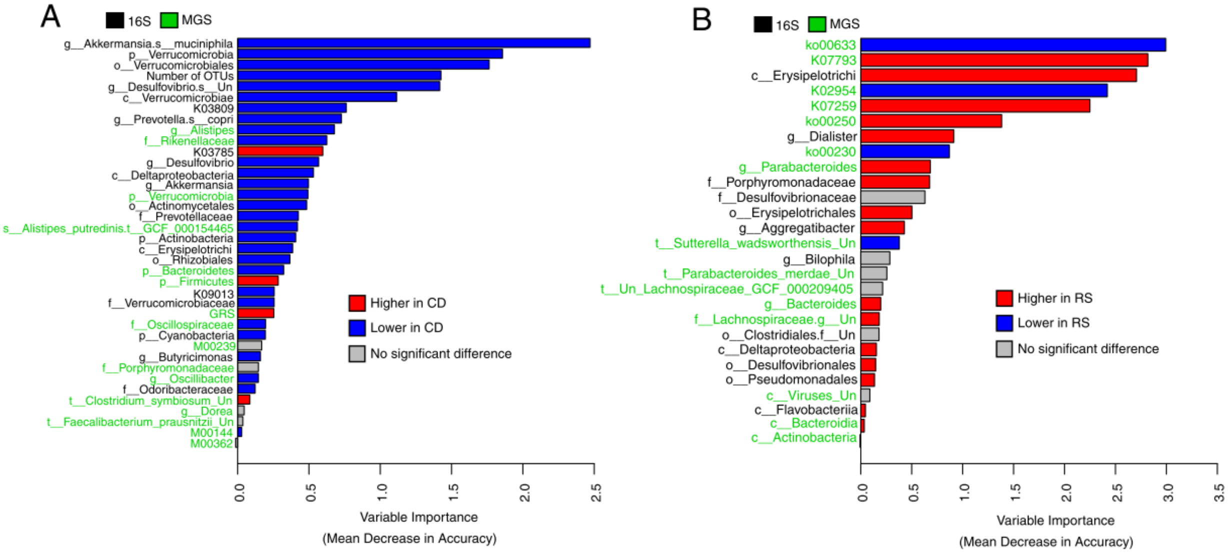


***Figure 3.10: Variable importance of features in combined random forest models for (A) disease state classification and (B) treatment response classification**. Red and blue are used to indicate which class has a higher mean standardized relative abundance. Features that did not significantly differ (P >= 0.05) between classes based on a two-tailed Mann-Whitney-Wilcoxon test are indicated in grey. Features in black and green font indicate 16S rRNA gene and shotgun metagenomics sequencing origins, respectively. "Un" stands for "Unclassified" when used in taxa names.*

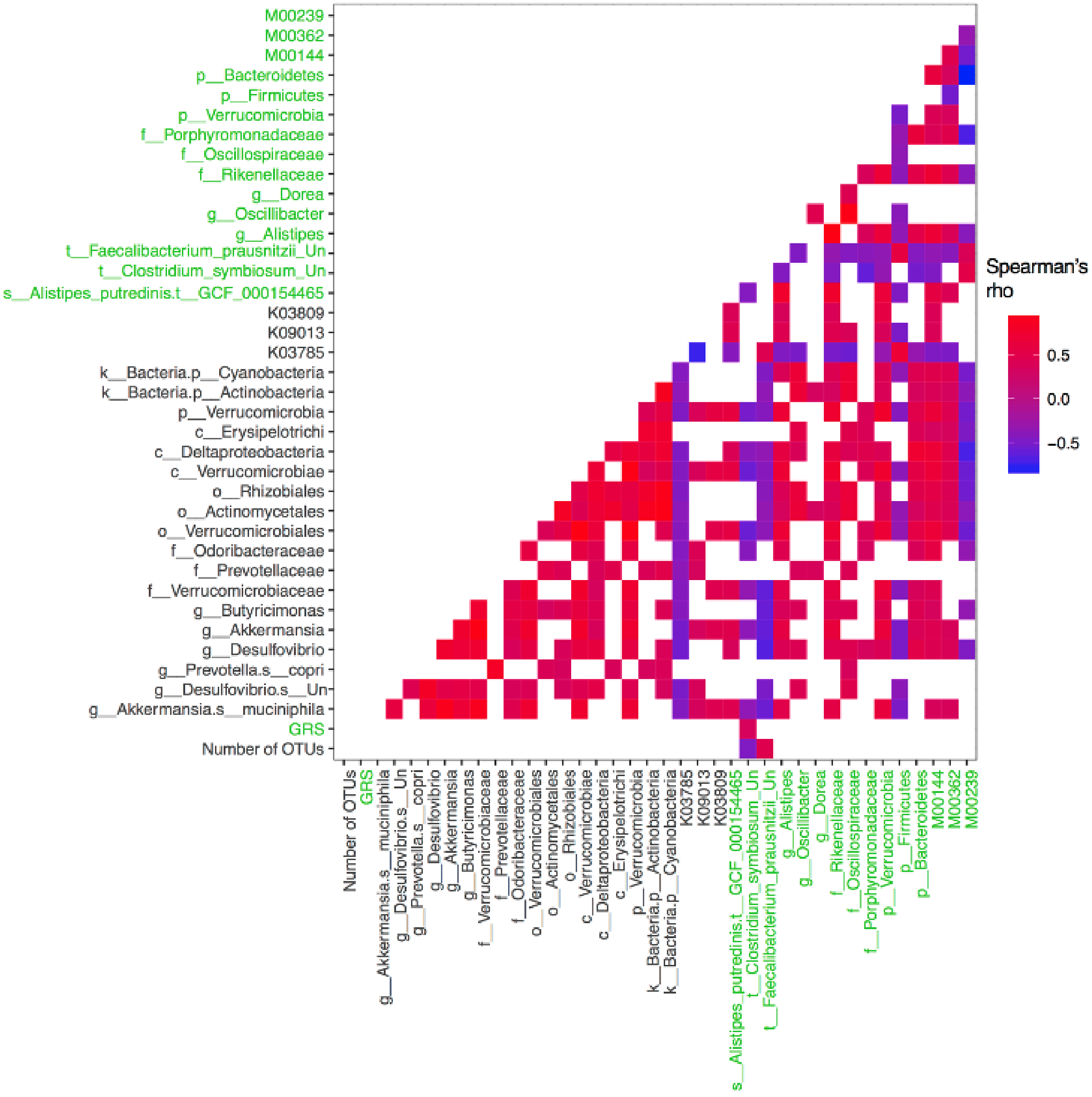


Figure 3.11: Heatmap of Spearman correlation coefficients for features in the combined disease random forest model that were significantly correlated (P < 0.05). *Metagenomics-identified feature names are coloured green. Only the bottom triangle is shown for simplicity. OTU: Operational Taxonomic Unit, GRS: Genetic Risk Score, Un: Unclassified.*

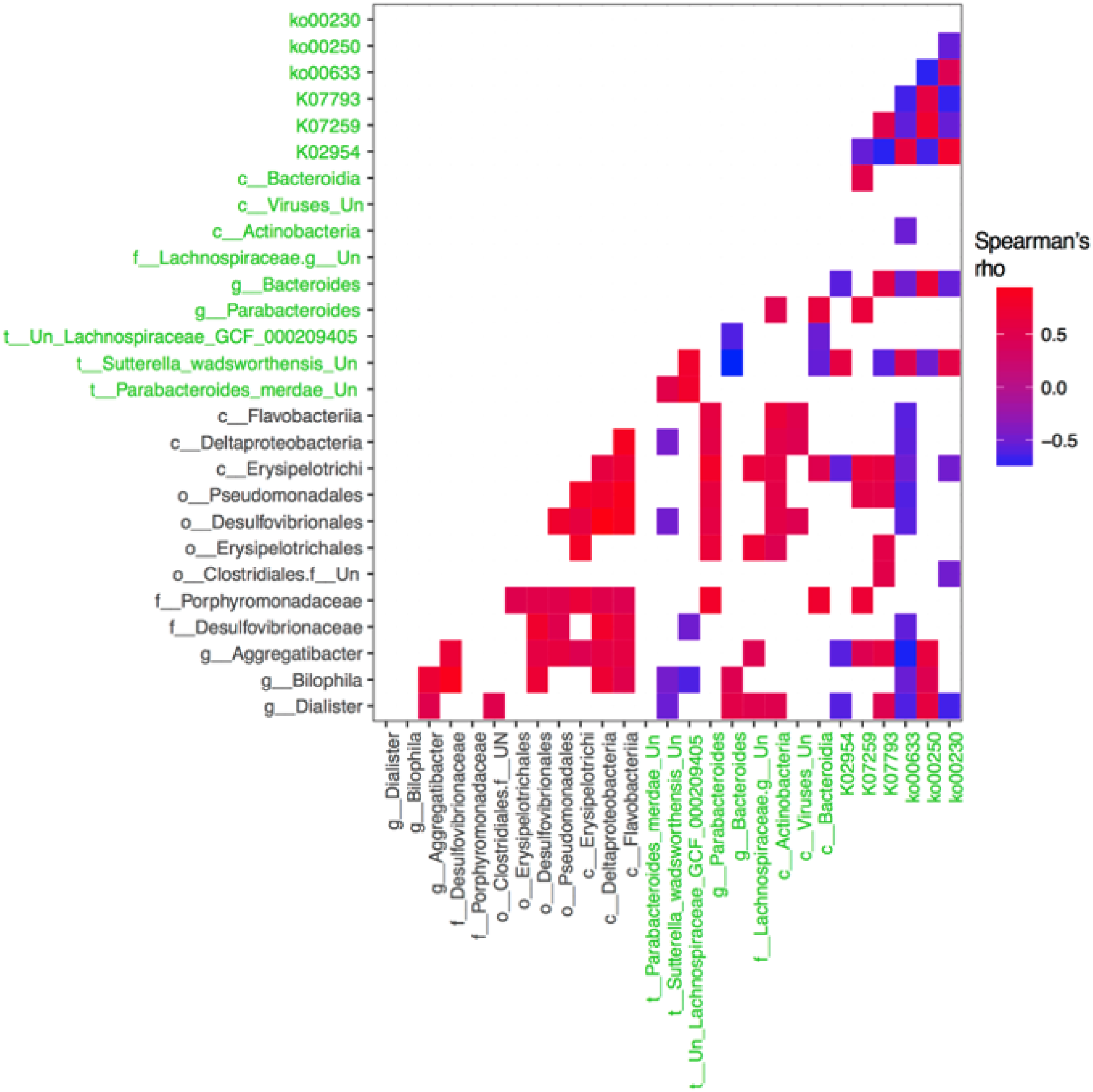


Figure 3.12: Heatmap of Spearman correlation coefficients for features in the combined treatment response random forest model that were significantly correlated (raw P < 0.05)*. Metagenomics-identified feature names are coloured green. Only the bottom triangle is shown for simplicity. "Un" stands for "Unclassified".*

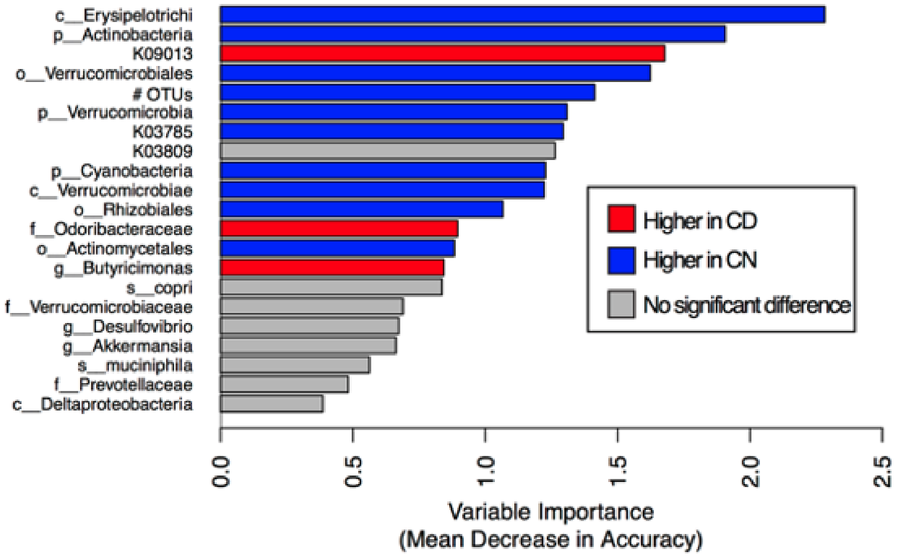


Figure 3.13: The top features from the disease state combined random forest model ranked by their variable importance in a new model trained on the RISK validation data*. Note that the 16S-identified unclassified species in Desulfovibrio was excluded since it was not present in the RISK data. All MGS features (including the genetic risk scores) were excluded from this analysis since MGS biopsy data was not available for this cohort. Features are coloured by whether they are significantly more abundant (raw P < 0.05) in Crohn's disease (CD; red) or normal colon control (CN; blue) samples, or not significantly different based on a two-tailed Mann-Whitney-Wilcoxon test.*

3.3.5 - Validating the Best 16S Disease Feature Rankings in an Independent Cohort

We validated the rankings of a subset of the 16S features (excluding the unclassified species in *Desulfovibrio*) used in the combined model for disease state by training a new model based on these features on the RISK cohort (Gevers et al. 2014), a large previously published dataset, that consisted of 16S data for 731 biopsy samples (444 CD and 287 CN) after processing. The goal of this analysis was to determine if the top features for classifying disease state would have similar relative importance ranks across both cohorts. Only 16S sequencing of biopsy samples is available in this dataset and so we excluded GRS and the MGS features from this analysis. The new RF model based on this subset of features and trained on the RISK dataset was highly significant although less accurate than what we observed in our data (Accuracy=73.2%, P<0.001). However, the relative ranking of these features was substantially different within the BISCUIT and RISK cohorts (Figure 3.10A and Figure 3.13). The top features in the RISK model were the class Erysipelotrichi, the phylum Actinobacteria, and the KO K09013. In addition, 8/21 16S features were not statistically different between CD and control patients (M-W-W test P >= 0.05). In particular, both *Desulfovibrio* and *Akkermansia,* did not significantly differ between CD and control patients within the RISK cohort.

3.4 - Discussion

In this study, we have classified treatment-naïve pediatric CD patients by both their disease state and treatment response with high accuracies with many different microbial datasets. Since these microbial profiles were taken from intestinal biopsy samples the main challenge of this study was to identify true microbial markers above the background of human DNA in the MGS data. Although we could identify microbial markers by generating much higher sequencing depth than is usual, the interpretation of analyses of this data come with the caveat that important rare taxa may have been below the detection threshold. For instance, although the RF models based on the MGS datasets were less accurate classifiers of disease state this likely was impacted by the fact that the most informative genera in the 16S data were undetected in many MGS samples. This observation suggests that the discrepancy between the 16S and MGS taxonomic classification accuracies could be partially due to a relatively greater taxonomic depth of 16S sequencing, currently cost-prohibitive for MGS of biopsy samples, which enabled rarer taxa to be identified.

Since the 16S data does not face these challenges, interpreting the analyses based on these datasets is more straightforward. Indeed, many of the top features in the significant 16S datasets used to classify disease state (see Additional File 4) have previously been associated with IBD. For instance, sulfur-reducing species within the *Desulfovibrio* genus have previously been positively linked to another form of IBD - ulcerative colitis (Rowan et al. 2010), and Mottawea *et al.* recently showed the importance of hydrogen sulfide producers in colonic CD (Mottawea et al. 2016). However, we found *Desulfovibrio* to be negatively associated with CD in our data, which could highlight a difference in microbiota between these two forms of IBD or merely reflect the different sampling strategies (stool, biopsy and MLI) between IBD studies to date. We also found *Akkermansia muciniphila* to have lower relative abundance in CD patients' biopsies, which has been previously observed (Dunn et al. 2016a). The top 16S-inferred KOs are also related to functions previously associated with CD symptoms. The lower proportion of K09013 (Fe-S cluster assembly ATP-binding protein) in CD patients is interesting to find since intestinal inflammation in general has been associated with the breakdown of Fe-S clusters (Schumann et al. 2012). Similarly, both K03809 (tryptophan repressor binding protein) and K03785 (3-dehydroquinate dehydratase I), which is involved in tryptophan and other amino acid biosynthesis, in CD patients could be interesting markers since lower serum tryptophan levels has previously been associated with CD (Gupta et al. 2012; Nikolaus et al. 2017). However, in this analysis these markers were both at higher levels in the unexpected direction (K03809 was lower in CD and K03785 was higher in CD).

The top MGS-identified features for classifying disease state also include several previously identified markers. The genus *Alistipes* is a known producer of short-chain fatty acids (SCFAs) (Brown et al. 2011). This genus was at lower relative abundance in CD patients, which could be related to lower levels of certain SCFAs that have long been a hallmark of IBD (Treem et al. 1994; Huda-Faujan et al. 2010; Morgan et al. 2012). In addition, although several key taxa identified by 16S sequencing appeared to be below the detection threshold in the MGS samples, both *Alistipes* and *Oscillibacter*, which has previously been negatively associated with CD (Mondot et al. 2011), were not identified in the 16S data. The absence of these informative taxa is likely related to how certain lineages cannot be identified with high-resolution based on 16S sequences. This difference highlights a trade-off in the MGS taxonomic results: improved taxonomic resolution at the cost of lower sensitivity, which has been discussed elsewhere (Tessler et al. 2017). The identification of the MGS-identified KEGG module M00144, which is involved in ATP synthesis, as being informative for classifying disease state is also interesting since IBD patients are known to have lower levels of intestinal ATP (Schürmann et al. 1999).

Similar to the RF models for disease state, many of the top features for classifying treatment response agreed with previous studies (see Additional File 5). For instance, the top 16S genus, *Dialister*, was at higher abundance in RS patients, which is consistent with previous work (Mondot et al. 2016). Similarly, the bacterial family Erysipelotrichaceae has been linked to human health in several ways (Kaakoush 2015). Although this taxon was not ranked highly, it is the only family within the top 16S-identified order, Erysipelotrichales, to pass pre-processing cut-offs. This order is found at higher relative abundance in RS patients. Erysipelotrichaceae are particularly of interest since they have been shown to decrease in abundance in CD patients given EEN therapy (Kaakoush et al. 2015) and species within this family are positively linked to inflammation (Dinh et al. 2015).

Several of the top MGS-identified KEGG functions also consistent with past work. The pathway ko00633, nitrotoluene degradation, has previously been identified as the most distinguishing pathway between EEN-treated CD patients and healthy controls (Dunn et al. 2016b). Similarly, microbial glutathione and purine biosynthesis have previously been positively and negatively associated with Crohn's disease respectively (Morgan et al. 2012). In our dataset, the pathway ko00250, glutamate and other amino acid metabolism, was found at higher relative abundance in RS patients whereas ko00230, purine metabolism, was found at lower relative abundance in RS patients. In addition, both the genera *Parabacteroides* and *Bacteroides* have previously been found at higher abundance in CD patients at the time of surgical resection who remain in remission (De Cruz et al. 2015), which is the same direction we find here. Our previous work in this cohort determined that *Sutterella wadsworthensis* is unlikely to be involved in IBD pathogenesis (Mukhopadhya et al. 2011). However, since this species was one of the best predictive features for treatment response and was found at lower abundance in RS patients it may still be clinically relevant. Although these results indicate that future CD treatments could be informed by the presence of these and other microbial markers, further work will be required to disentangle which markers are predictive of response to specific treatments.

The findings of *Akkermansia muciniphila* and the order and phylum (Verrucomicrobiales and Verrucomicrobia, respectively) that contain this species as the top three features for classifying disease state, highlights the importance of this taxon in our dataset. High levels of *A. muciniphila* in donor's stool has recently been found to be a strong predictor of remission in ulcerative colitis patients undergoing fecal microbiota transplantation treatment (Kump et al. 2018). This finding taken together with our and others' observation of lower *A. muciniphila* abundance in CD patients suggests that this species is a useful biomarker for gut health. Similarly, the relative importance of alpha diversity compared to genetic risk was also shown in this combined model. This finding illustrates the importance of microbial features in CD development, as compared with the weak contribution of genetic markers for CD development and the influence of the inherited variants on microbiome composition (Turpin et al. 2016). The top MGS-identified features largely performed worse than the 16S-identified functions in the combined RF model for classifying disease. One interesting exception is the genus *Alistipes* (and its corresponding family Rikenellaceae).

In the combined RF model for treatment response it is notable that MGS-identified functions were the most informative features. This observation could indicate that major metabolic shifts in the microbiome could be more informative for predicting treatment response than the presence of particular taxa, which is consistent with past results indicating that functions shift more consistently than taxa in CD patients (Morgan et al. 2012). Interestingly, functions were only found to be more informative for classifying patients by treatment response, and not by disease state. However, it is possible that with higher sequencing depth MGS-identified features may have been more informative. Note that patients' GRS were not significantly different between RS and NR samples, which is consistent with a recent study indicating that the genetic contributions to CD susceptibility are largely independent from the genetic contributions to CD prognosis (Lee et al. 2017).

Ideally the combined RF model trained on the top features from our cohort would also have been tested on the validation cohort. However, due to technical differences across the studies, such as different sampling protocols and different 16S variable regions sequenced, the same model cannot be implemented for both datasets. In addition, variation in pathophysiology due to geography as well as differential microbial profiles due to different distributions of patient age and sex across the cohorts could also result in differences in predictive markers across the two cohorts. This issue highlights that additional work in this area is needed to facilitate the comparison of microbiome datasets from different studies. Nonetheless, the independent validation cohort enabled the ranking of features within the combined model for disease state to be evaluated. The ranking of these features did differ in this cohort although the number of OTUs, Verrucomicrobiales, and Verrucomicrobia remain within the top 6 features (Figure 3.13). However, the genera *Desulfovibrio* and *Akkermansia* were not significantly different between CD and control patients within the RISK samples, which highlights the issue of comparing predictive features across different cohorts. Unfortunately, we were unable to validate the ranking of the top features for classifying treatment response on an independent dataset since there is no paired 16S and MGS dataset with adequate sample size available to our knowledge.

3.5 - Conclusions

Here, we have integrated human genetic data with 16S and MGS intestinal biopsy data to classify CD patients by disease state and treatment response for the first time. We found genera identified from 16S data to be the best classifiers of each outcome. One possible explanation for why 16S data was found to have higher performance than the MGS data could be that it enables much higher read depth for taxonomic assignment. This increased depth allows rare taxa to be identified, which was the case for the top 16S-identified genera. The biological importance of rare taxa in CD pathogenesis warrants further consideration, and indeed rarity may prove an important bias in culture-based studies of the IBD microbiome. Although we found alpha diversity to be a clear marker for disease state, GRS was relatively less informative. This result is perhaps not surprising since microbial shifts are likely causally related to disease onset, although the direction is unclear. In contrast, GRS has been developed as a metric for assessing disease risk at any point in a patient's life; including well before onset, but has not been of great influence in predicting onset or treatment stratification (Cleynen et al. 2016). The multi-genomics machine-learning approach presented in this study could be extended in the future to other diseases and to other data types such as transcriptomics and metabolomics to better understand the relative importance of each of these features. These models will provide new insights into the multifactorial nature of CD, helping highlight cohort-specific as well as fundamental contributors to disease pathophysiology, and may result in novel signatures to predict and guide personalized treatments.

3.6 - Declarations

3.6.1 - Ethics Approval and Consent to Participate

Ethical approval was granted by North of Scotland Research Ethics Service (09/S0802/24) and written informed consent was obtained from the parents of all subjects. Informed assent was also obtained from older children who were deemed capable of understanding the nature of the study. This study is publically registered on the United Kingdom Clinical Research Network Portfolio (9633). An ethics amendment allowed further review of the cohort participants' therapies and clinical response for the first year following diagnosis.

3.6.2 - Availability of Data and Material

All custom scripts used for this study are available at: https://github.com/LangilleLab/CD_RF_microbiome. The 16S rRNA gene and metagenomic sequencing data used in this study are available under accession PRJEB21933 at the European Nucleotide Archive.

Chapter 4 - PICRUSt2 for Prediction of Metagenome Functions

This chapter is a reproduction of my letter (and supplementary information) published in Nature Biotechnology (Douglas et al. 2020). The short published letter is presented first followed by the supplementary results section. This was done with permission from the journal and I was first author on this paper (see Appendices). I wrote the first draft and conducted all analyses and programming for this work. The other authors on this paper were: Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. The additional file and acknowledgements statement are available as part of the original publication on the Nature Biotechnology website. Note that the methods are presented separately in Chapter 2.

4.1 – Published Main Text

4.1.1 – Published Letter

To the editor - One limitation of microbial community marker-gene sequencing is that it does not provide information about the functional composition of sampled communities. PICRUSt (Langille et al. 2013) was developed in 2012 to predict the functional potential of a bacterial community based on marker gene sequencing profiles, and now we present PICRUSt2 (https://github.com/picrust/picrust2), which improves upon the original method. Specifically, PICRUSt2 contains an updated and larger database of gene families and reference genomes, provides interoperability with any OTU-picking or denoising algorithm, and enables phenotype predictions. Benchmarking shows that PICRUSt2 is more accurate than PICRUSt and other competing methods overall. PICRUSt2 also allows the addition of custom reference databases. We highlight these improvements and also important caveats regarding the use of predicted metagenomes.

The most common method for profiling bacterial communities is to sequence the conserved 16S rRNA gene. Functional profiles cannot be directly identified using 16S rRNA gene sequence data owing to strain variation so several methods have been developed to predict microbial community functions from taxonomic profiles (amplicon sequences) alone (Langille et al. 2013; Iwai et al. 2016; Jun et al. 2015; Aßhauer et al.

2015; Wemheuer et al. 2020). Shotgun metagenomic sequencing (MGS) which sequences entire genomes rather than marker genes can also be used to characterize the functions of a community, but does not work well if there is host contamination e.g. in a biopsy, or if there is very little community biomass.

PICRUSt (Langille et al. 2013) (hereafter "PICRUSt1") was the first tool developed for prediction of functions from 16S marker sequences, and is widely used but has some limitations. Standard PICRUSt1 workflows require input sequences to be operational taxonomic units (OTUs) generated from closed-reference OTU-picking against a compatible version of the Greengenes database (DeSantis et al. 2006). Due to this restriction to reference OTUs, the default PICRUSt1 workflow is incompatible with sequence denoising methods, which produce amplicon sequence variants (ASVs) rather than OTUs. ASVs have finer resolution, allowing closely related organisms to be more readily distinguished. Plus, the bacterial reference databases used by PICRUSt1 have not been updated since 2013 and lack thousands of recently added gene families.

We hypothesized that optimizing genome prediction would improve accuracy of functional predictions. Therefore, the PICRUSt2 algorithm (Figure 4.1a) includes steps that optimize genome prediction, including placing sequences into a reference phylogeny rather than relying on predictions limited to reference OTUs (Figure 4.1b); basing predictions on a larger database of reference genomes and gene families (Figure 4.1c); more stringent prediction of pathway abundance (Figure 4.2); enabling predictions of complex phenotypes and integration of custom databases.

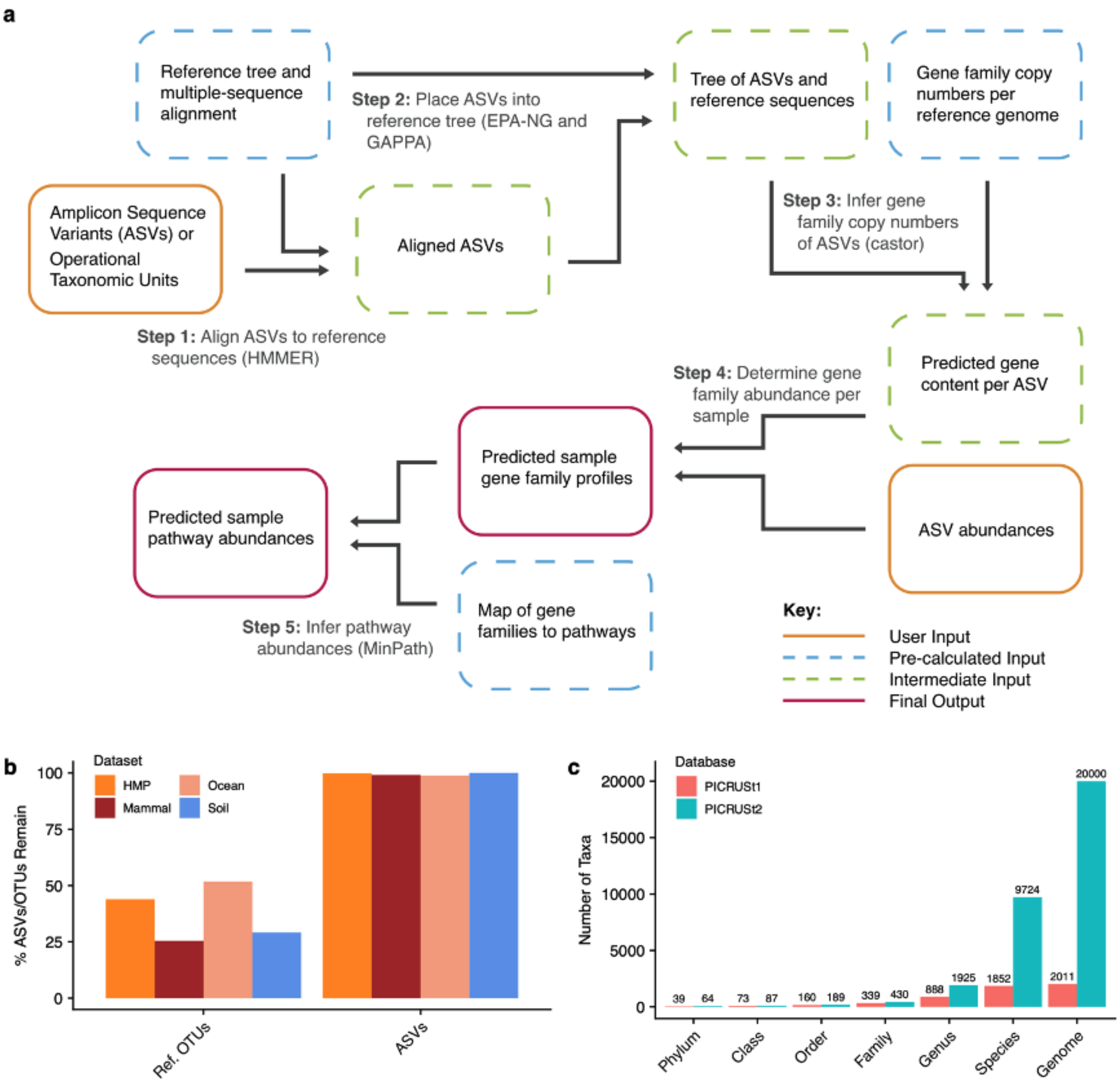

a
Reference tree and multiple-sequence alignment
Step 2: Place ASVs into reference tree (EPA-NG and GAPPA)
Tree of ASVs and reference sequences
Gene family copy numbers per reference genome
Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units
Aligned ASVs
Step 1: Align ASVs to reference sequences (HMMER)
Step 3: Infer gene family copy numbers of ASVs (castor)
Step 4: Determine gene family abundance per sample
Predicted gene content per ASV
Predicted sample gene family profiles
Predicted sample pathway abundances
ASV abundances
Map of gene families to pathways
Step 5: Infer pathway abundances (MinPath)
Key:
User Input
Pre-calculated Input
Intermediate Input
Final Output
b
Dataset
HMP
Ocean
Mammal
Soil
% ASVs/OTUs Remain
100
75
50
25
0
Ref. OTUs
ASVs
c
Database
PICRUSt1
PICRUSt2
Number of Taxa
20000
15000
10000
5000
0
39
64
73
87
160
189
339
430
888
1925
1852
9724
2011
20000
Phylum
Class
Order
Family
Genus
Species
Genome

***Figure 4.1: PICRUSt2 algorithm**. (a) The PICRUSt2 method consists of phylogenetic placement, hidden-state-prediction and sample-wise gene and pathway abundance tabulation. ASV sequences and abundances are taken as input, and gene family and pathway abundances are output. All necessary reference tree and trait databases for the default workflow are included in the PICRUSt2 implementation. (b) The default PICRUSt1 pipeline restricted predictions to reference operational taxonomic units (Ref. OTUs) in the Greengenes database. This requirement resulted in the exclusion of many study sequences across four representative 16S rRNA gene sequencing datasets. PICRUSt2 relaxes this requirement and is agnostic to whether the input sequences are within a reference or not, which results in almost all of the input amplicon sequence*

variants (ASVs) being retained in the final output. (c) An increase in the taxonomic diversity in the default PICRUSt2 database is observed compared to PICRUSt1.

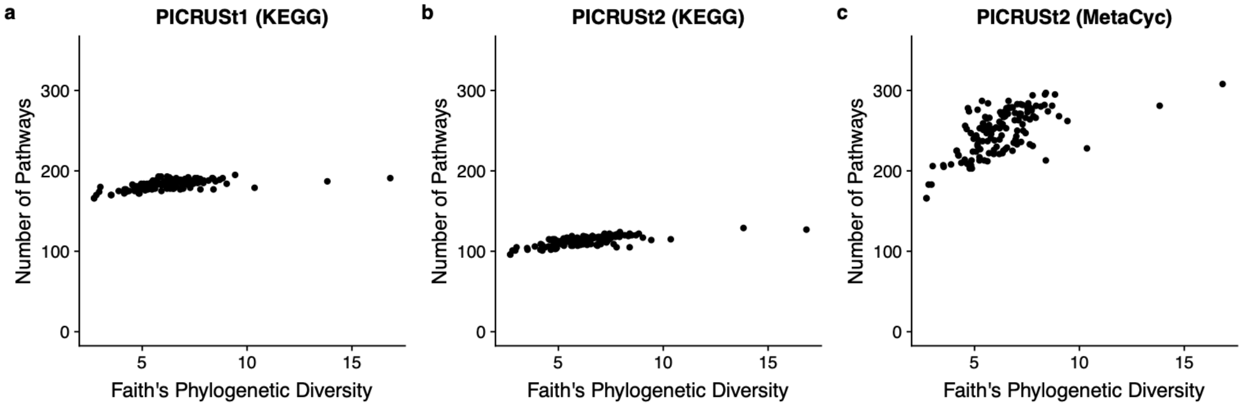


Figure 4.2: The number of predicted pathways as phylogenetic diversity varies in samples from the Human Microbiome Project. *A comparison of (a) KEGG pathways output by PICRUSt1, (b) KEGG pathways output by PICRUSt2, and (c) MetaCyc pathways output by the PICRUSt2 default pipeline. The number of KEGG pathways plateaus almost immediately whereas there is a much greater range in the MetaCyc pathways present. In addition, a mean of 1.6-fold more pathways are called as present in PICRUSt1 that are not called as present in PICRUSt2, which is due to the more stringent pathway pipeline intended to reduce false positives.*

PICRUSt2 integrates existing open-source tools to predict genomes of environmentally sampled 16S rRNA gene sequences. ASVs are placed into a reference tree, which is used as the basis of functional predictions. This reference tree contains 20,000 full 16S rRNA genes from bacterial and archaeal genomes in the Integrated Microbial Genomes (IMG) database (Markowitz et al. 2012). Phylogenetic placement in PICRUSt2 is based on running three tools: HMMER (www.hmmer.org) to place ASVs, EPA-ng (Barbera et al. 2019) to determine the optimal position of these placed ASVs in a reference phylogeny, and GAPPA (Czech & Stamatakis 2019) to output a new tree incorporating the ASV placements. This results in a phylogenetic tree containing both reference genomes and environmentally sampled organisms, which is used to predict individual gene family copy numbers for each ASV. This procedure is re-run for each input dataset, allowing users to utilize custom reference databases as needed, including those that may be optimized for the study of specific microbial niches.

As in PICRUSt1, hidden state prediction approaches are used in PICRUSt2 to infer the genomic content of sampled sequences. The castor R package (Louca & Doebeli 2018), which is substantially faster than the approach used in PICRUSt1, is used for core hidden state prediction functions. As in PICRUSt1, ASVs are corrected by their 16S rRNA gene copy number and then multiplied by their functional predictions to produce a predicted metagenome. PICRUSt2 also provides the ASV contribution of each predicted function allowing for taxonomy-informed statistical analyses to be conducted. Lastly, pathway abundances are inferred based on structured pathway mappings, which are more conservative than the 'bag-of-genes' approach used in PICRUSt1.

The PICRUSt2 default genome database is based on 41,926 bacterial and archaeal genomes from the IMG database (Markowitz et al. 2012) (November 8, 2017) which is a >20-fold increase over the 2,011 IMG genomes used by PICRUSt1. Many of the additional genomes are from strains of the same species and have identical 16S rRNA genes. We de-replicated the identical 16S rRNA genes across these genomes, which resulted in 20,000 final 16S rRNA gene clusters. The taxonomic diversity of the PICRUSt2 reference database is increased compared with PICRUSt1 (Figure 4.1c). The clearest increases in diversity is at the species and genus levels (5.3-fold and 2.2-fold increases respectively) but all taxonomic levels are more diverse including the phylum level where the coverage increased from 39 to 64 phyla (1.6-fold increase).

PICRUSt2 predictions based on several gene family databases are supported by default, including the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2012) (KEGG) orthologs (KO) and Enzyme Commission numbers (EC numbers) (Table 4.1). PICRUSt2 distinctly improves on PICRUSt1 by including gene families more recently added to the KEGG database. Specifically, the total number of KOs is 10,543 in PICRUSt2 compared to 6,909 in PICRUSt1, a 1.5-fold increase.

We validated PICRUSt2 metagenome predictions using samples from seven published datasets generated using both 16S rRNA marker-gene and shotgun metagenomics sequencing (MGS). We used three human-associated microbiome datasets: 57 stool samples from Cameroonian individuals, 91 stool samples from Indian individuals, and 137 samples spanning the human body (from the Human Microbiome Project [HMP]). We used four non-human-associated datasets including 77 non-human primate stool samples, eight mammalian stool samples, six ocean samples, and 22 bulk soil and blueberry rhizosphere samples. These datasets present a good variation of types of sequences and environments (Table 4.2).

Table 4.1: Summary statistics of PICRUSt2 trait reference databases

Database	# Categories	Mean	sd	Sparsity	Mean trait depth
16S rRNA gene	1	2.51	2.45	0.00	2.52
COG	4598	0.55	1.76	0.68	0.54
TIGRFAM	4287	0.31	0.98	0.77	0.48
EC	2913	0.33	0.95	0.78	0.43
PFAM	11089	0.40	2.10	0.82	0.40
KEGG	10543	0.19	0.65	0.85	0.32

Table 4.2: Descriptions of the paired 16S rRNA gene and shotgun metagenomics sequencing validation datasets used

Dataset	N*	16S pipeline	16S region	16S tech.	MGS tech.	# ASVs	Mean # 16S reads	Mean # MGS reads	16S read length	MGS read length
Cameroon	57	Deblur	V5-V6	Illumina MiSeq	Illumina HiSeq	4077	95354	42 x 10^6	228	101
India	91	Deblur	V3	Illumina NextSeq 500	Illumina NextSeq 500	2237	164847	6.5 x 10^6	130	151
HMP	137	DADA2	V4	Roche 454	Illumina HiSeq	1576	4592	30.0 x 10^6	245	101
Primate	77	Deblur (QIITA)	V4	Illumina MiSeq	Illumina HiSeq	7452	15536	7.5 x 10^6	150	160
Mammal	8	Deblur	V6-V8	Illumina MiSeq	Illumina MiSeq	323	2906	8.0 x 10^6	400	151
Ocean	6	Deblur	V4	Illumina MiSeq	Illumina HiSeq	1148	62498	64.1 x 10^6	250	101
Soil	22	Deblur	V6-V8	Illumina MiSeq	Illumina NextSeq	3333	5550	26.4 x 10^6	404	150

**Final number of samples overlapping between 16S rRNA gene and MGS datasets.*

PICRUSt2 KO predictions from 16S rRNA marker gene data were produced for each dataset. We compared these predictions to KO relative abundances profiled from the corresponding MGS metagenomes, which served as a gold-standard to evaluate prediction performance. We performed the same analyses with four alternative prediction pipelines: PICRUSt1, Piphillin (Iwai et al. 2016), PanFP (Jun et al. 2015) and Tax4Fun2 (Aßhauer et al. 2015; Wemheuer et al. 2020). We calculated Spearman correlation coefficients (hereafter "correlations") for matching samples between the predicted KO abundance and MGS KO abundance tables after filtering all tables to the 6,220 KOs that could be output by all tested databases (Figure 4.3). The correlation metric represents the similarity in rank ordering of KO abundances between the predicted and observed data. The correlations based on PICRUSt2 KO predictions ranged from a mean of 0.79 (standard deviation [SD] = 0.028; primate stool) to 0.88 (SD = 0.019; Cameroonian stool dataset). For all seven datasets, PICRUSt2 predictions were either better than or comparable with the best prediction method (paired-sample, two-tailed Wilcoxon tests [PTW] $P < 0.05$). Correlations based on PICRUSt2 predictions were substantially better for non-human associated datasets. This result could indicate an advantage of phylogenetic-based methods over non-phylogenetic-based methods, such as Piphillin, for environments poorly represented by reference genomes.

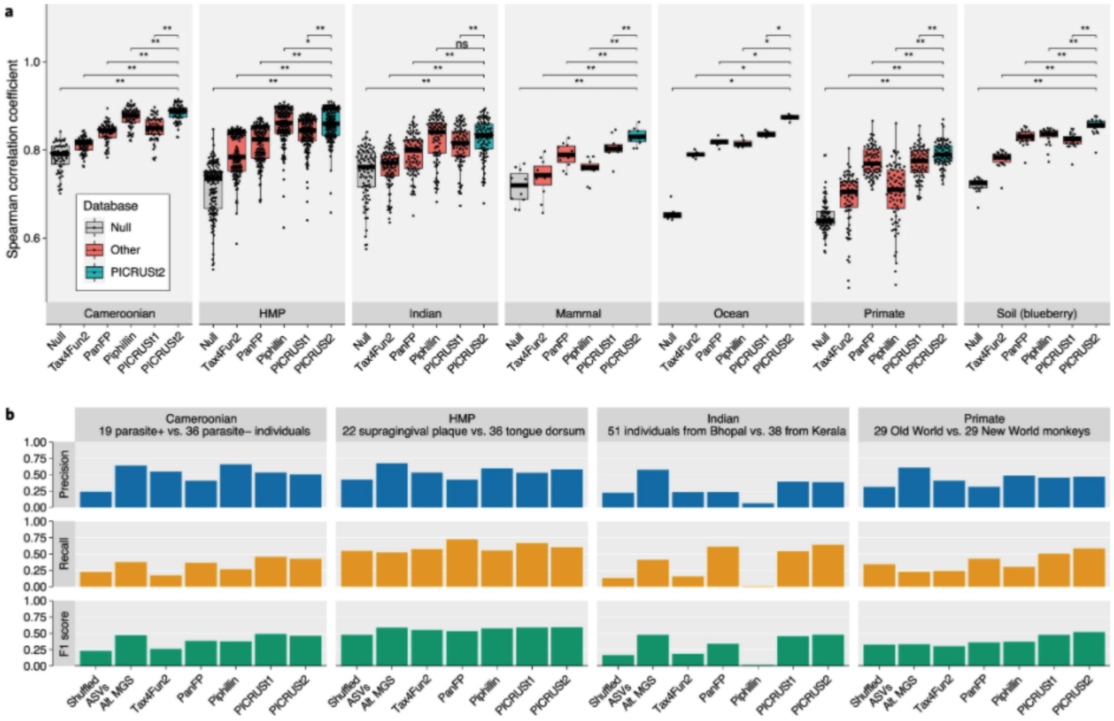


Figure 4.3: PICRUSt2 performance characteristics. *Validation of PICRUSt2 KEGG ortholog (KO) predictions comparing metagenome prediction performance against gold-standard shotgun metagenomic sequencing (MGS). (a) Boxplots of Spearman correlation coefficients observed in stool samples from Cameroonian individuals (n=57), the human microbiome project (HMP, n=137), stool samples from Indian individuals (n=91), non-human primate stool samples (n=77), mammalian stool (n=8), ocean water (n=6), and blueberry soil (n=22) datasets. The significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns correspond to P < 0.05, P < 0.001, and not significant respectively). (b) Comparison of significantly differentially abundant KOs between predicted metagenomes and MGS. Precision, recall, and F1 score are reported for each category compared to the MGS data. Precision corresponds to the proportion of significant KOs for that category also significant in the MGS data. Recall corresponds to the proportion of significant KOs in the MGS data also significant for that category. The F1 score is the harmonic mean of these metrics. The subsets of the four datasets compared are indicated above each panel (the Cameroonian parasite is Entamoeba). Wilcoxon tests were performed on the KO relative abundances after normalizing by the median number of universal single-copy genes per sample.*

Significance was defined at a false discovery rate < 0.05. The "Shuffled ASVs" category corresponds to PICRUSt2 predictions with ASV labels shuffled per dataset. The "Alt. MGS" category corresponds to an alternative MGS processing pipeline with reads aligned to the KEGG database rather than the default HUMAnN2 pipeline.

Gene families regularly co-occur within genomes, so the use of correlations to assess gene-table similarity may be limited by the lack of independence of gene families within a sample (Figure 4.4). To address this dependency, we compared the observed correlations between paired MGS and predicted metagenomes to correlations between MGS functions and a null reference genome, comprised of the mean gene family abundance across all reference genomes. For all datasets, PICRUSt2 metagenome tables were more similar to MGS values than the null (Figure 4.3a). However, this increase over the null expectation is predominately driven by each dataset's predicted genome content (rather than that of individual samples). This is demonstrated by the fact that these correlations are actually only slightly significantly higher than those observed when ASV labels are shuffled within a dataset (Figure 4.5). The observed correlations for the shuffled ASVs ranged from a mean of 0.77 (SD = 0.196; primate stool) to 0.84 (SD = 0.178; blueberry rhizosphere). Biologically these results are consistent with several patterns. First, gene families are correlated in copy number across diverse taxa (as captured by the 'Null' dataset). Second, these correlations are stronger within than between environments (as shown by the difference between the 'Null' and 'Shuffled ASV' results). Lastly, environment-to-environment differences tend to be larger than sample-to-sample differences within an environment (as shown by the differences between PICRUSt2 predictions and the 'Shuffled ASV' results).

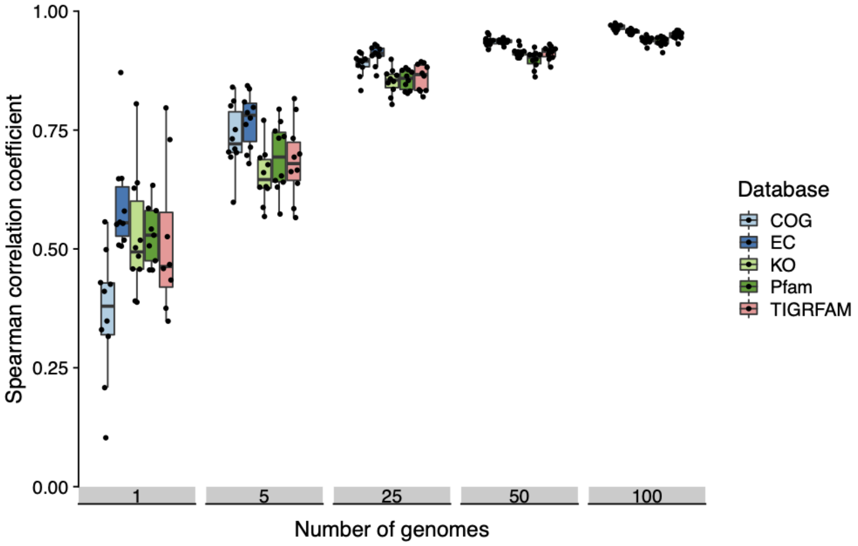


***Figure 4.4: Spearman correlations between number of gene families in two random subsets of genomes.** As the size of the random subsets increases the correlation coefficient between the two subsets approaches 1. Ten replicates are plotted for each number of genome subsets. These correlations are shown for all functional databases that can be used with PICRUSt2 by default. Clusters of Orthologous Genes (COG); Enzyme Commission Numbers (EC); Kyoto Encyclopedia of Genes and Genomes Ortholog (KO); Protein Families (Pfam); and The Institute for Genomic Research's database of protein FAMilies (TIGRFAM).*

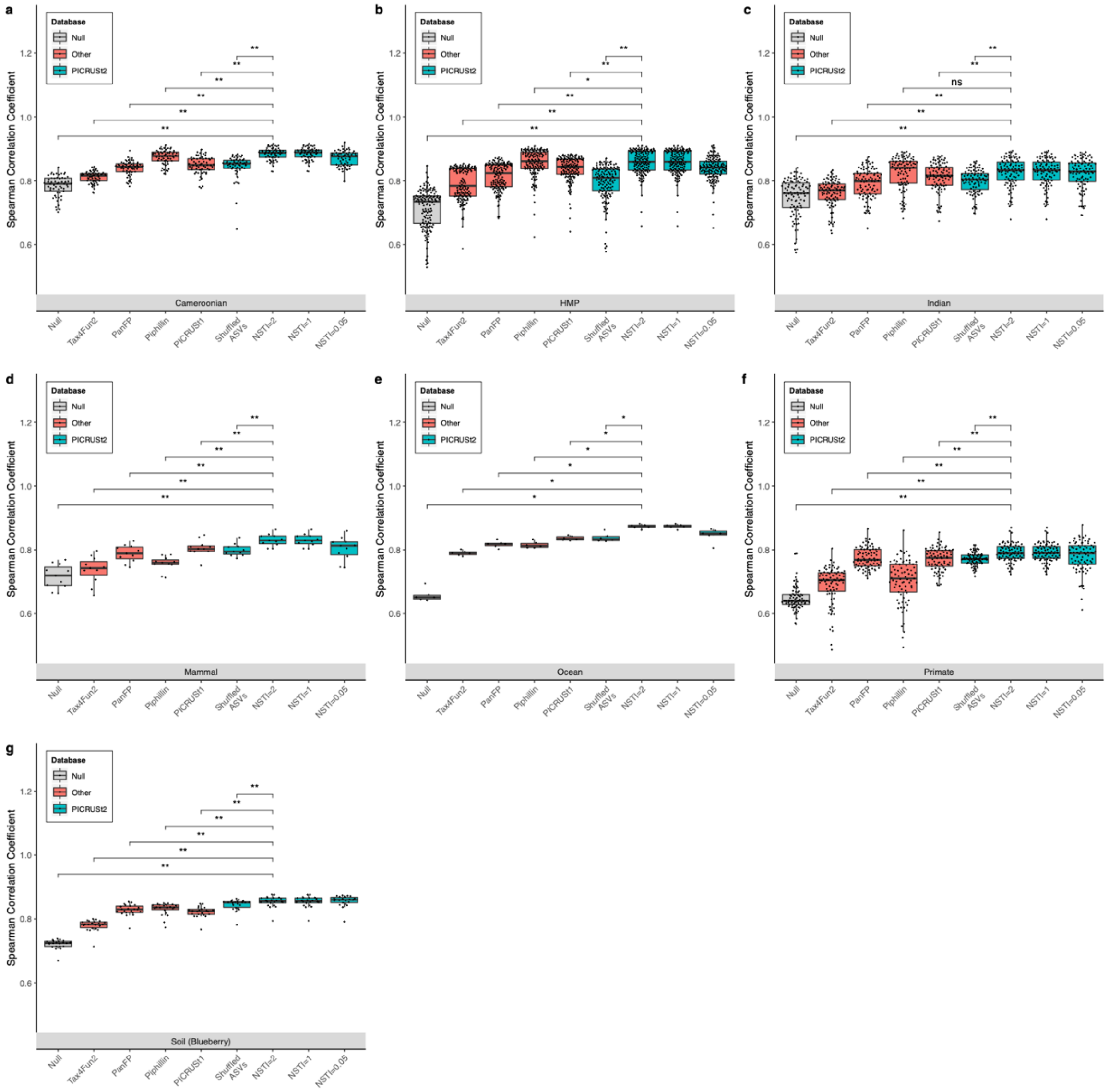


Figure 4.5: Full KEGG ortholog validation results of PICRUSt2 comparing metagenome prediction performance against gold-standard shotgun metagenomics sequencing. *The data shown here is the same as in Figure 4.3 except the performance at different nearest-sequenced taxon index (NSTI) cut-offs for the ASV input data and based on the 'Shuffled ASVs' data are also shown. HMP: Human Microbiome Project. Significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns correspond to P < 0.05, P < 0.001, and not significant respectively).*

A complementary approach for validating metagenome predictions is to compare the results of differential abundance tests on 16S-predicted metagenomes to MGS data. A recent analysis of Piphillin suggested that this tool out-performs PICRUSt2 based on this approach (Narayan et al. 2020). We similarly performed this evaluation on the KO predictions for four validation datasets (Figure 4.3b). Overall, PICRUSt2 displayed the highest F1 score, the harmonic mean of precision and recall, compared to other prediction methods (ranging from 0.46-0.59; mean=0.51; SD=0.06). However, all prediction tools displayed relatively low precision, the proportion of significant KOs that were also significant in the MGS data. In particular, precision ranged from 0.38-0.58 (mean=0.48; SD=0.08) for PICRUSt2 and 0.06-0.66 (mean=0.45; SD=0.27) for Piphillin. In all cases, PICRUSt2 predictions out-performed ASV-shuffled predictions, which ranged in precision from 0.22-0.42 (mean=0.30; SD=0.09). In addition, differential abundance tests performed on MGS-derived KOs from an alternative MGS-processing workflow resulted in only marginally higher precision (ranging from 0.57-0.67; mean=0.62; SD=0.04). Taken together, these results highlight the difficulty of reproducing microbial functional biomarkers with both predicted and actual metagenomics data.

MetaCyc pathway abundances are now the main high-level predictions output by PICRUSt2 by default. The MetaCyc database (Caspi et al. 2013) is an open-source alternative to KEGG and is also a major focus of the widely-used metagenomics functional profiler, HUMAnN2 (Franzosa et al. 2018). MetaCyc pathway abundances are calculated in PICRUSt2 through structured mappings of EC gene families to pathways. These pathway predictions performed better than the null distribution for all metrics overall (PTW $P < 0.05$; Figure 4.6a; Figure 4.7; Figure 4.8) compared to MGS-derived pathways. Similar to our previous analysis, shuffled ASV predictions representing overall functional structure within each dataset accounted for the majority of this signal (Figure 4.7). In addition, differential abundance tests on these pathways showed high variability in F1 scores across datasets and statistical methods with the ASV shuffled predictions contributing the majority of this signal (Figure 4.9; F1 scores ranged from 0.23-0.62 [mean=0.41; SD=0.17] and 0.22-0.60 [mean=0.34; SD=0.18] for the observed and ASV shuffled PICRUSt2 predictions, respectively). Again, these results suggest that identifying robust differentially abundant metagenome-wide pathways is difficult and highlights the challenge of analyzing microbial pathways in general.

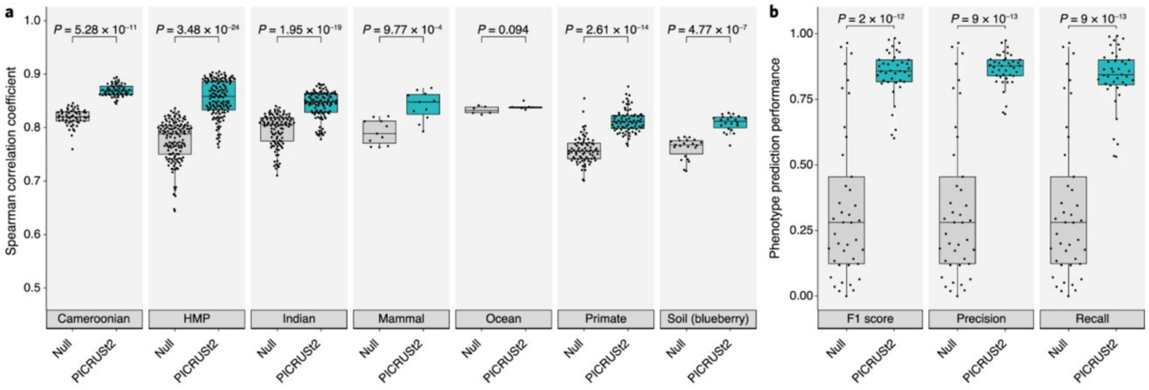


Figure 4.6: PICRUSt2 accurately predicts MetaCyc pathways and phenotypes for characterizing overall environments. *(a) Spearman correlation coefficients between PICRUSt2 predicted pathway abundances and gold-standard metagenomic sequencing (MGS). Results are shown for each validation dataset: stool from Cameroonian individuals, The Human Microbiome Project (HMP), stool from Indian individuals, mammalian stool, ocean water, non-human primate stool, and blueberry soil. These results are limited to the 575 pathways that could potentially be identified by PICRUSt2 and HUMAnN2. (b) Performance of binary phenotype predictions based on three metrics: F1 score, precision, and recall. Each point corresponds to one of the 41 phenotypes tested. Predictions assessed here are based on holding out each genome individually, predicting the phenotypes for that holdout genome, and comparing the predicted and observed values. The null distribution in this case is based on randomizing the phenotypes across the reference genomes and comparing to the actual values, which results in the same output for all three metrics. The P-values of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (* and ** correspond to P < 0.05 and P < 0.001, respectively). Note that in panel a the y-axis is truncated below 0.5 rather than 0 to better visualize small differences between categories. The sample sizes in panel a are 57 (Cameroonian), 137 (HMP), 91 (Indian), 8 (mammal), 6 (ocean), 77 (primate), and 22 (soil).*

Predictions for 41 microbial phenotypes, which are linked to IMG genomes (Chen et al. 2013), can also now be generated with PICRUSt2. These represent high-level microbial metabolic activities such as "Glucose utilizing" and "Denitrifier" that are annotated as present or absent within each reference genome. We performed a hold-out validation to assess the performance of PICRUSt2 phenotype predictions, which involved comparing the binary phenotype predictions to the expected phenotypes for each reference genome. Based on F1 score (mean=84.8%; SD=9.01%), precision (mean=86.5%; SD=6.21%), and recall (mean=83.5%; SD=11.4%), these predictions performed significantly better than the null expectation (Figure 4.6b; Wilcoxon tests $P < 0.05$).

There are two main criticisms of amplicon-based functional prediction. First, the predictions are biased towards existing reference genomes, which means that rare environment-specific functions are less likely to be identified. This limitation is decreasing over time as the number of high-quality available genomes continues to grow. PICRUSt2 also allows user-specified genomes to be used for generating predictions, which provides a flexible framework for studying particular environments. The second criticism is that amplicon-based predictions cannot provide resolution to distinguish strain-specific functionality. This is an important limitation of PICRUSt2 and any amplicon-based analysis, which can only differentiate taxa to the degree they differ at the amplified marker gene sequence.

PICRUSt2 provides improved accuracy and flexibility for marker gene metagenome inference. We have highlighted these improvements while also describing limitations with identifying consistent differentially abundant functions in microbiome studies. We hope that the expanded functionality of PICRUSt2 will continue to enable the identification of insights into functional microbial ecology from amplicon sequencing profiles.

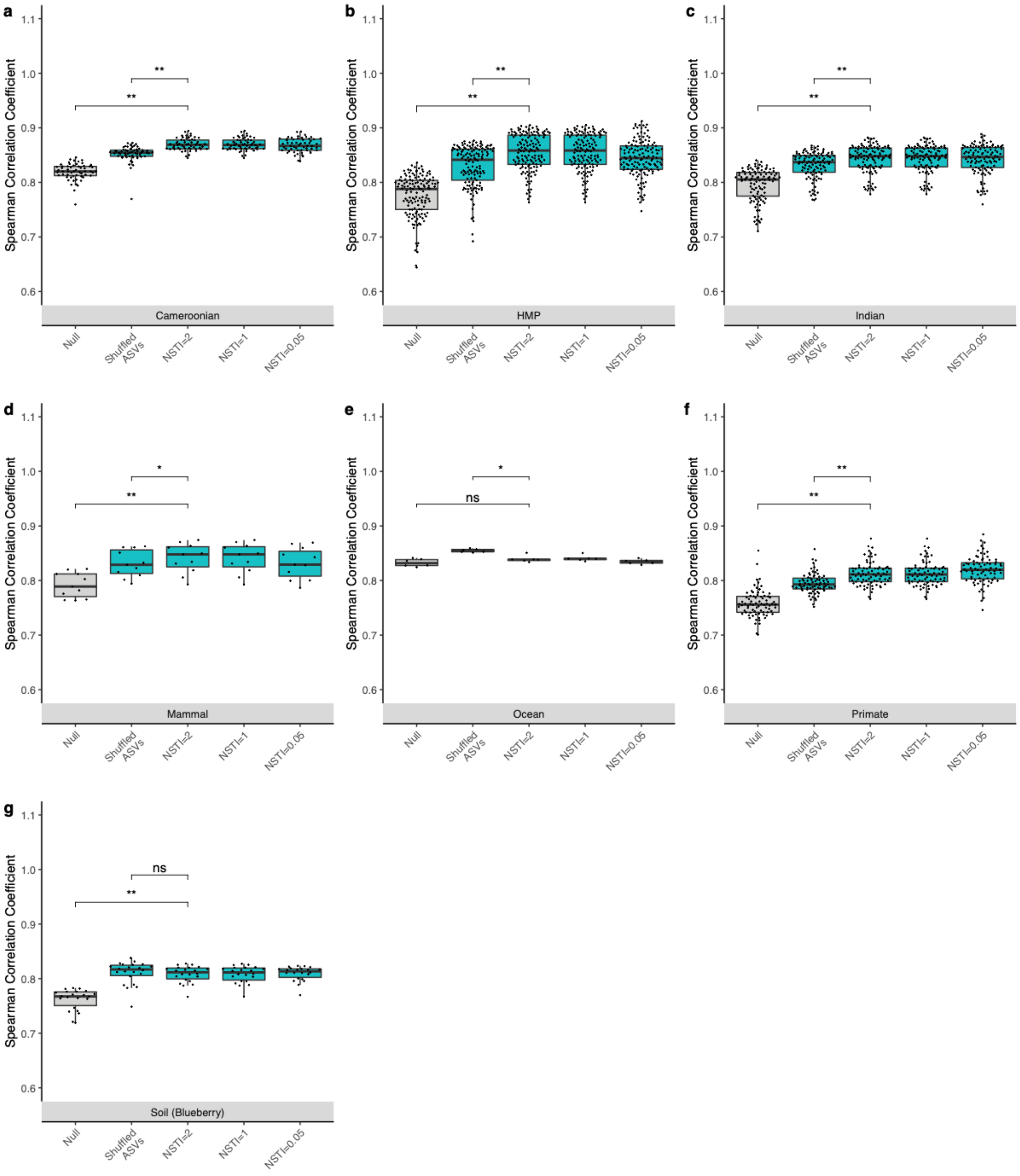


Figure 4.7: Spearman correlation coefficients of predicted MetaCyc pathway abundances compared to shotgun metagenomics on same samples *for (a) stool samples from Cameroonian individuals, (b) the human microbiome project (HMP), (c) stool samples from Indian individuals, (d) mammalian stool samples, (e) ocean samples, (f) non-human primate stool samples, and (g) soil samples. PICRUSt2 predictions based on varying nearest sequenced taxon index (NSTI) cut-offs are shown to illustrate how this parameter affects prediction performance. In addition, the 'Shuffled ASVs' group shows that the majority of the performance signal is driven by the predicted genomic content of*

the ASVs in each environment rather than within each sample. Significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (, **, and ns correspond to P < 0.05, P < 0.001, and not significant, respectively).*

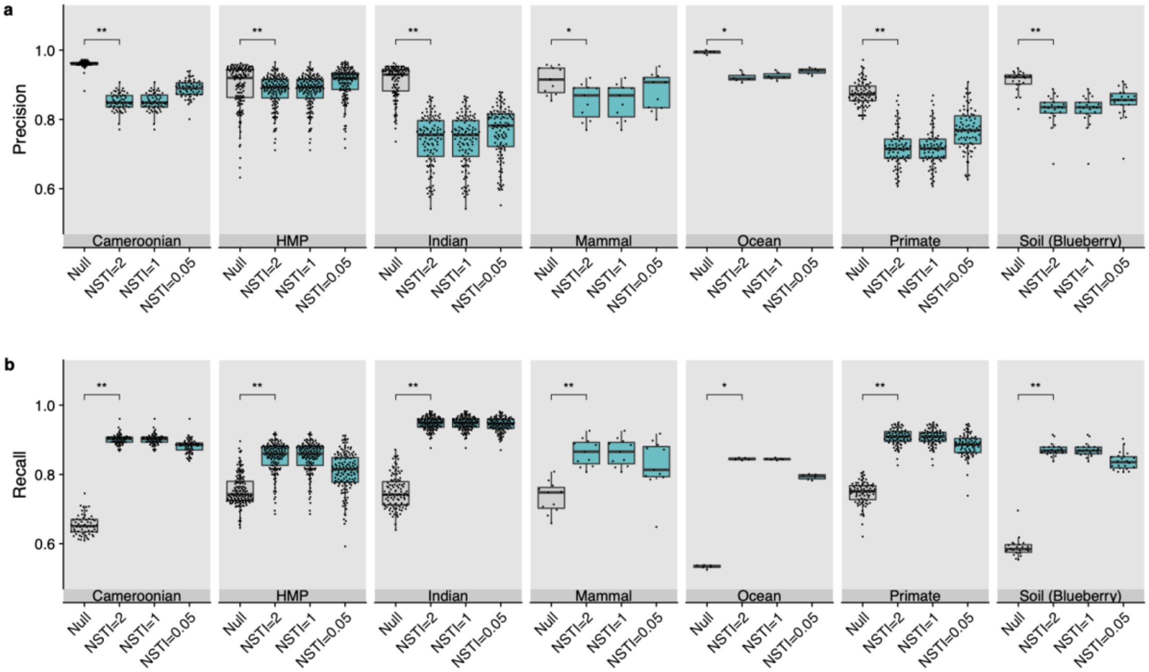


***Figure 4.8: (a) Precision and (b) recall of predicted MetaCyc pathway abundances compared to shotgun metagenomics** on the same samples from the seven validation datasets. PICRUSt2 predictions based on varying nearest sequenced taxon index (NSTI) cut-offs are shown to illustrate how this parameter affects prediction performance. Significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (* and ** correspond to P < 0.05 and P < 0.001, respectively).*

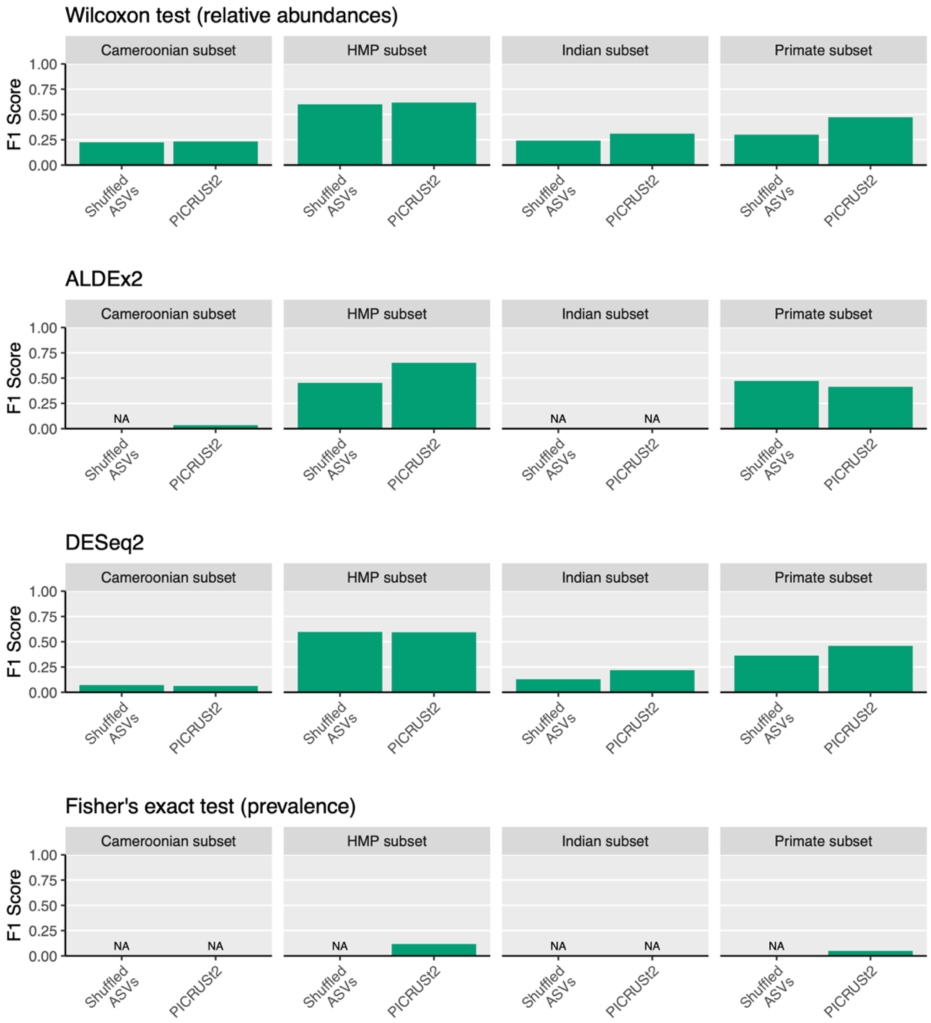


Figure 4.9: (a) Differential abundance and prevalence of predicted MetaCyc pathways vary across datasets and with statistical methods used and are similar to predictions based on shuffled ASVs*. The F1 score is plotted to summarize the agreement between the statistical testing of predicted pathways with pathways identified from HUMAnN2. All pathway abundances for this analysis were based on Enzyme Commission numbers over the entire metagenome sample (i.e. assuming a bag-of-genes model), rather than restricting pathway predictions to each Amplicon Sequence Variant's (ASV) predicted genome. Each of these statistical tests has a different null hypothesis, but the first three methods are all commonly used to perform differential abundance analyses on microbiome data. NA indicates cases where either the recall or precision could not be calculated (which means the F1 scores was not applicable). The sample groupings for*

these analyses are the same as for all the other differential testing results reported and are indicated by "subset" here simply to save space.

4.1.2 - Code and Data Availability

PICRUSt2 is available at: https://github.com/picrust/picrust2. The Python and R code used for the analyses and database construction described in this paper are available online at https://github.com/gavinmdouglas/picrust2_manuscript. This repository also includes the processed datafiles that can be used to re-generate the figures and findings in this paper. The accessions for all sequencing data used in this study are listed in Supplementary Results section below.

4.2 – Supplementary Results

The below results were published as part of the Supplementary Information to the above letter. The exception is Section 4.2.4 below, which was requested by my committee after we submitted this manuscript and so that section includes additional unpublished results.

4.2.1 - Paired 16S rRNA Gene and Shotgun Metagenomics Validations

We previously developed the nearest sequenced taxon index (NSTI) as a metric for summarizing a microbial taxonomic profile's novelty relative to isolate genomes (Langille et al. 2013). This measures the abundance-weighted distances in the phylogenetic tree between taxa (ASVs) from a community and the tips of the nearest sequenced neighbours. These NSTI distributions are calculated automatically in PICRUSt2 and differed significantly among the evaluation datasets, demonstrating their range of "unusualness" relative to sequenced isolates (Kruskal-Wallis χ^2=19,499, P < 2.2 x 10^{-16}; Figure 4.10, panels a and b). Datasets from more well-characterized communities have lower mean NSTI values overall as expected, ranging from 0.10 (SD: 0.11) in the Indian dataset to 0.51 (SD: 2.06) in the ocean dataset. A maximum NSTI cut-off of two is implemented by default in PICRUSt2, as a guideline to prevent unconsidered interpretation of overly speculative inferences, which resulted in a mean of 0.27% (SD: 0.4) of ASVs being excluded across these datasets. These excluded ASVs mainly correspond to either eukaryotic sequences or microbial phyla with no reference genomes available.

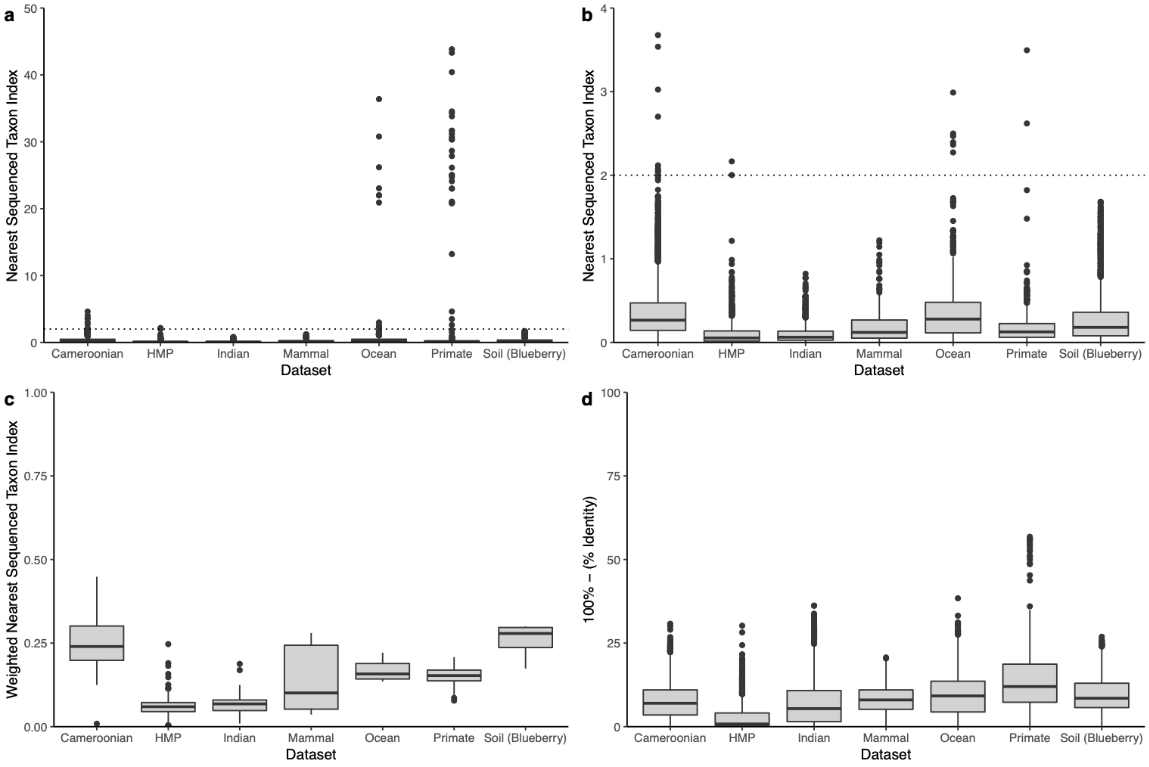


Figure 4.10: Reference-based quality metrics for amplicon sequence variants (ASVs) of validation datasets. *(a) Nearest sequenced taxon index (NSTI) for ASVs (branch length to nearest reference sequence). The grey horizontal dotted line at two indicates the default maximum NSTI value over which ASVs will be excluded. (b) Same as panel a except y-axis is truncated so visualizing these differences is easier. (c) Per-sample weighted NSTI values, which corresponds to the NSTI values in panels a and b under the maximum cut-off weighted by the abundance of each ASV. (d) Complement of percent identities of ASVs against reference database sequences*

In addition to the Spearman correlations reported in the main text, we also investigated metagenome predictions based on the presence and absence of output KOs by calling any function with non-zero abundance as present and assessing relative differences in precision, recall, and F1 score (see methods for definitions). PICRUSt2 and Piphillin exhibited the best (or non-statistically different) F1 score across 3/7 (HMP, mammalian stool, and ocean) and 4/7 (Cameroonian stool, HMP, Indian stool, and soil datasets) datasets, respectively (Figure 4.11); and were substantially better than the null (<0.6 in all datasets). Investigating precision and recall revealed that Piphillin tended to have higher precision scores (fewer false positives) while PICRUSt2 had better recall (fewer false negatives) (Figure 4.12). The slightly lower precision of PICRUSt2 is driven by falsely called KOs that are on average at significantly lower relative abundance in the PICRUSt2 results compared to the Piphillin results (PTW $P < 0.05$; mean of 0.0020% [SD: 0.0015] for PICRUSt2 vs mean of 0.0055% [SD: 0.0047] for Piphillin). However, it remains unclear what proportion of these disagreements between PICRUSt2 KO predictions and the MGS data are in fact false negatives in the MGS data due to low sequencing depth or annotation limitations from short MGS reads. Indeed, there is a significant positive relationship between the number of annotated MGS reads per sample and the observed precision of PICRUSt2 predictions (Spearman correlation coefficient across all datasets combined: 0.80; $P < 2.2 \times 10^{-16}$; mean coefficient of 0.49 [SD: 0.15] for individual datasets).

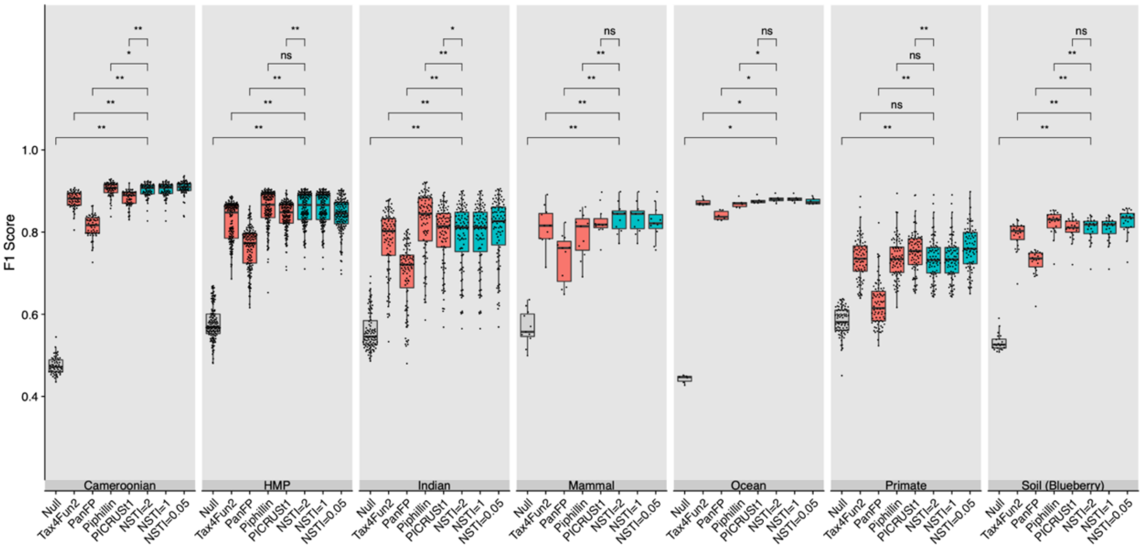


__Figure 4.11: KEGG ortholog prediction performance in terms of F1 score__, which is the harmonic mean of precision and recall. The datasets and prediction categories shown here are the same as in Figure 2 except the PICRUSt2 performance at different nearest-sequenced taxon index (NSTI) cut-offs for the ASV input data is also shown. Boxplot colours refer to three different category groupings: the null expectation (grey),

alternative prediction tools (red), and PICRUSt2 predictions (cyan). The significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (, **, and ns correspond to P < 0.05, P < 0.001, and not significant respectively). HMP: Human Microbiome Project.*

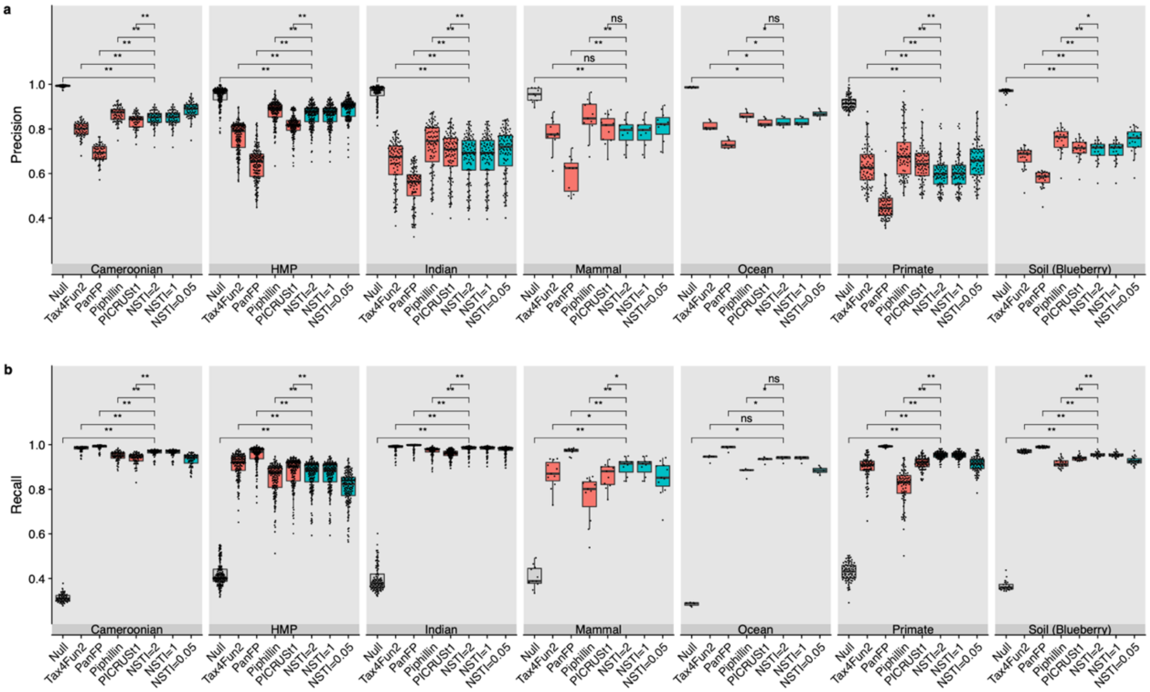


Figure 4.12: KEGG ortholog prediction performance in terms of (a) precision and (b) recall*. The datasets and prediction categories shown here are the same as in Figure 4.3 except the PICRUSt2 performance at different nearest-sequenced taxon index (NSTI) cut-offs for the ASV input data is also shown. Boxplot colours refer to three different category groupings: the null expectation (grey), alternative prediction tools (red), and PICRUSt2 predictions (cyan). The significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns correspond to P < 0.05, P < 0.001, and not significant respectively). HMP: Human Microbiome Project.*

We also evaluated the performance of the PICRUSt2 EC predictions by comparing with PAPRICA (Bowman & Ducklow 2015), which is another EC-based prediction tool. Correlations between EC predictions with the EC gene families observed in MGS data were significantly higher than PAPRICA for 4/7 validation datasets (PTW P < 0.05; Figure 4.13). In addition, the recall of PICRUSt2 predictions was significantly higher than PAPRICA although this came at a cost of precision (PTW P < 0.05; Figure 4.14).

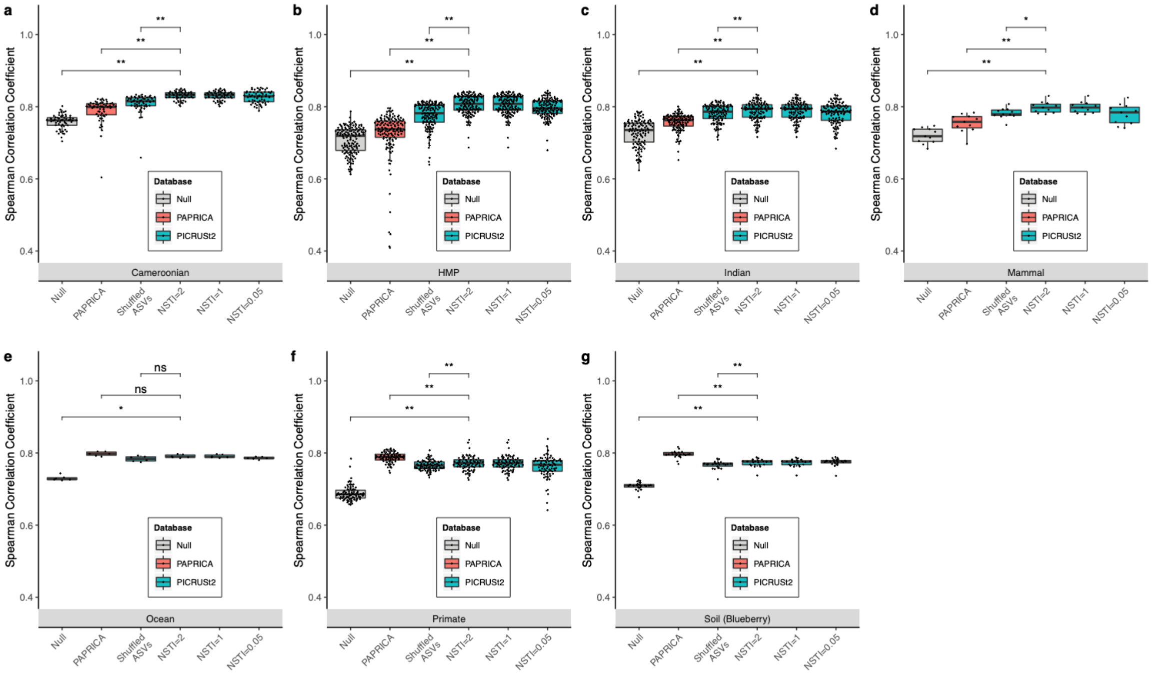


Figure 4.13: Spearman correlation coefficients of predicted Enzyme Commission (EC) number abundances compared to shotgun metagenomics on same samples *for (a) stool samples from Cameroonian individuals, (b) the human microbiome project (HMP), (c) stool samples from Indian individuals, (d) mammalian stool samples, (e) ocean samples, (f) non-human primate stool samples, and (g) soil samples. PICRUSt2 predictions based on varying nearest sequenced taxon index (NSTI) cut-offs are shown to illustrate how this parameter affects prediction performance. Significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns correspond to P < 0.05, P < 0.001, and not significant, respectively).*

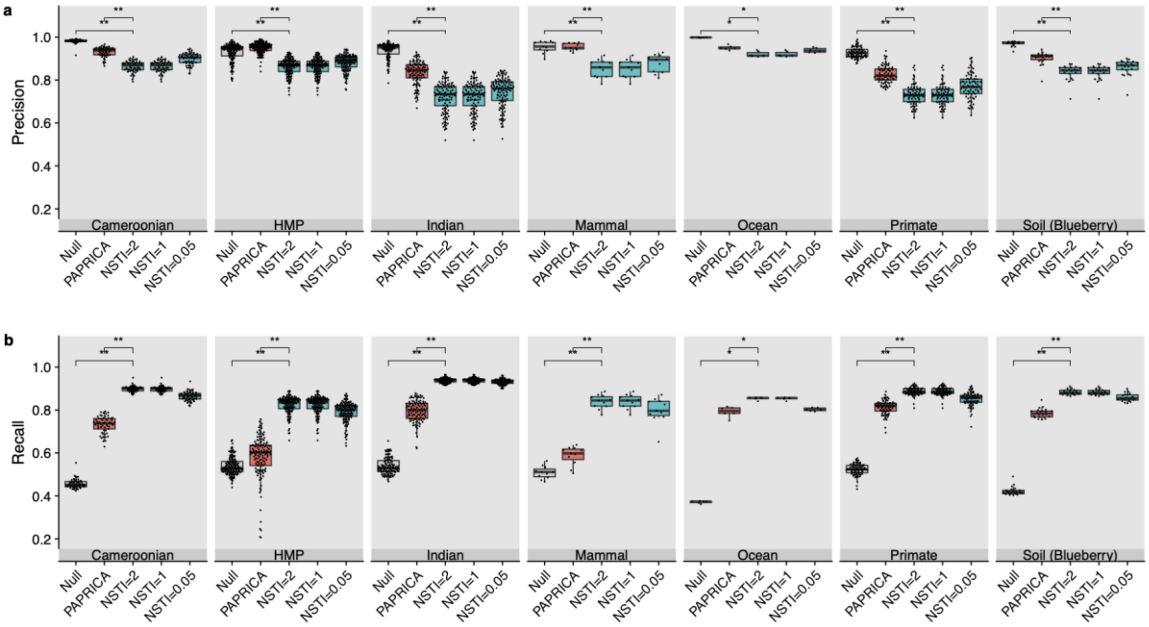


***Figure 4.14: (a) Precision and (b) recall of predicted Enzyme Commission (EC) numbers** (based on presence/absence) compared to shotgun metagenomics for all seven validation datasets. PICRUSt2 predictions based on varying nearest sequenced taxon index (NSTI) cut-offs are shown to illustrate how this parameter affects prediction performance. Significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (* and ** correspond to P < 0.05 and P < 0.001, respectively).*

The predicted MetaCyc pathways generated by PAPRICA drastically differ from the metagenomics pathway abundances (Figure 4.15) likely because PAPRICA implements a different approach for inferring pathway abundances. For this reason, evaluating the performance of the PAPRICA pathway predictions against HUMAnN2 pathways may not be appropriate.

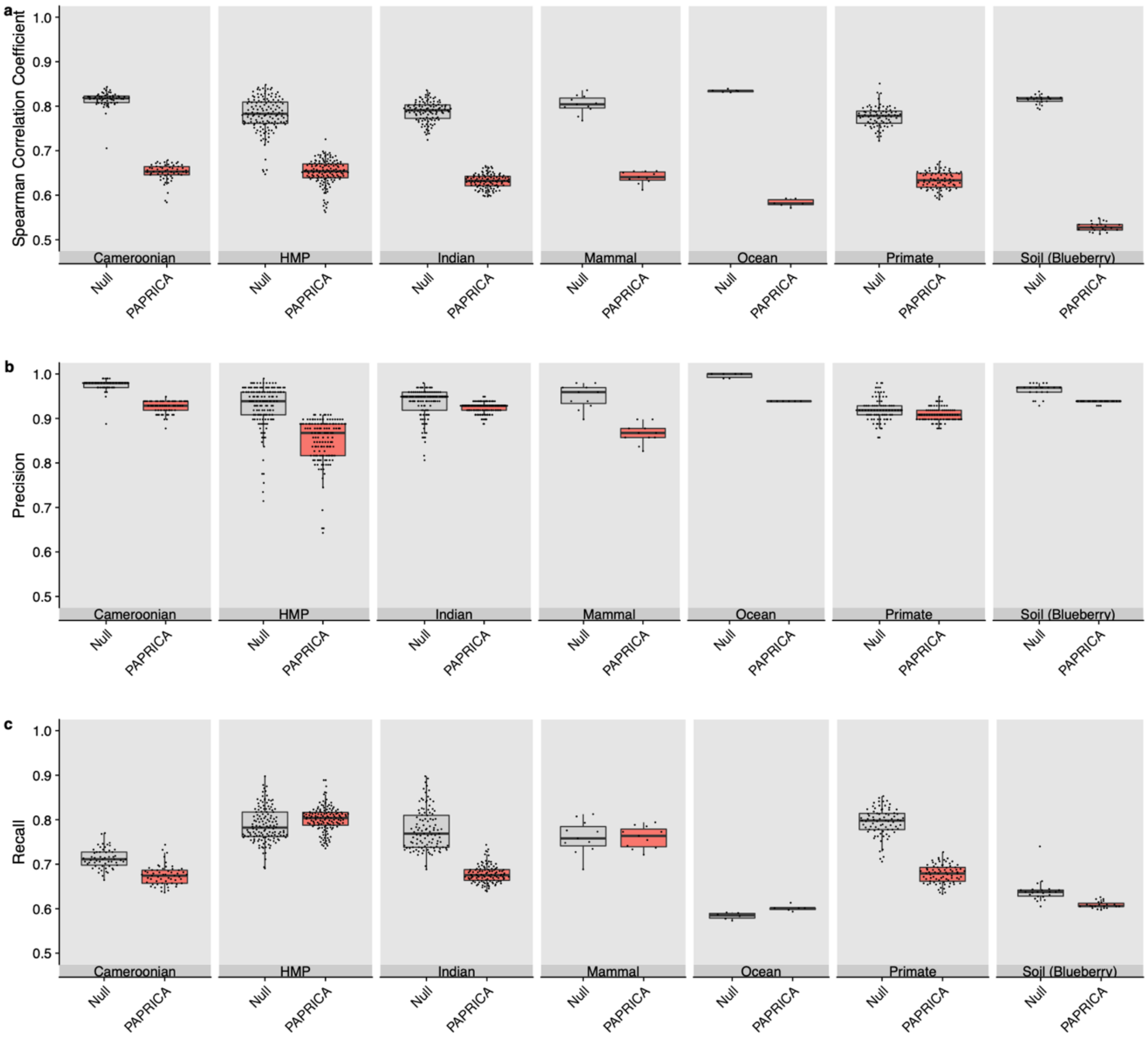


***Figure 4.15: Poor concordance of PAPRICA pathway abundance predictions compared to pathways identified by HUMAnN2 in shotgun metagenomics data** based on (a) Spearman correlation coefficients, (b) precision, and (c) recall. Only 193 MetaCyc pathways are considered here since this is the only set that could be identified by both PAPRICA and HUMAnN2. In contrast, 575 pathways could potentially be identified by both PICRUSt2 and HUMAnN2. The poor concordance shown here may be due to the differing approach for inferring pathway levels used by PAPRICA, which may mean that comparing these predictions with HUMAnN2 is not a fair evaluation.*

In addition to the differential abundance validations reported in the main text we also tested several other statistical methods. These methods included ALDEx2 and DESeq2 (with default options and options specifically recommended for microbiome data) as well as Fisher’s exact tests to test for differences in prevalence between samples. The raw results of these analyses are reported in Additional File 1 (available on Nature Biotechnology website). The DESeq2 results with both option selections were extremely similar, so we excluded the results based on non-default DESeq2 options from subsequent analyses. Overall, the different statistical methods result in different sets of significant functions for both the MGS data and the predicted metagenomes. In addition, the F1 scores based on the different differential abundance tools can vary substantially (Figure 4.16). Fisher’s exact tests for differential prevalence between sample groupings have especially low F1 scores (ranging from 0.02-0.29 for PICRUSt2). We believe that a more in-depth comparison of these statistical tests is beyond the scope of this manuscript, but this result does nonetheless highlight the challenge of reliably reproducing biomarkers from predicted metagenomes.

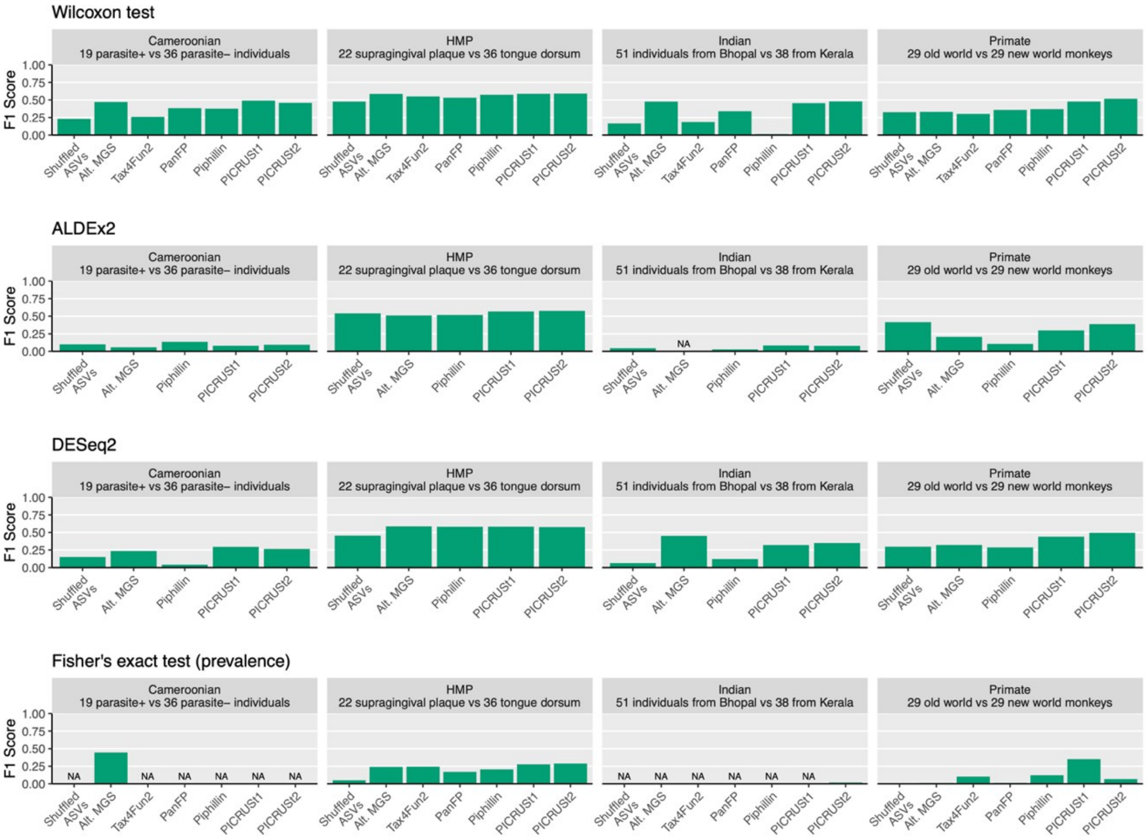


Figure 4.16: Differential abundance and prevalence of PICRUSt2 predicted KEGG orthologs (KOs) agree only marginally, but best overall, with shotgun metagenomics data*. The F1 score is plotted to summarize the agreement between the statistical testing of predicted KOs with KOs identified from HUMAnN2. PanFP and Tax4Fun2 are missing from the ALDEx2 and DESeq2 results because these statistical methods require count tables as input. Each of these statistical tests has a different null hypothesis, but the first three methods are all commonly used to perform differential abundance analyses on microbiome data. NA indicates cases where either the recall or precision could not be calculated (which means the F1 scores was not applicable). "Alt. MGS" refers to the alternative shotgun metagenomics processing pipeline.*

This challenge is especially apparent for the MetaCyc pathways (Figure 4.7), for which the PICRUSt2 predictions agreed only slightly better with the MGS pathways compared to the ASV shuffled datasets. The PICRUSt2 predicted pathways varied in precision and F1 score from 0.23-0.63 (mean=0.39; SD=0.19) and 0.23-0.62 (mean=0.41;

SD=17), respectively. The shuffled ASV-based pathway predictions had similar performance, ranging from 0.16-0.58 (mean=0.31; SD=0.19) and 0.22-0.60 (mean=0.34; SD=0.18) for the precision and F1 scores, respectively. These values are based on running Wilcoxon tests on the relative abundances of MetaCyc pathways. The alternative differential abundance (and prevalence) statistical methods resulted in similar concordances between the observed and ASV shuffled pathway predictions (Figure 4.9).

4.2.2 - Functional Profiling of Inflammatory Bowel Disease using PICRUSt2

To demonstrate the utility of PICRUSt2 in making functional inferences in a human health context where only amplicon sequencing is feasible, we profiled 27 ileal biopsy samples from subjects with Crohn's disease (CD) and 20 control subjects (non-IBD). These data are a subset of the Inflammatory Bowel Disease Multi'Omics Database (Lloyd-Price et al. 2019), which provides "multi-omics" data to identify host and microbial features associated with inflammatory bowel disease (IBD). Our analysis was based on 16S rRNA gene libraries collected from biopsy samples in this dataset. Importantly, MGS could not practically be performed on these (or any typical) biopsy samples due to the overwhelming predominance of human host DNA, which competes with microbial DNA for sequencing reads. As such, MGS data was produced only for subject stool samples for this study, which are analyzed here in addition to accompanying human RNA-seq transcriptional profiles (from the same biopsies) and metabolomic profiles from paired stool.

We first analyzed these data with a typical testing framework for identifying significant microbial features: testing for significantly differentially abundant ASVs clustered by taxonomy as well as inferred pathway abundances. Based on this standard approach we identified no pathways that significantly differed between CD and non-IBD subjects (FDR < 0.1). However, five taxa were identified with a differential relative abundance between non-IBD and CD subjects based on a lenient FDR q-value cut-off of 0.2. These included four taxa within the Clostridiales order, which were increased in relative abundance in control subjects, and the phylum Proteobacteria at higher relative abundance in CD subjects.

We next focused on the predicted MetaCyc pathways inferred by PICRUSt2 for ASVs underlying the significantly differentially abundant taxa: (1) 35 significant Clostridiales ASVs and (2) 192 Proteobacteria ASVs. For each predicted pathway in the community, we calculated the ratio of the abundance of that pathway contributed (i.e. potentially produced) by ASVs within the group of interest compared to the pathway's abundance contributed by all other ASVs. Based on this approach we identified three pathways significantly contributed by Proteobacteria (Figure 4.17a; Wilcoxon test FDR < 0.05). In addition, the relative contribution to 78 pathways by Clostridiales significantly differed between CD and non-IBD subjects (Wilcoxon test FDR < 0.05). These results demonstrate how PICRUSt2 stratified outputs allow integration of functional predictions with taxonomic findings as opposed to treating the two independently as in non-contributor stratified metagenomic analyses.

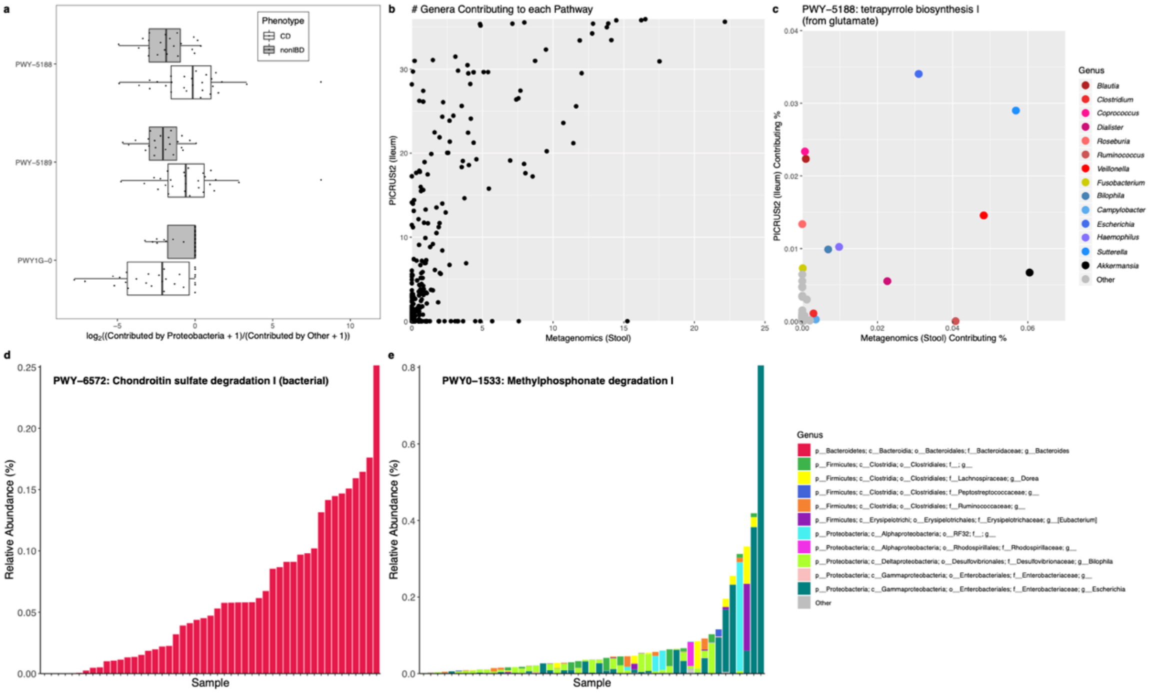


Figure 4.17: Applying PICRUSt2 to Crohn's disease cohort yields novel insights. *(a) Predicted MetaCyc pathways significantly contributed by Proteobacteria in the ileum of subjects with Crohn's disease (CD, n=27) compared to healthy controls (non-IBD, n=20) based on PICRUSt2 inference from 16S rRNA gene sequencing. These significant pathways are: PWY-5188 (tetrapyrrole biosynthesis I [from glutamate]), PWY-5189*

(tetrapyrrole biosynthesis II [from glycine]), and PWY1G-0 (mycothiol biosynthesis). (b) The mean number of classified genera contributing to each of the 313 MetaCyc pathways identified in either the ileum 16S rRNA gene sequencing (by PICRUSt2) or shotgun metagenomics sequencing (MGS) of the stool of the same CD subjects. (c) The top classified genera contributing to the relative abundance of PWY-5188 based on PICRUSt2 predictions of 16S rRNA gene sequencing in ileum tissue or MGS of the stool of the same subjects. Only the top 10 contributing genera are shown. Genera of the same phylum are shades of the same colour (Firmicutes and Proteobacteria are shades of red and blue, respectively). (d and e) Taxonomic breakdown of genera contributing to the predicted relative abundance of (d) PWY-6572 and (e) PWY0-1533 across all CD samples. Unclassified genera were included in these stacked bar charts, unlike in panel c. Only the top ten genera contributing to either PWY-6572 or PWY0-1533 are labelled.

We next investigated whether analysis of stool MGS data rather than ileal PICRUSt2 predictions resulted in substantially different conclusions, either due to methodology or body site. The number of classified genera contributing to each pathway within CD subjects differed strikingly depending on whether the contributors were identified through ileal 16S rRNA gene sequencing or stool MGS (mean difference: 7.3; SD: 10.2; Figure 4.17b). While a small number of pathways were uniquely identified by stool MGS, for most pathways a greater number of contributing taxa were identified by PICRUSt2. This is most likely due to the much greater taxonomic diversity accessible through amplicon databases than through reference genome isolates. This result could also be due to biological differences between the stool and ileal samples. However, we also identified a similar trend in the paired 16S rRNA gene-MGS datasets, although the magnitude of difference depended greatly on sequencing technology and sampling environment (Figure 4.18). Not just the number of taxa, but also their identities, differed between stool metagenomes versus biopsy inferences. For instance, for the tetrapyrrole biosynthesis I (from glutamate) pathway (PWY-5188), the top contributors differ between phylum Proteobacteria in biopsy 16S rRNA gene profiles, while *Akkermansia* (in phylum Verrucomicrobia) is the top contributor identified in the MGS data (Figure 4.17c).

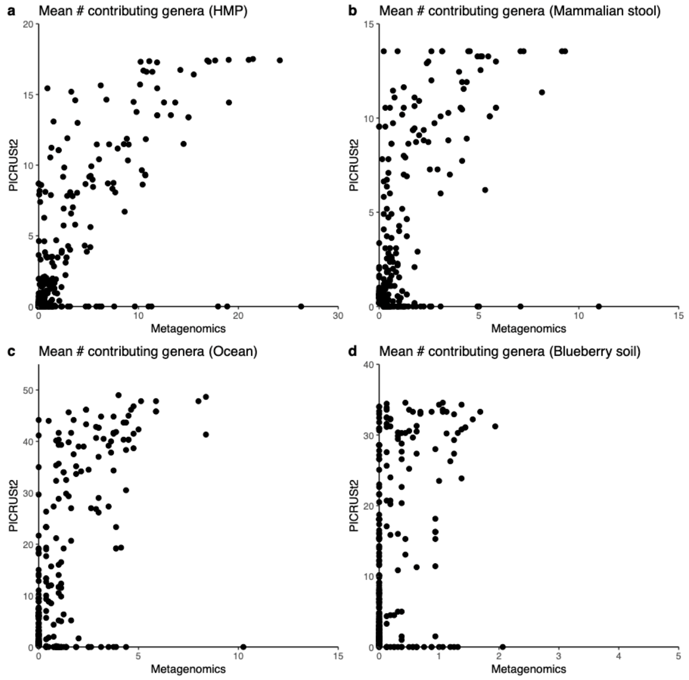
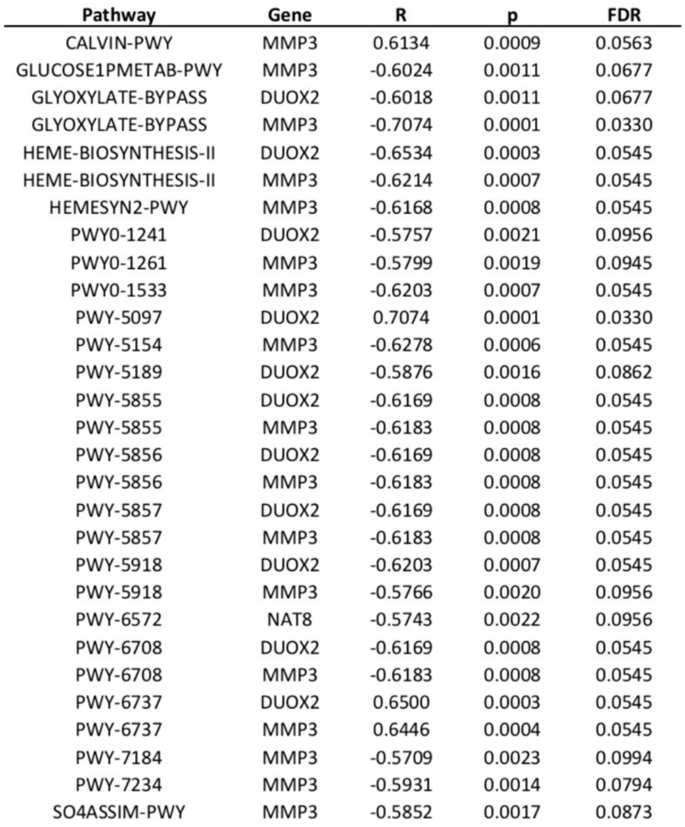


Figure 4.18: The mean number of classified genera contributing for every MetaCyc pathway identified in either the 16S rRNA gene sequencing (by PICRUSt2) or shotgun metagenomics sequencing (MGS) of four representative datasets*. (a) the Human Microbiome Project (HMP), (b) mammalian stool, (c) ocean, or (d) blueberry soil samples. This figure is meant to complement the analysis performed on the Crohn's disease biopsy data, because in this case the comparison is made between 16S rRNA gene sequencing and MGS on the same samples. Note that the number of characterized genera could be affected by sampling environment as labelled here, but importantly these datasets also differ in several technological ways as outlined in Table 4.2.*

Last, we tested whether the PICRUSt2 predictions give novel insights into CD biomarkers by associating 207 predicted pathways with both 583 metabolites from paired stool metabolomic profiles and the ileal transcription levels for six human host genes of interest. We identified no significant associations between predicted pathway and stool metabolite levels, but 29 associations between the predicted pathway and ileal transcript levels (FDR < 0.1; Table 4.3). Some of these significant associations are driven largely by individual taxa. For example, since the predicted relative abundance of chondroitin sulfate degradation is entirely contributed by *Bacteroides* (Figure 4.17d), the association between this pathway and NAT8 expression (partial R=-0.58) is trivially due to the relative abundance of this genus. However, not all significant associations are driven by individual taxa. For instance, there is no single taxon driving the association of the predicted relative abundance of the methylphosphonate degradation I pathway with MMP3 expression (partial R=-0.62; Figure 4.17e). This association is an example of PICRUSt2 predictions yielding potentially novel insights beyond those of the originating amplicon-based taxonomic profiles.

Table 4.3: *Predicted pathways significantly associated (based on partial Spearman correlation [R]) with gene expression levels of Crohn's disease biomarkers in ileal tissue of subjects with Crohn's disease*



Pathway	Gene	R	p	FDR
CALVIN-PWY	MMP3	0.6134	0.0009	0.0563
GLUCOSE1PMETAB-PWY	MMP3	-0.6024	0.0011	0.0677
GLYOXYLATE-BYPASS	DUOX2	-0.6018	0.0011	0.0677
GLYOXYLATE-BYPASS	MMP3	-0.7074	0.0001	0.0330
HEME-BIOSYNTHESIS-II	DUOX2	-0.6534	0.0003	0.0545
HEME-BIOSYNTHESIS-II	MMP3	-0.6214	0.0007	0.0545
HEMESYN2-PWY	MMP3	-0.6168	0.0008	0.0545
PWY0-1241	DUOX2	-0.5757	0.0021	0.0956
PWY0-1261	MMP3	-0.5799	0.0019	0.0945
PWY0-1533	MMP3	-0.6203	0.0007	0.0545
PWY-5097	DUOX2	0.7074	0.0001	0.0330
PWY-5154	MMP3	-0.6278	0.0006	0.0545
PWY-5189	DUOX2	-0.5876	0.0016	0.0862
PWY-5855	DUOX2	-0.6169	0.0008	0.0545
PWY-5855	MMP3	-0.6183	0.0008	0.0545
PWY-5856	DUOX2	-0.6169	0.0008	0.0545
PWY-5856	MMP3	-0.6183	0.0008	0.0545
PWY-5857	DUOX2	-0.6169	0.0008	0.0545
PWY-5857	MMP3	-0.6183	0.0008	0.0545
PWY-5918	DUOX2	-0.6203	0.0007	0.0545
PWY-5918	MMP3	-0.5766	0.0020	0.0956
PWY-6572	NAT8	-0.5743	0.0022	0.0956
PWY-6708	DUOX2	-0.6169	0.0008	0.0545
PWY-6708	MMP3	-0.6183	0.0008	0.0545
PWY-6737	DUOX2	0.6500	0.0003	0.0545
PWY-6737	MMP3	0.6446	0.0004	0.0545
PWY-7184	MMP3	-0.5709	0.0023	0.0994
PWY-7234	MMP3	-0.5931	0.0014	0.0794
SO4ASSIM-PWY	MMP3	-0.5852	0.0017	0.0873

4.2.3 - Validation of Fungal Metagenome Inference

We next assessed PICRUSt2's new capabilities for predicting metagenomes based on fungal amplicon sequencing of either the 18S rRNA gene or internal transcribed spacer (ITS) regions. Data used for these predictions included EC abundances from 294 fungal genomes from the 1000 Fungal Genomes Project that were publicly available as of November 16, 2018 and passed quality control criteria. Unlike the prokaryotic database, only a minority of 18S rRNA gene and ITS sequences were redundant across genomes (7.5% and 8.5%, respectively). A total of 7 and 8 phyla as well as 183 and 209 genomes are represented in the ITS and 18S rRNA gene databases, respectively (see Table 4.4 for the database counts at all taxonomic levels).

Table 4.4: *Counts of taxa in the tested internal transcribed spacer (ITS) and 18S rRNA gene databases*

Database	Phyla	Classes	Orders	Families	Genera	Species	Genomes
ITS	7	26	48	89	130	174	183
18S rRNA gene	8	27	52	102	151	193	209

We first evaluated the performance of PICRUSt2 18S rRNA gene and ITS metagenome predictions by leave-one-out cross-validation of individual genomes. Spearman correlations were calculated between the predicted EC abundance profiles in each held-out genome and the EC abundances in the known genome. For both the 18S rRNA gene (Spearman Rho mean=0.821; SD=0.141) and ITS databases (Spearman Rho mean=0.822; SD=0.135), the predictions were significantly better than the null expectation (Wilcoxon test P < 0.001; Figure 4.19). Similar to the 16S rRNA gene-based validations, genome prediction accuracy decreased as reference genomes were artificially held out of the training dataset at increasing taxonomic scale, suggesting that overall accuracy is hampered for those lineages without comprehensive representative genomes.

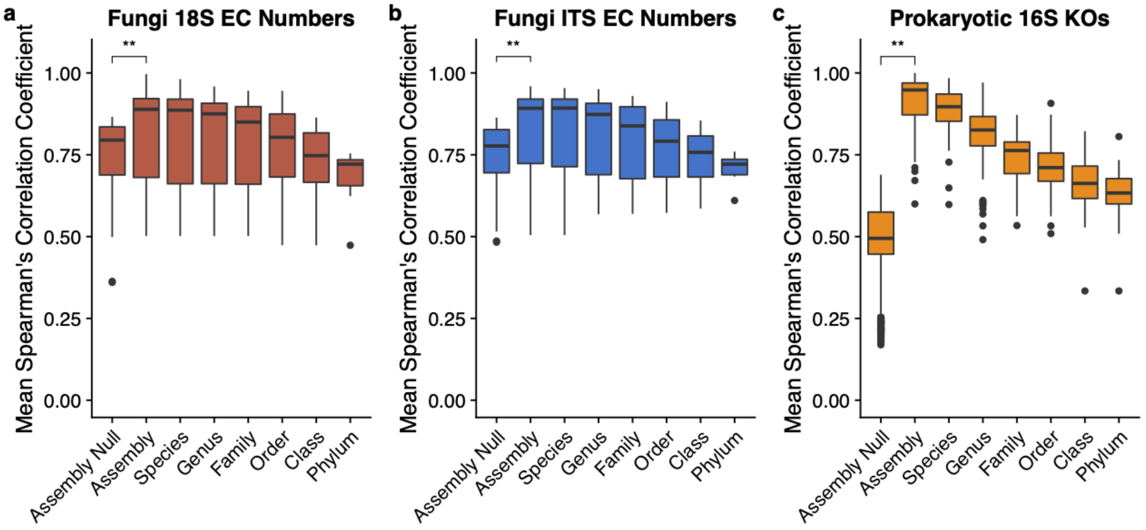


Figure 4.19: Validating test fungi databases and key prokaryotic database discussed in main-text with genome holdout analysis. *(a) fungi 18S rRNA gene-predicted Enzyme Commission (EC) Numbers, (b) fungi internal transcribed spacer (ITS)-predicted EC Numbers, and (c) 16S rRNA gene-predicted KEGG orthologs (KOs). The prokaryotic database is included here as a point of reference for the fungi databases. For each database shown above, predictions were made for all genomes within each clade at a given taxonomic level after pruning all those genomes from the reference tree. The mean Spearman correlation coefficient between the predicted and expected gene family abundances was then calculated for each clade. The "Assembly" level refers to individual genomes. The "Assembly Null" category corresponds to the correlation between the gene family abundances for each genome and the mean abundance of gene families across all genomes. The ** annotation indicates a significant Wilcoxon test (P < 0.001). Note that for the prokaryotic KO analysis that a maximum of 100 clades at each taxonomic level were selected randomly for this analysis to decrease computation time. In this plot the points correspond to outliers outside the boxplot whiskers only and all other points are not shown.*

Next, we evaluated the performance of fungal EC predictions on two amplicon sequencing datasets with paired MGS data using the same approach as with 16S rRNA gene predictions. The two validation datasets used were the same 22 blueberry soil samples described in the main-text, which also underwent 18S rRNA gene sequencing (Yurgel et al. 2017), and eight wine fermentation internal transcribed spacer (ITS1) sequencing samples (Sternes et al. 2017). EC predictions for both of these datasets were significantly more similar to the MGS gold-standard compared to the null expectation based on correlations (Figure 4.20a; P=9.5x10^{-7} and P=7.8x10^{-3} for the blueberry soil and wine fermentation datasets respectively). However, the correlations observed for these datasets was substantially lower than for the 16S rRNA gene-based validations, which is to be expected as these metagenomes include a substantial amount of functions from non-fungal origins. The mean correlations in the blueberry soil dataset were 0.340 (SD: 0.016) and 0.365 (SD: 0.019) for the null expectation and PICRUSt2 predictions, respectively. There was a larger mean difference for the wine fermentation dataset where the mean correlations were 0.492 (SD: 0.004) and 0.611 (SD: 0.028) for the null expectation and PICRUSt2 predictions, respectively. Interestingly, the correlations based on predicted MetaCyc pathway abundances were slightly lower than the null values for the blueberry soil 0.461 (SD: 0.009) and wine fermentation 0.501 (SD: 0.008) datasets (Figure 4.20b).

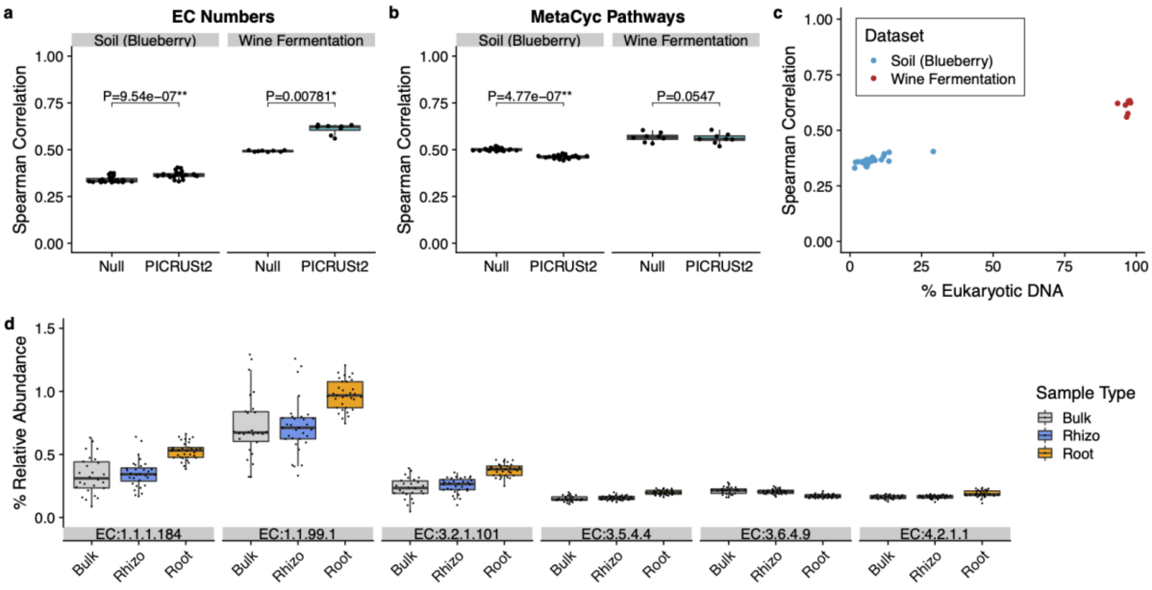


Figure 4.20: PICRUSt2 18S rRNA gene and internal transcribed spacer predictions exceed null prediction accuracy. *(a) Spearman correlation coefficients between amplicon predicted Enzyme Commission number abundances and gold-standard shotgun metagenomic (MGS) profiles from the same biological samples. (b) Spearman*

correlation coefficients between amplicon-predicted MetaCyc pathway abundances and MGS on the same biological samples. For panels a and b, the P-values of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (and ** correspond to P < 0.05 and P < 0.001, respectively). (c) The Spearman correlation coefficients as shown in panel a re-plotted against the percent of non-animal and non-plant eukaryotic DNA within each sample. The blueberry soil dataset consists of 22 18S rRNA gene sequencing samples and the wine fermentation dataset consists of eight internal transcribed spacer region one (ITS1) sequencing samples. (d) The relative abundance of significantly informative EC numbers (P < 0.001) in a Random Forest model for distinguishing blueberry soil and root samples by sample type. Only significant EC numbers with a mean relative abundance greater than 0.15% are shown. The EC numbers shown correspond to carbonyl reductase (NADPH) (EC:1.1.1.184), choline dehydrogenase (EC:1.1.99.1), mannan endo-1,6-alpha-mannosidase (EC:3.2.1.101), adenosine deaminase (EC:3.5.4.4), chaperonin ATPase (EC:3.6.4.9), and carbonate dehydratase (EC:4.2.1.1). There are 26, 33, and 32 samples for the bulk, rhizosphere (rhizo), and root environments, respectively.*

One potential factor affecting these results is the percent of eukaryotic DNA within the MGS data. A low percent of eukaryotic DNA would result in prokaryotes mainly contributing to gene family abundances. The percent of eukaryotic DNA (after excluding plant and animal DNA) within the MGS datasets differed dramatically between the blueberry soil (mean: 8.17%; SD: 5.82) and wine fermentation datasets (mean: 96.72%; SD: 1.44; Figure 4.20c). This low percent of eukaryotic DNA in the blueberry soil dataset could partially account for the relatively poor performance we observed.

To investigate whether the PICRUSt2 predictions for the blueberry soil dataset can nonetheless distinguish sample groupings, we ran PICRUSt2 on 18S rRNA gene sequencing data from additional blueberry soil samples with no matching MGS data as well as blueberry root samples from the same sampling location (Yurgel et al. 2018). We then generated a Random Forest model to identify the most informative predicted EC numbers that distinguish samples by whether they were taken from a bulk soil, rhizosphere, or root environment. This model resulted in a classification accuracy of

68%, which was substantially better than the random expectation (accuracy 36%) and identified 32 significantly informative EC numbers (P < 0.001; Figure 4.20d). Importantly, although this model was more accurate than random, this analysis does not prove that the predicted EC numbers themselves were accurately predicted. Instead, these combined analyses demonstrate a proof-of-concept that fungal metagenomes can be predicted more accurately than expected by chance. However, due to the low correlation values we observed, these predictions are unlikely to provide reliable biological insights in practice.

4.2.4 – Relative Contributions of PICRUSt2 Updates to Improved Performance

After submitting our manuscript describing PICRUSt2 my supervisory committee requested that I do an additional analysis to better compare the relative contributions to performance of the key improvements made in PICRUSt2. Although this was not included in the published manuscript, this analysis provides valuable insight into these improvements.

I re-ran PICRUSt2 on the seven 16S validation datasets used in the manuscript while varying three settings: (1) the reference database, (2) the HSP approach, and (3) the tree-building approach. The reference databases corresponded to the default PICRUSt2 database files and the PICRUSt1 database files formatted for PICRUSt2 (including the Greengenes phylogenetic tree and multiple-sequence alignment). The two HSP approaches compared were maximum parsimony (MP; the default in PICRUSt2) and phylogenetic independent contrasts (PIC; the default in PICRUSt1). Last, we compared the utility of running sequence placement with EPA-ng to re-building the tree for each dataset with FastTree. We compared the concordance between the MGS-based KO profiles and the predicted profiles output by PICRUSt2 based on Spearman correlations.

The PICRUSt1 database includes many fewer KOs compared to the updated database in PICRUSt2 so I compared the PICRUSt2 outputs based on the above input settings based on first the KOs that overlap between both tools and also KOs found in PICRUSt2. In both cases, the KOs evaluated were restricted by those that could be output by HUMAnN2 for the matching MGS data as well, as in the main text.

The predicted profiles based on the PICRUSt1 and PICRUSt2 database files differ in almost every case for both human (Figure 4.21) and non-human datasets (Figure 4.22). In particular, using the PICRUSt2 database provides a mean performance increase of 4.53% (SD=3.34%) across all samples and input settings. Although the concordance based on the null expectations differ in several cases, this is inconsistent: overall the null expectations are similar. Based on this set of restricted KOs that are present in each database and in HUMAnN2 (6,379), there is little shift in the concordance of the predicted profiles based on the HSP and tree-building settings.

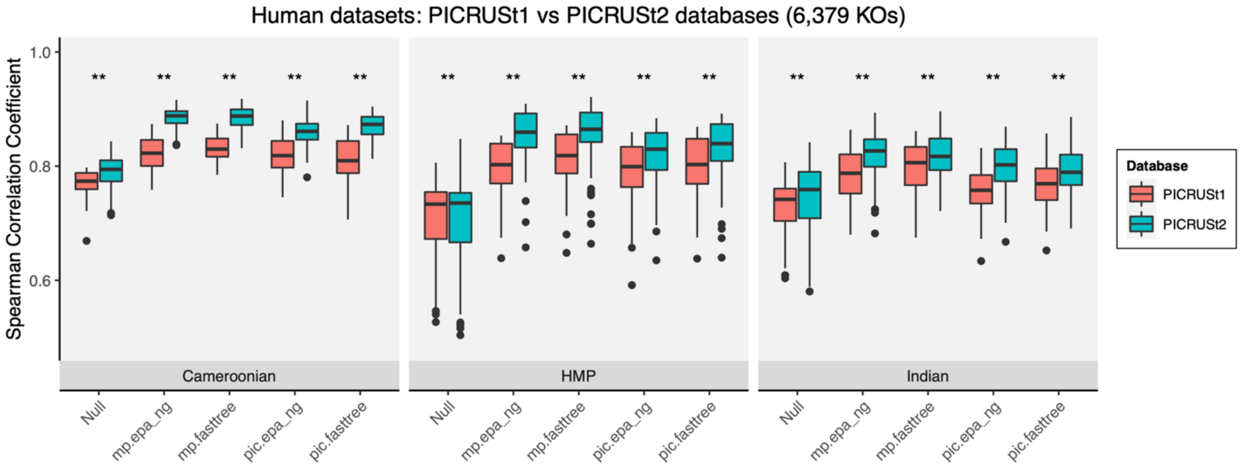


Figure 4.21: PICRUSt2 KO prediction performance on human-associated datasets based on PICRUSt1 and PICRUSt2 default databases*. Spearman correlations between predicted KEGG ortholog (KO) abundance profiles and shotgun metagenomics-based KO profiles computed with HUMAnN2. In contrast to the main text only PICRUSt2 is being compared in this figure. However, different database inputs are specified: PICRUSt1 (red) and PICRUSt2 (green). The null expectation was computed separately for each database. The hidden-state prediction and tree-building approaches are also varied across these results. The hidden-state prediction approach was either maximum parsimony (mp) or phylogenetic independent contrasts (pic). The tree-building approach was either sequence placement into the reference tree with EPA-ng (epa_ng) or building a de novo tree with FastTree (fasttree).*

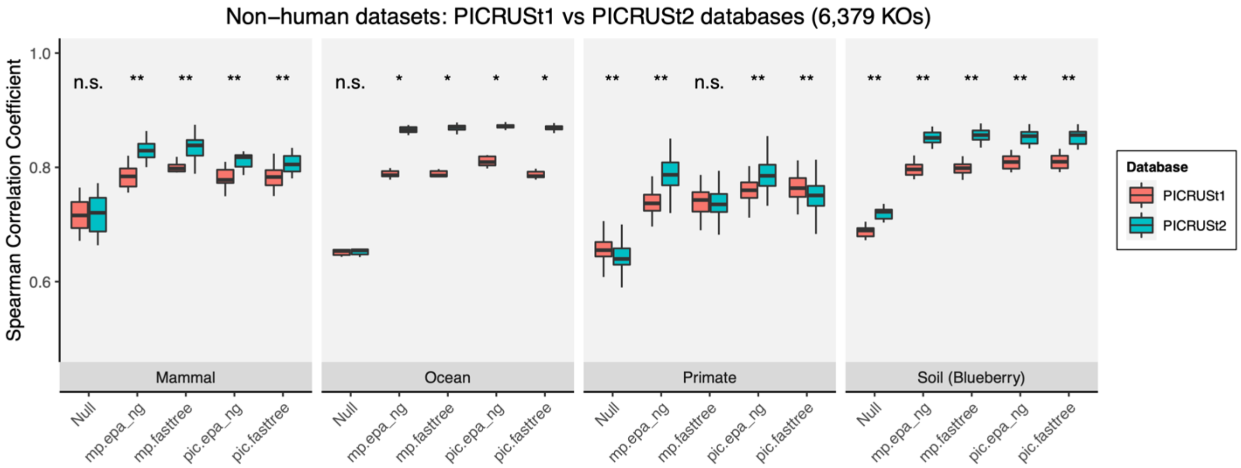


Figure 4.22: PICRUSt2 KO prediction performance on non-human-associated datasets based on PICRUSt1 and PICRUSt2 default databases*. These are the results for the non-human datasets based on the same comparisons described in Figure 4-21.*

Clearer differences are apparent when this comparison includes all 9,915 KOs that could be output by both PICRUSt2 and HUMAnN2. In particular, the MP hidden-state prediction approach results in higher correlations for both human (Figure 4.23) and non-human (Figure 4.24) associated datasets. The mean percent increase across all samples in the mp.epa_ng output profiles compared to pic.epa_ng profiles was 6.43% (SD=3.02%). This increased performance was also largely observed for predicted profiles based on FastTree (mean increase of 5.01% [SD=2.55%]). However, there were no consistent differences between MP-based predictions based on either EPA-ng or FastTree built phylogenetic trees (mean increase of 0.76% [SD=2.37]). This result highlights that based on these validation datasets that sequence placement did not provide any advantage over simply creating a de novo phylogenetic tree. In addition, the key improvement that provided a boost to performance, in addition to the new database, is the use of maximum parsimony for HSP by default in PICRUSt2.

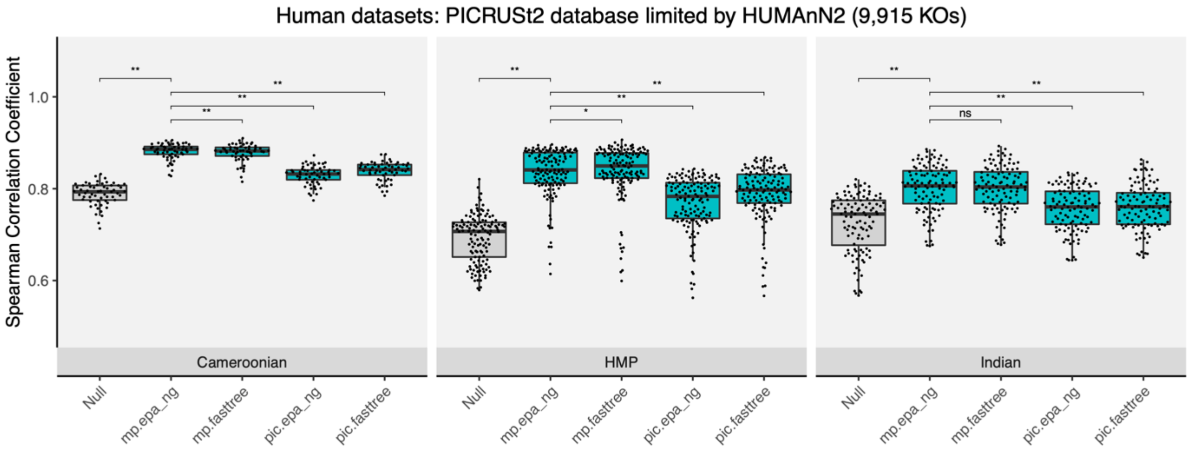


Figure 4.23: PICRUSt2 KO prediction performance on human-associated datasets based on varying parameter settings*. Spearman correlations between predicted KEGG ortholog (KO) abundance profiles and shotgun metagenomics-based KO profiles computed with HUMAnN2. The hidden-state prediction and tree-building approaches are varied across these results. The hidden-state prediction approach was either maximum parsimony (mp) or phylogenetic independent contrasts (pic). The tree-building approach was either sequence placement into the reference tree with EPA-ng (epa_ng) or building a de novo tree with FastTree (fasttree).*

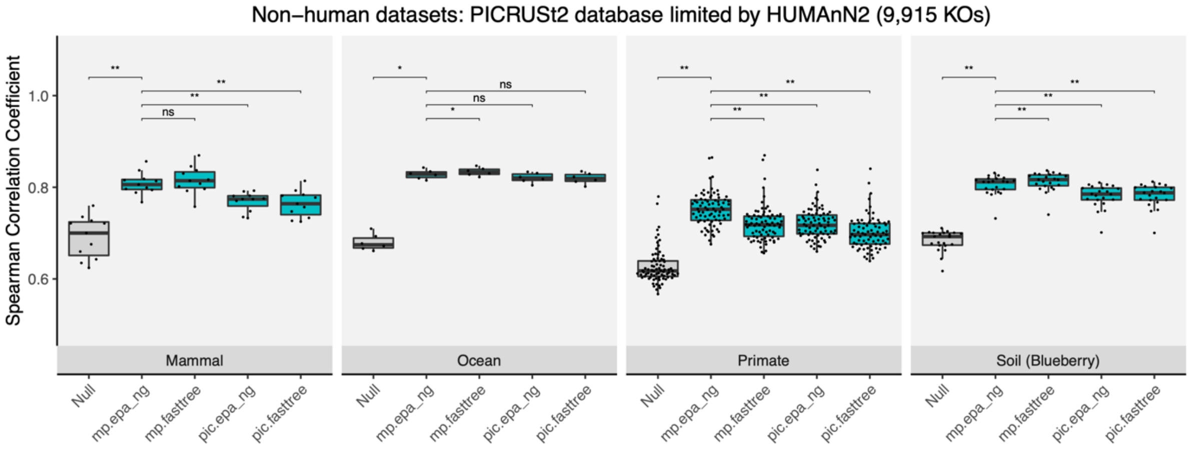


Figure 4.24: PICRUSt2 KO prediction performance on human-associated datasets based on varying parameter settings*. Spearman correlations between predicted KEGG*

ortholog (KO) abundance profiles and shotgun metagenomics-based KO profiles computed with HUMAnN2. These are the results for the non-human datasets based on the same comparisons described in Figure 4-23.

Chapter 5 – Phylogenetic Organization of Metagenome Signals

The below sections correspond to a draft manuscript soon to be submitted for publication. I am the lead author of this manuscript and conducted all analyses and programming. I was also involved with formulating the underlying ideas for this project in collaboration with the supervisors of this work: Dr. Elhanan Borenstein (Tel Aviv University) and Dr. Morgan Langille (Dalhousie University). The methods section for this manuscript is presented separately in Chapter 2.

5.1 - Abstract

Microbiome functional data are frequently analysed to identify associations between microbial gene families and sample groupings of interest. This is most commonly performed with approaches focused on the metagenome-wide relative abundance of microbial functions. Although this method can provide valuable insights, these differential abundance tools often provide drastically different profiles of significant associations. In addition, it is impossible to distinguish different possible explanations for variation in community-wide functional profiles by looking at functions alone. To help address these problems, we have developed a novel framework to expand taxonomic balance tree approaches to enable enriched functions to be more accurately identified. The key focus of our approach is on identifying functions that are consistently enriched in sample groupings amongst independent lineages of taxa. Our implementation of this framework is available in the R package POMS. Based on simulated data we demonstrate that POMS more accurately identifies gene families under selection compared to a representative differential abundance approach. POMS also identifies enriched gene families in several case-control metagenomics datasets that are putative targets of strong selection on the overall microbiome. This framework is unable to identify all likely functional enrichments, but the top enrichments are more interpretable and conservative than existing differential abundance methods. More generally, POMS is a novel method for exploring microbiome functional data, which could be used to complement standard analyses. POMS is freely available at: https://github.com/gavinmdouglas/POMS.

5.2 - Introduction

Microbiome sequencing has been applied to characterize myriad environments and is typically analyzed based on the relative abundance of microbial features. These features are split into taxa, the microbes present, and functions, the genes they encode and pathways that might be active. Both data types have been leveraged to make valuable observations, but they are typically analyzed independently of each other because joint analyses of these datatypes are challenging. However, linking these data types is often required to make coherent interpretations of microbiome data. Nonetheless, as previously noted (Manor & Borenstein 2017b), due to the complexity of these data they are often only linked anecdotally. In particular, mere co-occurrence between particular taxa and functions is often discussed as potential evidence of a direct link between them.

This approach is unreliable because different genic relative abundances between communities can be driven by numerous processes that cannot be distinguished by considering functional profiles alone. For example, increases in the relative abundance of *Escherichia coli* are associated with traveller's diarrhea (Nakamura et al. 2011). However, the corresponding microbial functions that increased in this case would be similar to those increased if a wider range of Enterobacteriaceae also increased. For instance, shifts in colonocyte metabolism have been associated with blooms of facultative anaerobes in general, including Enterobacteriaceae, at the expense of obligate anaerobes (Litvak et al. 2018). It would be difficult to distinguish these explanations for increased abundances of Enterobacteriaceae-related functions based on analyzing such functional profiles alone.

Instead, information on the taxonomic contributors to variable functions is needed to interpret these signals. One approach for addressing this issue is to identify the taxonomic contributors to functional shifts identified by differential abundances tests. FishTaco is a useful tool for generating this information (Manor & Borenstein 2017b). However, a limitation of FishTaco is that it is a post-hoc approach that is applied after identifying significant microbial functions. Ideally functional and taxonomic data would be integrated while testing for differential functions to better identify strong enrichment candidates. In particular, in cases where a small number of closely related taxa encode a function linked with a given context, it is challenging to identify a clear functional candidate driving the association. In other words, closely related taxa often share many functions in common and so it is difficult to hypothesize which functions are pertinent to the given context without extraneous information. In contrast, clearer interpretations of the data are possible by first organizing microbial functions by how they are linked with differentially abundant taxa.

One method that explicitly integrates functional and taxonomic data is phylogenize (Bradley et al. 2018; Bradley & Pollard 2020). This approach identifies functional associations based on the prevalence of taxa that encode each gene family. This is performed with phylogenetic linear models, which account for genetic similarity of co-occurring taxa due to their shared evolutionary history. Significant gene families and pathways identified by phylogenize as being contributed by a diverse set of taxa from within a given phylum. Although this approach is a key improvement over past methods, it also has limitations. In particular, it focuses solely on the binary presence/absence of gene families and does not incorporate taxa relative abundances.

This choice to focus on presence/absence profiles rather than abundances is largely because raw microbiome data cannot be used with standard statistics due to their compositionality. Fortunately, there is growing interest in improved compositional approaches for analyzing microbiome data. Analyzing ratios of microbiome feature relative abundances rather than independent features has recently been proposed as a solution to the compositionality problem inherent in sequencing data (Gloor et al. 2016; Morton et al. 2019). One major challenge of this approach is that it is unclear which features should be used for stably computing these ratios. Without a clear reference feature to compare to, one proposed solution has been to compare ratios of taxa on each side of every node in a phylogenetic tree (Silverman et al. 2017). These ratios are calculated based on the isometric log-ratio transformation between taxa on each side of a node. This general approach is now commonly referred to as analyzing balance trees (Morton et al. 2017). The downside of these approaches is that although they are statistically robust, they are less biologically satisfying because it is often unclear how to interpret differences in the ratio of taxon abundances. Nonetheless, it may be that inferences based on reference frames are the primary type of findings that can be reliably identified with microbiome sequencing data (Morton et al. 2019).

Here, we introduce the idea of testing for functional enrichment between reference frames, and specifically based on phylogenetic-based reference frames. This approach has the added benefit of providing improved interpretability over taxonomic reference frames alone. Our approach is focused on identifying consistent functional enrichments over phylogenetic balance trees, which is implemented in the Phylogenetic Organization of Metagenome Signals (POMS) R package. Although it is impossible to absolutely distinguish scenarios where functional enrichments are biologically relevant or occurring by chance, it is possible to leverage taxonomic data to give more weight to one scenario over another. In particular, it is possible to identify cases where multiple taxa that encode a given function are consistently associated in the same direction with a sample grouping. Such cases provide more support to the scenario where specific functions are under selection rather than just hitchhiking with taxa that might be blooming for other reasons.

POMS is primarily intended for identifying such cases, which are potential microbiome-wide targets of natural selection. The key assumption of this approach is that gene families and pathways that provide an advantage in a specific context are encoded by a range of phylogenetically disparate taxa. This means that many gene families with limited taxonomic breadth cannot be identified as highly enriched based on this approach. Therefore, POMS is not a replacement for current approaches, but is instead intended as a complementary tool. POMS also enables exploratory data analysis of all functional enrichments between sample groupings over a phylogenetic tree of taxa.

We demonstrate the POMS framework outperforms existing differential abundance tools in simulated datasets. In particular, POMS is a better approach for identifying gene families under strong selection across multiple lineages, which are challenging to identify using current 'bag-of-genes' approaches for metagenome analyses. POMS also identifies several gene families that are strongly associated with disease state in three case-control datasets. These top enriched gene families are reasonable candidates as potential targets for strong selection, which highlights that POMS provides improved interpretability over existing methods. Although our approach has limitations, it is a valuable proof-of-concept that integrating functional enrichments into reference frame analyses provides improved interpretation and novel insights.

5.3 – Results

5.3.1 – POMS Workflow

Phylogenetic Organization of Metagenome Signals (POMS) is a novel bioinformatics workflow for identifying microbial gene families linked with sample groupings of interest. The POMS R package is available at https://github.com/gavinmdouglas/POMS. The key assumption of this approach is that gene families and pathways that provide an advantage in a specific context are encoded by a range of phylogenetically disparate taxa.

The POMS workflow implements a modified balance tree framework for analyzing compositional data (Figure 5.1). The isometric log-ratios (ILR) of the relative abundances of all taxa on the left-hand and right-hand side of each node is first computed, which are referred to as balances (see Methods). Nodes with significantly different balances between sample groupings are then identified based on Wilcoxon rank-sum tests. Because the ILR cannot be reliably computed based on few taxa, by default we restrict the set of evaluated nodes to those with at least 10 tips (e.g. genomes) on both the left and right-hand of the node. Our extension to the general balance tree approach includes three steps after identifying these significant phylogenetic balances. First, for each significant node we determine whether the ILRs are significantly higher or lower for the focal sample group, which throughout this manuscript typically correspond to disease samples. We then compute Fisher's exact tests for each gene family based on the counts of taxa that either do or do not encode the gene family on either side of the significant node. This step also allows us to determine whether a significant gene family is either enriched or depleted in genomes that are at relatively higher abundance in the focal sample group. Last, we summarize this information over all significant nodes and gene families to categorize gene families that are more likely to specifically be under selection.

The key summary metric is the number of significant nodes where the gene family is either enriched or depleted from the perspective of the focal group. This is described as genes that are positively or negatively enriched, respectively, throughout this work. Significantly enriched gene families can be identified based on a pseudo-null distribution approach. This approach computes the probability of acquiring as equally extreme or greater enrichment pattern for a given gene family based on a random set of significant nodes (see Methods). This approach provides insight into highly enriched gene families but is imperfect (see Discussion). A simpler alternative approach, particularly for exploratory analyses, is to use hard cut-offs of enrichment for defining highly enriched gene families.

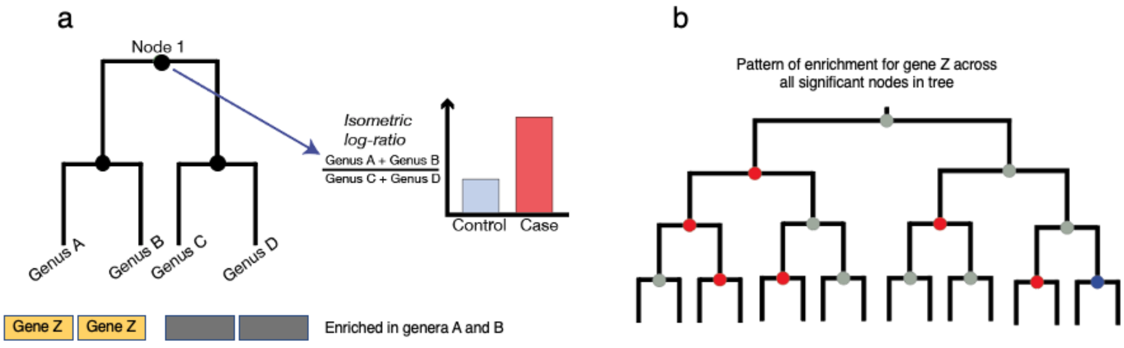


***Figure 5.1: POMS workflow overview**. (a) Significant nodes are first identified based on differing sample balances, which represent the isometric log-ratio of the abundances of all taxa between the left and right-hand side of each node. Genes enriched on one side of each significant node compare with the other side are then identified. The direction of this enrichment refers to whether the gene is enriched taxa that are relatively higher in one sample group over another. In the above example, gene Z is positively enriched in taxa that are at relatively higher levels in case samples (based on the balances at this node). (b) Patterns of enrichment for a given gene are then summarized across all significant nodes on the tree. This can be done by visualizing the enrichments on the tree itself or alternatively by analyzing summary metrics, such as the absolute enrichment, which are discussed in the main text. Our approach for visualizing this data throughout this manuscript is to colour nodes depending on the enrichment pattern for a given gene. Grey nodes are significantly different nodes that do not show differential enrichment of the gene. Blue and red nodes correspond to those that are enriched on the side of the node that is relatively higher in control and case samples, specifically.*

The POMS workflow is written modularly and is agnostic to the processing pipeline used to generate the input files. The key required input files are tables of taxonomic and functional abundances and a fully-resolved phylogenetic tree of the taxa. The functional abundance table could be based on MAGs, known genome annotations for strains identified in an environment, or alternative custom analyses. Metagenome predictions based on 16S rRNA gene sequencing could also be used as input. Because POMS is modular, users can bypass or use alternative approaches for custom analyses.

5.3.2 – Validating POMS with MAG-Based Simulated Data

Our first approach for validating POMS focused on in silico alterations to samples containing metagenome-assembled genomes (MAGs). The MAGs from MGS control samples, with corresponding relative abundances and phylogenetic tree, were taken from a large MGS meta-analysis (Almeida et al. 2019). These MAGs had previously been annotated with KEGG orthologs (KOs) (Kanehisa et al. 2016), which were the gene families we focused on for this analysis. We subsampled 704 control samples into two equally sized groups 500 times to create random test datasets. As expected, POMS detected extremely few significantly different nodes and functions across these replicates. This is reflected by the observation that only 0.14% of all tested KOs were enriched at any significant nodes, and only at a maximum of two nodes. Similarly, a more standard approach for performing differential abundance testing, the Wilcoxon rank-sum test, identified no significantly different functions based on these raw random datasets ($P < 0.05$).

We then conducted two straight-forward sets of simulations to compare the performance of POMS and Wilcoxon rank-sum tests (Figure 5.2). We first randomly selected a KO encoded by at least one genome for each of the previous 500 random test datasets, which we refer to as the focal genes. To simulate selection acting upon the genomes encoding a focal gene we multiplied the relative abundance of all genomes that encode the gene by ten-fold. This was performed in one sample group only and corresponds to the 'focal gene' replicates. We also conducted similar simulations where instead of genomes that encode the focal gene being under selection, random genomes instead underwent selection. For each matching focal gene replicate, the same number of random genomes were perturbed as encode the gene. These profiles are referred to as the 'random taxa' results below. In each case, we applied POMS and the Wilcoxon rank-sum test to identify KOs that differed between the two sample groupings for each simulation replicate.

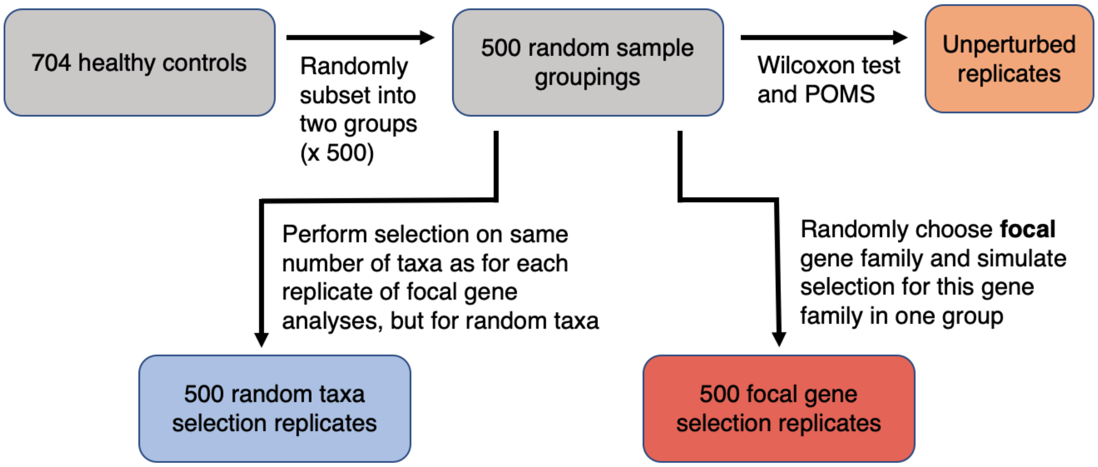

704 healthy controls
Randomly subset into two groups (x 500)
500 random sample groupings
Wilcoxon test and POMS
Unperturbed replicates
Perform selection on same number of taxa as for each replicate of focal gene analyses, but for random taxa
Randomly choose **focal** gene family and simulate selection for this gene family in one group
500 random taxa selection replicates
500 focal gene selection replicates

Figure 5.2: Workflow diagram for metagenome-assembled genome-based simulations*.* *Intermediate outputs are indicated in grey boxes and final simulation outputs are indicated in coloured boxes.*

For these analyses we identified enriched KOs based on hard cut-offs. Below we refer to functional enrichment interchangeably with statistical significance for ease of reading, but importantly the highly enriched genes output by POMS are based on identifying outliers above the selected cut-off rather than a statistical test. Due to variation in dataset characteristics the selection of specific cut-offs must be performed separately for each new dataset when using this approach. Based on the unperturbed datasets, POMS did not identify KOs that were consistently enriched in three nodes positively or three nodes negatively (and never in the opposite direction); therefore, we primarily used this cut-off for binning enriched genes for this analysis. However, we also tested two other cut-offs to ensure our results were robust to this selection. The first alternative cut-off was to classify enriched genes as those enriched in three or more nodes in one direction and at most one node in the opposite direction. The final alternative cut-off evaluated was to classify enrichment as KOs enriched in two or more nodes in one direction and never in the opposite direction.

Significant nodes were first identified and then, based on the above cut-offs, significant KOs were also identified for each simulation replicate. For the random taxon simulations there was no correlation between the number of significant nodes per replicate and the number of significant KOs (Spearman R=-0.055; P=0.26; Figure 5.3a).

Note that this correlation was computed only for replicates with at least five significant nodes, because there is a trivial positive association at lower numbers simply because at least some significant nodes are required for gene families to be identified as enriched with POMS. In contrast, there is negative relationship between the number of significant KOs and the number of significant nodes within the focal gene simulations (Spearman R=-0.54; P < $2.2*10^{-16}$; Figure 5.3b). In other words, under this simulation approach, replicates with the highest number of significant KOs were associated with fewer significant nodes. This observation is consistent with the number of significant KOs under the focal gene simulations increasing largely due to other factors besides the number of significant nodes. For instance, the proportion of significant KOs was also weakly correlated with the number of MAGs that encoded the focal gene per replicate (Spearman's correlation coefficient = 0.127; P=0.004). This overall finding of a negative association between the proportion of significant KOs and the number of significant MAGs was robust to alternative cut-offs, (Figure 5.3, panels c – f).

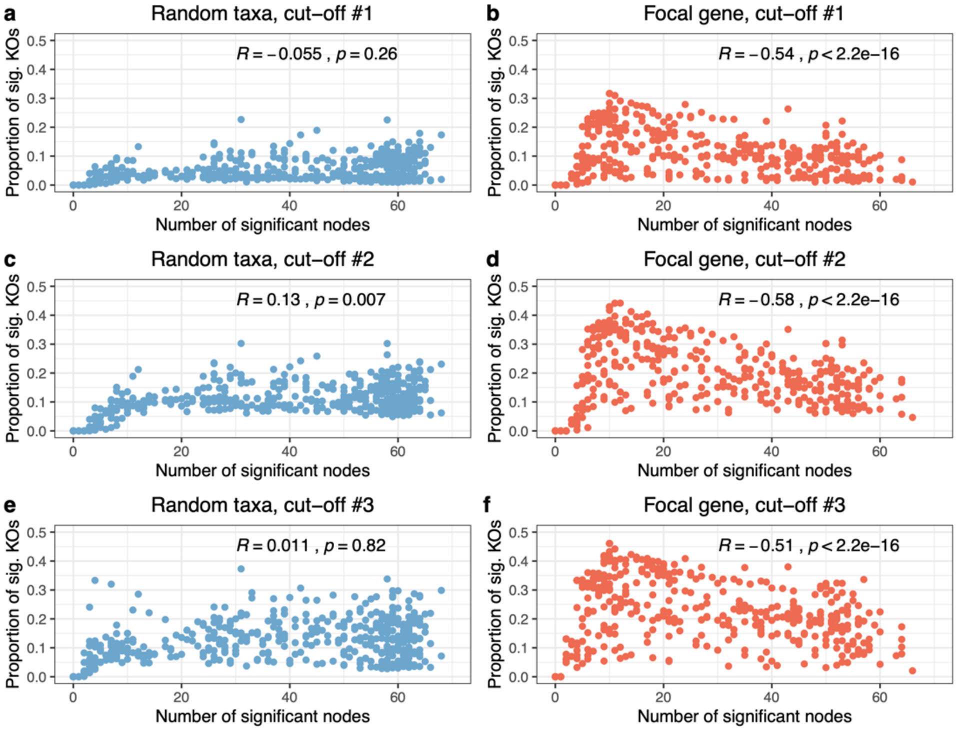


Figure 5.3: Proportion of significant gene families compared with the number of significant nodes in MAG-based simulations based on POMS cut-offs*. Results are*

shown separately for all (a, c, and e) taxa and (b, d, and f) gene-based simulation replicates. Each point corresponds to one of 500 replicates. Pearson correlation coefficient and P-values are indicated in each panel. Cut-off #1, the primary cut-off used for these evaluations, refers to defining enriched KOs as those enriched in at least three significant nodes in the same direction and never in the opposite direction. Cut-off #2 (panels a and b) refers to defining enriched KOs as those enriched in at least three significant nodes in the same direction and at most once in the opposite direction. Cut-off #3 (panels c and d) refers to defining enriched KOs as those enriched in at least two significant nodes in the same direction and never in the opposite direction. MAG: metagenome assembled genome; KOs: KEGG orthologs.

We next compared the proportion of significantly different KOs identified based on POMS and a standard Wilcoxon rank-sum test (hereafter referred to as Wilcoxon test). The Wilcoxon test is a representative of the most common framework for performing differential abundance testing on microbiome functional data, which is to focus on the relative abundance of genes from across the community. The proportion of significantly different KOs in the random taxa simulated profiles was ten-fold lower based on POMS (mean=0.044; standard deviation [SD] = 0.040) compared with the Wilcoxon test approach (mean: 0.441; SD=0.244) (Figure 5.4a). This overall trend also held for the focal gene simulated profiles as well (Figure 5.4b). However, the random taxa and focal gene simulated profiles resulted in substantially different POMS results, whereas there was little difference based on applying Wilcoxon tests. More specifically, the proportion of significant KOs in the focal gene simulated profiles for POMS (mean=0.109; SD=0.084) was significantly higher compared with the random taxa simulated profiles (mean=0.044; W=181,877; P < 10^{-15}), and represents a 148% increase. In contrast, this same proportion based on applying Wilcoxon tests to the focal gene simulated profiles (mean=0.428; SD=0.242) was not significantly higher compared with the random taxa simulated profiles (W=117,467; P=0.099). As above, these overall results were robust to the cut-off choice for identifying significant KOs with POMs (Figure 5.4, panels c-f).

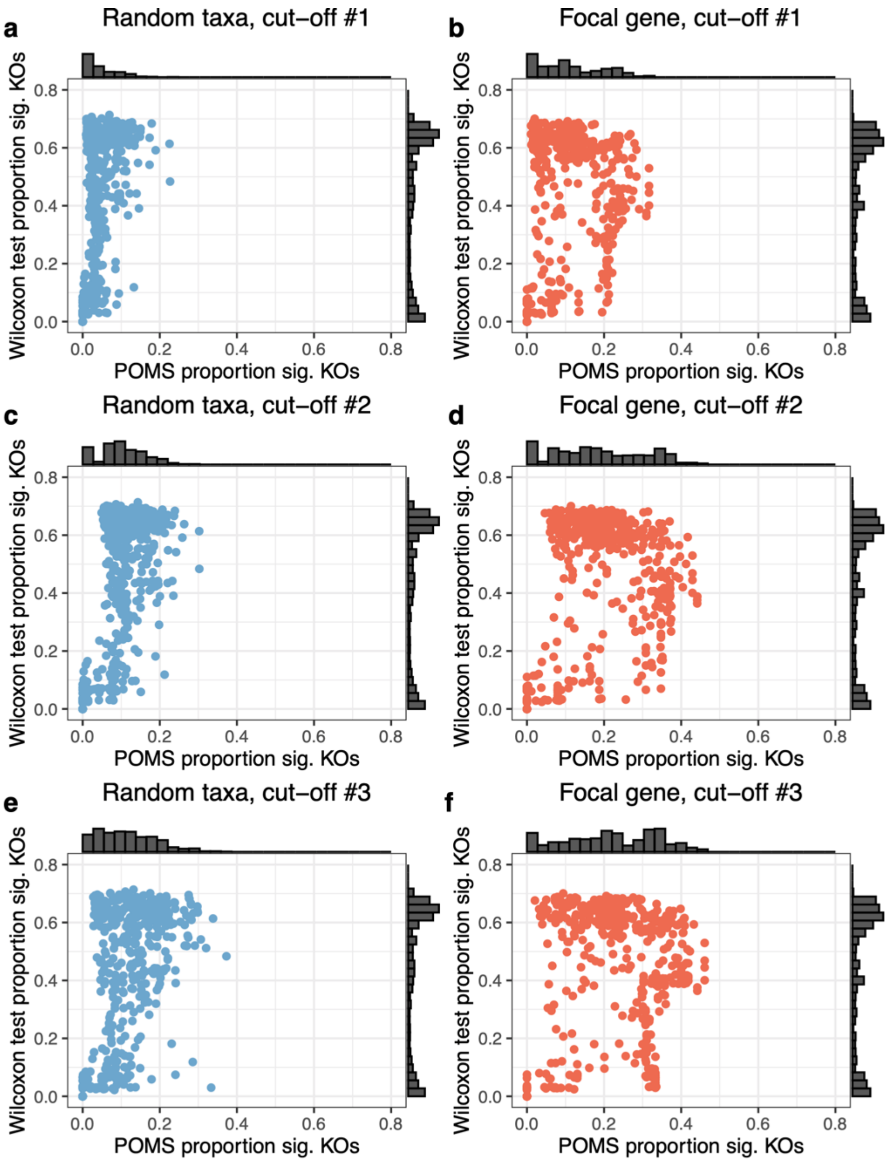


Figure 5.4: Proportion of significant gene families identified in MAG-based simulations based on POMS with primary cut-off compared with a Wilcoxon test*.* *Results are shown separately for all (a, c, and e) taxa and (b, d, and f) gene-based simulation replicates. Each point corresponds to one of 500 replicates. The marginal distributions are plotted on the x and y-axes. Cut-off #1, the primary cut-off used for these evaluations, refers to defining enriched KOs as those enriched in at least three significant nodes in the same direction and never in the opposite direction. Cut-off #2 refers to defining enriched KOs as those enriched in at least three significant nodes in the same direction and at most once in the opposite direction. Cut-off #3 refers to defining enriched KOs as those enriched in at least two significant nodes in the same direction*

and never in the opposite direction. MAG: metagenome assembled genome; KOs: KEGG orthologs.

We next evaluated the POMS output by determining the relative ranking in enrichment of the focal gene for each simulation replicate compared with all other KOs. Because under our straight-forward focal gene-based simulations the focal gene was the only direct target of selection, it would be expected to be highly ranked. KOs were ranked in the POMS output based on the absolute difference in the number of nodes positively and negatively enriched for each KO, which we refer to as the absolute enrichment. In the Wilcoxon test output, the KOs were ranked based on P-value. A drastic difference can be seen through ranking the resulting KOs based on these approaches (Figure 5.5, panels a and b). Specifically, the focal genes in the POMS output are ranked significantly higher (mean=0.204 percentile; SD=0.262 percentile; W=41,406; $P < 10^{-15}$) compared with the Wilcoxon test output (mean=0.535 percentile; SD=0.262 percentile). In addition, the variation across all replicates is largely associated with the number of MAGs that encode the focal gene (Figure 5.5, panels c and d). In the POMS results the focal genes are ranked highly overall with the exception of focal genes encoded by either very few or the majority of MAGs. For example, focal genes identified by POMS with relative rankings above the top 1% of KOs are encoded by a median of 38 MAGs. In addition, focal genes encoded by more than 2,300 MAGs are ranked qualitatively lower in the POMS output. These trends are partially reversed for the Wilcoxon test results; focal genes in the top 1% of KOs are encoded by a median of 22 MAGs. These relative ranking analyses, in addition to the supporting results presented above, indicate that POMS performs better in this simulated context compared to standard Wilcoxon tests.

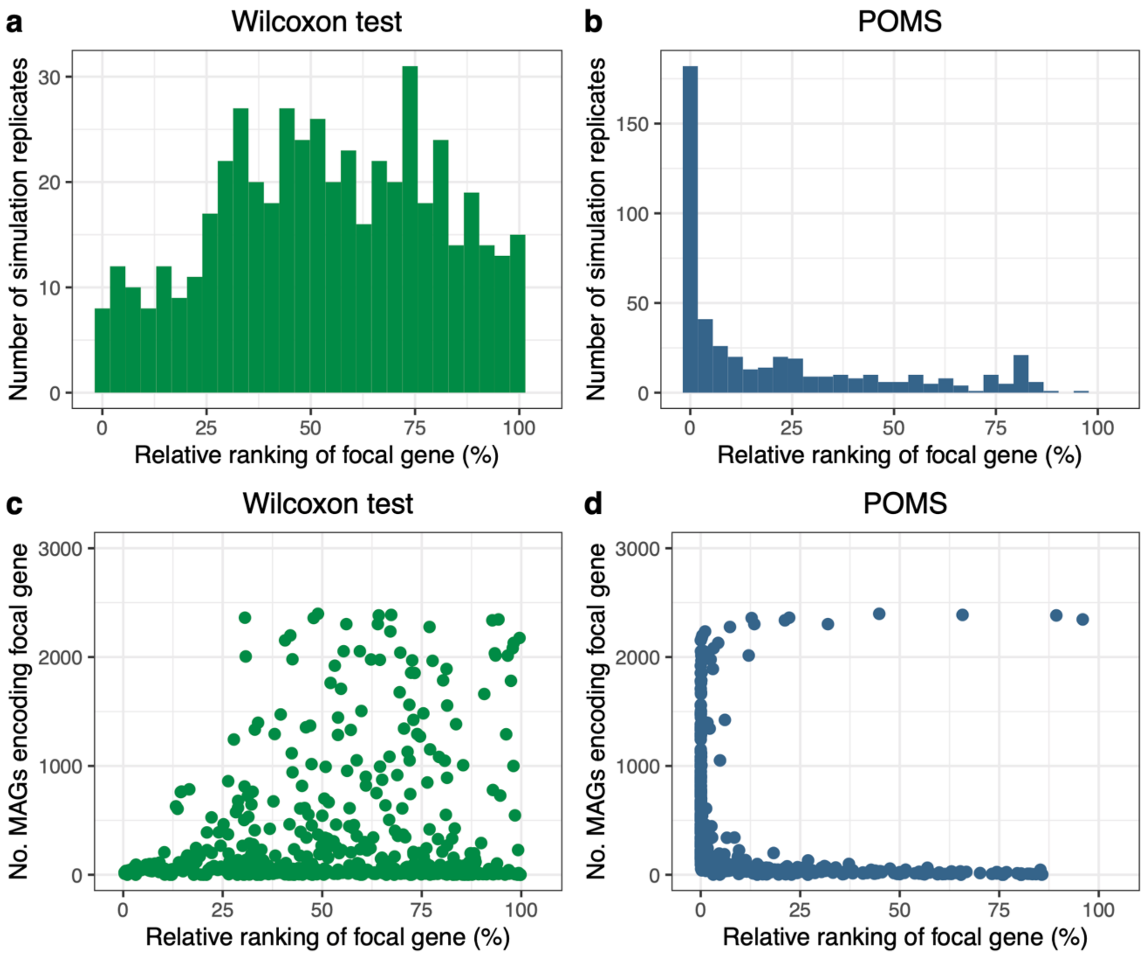


Figure 5.5: Ranking percentiles of KOs in both POMS and Wilcoxon test gene-based simulated profile outputs*. The focal gene was the KO under selection randomly selected for each simulation replicate. The ranking of all KOs was computed based on the enrichment effect size and P-value for POMS and the Wilcoxon test approach, respectively. The ranking of the focal gene is reported as its percentile in the distribution of all KOs (not just significant KOs) to illustrate where it lies on this distribution (panels a and b). Low ranking percentiles correspond to KOs that are identified as amongst the most enriched genes by the tool. Each point corresponds to a different simulation replicate, which were based on different focal genes. MAG: metagenome assembled genome; KOs: KEGG orthologs.*

5.3.3 – Validating POMS with Reference Genome-Based Simulated Data

The above observations based on the MAG-based simulations yielded valuable insights, but the quality of published MAGs is often questionable (Shaiber & Eren 2019). To ensure that misassembled MAGs were not driving our results, we repeated the key steps of our simulation analysis based on reference genomes. In particular, we investigated whether the drastic difference in rankings of the focal genes between POMS and the Wilcoxon test approach could be reproduced in an independent dataset of 3000 reference genomes. We simulated the abundance of these reference genomes based on a zero-inflated beta-distribution model for 1000 samples split into two equally sized groups. We created four such simulated datasets based on different parameter selections (see Methods for settings descriptions), which had a large impact on the sparsity and inter-sample overlap (Figure 5.6a). At one extreme (Setting 1), genomes are present at a mean of 385 samples (38.5%), while at the other extreme (Setting 4) they are present in an average of 2.35 samples (0.235%). These simulated datasets represent different partitions of the MAG-based abundance profiles analysed in the previous simulation analyses (Figure 5.6a), which had a mean MAG prevalence of 6.81%. We then repeated the focal gene-based simulation, including applying POMS and Wilcoxon tests, over 100 replicates for each of the four datasets. The ranking percentiles of focal genes varied substantially depending on the abundance table simulation approach (Figure 5.6b). In all cases the ranking percentiles of focal genes was lower in the POMS output (P < 0.0006), with one exception (Setting 2; W=4,563; P=0.285).

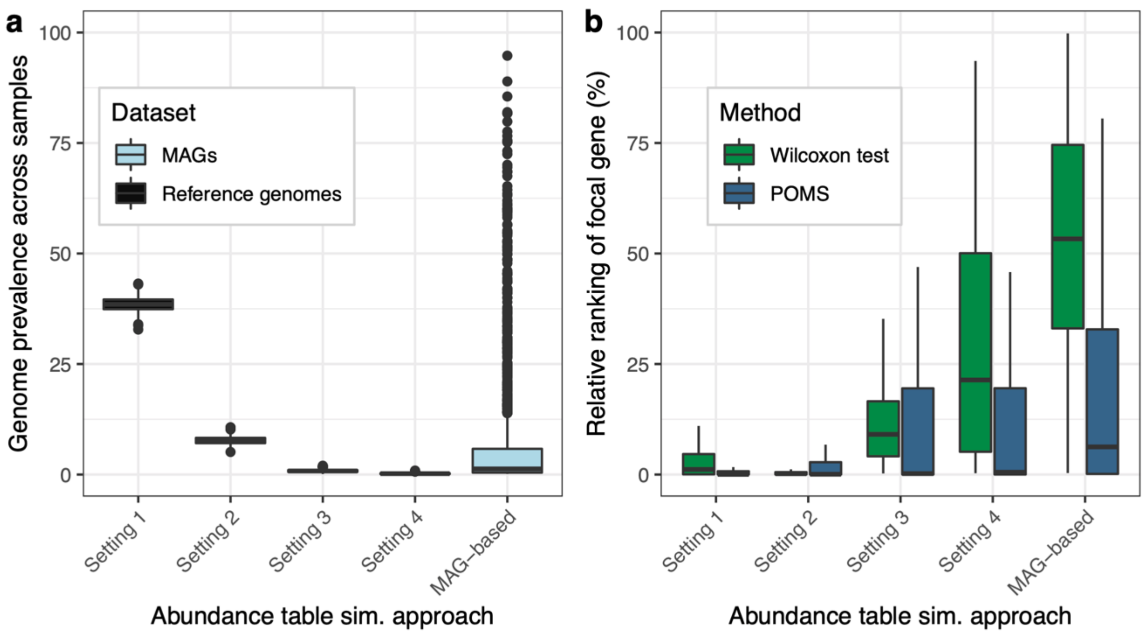


***Figure 5.6: Genome prevalence and ranking percentiles of focal genes varies across all simulated datasets**. (a) The prevalence (%) of each genome (or MAG) across all samples in a dataset. (b) The ranking percentiles for each simulated dataset setting, which differs depending on the data characteristics. The reference genome-based simulated datasets by four parameter settings (see Methods), which greatly affect genome prevalence. The "MAG-based sim." group corresponds to the (a) the original sample abundances leveraged for the MAG-based simulations and (b) the ranking percentile output based on the primary POMS cut-off. This category is displayed to enable clear comparisons with the MAG-based simulation results reported in the previous section. MAG: metagenome assembled genome.*

We next investigated the relationship between the number of genomes encoding the focal gene and the relative ranking of that gene. Similar to the MAG-based simulation results, focal genes at high ranking percentiles (i.e. not identified as highly enriched) in the POMS output were largely encoded by very few or almost all genomes in the dataset (Figure 5.7 and Figure 5.8). This effect became especially clear with simulated datasets with lower genome prevalence (e.g. Setting 4). In contrast, the focal gene ranking percentiles based on Wilcoxon test P-values displayed a distinct relationship with the number of encoding genomes at higher genome prevalence settings (Figure 5.7 and

Figure 5.8). In particular, these focal gene ranking percentiles were low overall with rare exceptions. For example, all focal gene were ranked lower than the 10th percentile under Setting 2 except for a subset of KOs encoded by more than 2,455 genomes (81.83% of all genomes). Notably, the Wilcoxon test focal gene ranking for Setting 2 (mean=2.5%; SD=9.7%) were at significantly lower percentiles compared with those at Setting 1 (mean=3.5%; SD=6.5%; W=6,677; P=$3.8*10^{-5}$), indicating that this method performs best with intermediate feature prevalence. For Settings 3 and 4, which represent highly sparse datasets, the focal gene ranking percentile showed a clear linear relationship with the number of encoding genomes. This relationship is particularly clear for the sparsest simulated dataset, which was under Setting 4, which is highly linearly correlated (Pearson R=0.968; $P < 10^{-15}$). Although this relationship is more simplistic, it captures the key signal represented by the matching analysis for the MAG-based simulations (Figure 5.5c). Overall, these reference genome-based simulation results are consistent with the key observations from the MAG-based simulations. In addition, the reference genome simulations based on tables of varying sparsity highlight how data characteristics can have substantial effects on the performance of both POMS and the Wilcoxon test.

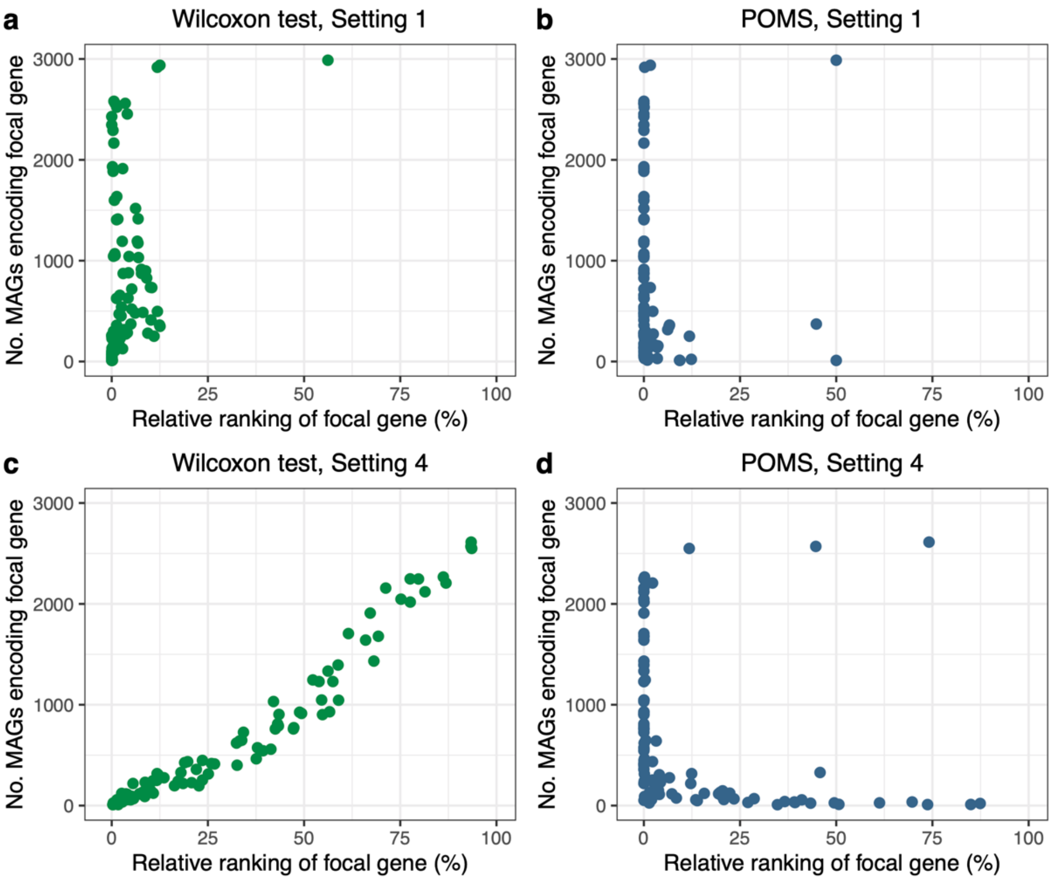


Figure 5.7: Ranking percentiles of focal gene based on most extreme reference genome-based simulation settings*. Visualization of the ranking percentiles for Setting 1 and 4 illustrates the overall shift for both methods. The ranking percentiles of focal genes based on Wilcoxon test P-values of all KOs are indicated in green (panels a and c). The ranking percentiles of focal genes based on the absolute enrichment all KOs in the POMS output are indicated in blue (panels b and d). KO: KEGG ortholog; MAG: metagenome assembled genome.*

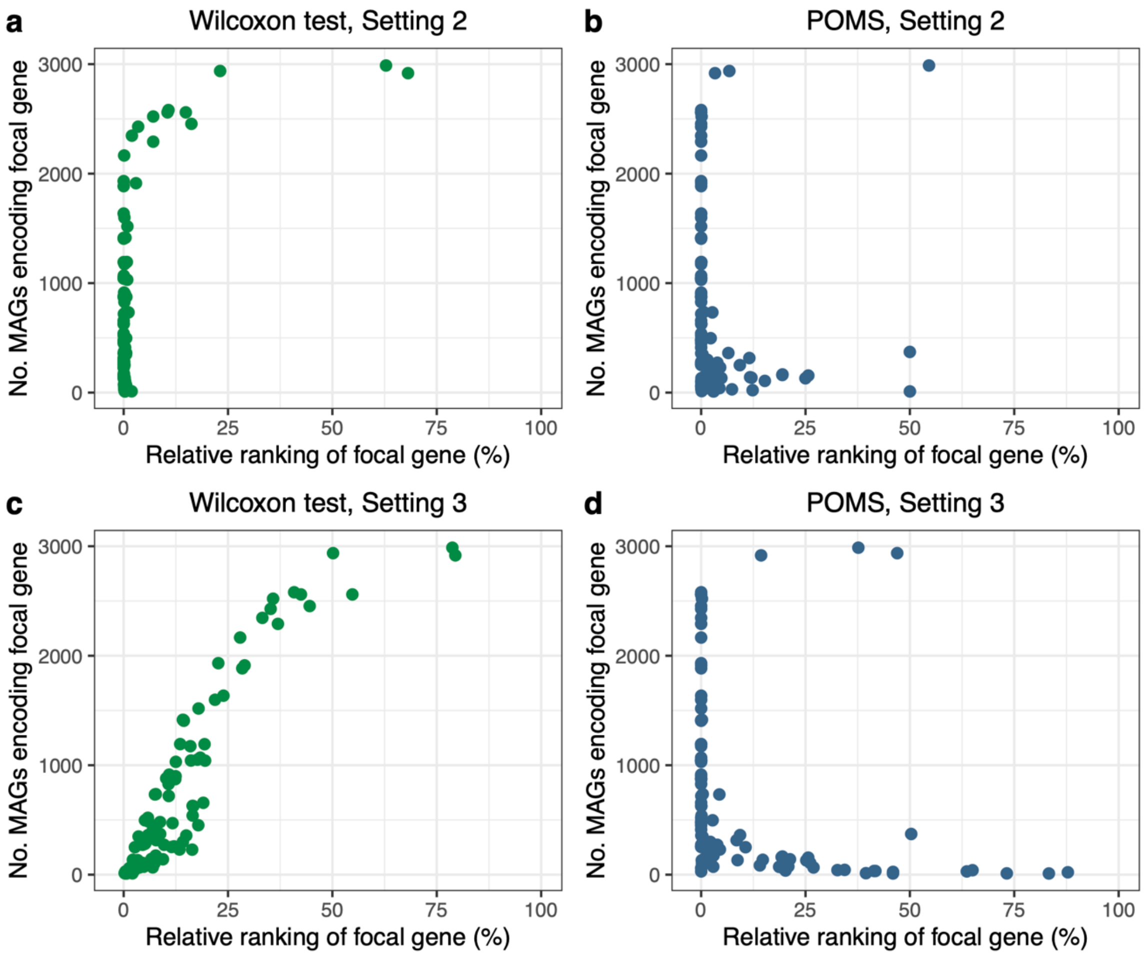


Figure 5.8: Ranking percentiles of focal gene based on two intermediate reference genome-based simulation settings*. Visualization of the rankings for Setting 2 and 3 helps illustrate the transition in relative ranking percentiles distributions depending on simulation approach. The ranking percentiles of focal genes based on Wilcoxon test P-values of all KOs are indicated in green (panels a and c). The ranking percentiles of focal genes based on the absolute enrichment all KOs in the POMS output are indicated in blue (panels b and d). KO: KEGG ortholog; MAG: metagenome assembled genome.*

5.3.4 – Exploring Human Gut Metagenome Assembled Genomes Dataset

We next investigated how POMS performs on actual metagenomics datasets. We focused on a large dataset of MAGs compiled from human-associated microbiomes that were published as part of a recent large-scale meta-analysis (Almeida et al. 2019). We used subsets of this large dataset corresponding to three disease datasets: two obesity datasets and one colorectal cancer dataset (see Methods).

The primary obesity dataset we analyzed included 477 obese and 257 control individuals that harbour a total of 1,401 MAGs in their stool microbiomes. We applied POMS to this dataset and identified 23 nodes with significantly different balances between obese and control individuals (Figure 5.9a; Benjamini-Yekutieli-corrected P-values (BY) < 0.05). These nodes correspond to a range of taxonomic separations, which would be difficult to interpret based on this information alone (Table 5.1). The consistency and number of node enrichments for each KO can be visualized to identify outliers (Figure 5.9b). The KO showing the strongest enrichment signal was K00941, which is involved in synthesis of the amino acid thiamine. This gene family was enriched positively in 15 nodes and never negatively enriched (Figure 5.10a). A major outlier in the opposite direction was K00091, which encodes a dihydroflavonol-4-reductase. This gene family was enriched in the direction of control samples (i.e. negatively enriched from the perspective of obese samples) at 11 significant nodes and never in the opposing direction.

To identify which gene families were significantly enriched, and not merely major outliers, we applied our pseudo-null approach. This approach confirmed that the major outliers we observed were significant. For example, the observed enrichment pattern for K00941 is entirely outside the computed pseudo-null distribution for that gene family (Figure 5.10b). Based on this approach, 21 KOs in total were identified as significantly enriched (Table 5.2). The KOs that were positively linked with obesity were also enriched for the phosphotransferase system KEGG pathway (ko02060; Fisher's exact test: odds ratio [OR] = 16.57; BY=$4.33*10^{-5}$).

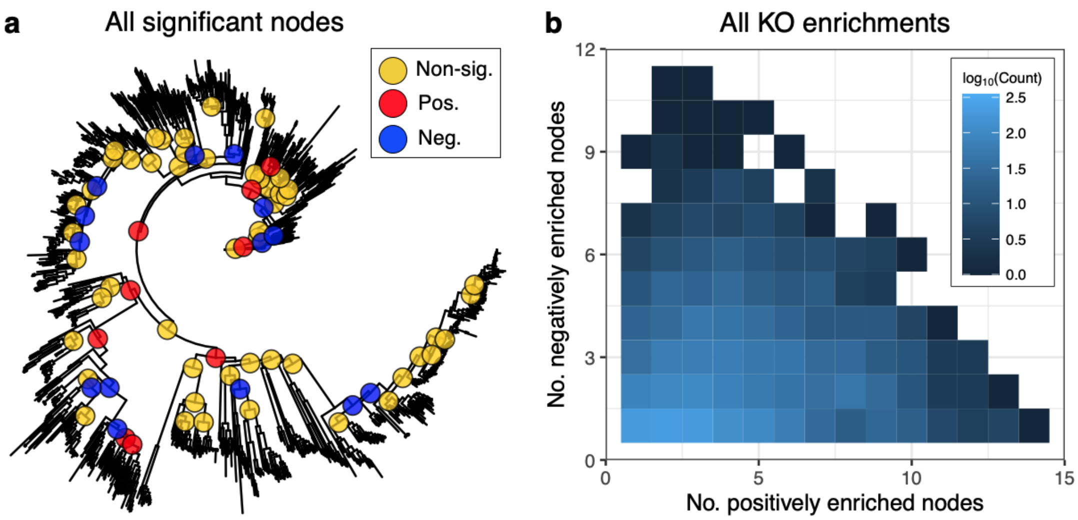


***Figure 5.9: Significant nodes and KO enrichments between the stool of obese and control individuals**. (a) Phylogenetic tree of all metagenome assembled genomes in the validation dataset. Red and blue nodes are those with significantly different balances in the direction of obese and control individuals, respectively. Yellow circles indicate tested nodes that were not significantly different. (b) Number of significant nodes (on log-scale) where each KEGG ortholog (KO) is enriched, either positively (in direction of obese samples) or negatively (in direction of control samples). Each value in this heatmap represents a single KO.*

Table 5.1: Taxonomic breakdown at significant nodes in primary obesity dataset

Node	Higher	Left-hand side taxa / Right-hand side taxa
n386	Control	Alphaproteobacteria (Class) / Proteobacteria (Phylum)
n459	Control	Bacteroidales (Order) / Bacteroidales (Order)
n914	Control	Clostridiales (Order) / *Blautia* (Genus)
n8	Control	Clostridiales (Order) / Clostridiales (Order)
n816	Control	Clostridiales (Order) / *Clostridium* (Genus)
n16	Control	Clostridiales (Order) / *Oscillibacter* (Genus)
n725	Control	*Clostridium* (Genus) / Clostridiales (Order)
n912	Control	Lachnospiraceae (Family) / Clostridiales (Order)
n913	Control	Lachnospiraceae (Family) / Lachnospiraceae (Family)
n246	Control	Mollicutes (Class) / *Solobacterium* (Genus)
n499	Control	Porphyromonadaceae (Family) / Bacteroidales (Order)
n170	Control	Tenericutes (Phylum) / Erysipelotrichaceae (Family)
n172	Control	Tenericutes (Phylum) / Tenericutes (Phylum)

Node	Higher	Left-hand side taxa / Right-hand side taxa
n67	Control	Ambiguous / Clostridiales (Order)
n298	Obese	Coriobacteriaceae (Family) / Ambiguous
n133	Obese	Lactobacillales (Order) / Tenericutes (Phylum)
n1248	Obese	*Ruminococcus* (Genus) / Ruminococcaceae (Family)
n74	Obese	Selenomonadales (Order) / Tenericutes (Phylum)
n173	Obese	Tenericutes (Phylum) / Tenericutes (Phylum)
n183	Obese	Tenericutes (Phylum) / Tenericutes (Phylum)
n5	Obese	Ambiguous / Clostridiales (Order)
n72	Obese	Ambiguous / Clostridiales (Order)
n68	Obese	Ambiguous / Ruminococcaceae (Family)

Table 5.2: Significant KOs in primary obesity dataset based on pseudo-null distribution

KO[a]	Pos[b]	Neg	KO Description
K00941	15	0	thiD; hydroxymethylpyrimidine/phosphomethylpyrimidine kinase
K13038	14	0	coaBC, dfp; phosphopantothenoylcysteine decarboxylase / phosphopantothenate---cysteine ligase
K11752	14	1	ribD; diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)uracil reductase
K02110	12	0	ATPF0C, atpE; F-type H+-transporting ATPase subunit c
K00091	0	11	Dihydroflavonol-4-reductase
K01933	11	0	purM; phosphoribosylformylglycinamidine cyclo-ligase
K02114	11	0	ATPF1E, atpC; F-type H+-transporting ATPase subunit epsilon
K07040	13	2	Uncharacterized protein
K00020	10	0	mmsB, HIBADH; 3-hydroxyisobutyrate dehydrogenase
K00526	9	0	E1.17.4.1B, nrdB, nrdF; ribonucleoside-diphosphate reductase beta chain
K02745	10	1	PTS-Aga-EIIB, agaV; PTS system, N-acetylgalactosamine-specific IIB comp.

KO	Pos	Neg	KO Description
K02746	10	1	PTS-Aga-EIIC, agaW; PTS system, N-acetylgalactosamine-specific IIC comp.
K02747	10	1	PTS-Aga-EIID, agaE; PTS system, N-acetylgalactosamine-specific IID comp.
K05846	10	1	opuBD; osmoprotectant transport system permease protein
K06956	9	0	Uncharacterized protein
K01232	8	0	glvA; maltose-6'-phosphate glucosidase
K02773	8	0	PTS-Gat-EIIA, gatA, sgcA; PTS system, galactitol-specific IIA component
K09773	8	0	ppsR; [pyruvate, water dikinase]-phosphate phosphotransferase kinase
K01878	7	0	glyQ; glycyl-tRNA synthetase alpha chain
K01879	7	0	glyS; glycyl-tRNA synthetase beta chain
K02822	7	0	PTS-Ula-EIIB, ulaB, sgaB; PTS system, ascorbate-specific IIB component

[a] *Green rows indicate KOs positively enriched in the direction of obese subjects while the orange row indicates the KO negatively enriched in obese subjects.*

[b] *Number of significant nodes where KO is enriched in the direction of obese samples.*

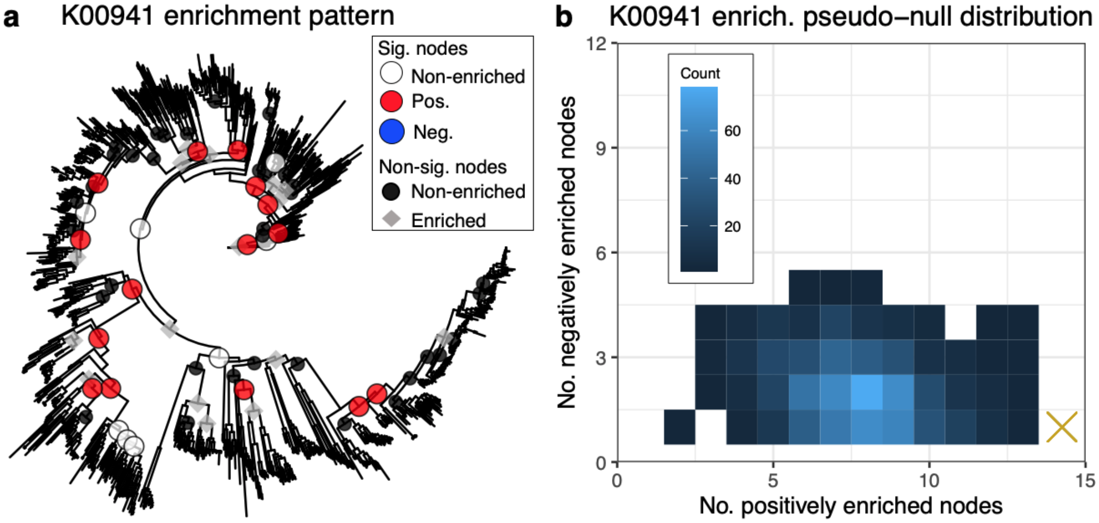


Figure 5.10: K00941 is the top enriched gene in the primary obesity dataset. *(a) Phylogenetic tree of all metagenome assembled genomes in the validation dataset. Red nodes are significant nodes with K00941 enriched in the direction of obese samples. Grey nodes are significant nodes based on balances where there is no significant enrichment of this gene family. (b) Pseudo-null distribution for K00941 showing the number of significant nodes where this gene family is enriched with randomly selected significant nodes. Each value in this heatmap represents one of 1000 replicates used to create the pseudo-null distribution. The actual observed enrichment values for K00941 are indicated by the gold x.*

We also applied POMS to a smaller dataset of 251 obese and 159 control samples from the same meta-analysis. Stool samples from these individuals contained 1,161 MAGs. There were 20 nodes with significantly different balances between obese and control individuals in this dataset (BY < 0.05). The top enriched KO was K13038 (Figure 5.11), which encodes an enzyme that interacts with cysteine and is involved in pantothenate and Coenzyme A biosynthesis. Based on computing P-values based on our pseudo-null approach, we identified six KOs positively linked with obesity (BY < 0.05; K00833, K01935, K03150, K07799, K13038, and K21498). Two of these KOs, K00833 and K01935, encode proteins involved in biotin biosynthesis and K03150 is involved in thiamine biosynthesis. We identified two KOs negatively linked with obesity based on this approach (BY < 0.05; K06374 and K18707).

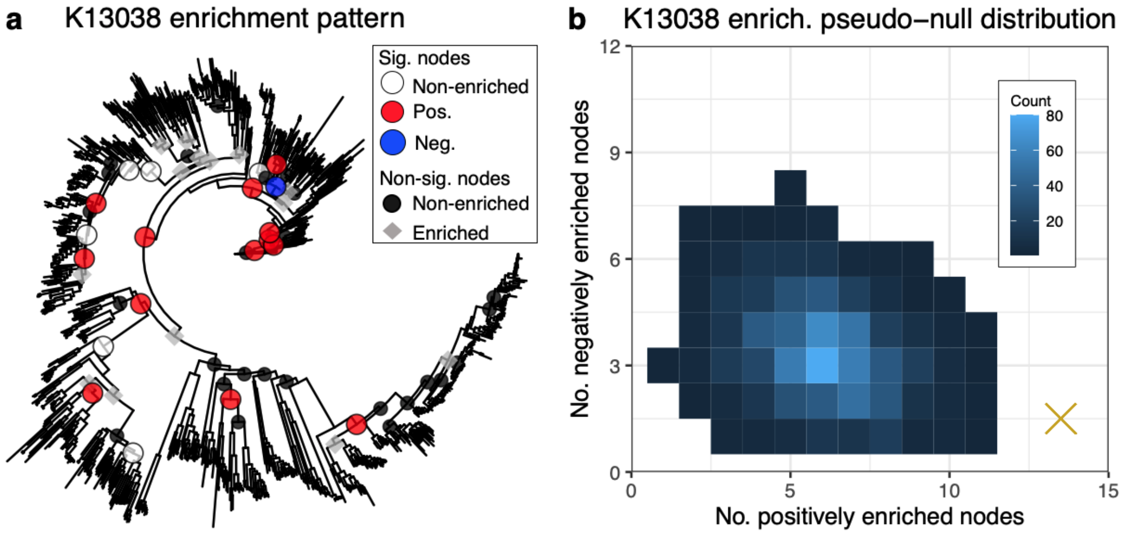


***Figure 5.11: K13038 is the top enriched gene in the secondary obesity dataset**. (a) Phylogenetic tree of all metagenome assembled genomes in the validation dataset. Red nodes are significant nodes with K13038 enriched in the direction of obese samples while the blue node represent an enrichment in the direction of control samples. Grey nodes are significant nodes based on balances where there is no significant enrichment of this gene family. (b) Pseudo-null distribution for K13038 showing the number of significant nodes where this gene family is enriched with randomly selected significant nodes. Each value in this heatmap represents one of 1000 replicates used to create the pseudo-null distribution. The actual observed enrichment values for K13038 are indicated by the gold x.*

Last, we applied POMS to the microbial profiles of stool samples from 75 colorectal cancer patients and 53 controls, which contained 1,187 MAGs. In this case, only five nodes were significantly different between case and control samples (BY < 0.05). Based on our pseudo-null distribution approach, two KOs were significantly linked with control samples in this dataset (K00702 and K09691), but none were linked with cancer samples. The top hit was K00702 (Figure 5.12), which encodes cellobiose phosphorylase: an enzyme that plays a key role in cellulose degradation. The other top hit was K09691, which encodes a lipopolysaccharide (LPS) transport system ATP-binding protein.

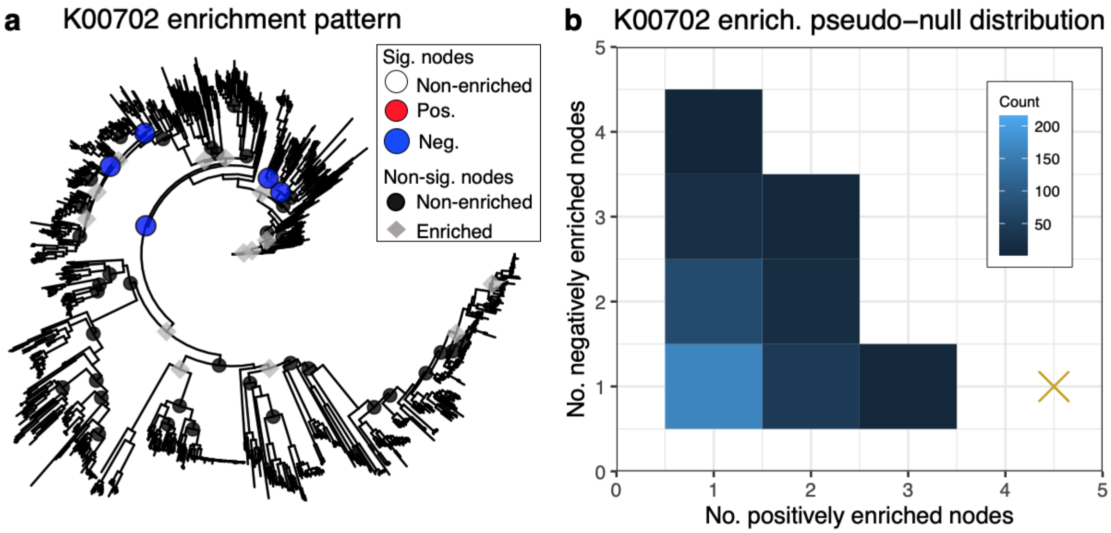


***Figure 5.12: K00702 is the top enriched gene in the colorectal cancer dataset**. (a) Phylogenetic tree of all metagenome assembled genomes in validation dataset. Red nodes are significant nodes with K00702 enriched in the direction of colorectal cancer samples. (b) Pseudo-null distribution for K00702 showing the number of significant nodes where this gene family is enriched with randomly selected significant nodes. Each value in this heatmap represents one of 1000 replicates used to create the pseudo-null distribution. The actual observed enrichment values for K00702 are indicated by the gold x.*

We next investigated the concordance of enriched KOs in these three datasets identified by POMS with common differential abundance tools. We tested three common approaches: Wilcoxon tests based on relative abundances, ALDEx2, and Limma-Voom. These approaches each test different hypotheses, but it is appropriate to compare these tools because they are often used interchangeably in the microbiome field. We were interested in determining how different the biological interpretation would be depending on the choice of tool.

The three datasets markedly differ in terms of the concordance between the tested methods (Figure 5.13). In the primary obesity dataset there was relatively high concordance between POMS and the alternative differential abundance methods (Figure 5.13a): 16/20 (80%) of significant KOs were also called as significant by at least one other method. In addition, all of the alternative methods agree well overall for this dataset.

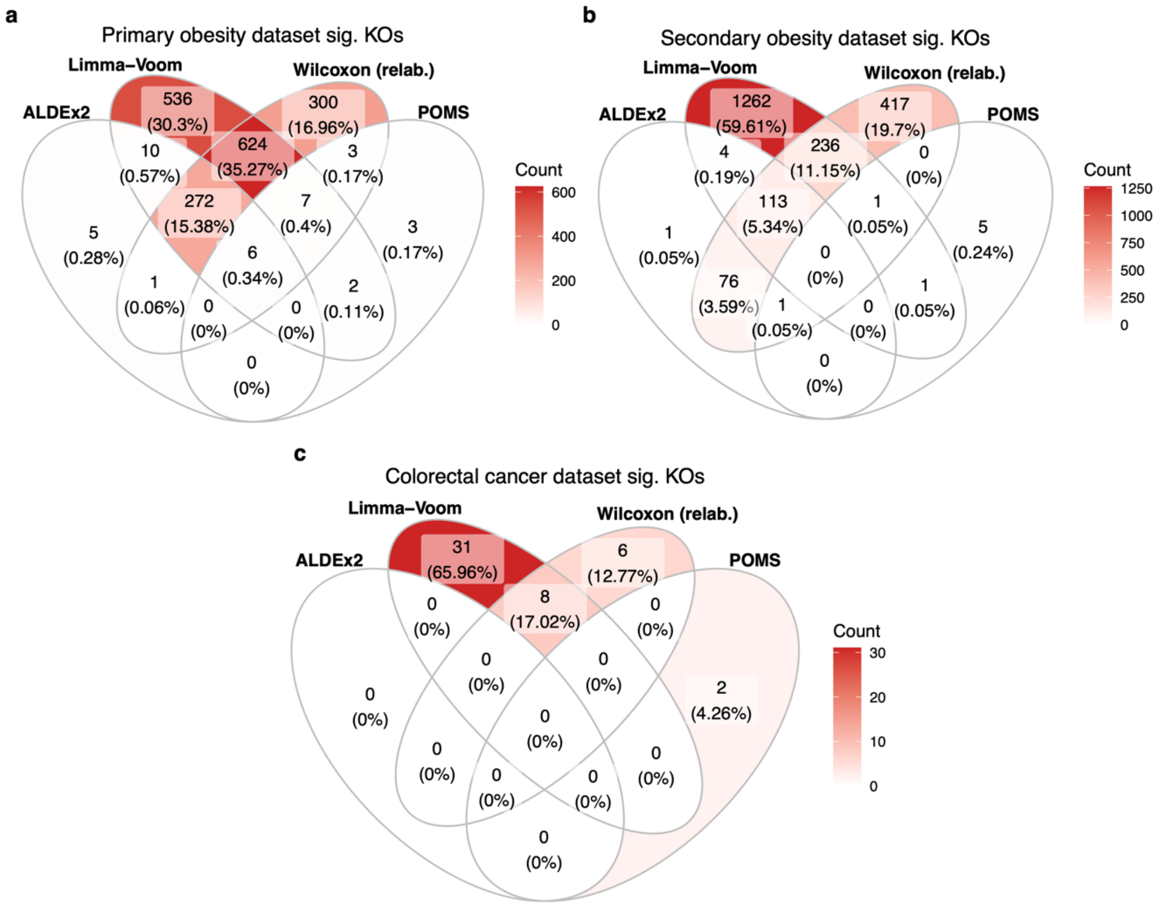


Figure 5.13: Overlap in significant KOs between common differential abundance tools and POMS. (a) Primary obesity dataset, (b) Secondary obesity dataset, and (c) colorectal cancer dataset.

The phosphotransferase pathway was positively enriched in the set of significant KOs in the output of all tools for the primary obesity dataset. Additional pathways were also identified by at least one tool in this dataset, including these obesity-associated pathways: ABC transporters (Limma-Voom), ascorbate and aldarate metabolism (Limma-Voom), fructose and mannose metabolism (ALDEx2), and galactose metabolism (ALDEx2 and Limma-Voom). These pathways were down-regulated in obese patients in this dataset based on both the Limma-Voom and Wilcoxon results: aminoacyl-tRNA biosynthesis, flagellar assembly (by ALDEx2 as well), homologous recombination, and ribosome. Bacterial chemotaxis was enriched in the KOs identified as downregulated in both the ALDEx2 and Wilcoxon test output.

For the two other datasets there is much lower concordance, particularly in terms of KOs identified as significant by POMS (Figure 5.13, panels a and b). Consistent with the overall consensus in significant KOs, the enriched pathways drastically differed for these two datasets across tools. In particular, in the secondary obesity dataset the only overrepresented pathways in obese individuals were identified based on the Wilcoxon test output as LPS biosynthesis and metabolic pathways. The only underrepresented pathways in this dataset were identified based on the Limma-Voom output to be bacterial chemotaxis and flagellar assembly. For the colorectal dataset, only these pathways were identified based on up-regulated KOs identified by Limma-Voom: arginine biosynthesis, lysine biosynthesis, 2-oxocarboxylic acid metabolism.

Last, we compared the effect sizes of the top enriched KOs identified by POMS compared with the mean differences in relative abundance (raw and centre log-ratio (CLR) transformed). As expected, these effect sizes are roughly positively correlated: KOs consistently positively enriched at nodes in case samples also tend to have higher relative abundance in those samples (Figure 5.14). However, there were many exceptions as several of the top enriched KOs identified by POMS showed displayed little or no difference in relative abundance between sample groupings. This included the top KO enrichments identified by POMS in both obesity datasets. These observations highlight that POMS is a qualitatively different approach than current bag-of-genes differential abundance tools.

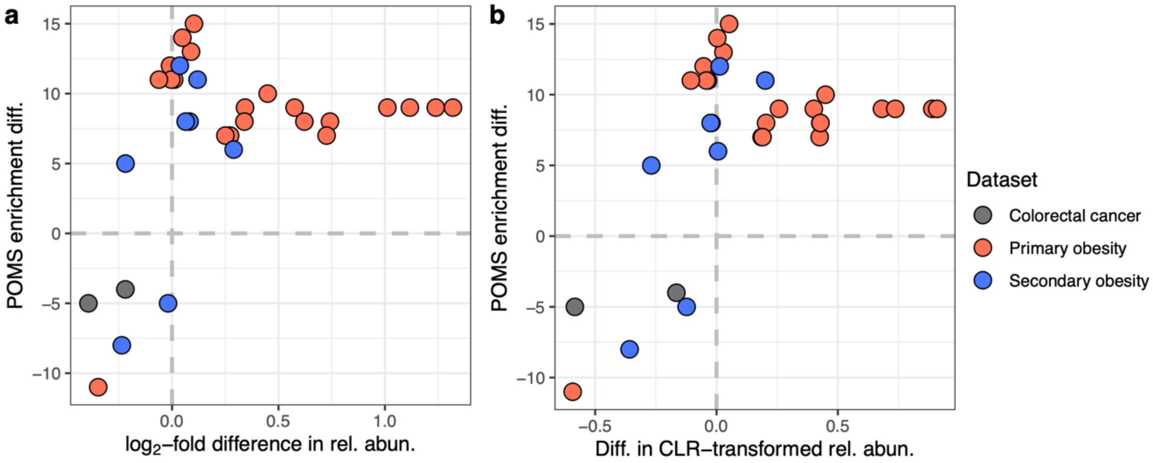


Figure 5.14: Concordance between POMS enrichment and difference in relative abundances for top KOs identified. (a) Log$_2$-fold difference of relative abundance and (b) the mean difference between centred log-ratio (CLR) transformed relative abundances. Mean differences higher than zero represent KOs higher in case samples. Similarly, the POMS enrichment difference corresponds to the difference in positively and negatively enriched nodes per KO (e.g. positive values indicate enrichment in case samples). The dotted lines indicate zero. KO: KEGG ortholog.

5.4 - Discussion

Herein we have presented the framework and validations for POMS, which is a novel approach for identifying functional microbiome biomarkers. Based on straight-forward simulations based on both MAGs and reference genomes, we have demonstrated that POMS can accurately identify widely encoded microbial functions that are under strong positive selection. The most convincing evidence was that focal genes in our simulations were extremely highly ranked in the POMS output, whereas this is not the case for the Wilcoxon test when applied to sparse microbiome data. In addition, as discussed below, the enriched gene families identified by POMS in actual case-control datasets are consistent with the tool identifying functions encoded by numerous taxa that are strongly linked with a given microbial environment.

In these case-control datasets we identified enriched gene families as outliers in the pseudo-null distributions. However, it is important to emphasize that this is an imperfect process for several reasons. First, it is based on shuffling significant nodes in a tree, which means that a user’s interpretation depends on the proportion of nodes that were originally significant. Throughout this manuscript, only a minority of nodes displayed significantly different balances between sample groupings. However, if the majority of nodes in a tree are significantly different then this pseudo-null approach would be overly conservative. This would simply be because most randomly sampled nodes would overlap with the original observed set. We do not anticipate that this situation will be common, but nonetheless it is an important caveat that users should keep in mind. Conversely, the POMS approach for identifying functional enrichments requires significant nodes to be identified in the phylogenetic tree of taxa. Without a sufficient number of these nodes to test for consistent patterns in gene enrichments this approach would not be useful. In addition, the approach of shuffling significant nodes to identify significantly enriched gene families assumes that every combination of nodes is equally possible to be significant by chance. This is partially invalid because different taxonomic groups are more likely to vary across individuals than others and taxa co-occurrence can occur even at long evolutionary distances (Ma et al. 2020).

For these reasons, POMS should predominately be considered an exploratory approach; it can be used to identify putative gene enrichments in particular sample groupings or environments, but how to clearly distinguish enriched and non-enriched genes is not always apparent. In addition, researchers must investigate the enrichment patterns of genes within the context of a specific dataset. The chosen cut-offs for identifying enriched genes will differ depending on characteristics of the dataset such as the overall number of significant nodes and the inter-sample taxonomic variability. In other words, it would be inappropriate to apply the same enrichment cut-off, such as calling genes enriched at three nodes consistently as significantly enriched, to all datasets. Application of the pseudo-null distribution approach can help address this problem, but this must be done with appreciation of the related caveats, as described above. Either way, gene enrichment effect sizes must be reported in the context of the entire dataset. Researchers should also appreciate that the set of enriched genes can substantially change depending on the choice of cut-off, as highlighted by the shifting distributions depending on the cut-off used for the MAG-based simulation analyses.

There are also limitations in terms of the characteristics of enriched genes that POMS can identify. These genes must be adequately dispersed so that they can be enriched at multiple nodes but also be sufficiently conserved between closely related taxa. This requirement eliminates many possible gene enrichments. This limitation is especially clear for important functions encoded by keystone taxa. For instance, *Methanobrevibacter smithii* is the predominant gut symbiont that removes several fermentation products from other bacteria through methanogenesis (Samuel et al. 2007). A shift in methanogenesis between different human gut sample groupings would be challenging to identify with POMS due to the trait's limited taxonomic breadth across the human gut microbiota. Similarly, not all taxa that encode a selected gene family will necessarily experience the same benefit. For instance, competition between strains through clonal interference has been well described experimentally (Lässig et al. 2017), but it is unclear how commonly this occurs in microbiomes (Garud & Pollard 2020). These limitations highlight that POMS will often miss putatively enriched genes.

Despite these limitations, the POMS framework represents a valuable alternative to existing differential abundance approaches for microbial functions. These approaches predominately consider this data type as a bag-of-genes, meaning that analyses are performed on the relative abundances of gene families across an entire community. Although this is the predominant method for analyzing functional data, it is fundamentally flawed (see Section 5.2). In addition to issues of interpretability, these methods also provide notoriously inconsistent outputs. This has largely been shown based on taxonomic data (Weiss et al. 2017), where in some contexts false discovery rates can be above 50% (Hawinkel et al. 2019). In addition, compositional approaches have not resolved this issue: a recent evaluation found that several non-compositional approaches produced more consistent results across a range of datasets (Calgaro et al. 2020).

Our application of standard differential approaches largely agrees with these assessments: the choice of method resulted in substantial differences in the number of significant KOs on the tested case-control datasets. In addition, focal genes under selection in our straight-forward MAG-based simulations were often mid-ranked amongst all KOs based on Wilcoxon tests. In other words, it would be impossible for a researcher to accurately identify the focal genes under strong selection with this test, even under our straight-forward simulation framework. In contrast, POMS accurately identified the focal gene amongst the top ranked KOs.

Our simulations based on reference genomes provide insight on how these trends can change depending on the genome abundance profiles. In particular, the Wilcoxon test can accurately identify the focal gene in non-sparse data (Settings 1 and 2 in that section). In fact, the Wilcoxon test approach outperforms POMS based on these rankings specifically when genomes around found across a mean of 7.7% of samples (Setting 2). However, the actual MAG datasets we evaluated in this study exhibited high variance and sparsity, which are a common characteristics of microbiome data and likely explain why the standard differential abundance tests performed poorly overall on real data.

Given these troubling observations on the performance of standard differential abundances tests on sparse data, it is important to reconsider how standard differential abundance tools should be used. Although these methods provide P-values to enable significance to be determined, clearly these frameworks are flawed if biological interpretations largely vary depending on the choice of tool. Although granted, POMS provides only a limited statistical framework for identifying enriched genes based on P-values, the effect sizes output by POMS are much more interpretable than looking at fold-changes in relative abundance.

This improved interpretability is also reflected in the top enriched KOs identified by POMS in the validation case-control datasets. In particular, these KOs are consistent with functions that could confer a strong adaptive benefit in the human gut depending on disease status. Although enrichment of these KOs is not proof of selection acting upon these gene families, it does enable more precise hypothesis generation than would be possible based on a bag-of-genes approach. For instance, in the primary obesity dataset the significant KOs identified based on POMS were enriched for one KEGG pathway: phosphotransferase system (PTS). These bacterial transport systems are involved in carbohydrate uptake and have previously been identified at higher levels in obese individuals (Greenblum et al. 2012). Higher PTS levels have also been observed in mice given a Western diet (Turnbaugh et al. 2009b). Importantly, PTS was also enriched based on all three tested differential abundance tools with obesity. Several other pathways were enriched in the output of these tools as well with clear links to obesity, such as galactose metabolism. However, there were also pathways identified with unclear connections, such as bacterial chemotaxis. This highlights that the enriched pathways identified by POMS are highly conservative, but that our approach likely will miss enriched pathways identified by more common approaches.

Nonetheless, POMS is valuable for identifying specific gene families that are linked with overall shifts in environment. In particular, many of the top enriched KOs in the case-control datasets identified by POMS are involved with resource limitation in the case gut environment. For instance, multiple KOs identified by POMS as enriched in obese patients are involved with thiamine biosynthesis. Thiamine (vitamin B1) is an essential micronutrient involved in glucose metabolism. More generally, it is a co-factor in several metabolic pathways, including the Krebs cycle and the pentose phosphate pathway. Thiamine levels are often recorded before and after bariatric surgery due to the high risk of micronutrient deficiencies following bariatric surgery (Kazemi et al. 2010). Although thiamine deficiency is quite common after surgery, it has also been reported in 12-15.5% of obese patients prior to bariatric surgery (Carrodeguas et al. 2005; Coupaye et al. 2009). It is unclear whether these percentages reflect the overall obese population (Kerns et al. 2015). In particular, thiamine deficiency rates could be higher for obese individuals in general because attempts at alternative weight-loss approaches, such as dieting, are typically required before turning to surgery. In addition, individuals with thiamine deficiency can display a wide range of ambiguous symptoms, which can make it difficult to diagnose (Kerns et al. 2015). Accordingly, there is reason to believe that thiamine availability might be limited in the gut of obese individuals, which could result in an adaptive benefit to microbiota that produce their own thiamine.

A similar mechanism could explain the enrichment gene families involved with producing other micronutrients as well. In particular, genes families involved with biotin synthesis were also linked with obesity. The biotin levels of obese individuals have not been well-studied, but biotin deficiency has been linked with higher blood glucose levels and insulin resistance (Via 2012). These functional shifts could be enabled by environmental changes in the gut of obese individuals. For instance, one of the significant KOs involved with thiamine biosynthesis and linked with obesity, K03150, metabolizes tyrosine, which in turn is known to be at higher levels in the serum of obese individuals (Adams 2011).

Similarly, higher levels of cysteine have been linked with obesity in numerous epidemiological studies and animal experiments (Elshorbagy et al. 2012a). In fact, it has been argued that cysteine may be the only obesogenic amino acid (Elshorbagy et al. 2012a, 2012b), meaning that it may partially cause obesity independent of confounding factors. Accordingly, it is highly relevant that the top enriched KO in the secondary obesity dataset we analyzed, K13038, is a cysteine ligase. This gene family acts downstream of cysteine in the pantothenate and Coenzyme A biosynthesis pathway. The observed enrichment of this gene leads to the clear hypothesis that gut microbiota may gain an advantage by increased utilization of cysteine specifically in obese individuals.

The two gene families negatively linked with colorectal cancer patients by POMS were a cellobiose phosphorylase, K00702, and an LPS transport system ATP-binding protein, K09691. The enrichment of the cellobiose phosphorylase gene family is noteworthy because cellulose has specifically been shown to be protective against colonic tumours in rat experiments (Nakaji et al. 2004). More generally, dietary fiber intake is moderately associated with decreased risk of developing colorectal cancer (Park et al. 2005). In contrast, higher LPS levels have been associated with colorectal cancer (de Waal et al. 2020), which makes the negative enrichment of the LPS transport system ATP-binding protein unexpected. However, this protein was enriched at nodes splitting predominately Gram-positive lineages (largely Firmicutes), which do not encode LPS. It is possible that this gene family is involved in transporting other membrane proteins instead of LPS, such as teichoic acid. In any case, this example highlights that an understanding of the taxonomic contributors to a functional signal can help elucidate problematic cases.

These observed top enriched gene families are encouraging overall, but there is room for further development of our framework. Although POMS is a valuable tool for exploratory data analysis, the key contribution of our work is the proof-of-concept that this framework can identify gene families highly enriched in specific environments. This general framework could be expanded to incorporate other advances in this area, such as more sophisticated approaches for analyzing balances across phylogenetic trees

(Silverman et al. 2017; Washburne et al. 2019). In particular, it would be straight-forward to apply our framework with different methods for computing balances (Silverman et al. 2017) and/or phylogenetic reference frames (Washburne et al. 2019). In addition, POMS is intended to be used for two-group comparisons, but the overall framework could be expanded for multiple group comparisons as well.

To conclude, we have presented a novel framework for analyzing microbial functions with phylogenetic balance trees. This approach is implemented in the R package POMS, which can be used to identify gene families strongly enriched in one environment over another. Although this approach has limitations, it represents a valuable step towards more interpretable differential abundance testing with functional microbiome data.

Chapter 6 – Discussion

The above investigations have explored integrating taxa-function data from three perspectives: classification models, metagenome prediction, and identifying functional biomarkers. These projects have provided valuable and novel insights. For instance, we identified that taxonomic data types perform best for classification on their own, but that MGS functions often are most informative in combined classification models. In addition, our work developing an improved metagenome prediction method highlighted the difficulty of evaluating functions predicted from taxonomic data. And last, our method POMS is a novel analysis framework that we demonstrated can identify intriguing enriched functions in case-control samples that would not be possible with standard tools. These and other specific observations have been discussed in each respective chapter, and for the most part are not the focus of this closing section.

Instead, this chapter will touch on the high-level observations and broad parallels between these chapters. First, a repeated observation throughout this thesis is that there is a lack of consistency and standardization when performing microbiome data analysis. This was true for virtually all of the analyses presented in this thesis and the implications of this problem are largely unappreciated by microbiome researchers. Second, since publishing our Crohn's disease classification project in 2018 there have been subsequent developments and improvements to machine learning applications to microbiome data. However, there remain numerous open questions, and particularly the best approach for integrating data types in these models remains unclear. Next, our work on metagenome predictions raised numerous questions regarding the usefulness of this data type. Nonetheless, despite these major caveats, I believe predicted data types will remain relevant and likely will become more accurate as sequencing technology continues to improve. Last, our work on POMS represents a valuable alternative approach for identifying enriched microbial functions compared with existing tools. However, POMS has several important caveats and there remain several important issues to be addressed for improved joint taxa-function analysis in general. I will discuss these and other areas that I believe would benefit from further research.

6.1 – Biologically Interpreting Microbiome Data

My thesis work was motivated by biological questions, such as how to best biologically interpret microbiome sequencing data. Unfortunately, a common thread through this thesis has been that technical variation in microbiome data analyses means that making robust biological inferences, particularly regarding specific microbial features, is challenging. Indeed, the lack of standardization in microbiome data analysis has previously been strongly criticized. An assessment of numerous papers attempting to define standard pipelines concluded that there was disturbingly little consensus (Pollock et al. 2018). This is true for many steps related to the processing, sequencing, and analysis of microbiome data. For instance, there have been contradictory results regarding the efficacy of different extraction protocols (Salonen et al. 2010). In particular, underrepresentation of Gram-positives has been observed (Maukonen et al. 2012), which may be partially resolved by using bead-beating extraction protocols (Guo & Zhang 2013). There is also substantial technical variation related to bioinformatics choices, which represent the final steps of a microbiome project. For example, as discussed at length in the Introduction, the bioinformatics choices made when performing differential abundance testing on microbiome data can have severe impacts on any interpretations (Thorsen et al. 2016; Hawinkel et al. 2019).

This general lesson was reinforced during my analysis of the pediatric Crohn's disease patients' microbiome profiles (Chapter 3). An important characteristic of these data was that 98% of the sequenced reads mapped to the human genome (Douglas et al. 2018). This characteristic made taxonomic profiling of these data especially prone to false positives. In particular, an initial draft of our manuscript was based on profiles that included large proportions of viral-identified DNA and matches to certain eukaryotic parasites. We were initially excited about these observations, because the abundances of these non-prokaryotic taxa were discriminative for classifying patient disease state and treatment response. However, the exact taxa identified were peculiar: they were predominately represented by a range of plant-associated viruses and the eukaryotic genus *Plasmodium*, which is best known as including the causative agent for malaria, *Plasmodium falciparum*. Upon closer investigation it became clear that this signal was driven entirely by a difference in how reads were mapped to lineage-specific marker genes. Altering the parameter choice from local to global mapping entirely removed these taxa. This relatively small difference in parameter choice appeared to only affect our data and not more typical microbiome datasets, which we believe was due to the high proportion of human DNA in our data.

Although this error was moderately embarrassing, it was more importantly an example of how easily a single parameter setting can result in starkly different biological interpretations. In this case the difference was driven by an option used for a single bioinformatics tool, MetaPhlAn2. I highlighted similar issues in Chapter 4 and 5 of this thesis. In particular, our comparisons of differential testing on predicted metagenomes highlighted a range of performance in terms of concordance with matching MGS data (Chapter 4). In addition, the concordance in significant gene families was also relatively low in a comparison of differential abundance tests applied to MGS data generated by two different workflows. These workflows are both commonly performed and simply represent different ways of identifying gene families with HUMAnN2. In one workflow KEGG orthologs were mapped to directly and in the other UniRef gene families were first mapped to and then regrouped to KEGG orthologs using known links between the gene family databases. It is highly troubling that the biological interpretations can starkly differ depending on which of these common workflows is followed.

Similarly, in Chapter 5 we demonstrated a representative bag-of-genes approach for conducting functional differential abundance testing (the Wilcoxon test) fails under straight-forward simulation conditions. More specifically, this tool was unable to identify gene families under strong selection as highly ranked in the resulting output. This observation is important, because it highlights that it is non-trivial to biologically interpret the top hits in a standard differential abundance analysis. This appeared largely true in highly sparse datasets as gene families under selection could be identified more clearly in less sparse simulated datasets (Chapter 5). However, a key characteristic of common microbiome sequencing datasets, such as those representing the human gut, is that they are highly variable and sparse across individuals (see Introduction). Accordingly, our troubling observations could be relevant to how data analysis is performed on the majority of existing microbiome datasets.

Such inconsistencies in microbiome analyses have previously been identified and been shown to make meaningful comparisons across studies challenging. For instance, associations between obesity and the human microbiome are commonly discussed as support for the utility of considering microbial links with human disease, despite inconsistencies across studies (Castaner et al. 2018; Muscogiuri et al. 2019). These inconsistencies are typically explained due to confounding variables that may differ between patient cohorts. Although this is a valid explanation, it is likely that technical variation, including in terms of bioinformatics analyses, also drives these inconsistencies. For instance, a meta-analysis of ten obesity human microbiome datasets identified only extremely weak signals when re-analyzing all datasets with a standardized approach (Sze & Schloss 2016). This finding greatly contrasts with how these studies were originally presented and again highlights how variation in bioinformatics can greatly affect how to biologically interpret microbiome data.

Similarly lower alpha diversity in stool microbiomes has been frequently linked with disease states (Mosca et al. 2016). These observations are intuitively reasonable as reduced alpha diversity could enable pathogens to bloom (Vincent et al. 2013) or represent differences in resource availability (Turnbaugh et al. 2009a). However a re-analysis of data from 28 studies representing ten diseases was unable to identify evidence for links between alpha diversity and disease states (Duvallet et al. 2017). The exceptions were diarrheal diseases and inflammatory bowel diseases.

Such inconsistencies across analyses on the same data are gradually coming to the forefront of the microbiome field (Allaband et al. 2019). Indeed, a recent plea for improved standardization has been made to enable better comparisons across studies (Hill 2020). This is a commendable goal, but given the diversity of opinions regarding best-practices (Callahan et al. 2016b; Knight et al. 2018; Schloss 2020), it is difficult to coherently recommend a single workflow for analyses at the moment. Accordingly, further work and benchmarking of different bioinformatics is needed to convincingly argue for best practices in microbiome data analysis.

Until a clear consensus is reached it is the responsibility of microbiome researchers to make the caveats and challenges facing this area clear to readers and newcomers to the field. This is crucial given the widespread interest in studying microbiomes through DNA sequencing; the number of microbiome sequencing-related publications continues to rapidly grow. This is in tandem with funding for these projects, which has steadily increased in the USA from at least 2007 to 2016 (NIH 2019). According to the US National Health Institute, there was US$766 million dollars invested in microbiome research in 2019 (https://report.nih.gov/categorical_spending.aspx), which was the 63rd most highly funded health-related research category out of 291. Although comparing across research categories of varying granularity is difficult, it is noteworthy that microbiome research was more highly funded than both breast cancer and Alzheimer's disease research. Importantly, an increased interest in microbiome research is warranted: recent technological developments are enabling improved investigations into this aspect of animal biology. However, as the monetary investment and research hours dedicated to microbiome research grows, it is crucial that scientists ensure the best use of these resources. Open discussions regarding that many aspects of microbiome data analysis are contentious and currently works in-progress would help with this issue. Indeed, such clarifications by leaders in the microbiome field are starting become more common (Allaband et al. 2019). Although these contributions are valuable, they do not adequately address the problem. In particular, instead of mentioning these issues in passing, they should be emphasized more clearly for the benefit of the uninitiated.

Another practical improvement would be to normalize, and potentially require, explicit summaries of the effects of technical variation on any biological interpretations reported in microbiome studies. This is impossible to capture entirely, but it could be done by comparing how key results change depending on a subset of representative bioinformatics choices. For instance, researchers could compare how insights change depending on the combinations of denoising tools and differential abundance methods that they have applied when analyzing 16S data. Although these changes would result in increased workloads when conducting analyses and when communicating results, they would help ensure that any major biological findings are at least robust to a representative set of bioinformatics choices.

6.2 - Microbiome-Based Classification Models

Our work applying disease classification models to Crohn's disease patients was one of numerous recent valuable investigations into the utility of leveraging microbiome data in machine learning models (Zhou & Gallins 2019). As described in Chapter 1, these investigations have successfully classified a range of diseases based on the microbiome, such as colorectal cancer (Wirbel et al. 2019), asthma (Saglani & Custovic 2019), and a range of others (Pasolli et al. 2016; Duvallet et al. 2017). However, it remains contentious how generalizable these disease classification models are across cohorts. Consistent with our results in Chapter 3, key classification models have been shown to have poor accuracy when applied to independent cohorts (Sze & Schloss 2016). But there are important exceptions. In particular, independently trained classifiers for colorectal cancer performed well with independent microbiome datasets (Wirbel et al. 2019). A global signal for non-specific disease state has also been hypothesized, as classifiers developed for certain diseases perform reasonably well when applied to cohorts with different diseases (Duvallet et al. 2017; Gupta et al. 2020). Accordingly, the generalizability of microbiome-based disease classification models remains an open area of investigation and further work should aim to address the underlying reasons for these inconsistent results.

Many other questions also remain to be answered regarding the use of microbiome data for disease classification, and more generally for classifying any arbitrary sample groupings. Several of these questions have parallels with other aspects of microbiome data analysis, in that the appropriateness of different data transformations and machine learning models remains unclear. For instance, when conducting our work on the pediatric Crohn's disease profiles, we decided to convert all relative abundance values to standard scores within samples (i.e. to scale them). We also compared how our inferences would change if centred-log ratio transformation of the data was performed instead. In this case, performing this transformation resulted in lower classification accuracies compared to our original approach (Chapter 3).

A recent high-profile project focusing on performing classification with microbiome data opted to perform arcsine square root transformation of the raw count data instead (Lloyd-Price et al. 2019). This transformation has previously been shown to yield higher classification performance with highly sparse shotgun metagenomics data (Liu et al. 2011), although it has not yet been ubiquitously accepted. Similarly, the application of different machine learning approaches remains inconsistent in the microbiome field (Zhou & Gallins 2019). Our preference for analyzing Random Forest models in Chapter 3 was largely motivated by the relatively straight-forward interpretation of the output of these models (Breiman 2001). Indeed, this approach has an intermediate level of complexity compared to other possible methods. For instance, basic linear regressions often only perform slightly worse than Random Forests and have extremely clear interpretations (Prifti et al. 2020). In contrast, deep learning and neural network approaches theoretically may allow heighted classification performance at the cost of interpretability (Namkung 2020).

A more biological open question is regarding the relative utility of functional and taxonomic data types for classification. Since publishing our work (Chapter 3) there have been few comparisons of these data types. However, these comparisons have consistently identified relatively equal classification performance with both data types (see Introduction). This recurrent observation is inconsistent with the hypothesis that environmental conditions should be more strongly associated with microbial functions than taxa (Doolittle & Booth 2017). However, more work is needed in this area to confirm that these observations are generalizable. It is possible that this may strongly depend on the environment of interest; for example, marine conditions have previously been shown to be strongly associated with functional, but not taxonomic, groupings in the ocean microbiome (Louca et al. 2016).

A similar open area of research is regarding the utility of directly integrating data types into the same classification model. Our work in Chapter 3 represents one of the only attempts to explicitly integrate multiple data types in the same model (see Introduction). I believe that this work represents an important step towards leveraging the most information out of microbiome profiles. However, our approach had several weaknesses. In particular, our integrated models included redundant features, which increased the complexity and difficulty of interpreting the output variable importance measures. I have several hypothesized solutions to this and other problems with integrating data types in classification models, which I outline in detail below (see Section 6.4).

6.3 - Metagenome Predictions

Our work on developing an improved metagenome prediction approach, PICRUSt2, was motivated by several updates that we hypothesized would increase performance. The key hypothesized improvements included: the expansion of the database, the use of sequence placement to enable de novo query sequences, and more conservative hidden-state prediction and pathway reconstruction algorithms. Although we believe these improvements are useful, they provided only a moderate increase in prediction performance compared to PICRUSt1 and other tools (Chapter 4). It is unclear to what extent this relatively minor increase in the performance metrics (e.g. the Spearman correlation coefficient) is biologically significant. This is particularly due to the unreliable nature of using MGS data as a gold-standard for comparison: if that data is imperfect then perhaps it is impossible to reach perfect concordance.

However, it is important to recognize that this increased performance was even smaller when comparing PICRUSt2 with alternative methods. In particular, Piphillin performed only slightly worse than PICRUSt2 based on our validations overall and in some cases performed better. Piphillin is a much simpler approach as it is based on taking the predicted genome annotation for input 16S sequence as the nearest neighbour in the reference database. This has the advantage of being substantially faster than PICRUSt2 because little computation is required. Another potential advantage is that it is easier to determine precisely why a predicted genome annotation was output with Piphillin. This is more challenging with PICRUSt2 because predictions for individual gene families are computed independently and are based on multiple sources of information: the annotations of all reference genomes and the inferred pattern of gain and loss in the reference phylogenetic tree.

Since PICRUSt2 was made available it has been independently used on many occasions. Most relevantly several of these independent users have conducted evaluations of our new approach compared to other metagenome prediction tools. The Piphillin authors determined that their tool performs better compared to PICRUSt2 on a small dataset of oral samples from control and cancer patients (Narayan et al. 2020). In particular, they determined that Piphillin had 54% higher precision compared to PICRUSt2 in terms of the concordance of differential abundance results with matching MGS data. These results were the key motivation for us to implement our independent comparisons based on differential abundance testing in our manuscript. As I showed in Chapter 4, there are often variable results in terms of concordance based on differential testing and so inferences are best made over a larger set of validation datasets. Based on my analysis the performance of both tools was quite similar overall based on this approach, although PICRUSt2 performed slightly better.

Another research group compared PICRUSt1, PICRUSt2, and Tax4Fun2 across seven paired MGS and 16S datasets and were unable to find a single tool that clearly performed best (Sun et al. 2020). However, they suggested that metagenome predictions are mainly useful for human-associated datasets as they observed decreased accuracy in five non-human datasets compared to two human datasets. These inferences were largely based on the concordance of differential abundance testing as well. This is an interesting observation, which will require larger numbers of test datasets to validate. However, it is noteworthy that the primate stool dataset I tested using a similar approach performed similarly to the human-associated datasets (Chapter 4).

These independent evaluations highlight that the inferred performance of metagenome prediction tools can highly vary, which could be due to variation across datasets, but also due to the exact workflow for evaluating performance. Nonetheless, although the hidden-state prediction approaches implemented in the PICRUSt tools may not necessarily always result in the highest performance accuracy, they still fundamentally are capable of predicting functional patterns across genomes that other tools cannot. In particular, the likelihood of different possible predictions based on existing taxa can be computed with hidden-state prediction approaches (Zaneveld & Thurber 2014).

These approaches, and specifically PICRUSt2, will likely perform better as higher-quality data is used as input. In particular, full-length ASV sequencing is rapidly becoming more common which enables higher resolution to distinguish closely related taxa (Callahan et al. 2019). It is foreseeable that additional marker genes or larger single regions could also be profiled that would enable strains to be better distinguished. Similarly, as the number of reference genomes continues to grow, it will be possible to create more environment-specific genome databases for more targeted prediction (Wilkinson et al. 2018).

A specific advantage of hidden-state prediction approaches is that a wide variety of algorithms can be implemented for conducting these predictions. For example, Markov models could be constructed for each individual gene family in a database (Louca & Doebeli 2018). This approach might better capture the specific pattern of gain and loss for given gene families than maximum parsimony. I attempted to implement this approach while developing PICRUSt2, but it currently takes a computationally prohibitively amount of time to run. However, as improved algorithms and computing resources continue to grow it is foreseeable that implementing individualized models for predicting gene families will soon be feasible, which would allow for improved prediction performance.

Similarly, further improvements to genome reference databases and construction of phylogenetic trees may make the benefits of phylogenetic placement used with PICRUSt2 clearer. This approach was expected to perform better than simply re-computing the phylogenetic tree of query and reference sequences for each dataset. However, based on the Spearman correlation coefficient validations, using phylogenetic placement had no consistent impact on the performance of PICRUSt2 compared with re-computing trees with FastTree (Chapter 4). We anticipate this will change as data quality improves over time as phylogenetic placement has previously been shown to yield more robust inferences for standard 16S analyses compared to creating de novo phylogenetic trees (Janssen et al. 2018).

A final important discussion point regarding metagenome prediction, is that it is frequently observed that metagenome predictions are becoming irrelevant given the availability of cost-effective MGS data (Hillmann et al. 2018). Although MGS data is indeed becoming more readily available, I believe metagenome predictions will remain relevant for two key reasons. First, marker-gene sequencing is currently the only feasible approach for profiling low-biomass and host-DNA contaminated samples. Although it is possible to apply alternative approaches like MGS to these sample types (as we did in

Chapter 3), it typically is not cost-effective. Accordingly, metagenome predictions of this marker-gene sequencing data will likely remain the predominant method for inferring microbial functions for these sample types. Second, a major advantage of metagenome prediction is that it enables complete taxa-function links to be generated. As discussed in Chapter 1, generating these links is uncommonly performed with MGS data. Although this is changing, for the time being metagenome predictions are, perhaps counter-intuitively, a straight-forward method to generate complete taxa-function links.

6.4 - Joint Taxa-Function Analysis

As described in Chapters 1 and 5, joint analyses of functional and taxonomic data greatly increase the interpretability of microbiome data. The framework I investigated, as implemented in POMS, is a useful approach for identifying functional microbial biomarkers based on phylogenetic balances. It is an imperfect approach as discussed in detail in Chapter 5. Nonetheless, it also represents a qualitatively different framework compared to standard approaches for conducting differential abundance testing. Namely, the POMS framework is useful for identifying gene families that are putatively under selection. This was shown based on our simulations where POMS performed well at identifying selection acting on gene families encoded by at least several MAGs. In terms of high-level gene families this likely captures the vast majority of the segregating genetic variation: as stated in Chapter 1, KEGG orthologs are encoded by a mean of 184.3 species (Inkpen et al. 2017).

The POMS framework is based on phylogenetic balances, because this approach provides a convenient and sensible way of computing ratios between groups of taxa (Silverman et al. 2017). However, there is a diverse range of other reference frame-based analyses that can be conducted with microbiome data (Morton et al. 2019). It remains contentious which is the best approach for determining appropriate sets of taxa to consider for reference frames. The basic idea of the POMS framework, to test for consistent functional enrichments across independent reference frames, could be expanded to these alternative approaches as well.

Improved integration of microbiome data types for biomarker identification is clearly needed, which POMS partially addresses. However, there are many other areas where further development of integrated analyses is needed. One area where this is particularly needed is in the context of classification models, as discussed above. Recent work has involved applying classification models based on gene families and then subsequently identifying metagenome assembled genomes within a given dataset enriched for the top genes (Rahman et al. 2018). This still relies on follow-up analyses rather than integrating the data types. Instead, an improved approach could leverage the explicit hierarchical nature of microbiome data types. Functional and taxonomic data types form clear hierarchical structures independently (e.g. Phylum - Class - Order, etc.). The connection between taxa and gene families and pathways is more complex, but nonetheless, links between groups of strains or ASVs and microbial functions can be defined. A modified machine learning framework that explicitly accounted for these relationships could result in more interpretable outputs.

For example, a Random Forest model could potentially be modified to account for these relationships. In each decision tree within a given Random Forest model a subset of features are randomly sampled at each node and the feature providing the best split of the sample groupings at that point is identified. It is possible that this could be modified so that the most informative features in a given hierarchy could be identified and that the least granular feature that was most informative would be selected. For instance, if the relative abundance of a phylum and a gene family were equally informative then it would be more conservative to identify the phylum as providing the best split at that node. This would provide the benefit that any interpretations regarding specific functions would require them to be more informative than any single taxonomic contributor. In other words, biological interpretations regarding functions would require stronger evidence.

Although I believe this general framework is promising it would still have important caveats. In particular, it is unclear how features would be sampled for testing in each individual decision tree to ensure that there was not a bias towards certain data type ranks simply due to differences in the total number of features. Similarly, the best approach for controlling for the compositionality of the data would need to be determined while using this approach. Despite these major caveats, I believe this approach, or a similar hierarchical modification of standard machine learning models, would greatly improve the interpretability of microbiome analyses.

Regardless of whether POMS or this proposed alternative method eventually becomes widely adopted, an increased focus on integrating microbiome data types is needed. As stated at the beginning of this work, it is odd to distinguish between functional and taxonomic datatypes: they are inextricably linked after all. The term "metagenome" itself is in some ways unfortunate as it implies that the genetic information for all organisms in a community can be simultaneously analyzed in a coherent way. This may be valid for high-level pathways (as discussed in Chapter 5), but for generating hypotheses regarding specific gene families it is too often misleading. This perspective is becoming more common, particularly as the availability of metagenome-assembled genomes increases (Frioux et al. 2020).

Despite these open questions, my contributions to the improved understandings of classification models, metagenome prediction, and functional biomarker identification have provided useful insights that will enable better informed taxa-function analyses moving forward. A common thread throughout my work has been the recurrent observation of inconsistencies across bioinformatics tools. As I have discussed at length, better integration of taxonomic and functional data types could enable more conservative hypotheses to be generated based on microbiome data, which would help reduce the burden of false positives. However, such improvements do not only yield statistical benefits: they also make microbiome data easier to interpret.

References

1000 Genome Project Consortium. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Abu-Ali GS et al. 2018. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat. Microbiol. 3:356–366.

Abubucker S et al. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLOS Comput. Biol. 8:e1002358.

Adams SH. 2011. Emerging perspectives on essential amino acid metabolism in obesity and the insulin-resistant state. Adv. Nutr. 2:445–456.

Aitchison J. 1982. The Statistical Analysis of Compositional Data. J. R. Stat. Soc. Ser. B. 44:139–177.

Alemzadeh N et al. 2002. Adult height in patients with early onset of Crohn’s disease. Gut. 51:26–29.

Allaband C et al. 2019. Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. Clin. Gastroenterol. Hepatol. 17:218–230.

Almeida A et al. 2019. A new genomic blueprint of the human gut microbiota. Nature. 568:499–504.

Alneberg J et al. 2014. Binning metagenomic contigs by coverage and composition. Nat. Methods. 11:1144–1146.

Amato KR et al. 2019. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. ISME J. 13:576–587.

Ambler RP et al. 1979. Cytochrome C2 sequence variation among the recognised species of purple nonsulphur photosynthetic bacteria. Nature. 278:659–660.

Amir A et al. 2017. Deblur Rapidly Resolves Single- Nucleotide Community Sequence Patterns. mSystems. 2:e00191-16.

Ananthakrishnan AN. 2015. Epidemiology and risk factors for IBD. Nat Rev Gastroenterol Hepatol. 12:205–217.

Angly FE et al. 2014. CopyRighter: A rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome. 2:11.

Apweiler R et al. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32:D115-119.

Arboleya S et al. 2018. Gene-trait matching across the *Bifidobacterium longum* pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. BMC Genomics. 19:33.

Armour CR, Nayfach S, Pollard KS, Sharpton TJ. 2019. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. mSystems. 4:e00332-18.

Arrieta M-C et al. 2015. Early infancy microbial and metabolic alterations affect risk of childhood asthma. Sci. Transl. Med. 7:307ra152.

Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. 2015. Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 31:2882–2884.

Atarashi K et al. 2013. Treg induction by a rationally selected mixture of *Clostridia* strains from the human microbiota. Nature. 500:232–236.

Van der Auwera GA et al. 2013. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc.

Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. Brief. Bioinform. 00:1–11.

Bäckhed F et al. 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 17:690–703.

Barbera P et al. 2019. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Syst. Biol. 68:365–369.

Barott KL et al. 2012. Microbial to reef scale interactions between the reef-building coral *Montastraea annularis* and benthic algae. Proc. R. Soc. B Biol. Sci. 279:1655–1664.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. USA. 102:14332–14337.

Benchimol EI et al. 2017. Trends in Epidemiology of Pediatric Inflammatory Bowel Disease in Canada: Distributed Network Analysis of Multiple Population-Based Provincial Health Administrative Databases. Am. J. Gastroenterol. 112:1120–1134.

Bengtsson-Palme J et al. 2013. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol. Evol. 4:914–919.

Bengtsson-Palme J et al. 2015. METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Mol. Ecol. Resour. 15:1403–1414.

Benjamini Y, Yekutieli D. 2001. The Control of the False Discovery Rate in Multiple Testing Under Dependency. Ann. Stat. 29:1165–1188.

Bernell O, Lapidus A, Hellers G. 2000. Risk factors for surgery and postoperative recurrence in Crohn's disease. Ann. Surg. 231:38–45.

Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta: Scalable de novo metagenome assembly and profiling. Genome Biol. 13:R122.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Bolyen E et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37:852–857.

Bonder MJ et al. 2016. The effect of host genetics on the gut microbiome. Nat. Genet.

48:1407–1415.

Bonovas S et al. 2016. Biologic Therapies and Risk of Infection and Malignancy in Patients With Inflammatory Bowel Disease: A Systematic Review and Network Meta-analysis. Clin. Gastroenterol. Hepatol. 14:1385–1397.

Bowers RM et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35:725–731.

Bowman JS, Ducklow HW. 2015. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the coastal West Antarctic Peninsula. PLOS One. 10:e0135868.

Bradley PH, Nayfach S, Pollard KS. 2018. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. PLOS Comput. Biol. 14:e1006242.

Bradley PH, Pollard KS. 2020. Phylogenize: Correcting for phylogeny reveals genes associated with microbial distributions. Bioinformatics. 36:1289–1290.

Brant SR. 2011. Update on the heritability of inflammatory bowel disease: The importance of twin studies. Inflamm. Bowel Dis. 17:1–5.

Breiman L. 2001. Random Forests. Mach. Learn. 45:5–32.

Brenner DJ. 1973. Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. Int. J. Syst. Bacteriol. 23:298–307.

Brown CT et al. 2011. Gut Microbiome Metagenomics Analysis Suggests a Functional Model for the Development of Autoimmunity for Type 1 Diabetes. PLOS One. 6:e25792.

Buchfink B, Xie C, Huson DH. 2015. Fast and Sensitive Protein Alignment using DIAMOND. Nat. Methods. 12:59–60.

Bukin YS et al. 2019. The effect of 16s rRNA region choice on bacterial community metabarcoding results. Sci. Data. 6:190007.

Burke C, Steinberg P, Rusch DB, Kjelleberg S, Thomas T. 2011. Bacterial community assembly based on functional genes rather than species. Proc. Natl. Acad. Sci. USA. 108:14288–14293.

Byndloss MX, Pernitzsch SR, Bäumler AJ. 2018. Healthy hosts rule within: ecological forces shaping the gut microbiota. Mucosal Immunol. 11:1299–1305.

Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. 2020. Assessment of single cell RNA-seq statistical methods on microbiome data. Genome Biol. 191.

Callahan BJ et al. 2016a. DADA2: High resolution sample inference from amplicon data. Nat. Methods. 13:581–583.

Callahan BJ et al. 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 47:e103.

Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. 2016b. Bioconductor workflow for microbiome data analysis: From raw reads to community analyses. F1000 Res. 5:1492.

Carrodeguas L, Kaidar-Person O, Szomstein S, Antozzi P, Rosenthal R. 2005. Preoperative thiamine deficiency in obese population undergoing laparoscopic bariatric surgery. Surg. Obes. Relat. Dis. 1:517–522.

Casati J, Toner BB. 2000. Psychosocial aspects of inflammatory bowel disease. Biomed. Pharmacother. 54:388–393.

Caspi R et al. 2013. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 42:D459–D471.

Castaner O et al. 2018. The gut microbiome profile in obesity: A systematic review. Int. J. Endocrinol. 2018:4095789.

Chaffron S, Rehrauer H, Pernthaler J, Von Mering C. 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 20:947–959.

Chen IMA et al. 2013. Improving Microbial Genome Annotations in an Integrated Database Context. PLOS One. 8:e54859.

Chen PE et al. 2010. Genomic characterization of the *Yersinia* genus. Genome Biol. 11:R1.

Chen Z et al. 2019. Impact of Preservation Method and 16S rRNA Hypervariable Region on Gut Microbiota Profiling. mSystems. 4:e00271-18.

Chin CS et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods. 10:563–569.

Cho JH. 2008. The genetics and immunopathogenesis of inflammatory bowel disease. Nat. Rev. Immunol. 8:458–466.

Clarke G et al. 2019. Gut reactions: Breaking down xenobiotic–microbiome interactions. Pharmacol. Rev. 71:198–224.

Cleynen I et al. 2016. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. Lancet. 387:156–167.

Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems. 2:e00127-16.

Comeau AM, Lagunas MG, Scarcella K, Varela DE, Lovejoy C. 2019. Nitrate Consumers in Arctic Marine Eukaryotic Communities: Comparative Diversities of 18S rRNA, 18S rRNA Genes, and Nitrate Reductase Genes. Appl. Environ. Microbiol. 85:e00247-19.

Cosnes J, Gowerrousseau C, Seksik P, Cortot A. 2011. Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology. 140:1785–1794.

Coupaye M et al. 2009. Nutritional consequences of adjustable gastric banding and gastric bypass: A 1-year prospective study. Obes. Surg. 19:56–65.

Critch J et al. 2012. Use of enteral nutrition for the control of intestinal inflammation in pediatric crohn disease. J. Pediatr. Gastroenterol. Nutr. 54:298–305.

De Cruz P et al. 2015. Association between specific mucosa-associated microbiota in Crohn's disease at the time of resection and subsequent disease recurrence: A pilot study. J. Gastroenterol. Hepatol. 30:268–278.

Czech L, Stamatakis A. 2019. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. PLOS One. 14:e0217050.

Danecek P et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

Darling AE et al. 2014. PhyloSift: Phylogenetic analysis of genomes and metagenomes. PeerJ. 2014:243.

Davenport M et al. 2014. Metabolic alterations to the mucosal microbiota in inflammatory bowel disease. Inflamm. Bowel Dis. 20:723–731.

Day AS et al. 2006. Exclusive enteral feeding as primary therapy for Crohn's disease in Australian children and adolescents: A feasible and effective approach. J. Gastroenterol. Hepatol. 21:1609–1614.

Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. Nat. Methods. 9:179–81.

DePristo MA et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.

DeSantis TZ et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069–5072.

Devereux R et al. 1990. Diversity and origin of *Desulfovibrio* species: Phylogenetic definition of a family. J. Bacteriol. 172:3609–3619.

Dhakan DB et al. 2019. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. Gigascience. 8:1–20.

Dinh DM et al. 2015. Intestinal Microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. J. Infect. Dis. 211:19–27.

Doolittle WF, Booth A. 2017. It's the song, not the singer: an exploration of holobiosis and evolutionary theory. Biol. Philos. 32:5–24.

Douglas GM et al. 2018. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 6:13.

Douglas GM et al. 2020. PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. 38:685–688.

Douglas GM, Langille MGI. 2019. Current and promising approaches to identify horizontal gene transfer events in metagenomes. Genome Biol. Evol. 11:2750–2766.

Drouin G, Godin J-R, Pagé B. 2011. The genetics of vitamin C loss in vertebrates. Curr. Genomics. 12:371–8.

Dunn KA et al. 2016a. Early Changes in Microbial Community Structure Are Associated with Sustained Remission After Nutritional Treatment of Pediatric Crohn's Disease. Inflamm. Bowel Dis. 22:2853–2862.

Dunn KA et al. 2016b. The Gut Microbiome of Pediatric Crohn's Disease Patients Differs from Healthy Controls in Genes That Can Influence the Balance Between a Healthy and Dysregulated Immune Response. Inflamm. Bowel Dis. 22:2607–2618.

Durack J, Lynch S V. 2019. The gut microbiome: Relationships with disease and opportunities for therapy. J. Exp. Med. 216:20–40.

Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. 8:1784.

Edgar RC. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. https://doi.org/10.1101/081257.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 27:2194–2200.

Egozcue JJ, Pawlowsky-Glahn V. 2011. Exploring compositional data with the CoDa-dendrogram. AUSTRIAN J. Stat. 40:103–113.

Ekbom A, Helmick C, Zack M, Adami HO. 1991. The epidemiology of inflammatory bowel disease: A large, population-based study in Sweden. Gastroenterology. 100:350–358.

Elshorbagy AK, Kozich V, David Smith A, Refsum H. 2012a. Cysteine and obesity: Consistency of the evidence across epidemiologic, animal and cellular studies. Curr. Opin. Clin. Nutr. Metab. Care. 15:49–57.

Elshorbagy AK, Valdivia-Garcia M, Refsum H, Butte N. 2012b. The Association of Cysteine with Obesity, Inflammatory Cytokines and Insulin Resistance in Hispanic Children and Adolescents. PLOS One. 7:e44166.

Eng A, Borenstein E. 2018. Taxa-function robustness in microbial communities. Microbiome. 6:45.

Farrelly V, Rainey FA, Stackebrandt E. 1995. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl. Environ. Microbiol. 61:2798–2801.

Faust K et al. 2012. Microbial co-occurrence relationships in the Human Microbiome. PLOS Comput. Biol. 8.

Feeney MA et al. 2002. A case-control study of childhood environmental risk factors for the development of inflammatory bowel disease. Eur. J. Gastroenterol. Hepatol. 14:529–534.

Felsenstein J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1–15.

Fernandes AD et al. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2:15.

Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. 2013. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLOS One. 8:e67019.

Finlayson-Trick ECL et al. 2017. Taxonomic differences of gut microbiomes drive cellulolytic enzymatic potential within hind-gut fermenting mammals. PLOS One. 12:e0189404.

Finn RD et al. 2014. Pfam: The protein families database. Nucleic Acids Res. 42:D222–D230.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279–284.

Flemer B et al. 2018. The oral microbiota in colorectal cancer is distinctive and predictive. Gut. 67:1454–1463.

Francke C, Siezen RJ, Teusink B. 2005. Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol. 13:550–558.

Frank DN et al. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc. Natl. Acad. Sci. USA. 104:13780–13785.

Franzosa EA et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods. 15:962–968.

Friedman J, Alm EJ. 2012. Inferring Correlation Networks from Genomic Survey Data. PLOS Comput. Biol. 8:e1002687.

Frioux C, Singh D, Korcsmaros T, Hildebrand F. 2020. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput. Struct. Biotechnol. J. 18:1722–1734.

Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin E V. 2019. Microbial genome analysis: The COG approach. Brief. Bioinform. 20:1063–1070.

Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 43:D261–D269.

Gao C-H. 2019. ggVennDiagram: A ‘ggplot2’ Implement of Venn Diagram.

Garud NR, Pollard KS. 2020. Population Genetics in the Human Microbiome. Trends Genet. 36:53–67.

Geuking MB et al. 2011. Intestinal Bacterial Colonization Induces Mutualistic Regulatory T Cell Responses. Immunity. 34:794–806.

Gevers D et al. 2014. The treatment-naive microbiome in new-onset Crohn’s disease. Cell Host Microbe. 15:382–392.

Gillies LE, Thrash JC, deRada S, Rabalais NN, Mason OU. 2015. Archaeal enrichment in the hypoxic zone in the northern Gulf of Mexico. Environ. Microbiol. 17:3847–3856.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: And this is not optional. Front. Microbiol. 8:2224.

Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. 2016. It's all relative: analyzing microbiome data as compositions. Ann. Epidemiol. 26:322–329.

Gogokhia L et al. 2019. Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. Cell Host Microbe. 25:285–299.

Gonzalez A et al. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods. 15:796–798.

Goodhand J et al. 2010. Inflammatory bowel disease in young people: The case for transitional clinics. Inflamm. Bowel Dis. 16:947–952.

Goodrich JK et al. 2014. Conducting a microbiome study. Cell. 158:250–262.

Graspeuntner S, Loeper N, Künzel S, Baines JF, Rupp J. 2018. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. Sci. Rep. 8:9678.

Greenblum S, Chiu HC, Levy R, Carr R, Borenstein E. 2013. Towards a predictive systems-level model of the human microbiome: Progress, challenges, and opportunities. Curr. Opin. Biotechnol. 24:810–820.

Greenblum S, Turnbaugh PJ, Borenstein E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Natl. Acad. Sci. USA. 109:594–599.

Guo F, Zhang T. 2013. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. Appl. Microbiol. Biotechnol. 97:4607–4616.

Gupta NK et al. 2012. Serum analysis of tryptophan catabolism pathway: Correlation with Crohn's disease activity. Inflamm. Bowel Dis. 18:1214–1220.

Gupta V et al. 2015. RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. Gigascience. 4:5.

Gupta VK et al. 2020. A predictive index for health status using species-level gut microbiome profiling. Nat. Commun. 11:4635.

Haberman Y et al. 2014. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J. Clin. Invest. 124:3617–3633.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31:371–373.

Haider B et al. 2014. Omega: An Overlap-graph de novo Assembler for Metagenomics. Bioinformatics. 30:2717–2722.

Halfvarson J et al. 2006. Environmental factors in inflammatory bowel disease: A co-twin control study of a Swedish-Danish twin population. Inflamm. Bowel Dis. 12:925–933.

Halfvarson J, Bodin L, Tysk C, Lindberg E, Järnerot G. 2003. Inflammatory bowel disease in a Swedish twin cohort: A long- term follow-up of concordance and clinical characteristics. Gastroenterology. 124:1767–1773.

Hallam SJ, Girguis PR, Preston CM, Richardson PM, DeLong EF. 2003. Identification of Methyl Coenzyme M Reductase A (mcrA) Genes Associated with Methane-Oxidizing Archaea. Appl. Environ. Microbiol. 69:5483–5491.

Halme L et al. 2006. Family and twin studies in inflammatory bowel disease. World J. Gastroenterol. 12:3668–3672.

Hansen R et al. 2012. Microbiota of de-novo pediatric IBD: increased Faecalibacterium prausnitzii and reduced bacterial diversity in Crohn's but not in ulcerative colitis. Am. J. Gastroenterol. 107:1913–22.

Hansen R et al. 2013. The Microaerophilic Microbiota of De-Novo Paediatric Inflammatory Bowel Disease: The BISCUIT Study. PLOS One. 8:e58825.

Hao X, Chen T. 2012. OTU Analysis Using Metagenomic Shotgun Sequencing Data. PLOS One. 7:e49785.

Hauben L, Vauterin L, Moore ERB, Hoste B, Swings J. 1999. Genomic diversity of the genus *Stenotrophomonas*. Int. J. Syst. Bacteriol. 49:1749–1760.

Hauben L, Vauterin L, Swings J, Moore ERB. 1997. Comparison of 16S Ribosomal DNA Sequences of All *Xanthomonas* Species. Int. J. Syst. Bacteriol. 47:328–335.

Hawinkel S, Mattiello F, Bijnens L, Thas O. 2019. A broken promise: Microbiome differential abundance methods do not control the false discovery rate. Brief. Bioinform. 20:1–12.

Henderson P et al. 2012. Rising incidence of pediatric inflammatory bowel disease in Scotland. Inflamm. Bowel Dis. 18:999–1005.

Hill C. 2020. You have the microbiome you deserve. Gut Microbiome. 1:1–4.

Hillman ET, Lu H, Yao T, Nakatsu CH. 2017. Microbial ecology along the gastrointestinal tract. Microbes Environ. 32:300–313.

Hillmann B et al. 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 3:e00069-18.

HMP-consortium. 2013. Structure, Function and Diversity of the Healthy Human Microbiome. Nature. 486:207–214.

Ho F, Khalil H. 2015. Crohn's disease: A clinical update. Therap. Adv. Gastroenterol. 8:352–359.

Hou J, Abraham B, El-Serag H. 2011. Dietary Intake and Risk of Developing Inflammatory Bowel Disease: A Systematic Review of the Literature. Am. J. Gastroenterol. 106:563–573.

Howie B, Marchini J, Stephens M. 2011. Genotype Imputation with Thousands of Genomes. G3. 1:457–470.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLOS Genet. 5:e1000529.

Huda-Faujan N et al. 2010. The Impact of the Level of the Intestinal Short Chain Fatty Acids in Inflammatory Bowel Disease Patients Versus Healthy Subjects. Open Biochem. J. 4:53–58.

Hufnagl K, Pali-Schöll I, Roth-Walter F, Jensen-Jarolim E. 2020. Dysbiosis of the gut and lung microbiome has a role in asthma. Semin. Immunopathol. 42:75–93.

Hugot J et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. Nature. 411:599–603.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Res. 17:377–386.

Huttenhower C et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature. 486:207–214.

Ibrahim A, Goebel BM, Liesack W, Griffiths M, Stackebrandt E. 1993. The phylogeny of the genus Yersinia based on 16S rDNA sequences. FEMS Microbiol. Lett. 114:173–177.

Inkpen AI et al. 2017. The Coupling of Taxonomy and Function in Microbiomes. Biol. Philos. 32:1225–1243.

Ivanov II et al. 2009. Induction of Intestinal Th17 Cells by Segmented Filamentous Bacteria. Cell. 139:485–498.

Iwai S et al. 2016. Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. PLOS One. 11:e0166104.

Jabandziev P et al. 2020. Regional Incidence of Inflammatory Bowel Disease in a Czech Pediatric Population: 16 Years of Experience (2002-2017). J. Pediatr. Gastroenterol. Nutr. 70:586–592.

Jackson DA. 1997. Compositional data in community ecology: The paradigm or peril of proportions? Ecology. 78:929–940.

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J. Clin. Microbiol. 45:2761–2764.

Janssen S et al. 2018. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. mSystems. 3:e00021.

Jensen LJ et al. 2008. eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36:D250–D254.

Johnson JS et al. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat. Commun. 10:5029.

Jones MB et al. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc. Natl. Acad. Sci. USA. 112:14024–14029.

Jost L. 2006. Entropy and diversity. Oikos. 113:363–375.

Jostins L et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 491:119–24.

Jostins L, Levine AP, Barrett JC. 2013. Using Genetic Prediction from Known Complex Disease Loci to Guide the Design of Next-Generation Sequencing Experiments. PLOS One. 8:e76328.

Jun SR, Robeson MS, Hauser LJ, Schadt CW, Gorin AA. 2015. PanFP: Pangenome-based functional profiles for microbial communities. BMC Res. Notes. 8:479.

Kaakoush NO et al. 2015. Effect of Exclusive Enteral Nutrition on the Microbiota of Children With Newly Diagnosed Crohn's Disease. Clin. Transl. Gastroenterol. 6:e71.

Kaakoush NO. 2015. Insights into the Role of Erysipelotrichaceae in the Human Host. Front. Cell. Infect. Microbiol. 5:84.

Kalan LR et al. 2019. Strain- and Species-Level Variation in the Microbiome of Diabetic Wounds Is Associated with Clinical Outcomes and Therapeutic Efficacy. Cell Host Microbe. 25:641–655.

Kallonen T et al. 2017. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. 27:1437–1449.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28:27–30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40:109–114.

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44:D457–D462.

Kang CH et al. 2007. Relationship between genome similarity and DNA-DNA hybridization among closely related bacteria. J. Microbiol. Biotechnol. 17:945–951.

Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 3:e1165.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kazemi A, Frazier T, Cave M. 2010. Micronutrient-related neurologic complications following bariatric surgery. Curr. Gastroenterol. Rep. 12:288–295.

Keinan A, Sandbank B, Hilgetag CC, Meilijson I, Ruppin E. 2004. Fair attribution of functional contribution in artificial and biological networks. Neural Comput. 16:1887–1915.

Kerns JC, Arundel C, Chawla LS. 2015. Thiamin Deficiency in People with Obesity. Adv. Nutr. 6:147–153.

Keshavarzian A et al. 2003. Increases in free radicals and cytoskeletal protein oxidation and nitration in the colon of patients with inflammatory bowel disease. Gut. 52:720–728.

Keswani J, Whitman WB. 2001. Relationship of 16S rRNA sequence similarity to DNA hybridization in prokaryotes. Int. J. Syst. Evol. Microbiol. 51:667–678.

Khan I et al. 2019. Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome. Pathogens. 8:126.

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and accurate classificaton of metagenomic sequences. Genome Res. 26:1721–1729.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol. Biol. Evol. 28:1057–1074.

Knight R et al. 2018. Best practices for analysing microbiomes. Nat. Rev. Microbiol. 16:410–422.

Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. 2011. Human-associated microbial signatures: Examining their predictive value. Cell Host Microbe. 10:292–296.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA. 102:2567–2572.

Koonin E V, Galperin MY. 2003. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic: Boston.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 28:3211–3217.

Koutsovoulos G et al. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc. Natl. Acad. Sci. USA. 113:5053–5058.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 1–3.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28:1–26.

Kump P et al. 2018. The taxonomic composition of the donor intestinal microbiota is a major factor influencing the efficacy of faecal microbiota transplantation in therapy refractory ulcerative colitis. Aliment. Pharmacol. Ther. 47:67–77.

Kurtz ZD et al. 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Comput. Biol. 11:e1004226.

Langille MGI. 2018. Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. mSystems. 3:e00163-17.

Langille MGI et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31:814–821.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–9.

Laserna-Mendieta EJ et al. 2018. Determinants of reduced genetic capacity for butyrate synthesis by the gut microbiome in Crohn's disease and ulcerative colitis. J. Crohn's Colitis. 12:204–216.

Lässig M, Mustonen V, Walczak AM. 2017. Predicting evolution. Nat. Ecol. Evol. 1:0077.

Lau JT et al. 2016. Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. Genome Med. 8:72.

Law CW, Chen Y, Shi W, Smyth GK. 2014. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15:R29.

Lee JC et al. 2017. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. Nat. Genet. 49:262–268.

Lee MD, Ponty Y. 2019. GToTree: A user-friendly workflow for phylogenomics. Bioinformatics. 35:4162–4164.

Levine A et al. 2011. Pediatric modification of the Montreal classification for inflammatory bowel disease: The Paris classification. Inflamm. Bowel Dis. 17:1314–1321.

Lewis JD et al. 2015. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe. 18:489–500.

Ley RE et al. 2005. Obesity alters gut microbial ecology. Proc. Natl. Acad. Sci. USA. 102:11070–11075.

Ley RE, Peterson DA, Gordon JI. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 124:837–848.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31:1674–1676.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News. 2:18–22.

Litvak Y, Byndloss MX, Bäumler AJ. 2018. Colonocyte metabolism shapes the gut microbiota. Science. 362:eaat9076.

Liu J, Yu Y, Cai Z, Bartlam M, Wang Y. 2015a. Comparison of ITS and 18S rDNA for estimating fungal diversity using PCR–DGGE. World J. Microbiol. Biotechnol. 31:1387–1395.

Liu JZ et al. 2015b. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47:979–989.

Liu JZ, Anderson CA. 2014. Genetic studies of Crohn's disease: Past, present and future. Best Pract. Res. Clin. Gastroenterol. 28:373–386.

Liu Z, Hsiao W, Cantarel BL, Drábek EF, Fraser-Liggett C. 2011. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. Bioinformatics. 27:3242–3249.

Lloyd-Price J et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 569:655–662.

Lloyd-Price J et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 550:61–66.

Lokmer A et al. 2019. Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLOS One. 14:e0211139.

Louca S et al. 2018a. Function and functional redundancy in microbial systems. Nat. Ecol. Evol. 2:936–943.

Louca S, Doebeli M. 2018. Efficient comparative phylogenetics on large trees. Bioinformatics. 34:1053–1055.

Louca S, Doebeli M. 2017. Taxonomic variability and functional stability in microbial communities infected by phages. Environ. Microbiol. 19:3863–3878.

Louca S, Doebeli M, Parfrey LW. 2018b. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 6:41.

Louca S, Parfrey LW, Doebeli M. 2016. Decoupling function and taxonomy in the global ocean microbiome. Science. 353:1272–1277.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl. Environ. Microbiol. 71:8228–8235.

Lozupone CA et al. 2008. The convergence of carbohydrate active gene repertoires in human gut microbes. Proc. Natl. Acad. Sci. USA. 105:15076–15081.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput. Sci. 3:e104.

Ma B et al. 2020. Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. Microbiome. 8:82.

MacLellan A et al. 2017. The impact of exclusive enteral nutrition (EEN) on the gut microbiome in Crohn's disease: A review. Nutrients. 9:447.

Makarova KS, Wolf YI, Koonin E V. 2015. Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. Life. 5:818–840.

Mandal S et al. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Heal. Dis. 26:27663.

Mandel M. 1966. Deoxyribonucleic Acid Base Composition in the Genus *Pseudomonas*. J. Gen. Microbiol. 43:273–292.

Manichanh C et al. 2006. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 55:205–11.

Manor O, Borenstein E. 2015. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biol. 16:53.

Manor O, Borenstein E. 2017a. Revised computational metagenomic processing uncovers hidden and biologically meaningful functional variation in the human microbiome. Microbiome. 5:19.

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

Markowitz VM et al. 2012. IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40:115–122.

Martin BD, Witten D, Willis AD. 2020. Modeling microbial abundances and dysbiosis with beta-binomial regression. Ann. Appl. Stat. 14:94–115.

Martiny AC. 2019. High proportions of bacteria are culturable across major biomes. ISME J. 13:2125–2128.

Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 7:830–838.

Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 11:538.

Maukonen J, Simões C, Saarela M. 2012. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. FEMS Microbiol. Ecol. 79:697–708.

McDonald D et al. 2019. redbiom: a Rapid Sample Discovery and Feature Characterization System. mSystems. 4:e00215-19.

McGovern DPB, Kugathasan S, Cho JH. 2015. Genetics of Inflammatory Bowel Diseases. Gastroenterology. 149:1163–1176.

McIntyre ABR et al. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 18:182.

McKenna A et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McMahon K. 2015. 'Metagenomics 2.0'. Environ. Microbiol. Rep. 7:38–39.

Meyer F et al. 2018. AMBER: Assessment of Metagenome BinnERs. Gigascience. 7:1–8.

Meyer F et al. 2008. The metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC

Bioinformatics. 9:386.

Meyer F, Overbeek R, Rodriguez A. 2009. FIGfams: Yet another set of protein families. Nucleic Acids Res. 37:6643–6654.

Miossec MJ et al. 2020. Evaluation of computational methods for human microbiome analysis using simulated data. PeerJ. 8:e9688.

Molodecky NA et al. 2012. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 142:46–54.

Mondot S et al. 2011. Highlighting new phylogenetic specificities of Crohn's disease microbiota. Inflamm. Bowel Dis. 17:185–192.

Mondot S et al. 2016. Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery. Gut. 65:954–62.

Morain CO, Segal AW, Levi AJ. 1984. Elemental diet as primary treatment of acute Crohn's disease: A controlled trial. Br. Med. J. 288:1859–1862.

Moreira LO, Zamboni DS. 2012. NOD1 and NOD2 signaling in infection and inflammation. Front. Immunol. 3.

Morgan XC et al. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 13:R79.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. MBio. 3:e00036-12.

Morton ER et al. 2015. Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by *Entamoeba* and Subsistence. PLOS Genet. 11:e1005658.

Morton JT et al. 2017. Balance Trees Reveal Microbial Niche Differentiation. mSystems. 2:e00162-16.

Morton JT et al. 2019. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 10:2719.

Mosca A, Leclerc M, Hugot JP. 2016. Gut microbiota diversity and human diseases: Should we reintroduce key predators in our ecosystem? Front. Microbiol. 7:455.

Mottawea W et al. 2016. Altered intestinal microbiota-host mitochondria crosstalk in new onset Crohn's disease. Nat. Commun. 7:13419.

Moya A, Ferrer M. 2016. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. Trends Microbiol. 24:402–413.

Muegge BD et al. 2011. Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. Science. 332:970–974.

Mukhopadhya I et al. 2011. A comprehensive evaluation of colonic mucosal isolates of sutterella wadsworthensis from inflammatory bowel disease. PLOS One. 6:e27076.

Mukhopadhya I et al. 2015. The fungal microbiota of de-novo paediatric inflammatory bowel disease. Microbes Infect. 17:304–310.

Murphy MA et al. 2016. Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. Ecology. 91:252–261.

Muscogiuri G et al. 2019. Gut microbiota: a new path to treat obesity. Int. J. Obes. Suppl. 9:10–19.

Mysara M et al. 2017. Reconciliation between operational taxonomic units and species boundaries. FEMS Microbiol. Ecol. 93:fix029.

Al Nabhani Z, Dietrich G, Hugot JP, Barreau F. 2017. Nod2: The intestinal gate keeper. PLOS Pathog. 13:e1006177.

Naeem S, Kawabata Z, Loreau M. 1998. Transcending boundaries in biodiversity research. Trends Ecol. Evol. 13:134–135.

Nakaji S et al. 2004. The Prevention of Colon Carcinogenesis in Rats by Dietary Cellulose Is Greater than the Promotive Effect of Dietary Lard As Assessed by Repeated Endoscopic Observation. J. Nutr. 134:935–939.

Nakamura S, Nakaya T, Iida T. 2011. Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing. Exp. Biol. Med. 236:968–971.

Namkung J. 2020. Machine learning methods for microbiome studies. J. Microbiol. 58:206–216.

Narayan NR et al. 2020. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. BMC Genomics. 21:56.

Nearing JT, Douglas GM, Comeau AM, Langille MGI. 2018. Denoising the Denoisers: An independent evaluation of microbiome sequence error- correction approaches. PeerJ. 2018:e5364.

Nejman D et al. 2020. The human tumor microbiome is composed of tumor type-specific intra-cellular bacteria. Science. 980:973–980.

Neovius M, Arkema E, Blomqvist P, Ekbom A, Smedby KE. 2013. Patients with ulcerative colitis miss more days of work than the general population, even following colectomy. Gastroenterology. 144:536–543.

Ng SC, Woodrow S, Patel N, Subhani J, Harbord M. 2012. Role of genetic and environmental factors in British twins with inflammatory bowel disease. Inflamm. Bowel Dis. 18:725–736.

Nguyen LH et al. 2020. Antibiotic use and the development of inflammatory bowel disease: a national case-control study in Sweden. Lancet Gastroenterol. Hepatol. 1253:1–9.

NIH. 2019. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome. 7:31.

Nikolaus S et al. 2017. Increased Tryptophan Metabolism is Associated With Activity of Inflammatory Bowel Diseases. Gastroenterology. 153:1504–1516.

Ning J, Beiko RG. 2015. Phylogenetic approaches to microbial community classification. Microbiome. 3:47.

Norman JM et al. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 160:447–460.

Nørreslet LB, Agner T, Clausen ML. 2020. The Skin Microbiome in Inflammatory Skin Diseases. Curr. Dermatol. Rep. 9:141–151.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27:824–834.

Oberhardt MA, Puchałka J, Fryer KE, Martins Dos Santos VAP, Papin JA. 2008. Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1. J. Bacteriol. 190:2790–2803.

Ogura Y et al. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 411:603–606.

Oh J et al. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature. 514:59–64.

Olson ND et al. 2019. Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief. Bioinform. 20:1140–1150.

Omelchenko M V., Galperin MY, Wolf YI, Koonin E V. 2010. Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. Biol. Direct. 5:31.

Orholm M, Binder V, Sorensen TIA, Rasmussen LP, Kyvik KO. 2000. Concordance of inflammatory bowel disease among Danish twins: Results of a nationwide study. Scand. J. Gastroenterol. 35:1075–1081.

Palarea-Albaladejo J, Martín-Fernández JA. 2015. zCompositions - R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 143:85–96.

Palau M et al. 2020. Detection of helicobacter pylori microevolution and multiple infection from gastric biopsies by housekeeping gene amplicon sequencing. Pathogens. 9:1–13.

Palau M et al. 2016. Usefulness of Housekeeping Genes for the Diagnosis of Helicobacter pylori Infection, Strain Discrimination and Detection of Multiple Infection. Helicobacter. 21:481–487.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20:289–290.

Park Y et al. 2005. Dietary fiber intake and risk of colorectal cancer: A pooled analysis of prospective cohort studies. J. Am. Med. Assoc. 294:2849–2857.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043–55.

Pascal V et al. 2017. A microbial signature for Crohn's disease. Gut. 66:813–822.

Pasolli E et al. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 176:649–662.

Pasolli E, Truong T, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLOS Comput. Biol. 12:e1004977.

Paulson JN, Colin Stine O, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods. 10:1200–1202.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 28:1420–1428.

Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Appl. Environ. Microbiol. 84:e02627-17.

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Curr. Opin. Microbiol. 14:615–623.

Prakash T, Taylor TD. 2012. Functional assignment of metagenomic data: challenges and applications. Brief. Bioinform. 13:711–727.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLOS One. 5:e9490.

Prifti E et al. 2020. Interpretable and accurate prediction models for metagenomics data. Gigascience. 9:1–11.

Prodan A et al. 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLOS One. 15:e0227434.

Punta M et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40:D290–D301.

Purcell S et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Quast C et al. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41:590–596.

R Core Team. 2019. R: A Language and Environment for Statistical Computing.

Raftery T, O'Sullivan M. 2015. Optimal vitamin D levels in Crohn's disease: A review. Proc. Nutr. Soc. 74:56–66.

Rahman SF, Olm MR, Morowitz MJ, Banfield JF. 2018. Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome. mSystems. 3:e00123-17.

Rasko DA et al. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol.

190:6881–6893.

Riley M. 1993. Functions of the gene products of *Escherichia coli*. Microbiol. Rev. 57:862–952.

Rivera-Chávez F et al. 2016. Depletion of Butyrate-Producing <i>Clostridia<\i> from the Gut Microbiota Drives an Aerobic Luminal Expansion of Salmonella. Cell Host Microbe. 19:443–454.

Rivera-Chávez F, Lopez CA, Bäumler AJ. 2017. Oxygen as a driver of gut dysbiosis. Free Radic. Biol. Med. 105:93–101.

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

Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M. 2016. Challenges in the analysis of viral metagenomes. Virus Evol. 2:vew022.

Rowan F et al. 2010. *Desulfovibrio* Bacterial Species Are Increased in Ulcerative Colitis. Dis. Colon Rectum. 53:1530–1536.

Saglani S, Custovic A. 2019. Childhood asthma: Advances using machine learning and mechanistic studies. Am. J. Respir. Crit. Care Med. 199:414–422.

Salonen A et al. 2010. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J. Microbiol. Methods. 81:127–134.

Samuel BS et al. 2007. Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. Proc. Natl. Acad. Sci. 104:10643–10648.

Saroj DB, Dengeti SN, Aher S, Gupta AK. 2015. ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. World J. Microbiol. Biotechnol. 31:189.

Sartor RB. 2008. Microbial Influences in Inflammatory Bowel Diseases. Gastroenterology. 134:577–594.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics. 27:592–593.

Schloss PD. 2020. Reintroducing mothur: 10 Years Later. Appl. Environ. Microbiol. 86:e02343-19.

Schoch CL et al. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. USA. 109:6241–6246.

Schumann S, Alpert C, Engst W, Loh G, Blaut M. 2012. Dextran sodium sulfate-induced inflammation alters the expression of proteins by intestinal Escherichia coli strains in a gnotobiotic mouse model. Appl. Environ. Microbiol. 78:1513–1522.

Schürmann G et al. 1999. Transepithelial transport processes at the intestinal mucosa in inflammatory bowel disease. Int. J. Colorectal Dis. 14:41–46.

Schwager E, Mallick H, Ventz S, Huttenhower C. 2017. A Bayesian method for detecting pairwise associations in compositional data. PLOS Comput. Biol. 13:e1005852.

Segata N et al. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12:R60.

Segata N et al. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods. 9:811–814.

Seksik P et al. 2003. Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut. 52:237–242.

Sewitch MJ et al. 2001. Psychological distress, social support, and disease activity in patients with inflammatory bowel disease. Am. J. Gastroenterol. 96:1470–1479.

Shade A. 2017. Diversity is the question, not the answer. ISME J. 11:1–6.

Shaiber A, Eren AM. 2019. Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. MBio. 10:e00725.

Shapley LS. 1953. A value for n-person games. In: Contributions to the Theory of Games, 2. Kuhn, HW & Tucker, W, editors. Princeton University Press: Princeton, NJ pp. 307–317.

Shaw SY, Blanchard JF, Bernstein CN. 2010. Association Between the Use of Antibiotics in the First Year of Life and Pediatric Inflammatory Bowel Disease. Am. J. Gastroenterol. 105:2687–2692.

Sieber CMK et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 3:836–843.

Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 6:e21887.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Singh V et al. 2015. Interplay between enterobactin, myeloperoxidase and lipocalin 2 regulates *E. coli* survival in the inflamed gut. Nat. Commun. 6:7113.

Sokol H et al. 2017. Fungal microbiota dysbiosis in IBD. Gut. 66:1039–1048.

Sokol H, Seksik P. 2010. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. Curr. Opin. Gastroenterol. 26:327–331.

De Souza HSP, Fiocchi C. 2016. Immunopathogenesis of IBD: Current state of the art. Nat. Rev. Gastroenterol. Hepatol. 13:13–27.

Sperling JL et al. 2017. Comparison of bacterial 16S rRNA variable regions for microbiome surveys of ticks. Ticks Tick. Borne. Dis. 8:453–461.

Sprockett D et al. 2019. Treatment-specific composition of the gut microbiota is associated with disease remission in a pediatric crohn's disease cohort. Inflamm. Bowel Dis. 25:1927–1938.

Stackebrandt E, Goebel BM. 1994. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44:846–849.

Staley J, Konopka A. 1985. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annu. Rev. Microbiol. 39:321–346.

Stamatiades GA, Ioannou P, Petrikkos G, Tsioutis C. 2018. Fungal infections in patients with inflammatory bowel disease: A systematic review. Mycoses. 61:336–376.

Stasinopoulos M, Rigby R. 2020. gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape.

Stearns JC et al. 2011. Bacterial biogeography of the human digestive tract. Sci. Rep. 1:170.

Stein JL, Marsh TL, Wu KY, Shizuya H, Delong EF. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178:591–599.

Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. Nat. Commun. 9:2542.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35:1026–1028.

Sternes PR, Lee D, Kutyna DR, Borneman AR. 2017. A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation. Gigascience. 6:1–10.

Suardana IW. 2014. Analysis of Nucleotide Sequences of the 16S rRNA Gene of Novel *Escherichia coli* Strains Isolated from Feces of Human and Bali Cattle. J. Nucleic Acids. 2014:475754.

Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl. Environ. Microbiol. 79:5962–5969.

Sun S, Jones RB, Fodor AA. 2020. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. Microbiome. 8:46.

Sze MA, Schloss PD. 2016. Looking for a signal in the noise: Revisiting obesity and the microbiome. MBio. 7:e01018-16.

Takahashi K et al. 2016. Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease. Digestion. 93:59–65.

Tange O. 2011. GNU Parallel: the command-line power tool. ;login USENIX Mag. 36:42–47.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

Tedjo DI et al. 2016. The fecal microbiota as a biomarker for disease activity in Crohn's disease. Sci. Rep. 6:35216.

Tessler M et al. 2017. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci. Rep. 7:6589.

Tettelin H et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome’. Proc. Natl. Acad. Sci. USA. 102:3950–13955.

Thia KT, Sandborn WJ, Harmsen WS, Zinsmeister AR, Loftus E. 2010. Risk factors associated with progression to intestinal complications of Crohn’s disease in a population-based cohort. Gastroenterology. 139:1147–1155.

Thompson LR et al. 2017. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature. 551:457–463.

Thorsen J et al. 2016. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 4:62.

Treem W, Ahsan N, M S, Hyams J. 1994. Fecal Short-Chain Fatty Acids in Children with Inflammatory Bowel Disease. J. Pediatr. Gastroenterol. Nutr. 18:159–164.

Truong DT et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods. 12:902–903.

Turnbaugh PJ et al. 2009a. A core gut microbiome in obese and lean twins. Nature. 457:480–484.

Turnbaugh PJ et al. 2009b. The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. Sci. Transl. Med. 1:6ra14.

Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. 2008. Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. Cell Host Microbe. 3:213–223.

Turpin W et al. 2016. Association of host genome with intestinal microbial composition in a large healthy cohort. Nat. Genet. 48:1413–1417.

Ungara F et al. 2019a. Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease. Gut Microbes. 10:149–158.

Ungara F, Massimino L, D’Alessio S, Danese S. 2019b. The gut virome in inflammatory bowel disease pathogenesis: From metagenomics to novel therapeutic approaches. United Eur. Gastroenterol. J. 7:999–1007.

Vandeputte D et al. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 551:507–511.

Vatanen T et al. 2018. Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. Nat. Microbiol. 4:470–479.

Velázquez OC, Lederer HM, Rombeau JL. 1997. Butyrate and the colonocyte. Production, absorption, metabolism, and therapeutic implications. Adv. Exp. Med. Biol. 427:123–34.

Venegas DP et al. 2019. Short chain fatty acids (SCFAs) mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. Front. Immunol. 10:277.

Venter JC et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 304:66–74.

Verster AJ, Borenstein E. 2018. Competitive lottery-based assembly of selected clades in the human gut microbiome. Microbiome. 6:186.

Větrovský T, Baldrian P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLOS One. 8:e57923.

Via M. 2012. The Malnutrition of Obesity: Micronutrient Deficiencies That Promote Diabetes. ISRN Endocrinol. 2012:103472.

Vincent C et al. 2013. Reductions in intestinal Clostridiales precede the development of nosocomial *Clostridium difficile* infection. Microbiome. 1:18.

Vollmers J, Wiegand S, Kaster A-K. 2017. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLOS One. 12:e0169662.

de Waal GM, de Villiers WJS, Forgan T, Roberts T, Pretorius E. 2020. Colorectal cancer is associated with increased circulating lipopolysaccharide, inflammation and hypercoagulability. Sci. Rep. 10:8777.

Wall CL, Day AS, Gearry RB. 2013. Use of exclusive enteral nutrition in adults with Crohn's disease: A review. World J. Gastroenterol. 19:7652–7660.

Wang M-H et al. 2013a. A novel approach to detect cumulative genetic effects and genetic interactions in Crohn's disease. Inflamm. Bowel Dis. 19:1799–808.

Wang W et al. 2014. Increased proportions of Bifidobacterium and the Lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease. J. Clin. Microbiol. 52:398–406.

Wang Y, Qian P, Ya S. 2013b. Conserved Regions in 16S Ribosome RNA Sequences and Primer Design for Studies of Environmental Microbes. Encycl. Metagenomics. https://doi.org/10.1007/978-1-4614-6418-1_772-1.

Ward T et al. 2017. BugBase predicts organism-level microbiome phenotypes. bioRxiv.

Washburne AD et al. 2019. Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. Ecol. Monogr. 89:e01353.

Watson EJ, Giles J, Scherer BL, Blatchford P. 2019. Human faecal collection methods demonstrate a bias in microbiome composition by cell wall structure. Sci. Rep. 9:16831.

Weiss S et al. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 5:27.

Wemheuer F et al. 2020. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. Environ. Microbiome. 15:11.

Wheeler NE, Barquist L, Kingsley RA, Gardner PP. 2016. A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes.

Bioinformatics. 32:3566–3574.

Whitten KE, Rogers P, Ooi CKY, Day AS. 2012. International survey of enteral nutrition protocols used in children with Crohn’s disease. J. Dig. Dis. 13:107–112.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham H. 2007. Reshaping Data with the {reshape} Package. J. Stat. Softw. 21:1–20.

Wickham H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations.

Wickham H. 2011. The Split-Apply-Combine Strategy for Data Analysis. J. Stat. Softw. 40:1–29.

Wilke CO. 2019. cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2’.

Wilkinson TJ et al. 2018. CowPI: A rumen microbiome focussed version of the PICRUSt functional inference software. Front. Microbiol. 9:1095.

Willing B et al. 2009. Twin studies reveal specific imbalances in the mucosa-associated microbiota of patients with ileal Crohn’s disease. Inflamm. Bowel Dis. 15:653–660.

Willis C, Desai D, Laroche J. 2019. Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. FEMS Microbiol. Lett. 366:fnz152.

Willis JR, Gabaldón T. 2020. The human oral microbiome in health and disease: From sequences to ecosystems. Microorganisms. 8:308.

Wilson GA et al. 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology. 151:2499–2501.

Wirbel J et al. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med. 25:679–689.

Woese CR. 1987. Bacterial evolution. Microbiol. Rev. 51:221–271.

Woese CR et al. 1980. Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. 8:2275–2294.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. USA. 74:5088–5090.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20:257.

Wright EK et al. 2015. Recent advances in characterizing the gastrointestinal microbiome in Crohn’s disease: a systematic review. Inflamm Bowel Dis. 21:1219–1228.

Wu D, Jospin G, Eisen JA. 2013. Systematic Identification of Gene Families for Use as ‘Markers’ for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. PLOS One. 8:e77033.

Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 32:605–607.

Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification Simon. Cell. 178:779–794.

Ye Y, Doak TG. 2011. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. PLOS Comput. Biol. 5:e1000465.

Yu G. 2020. Using ggtree to Visualize Data on Tree-Like Structures. Curr. Protoc. Bioinforma. 69:e96.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. andEvolution. 8:28–36.

Yurgel SN et al. 2017. Variation in Bacterial and Eukaryotic Communities Associated with Natural and Managed Wild Blueberry Habitats. Phytobiomes. 1:102–113.

Yurgel SN, Douglas GM, Dusault A, Percival D, Langille MGI. 2018. Dissecting community structure in wild blueberry root and soil microbiome. Front. Microbiol. 9:1187.

Yurgel SN, Nearing JT, Douglas GM, Langille MGI. 2019. Metagenomic Functional Shifts to Plant Induced Environmental Changes. Front. Microbiol. 10:1682.

Zaneveld JR, Lozupone C, Gordon JI, Knight R. 2010. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res. 38:3869–3879.

Zaneveld JRR, Thurber RLV. 2014. Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. Front. Microbiol. 5:431.

Zeller G et al. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. 10:766–766.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 30:614–620.

Zhou J et al. 2015. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. MBio. 6:e02288-14.

Zhou W et al. 2019. Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature. 569:663–671.

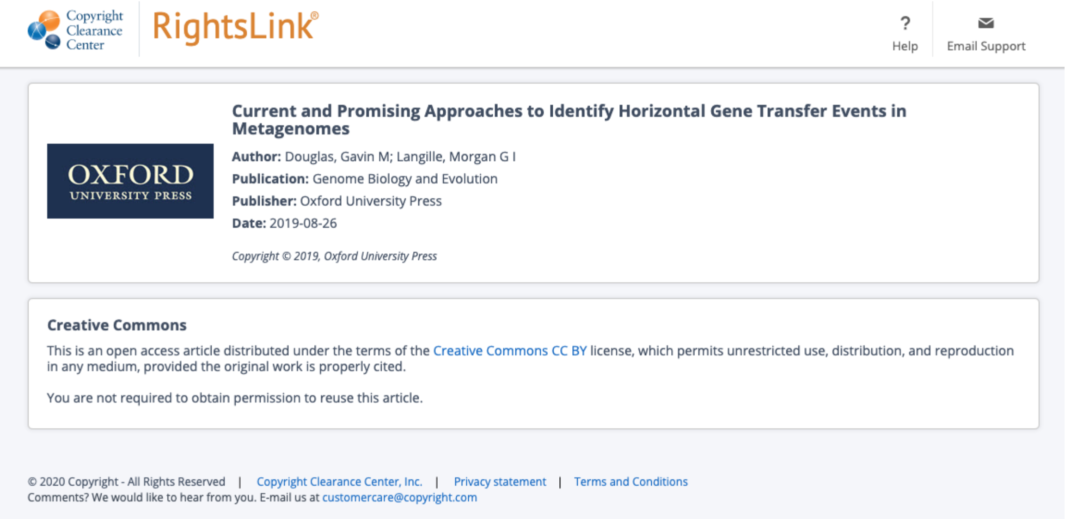
Zhou YH, Gallins P. 2019. A review and tutorial of machine learning methods for microbiome host trait prediction. Front. Genet. 10:579.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of history. J. Theor. Biol. 8:357–366.

Zupančič K et al. 2016. Multi-locus genetic risk score predicts risk for Crohn’s disease in Slovenian population. World J. Gastroenterol. 22:3777–3784.

Appendices

Copyright Permissions



Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes

Author: Douglas, Gavin M; Langille, Morgan G I

Publication: Genome Biology and Evolution

Publisher: Oxford University Press

Date: 2019-08-26

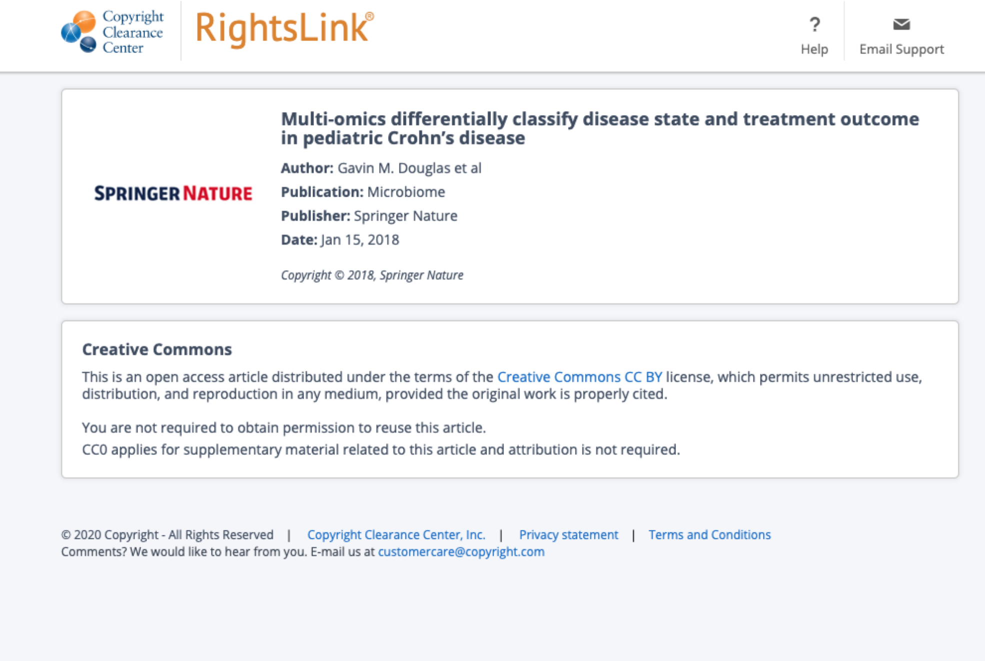
Copyright © 2019, Oxford University Press

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com



Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease

Author: Gavin M. Douglas et al

Publication: Microbiome

Publisher: Springer Nature

Date: Jan 15, 2018

Copyright © 2018, Springer Nature

Creative Commons

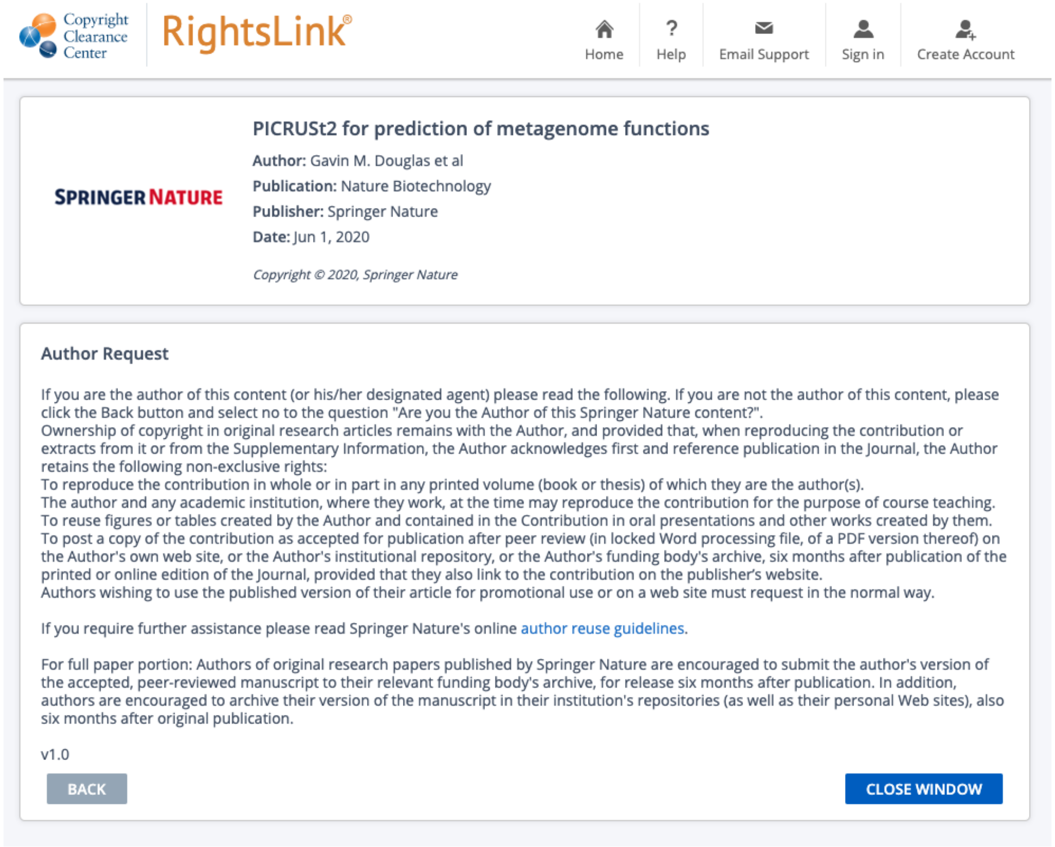
This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

CC0 applies for supplementary material related to this article and attribution is not required.

© 2020 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions

Comments? We would like to hear from you. E-mail us at customercare@copyright.com



PICRUSt2 for prediction of metagenome functions

Author: Gavin M. Douglas et al

Publication: Nature Biotechnology

Publisher: Springer Nature

Date: Jun 1, 2020

Copyright © 2020, Springer Nature

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0