EXPLORING THE HUMAN GUT MICROBIOME: STATISTICAL METHODS, COMPUTATION, AND APPLICATIONS IN METAGENOMICS

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Felicia Nicole New
December 2021

EXPLORING THE HUMAN GUT MICROBIOME: STATISTICAL METHODS, COMPUTATION, AND APPLICATIONS IN METAGENOMICS

Felicia Nicole New, Ph.D., M.S.
Cornell University 2021

The totality of microbial species and their associated genomes living within the human gastrointestinal tract are known collectively as the human gut microbiome. The human gut microbiome is an integral part of human health. There is some evidence that human genomic variation is associated with differences in the composition of the gut microbiome, leading to potential health effects. For example, mutations in NOD2, a gene associated with Crohn's disease, and mutations in MEFV, a gene causing Mediterranean fever, are associated with compositional shifts in certain bacterial phyla. By jointly analyzing the genomes and the metagenomes of individuals in a population, we can uncover the connection between the two, and how they relate to health outcomes using health or phenotype data. To investigate these questions, I used the shotgun metagenomic sequencing data, along with genotype and phenotype information, for 250 adult female twins from TwinsUK.

To understand the link between the gut microbiome's composition and functions with human health outcomes, I apply classical statistical and machine learning methods to identify features of the gut microbiome that can predict host diseases and phenotypes. I find interesting results for anxiety symptoms within twin pairs who are discordant for anxiety. Specifically, 175 genes were found to be enriched in the twins without anxiety and absent in those with anxiety. Using strain-level metagenomic analyses, I identify the source of these genes as a species within the genus *Azospirillum*.

Studies of the impact of host genetics on the gut microbiome composition have mainly focused on the impact of individual host variants, without considering their collective impact or the specific functions of the gut microbiome. To assess the aggregate role of human genetics on the gut microbiome composition and function, I apply both the Tweedie distribution, for modeling gene and species abundances in metagenomic data, and the multivariate data integration method known as sparse canonical correlation analysis to the challenge of identifying correlations between overall host genetics and the composition of the gut microbiome or its composite functions.

BIOGRAPHICAL SKETCH

Felicia New is from Gainesville, Florida. She completed her Bachelor of Science degree at the University of Florida in 2013 in biology. She then began working as a bioinformatics lab technician for Dr. Lauren McIntyre the day after graduation. She worked on various computational analyses involving many species and data types and grew to love computational genetics. In 2014, she entered the Master of Medical Sciences program at the University of Florida. She studied the population genomic patterns of a new *Drosophila simulans* population panel and earned the degree of Master of Science in December 2015. After working as a computational biologist in the McIntyre lab from 2013 to 2016, she decided to pursue a Ph.D. in genetics. In 2016, she joined the Genetics, Genomics, and Development doctoral program at Cornell University. At Cornell, Felicia joined the lab of Dr. Ilana Brito where she would study the genetics of the human gut microbiome and the field of metagenomics. In addition to research, she spent the following years pursuing her passion in leadership and outreach activities including, but not limited to, co-founding the First Generation and Low Income Graduate Student Association at Cornell, volunteering as a code instructor for young girls, and organizing and hosting several microbiome-themed data hackathons for students in the Ithaca area. Her doctoral work has culminated in the publication of a first author review paper in *Annual Reviews Microbiology*, a first author publication currently in review, a co-authored paper in *eLife*, and this dissertation.

This dissertation is dedicated to the four most important teachers in my life:
To my grandma for teaching me the importance of education and the natural world.
To my grandpa for teaching me the importance of reading and writing.
To my dad for teaching me to be curious.
And to my mom for teaching me empathy and compassion.

ACKNOWLEDGMENTS

I want to begin by thanking my graduate advisor, Prof. Ilana Brito for making this work possible. We started at Cornell University at the same time, and I am so grateful that you accepted me into your new lab.

Thank you to Prof. Andrew Clark and Prof. Philipp Messer for serving on my thesis committee and providing advice and feedback throughout the years.

I would like to thank my colleagues in the Brito Lab for the conversations, lunches, and fun throughout the years. Some of my favorite times from graduate school are from our lab's hackathons and I am very grateful to those of you who helped me with those events.

Next, I would like to thank Ben Baer who taught me so much over the years. Our collaboration, paper discussions, and brainstorming sessions meant a lot to me during graduate school and I hope we can collaborate again in the future.

Thank you to my family and friends for everything. To my mom, for all her sacrifices and hard work that allowed me to achieve this. To Natya and Alex, for your unwavering support from a distance. And thank you, especially, to Brad for your love and support throughout my academic career.

Finally, I'd like to acknowledge my funding sources over the years including the NIH NIDDK for funding our work on the TwinsUK project and my diversity fellowship. I want to thank the Cornell Graduate school for awarding me a Graduate Dean's Scholarship and to the State of New York for awarding me the SUNY Diversity fellowship.

TABLE OF CONTENTS

CHAPTER 1: Introduction and Review

This part of the dissertation is adapted from the review article by Felicia New and Ilana Brito in 2020. It is published in Annual Reviews Microbiology. New FN, Brito IL. (2020). "What Is Metagenomics Teaching Us, and What Is Missed?" *Annual Reviews Microbiology*. 74:117-35.

Abstract

Shotgun metagenomic sequencing has revolutionized our ability to detect and characterize the diversity and function of complex microbial communities. In this review, we highlight the benefits of using metagenomics as well as the breadth of conclusions that can be made using currently available analytical tools, such as greater resolution of species and strains across phyla and functional content, while highlighting challenges of metagenomic data analysis. Major challenges remain in annotating function, given the dearth of functional databases for environmental bacteria compared to model organisms, and the technical difficulties of metagenome assembly and phasing in heterogeneous environmental samples. In the future, improvements and innovation in technology and methodology will lead to lowered costs. Data integration using multiple technological

platforms will lead to a better understanding of how to harness metagenomes. Subsequently, we will be able to not only characterize complex microbiomes but also able to manipulate communities to achieve prosperous outcomes for health, agriculture, and environmental sustainability.

Introduction

The tools of microbiology—microscopy, culturing, and genetic engineering—have allowed researchers to observe, grow, and experiment on a small number of well-studied organisms, revealing insights into their biological, ecological, and evolutionary capacities. Yet microbes live nearly everywhere on Earth, have vast influence over ecosystem services and host health, and are dominant members of all three domains of life. Despite scientific awareness of this diversity, it has remained unexplored until recently. The advent of high-throughput sequencing platforms has rapidly enhanced our ability to understand the diversity of species in microbiomes by coupling physiological data with the underlying genetic data. Though metagenomes have given us glimpses into the diversity and function of complex microbiomes, this data can itself be incomplete, biased, and challenging. It is thus important to be extremely critical of and to understand

the limitations of metagenomic data as the scientific community continues to embrace this technology.

Due to the cost, computational footprint, and analytical hurdles of whole-microbiome shotgun sequencing, amplicon sequencing of the 16S rRNA gene is used broadly to determine the taxonomic identities of members of a microbial community. The 16S rRNA gene, ubiquitous in all bacteria and archaea, was chosen as a genetic marker for taxonomic identification for several reasons. Barring some exceptions, this gene is evolutionarily stable, meaning that it has gone through little horizontal gene transfer, follows a molecular clock, and has regions of conservation and regions of divergence. For most microbiomes, several tens of thousands of sequences are adequate to assess the diversity in a sample (55a), and as of 2020, the cost of DNA extraction, library preparation, and sequencing would cost less than $25–50 per sample, depending on the number and sample type, making this data type the most broadly accessible. With this accessibility has come streamlined analytical platforms, such as QIIME (13), UCHIME2 (37), mothur (97), and dada2 (25), using reference-based assignment of taxonomies and/or de novo sequence clustering.

Yet, sequencing this single gene to determine community composition results in several complications, arising from the facts that organisms may

carry 1–15 genetically dissimilar, and occasionally fairly distant, copies; that artifacts, such as chimeras and jackpot effects, arise during the amplification process; and that the resolution of taxa varies depending on the branch of the bacterial tree of life. Despite innovations to solve these problems, both experimentally and computationally, with programs to remove chimeras and deal with errors inherent to the platform (25), significant biases may remain. As a result, the studies exploit the low cost and streamlined computational analyses of 16S rRNA data sets either to cover large dense time courses where increased coverage can mitigate the effects of noise, or diverse ecologies where the differences are more robust.

With the increasing appreciation for dramatic phenotypic differences arising from differences in the genomic content of organisms and strains, there has been a movement toward using higher-resolution differences in 16S sequences, called amplicon sequence variants (ASVs). Previously, DNA sequences would be clustered at 97% sequence identity, a cutoff meant to distinguish between species while masking the effect of PCR (polymerase chain reaction) or sequencing errors. Among the benefits of using ASVs rather than clustered 16S sequences are a greater comparability across studies, greater reproducibility, and lack of reliance on previously curated reference libraries (24).

As part of this trend to obtain greater resolution, whole-microbiome shotgun sequencing, hereafter referred to as metagenomics, is becoming an important data type for many studies aiming to understand the mechanisms driving microbiome-associated traits. Rather than amplifying a single gene, all the DNA within a sample is sequenced, regardless of whether it originated from bacteria. DNA is simply extracted, made into libraries, and sequenced either on a short-read platform (e.g., sequencing by synthesis, such as Illumina's platform) or on a long-read platform [e.g., single-molecule real-time (SMRT) sequencing, used by PacBio, or nanopore, used by Oxford Nanopore]. Recently, with the decreased costs of DNA sequencing and library preparation, studies have grown in breadth and scope.

This review focuses mainly on the benefits of using metagenomics and outlining the breadth of conclusions that can be made using currently available analytical tools, such as greater resolution of species and strains across phyla and functional content, while highlighting challenges of metagenomic data analysis (Figure 1). These major challenges include annotating function, given the relative lack of functional databases for environmental bacteria compared to model organisms, and the technical challenges of assembly and phasing in heterogeneous environmental samples.

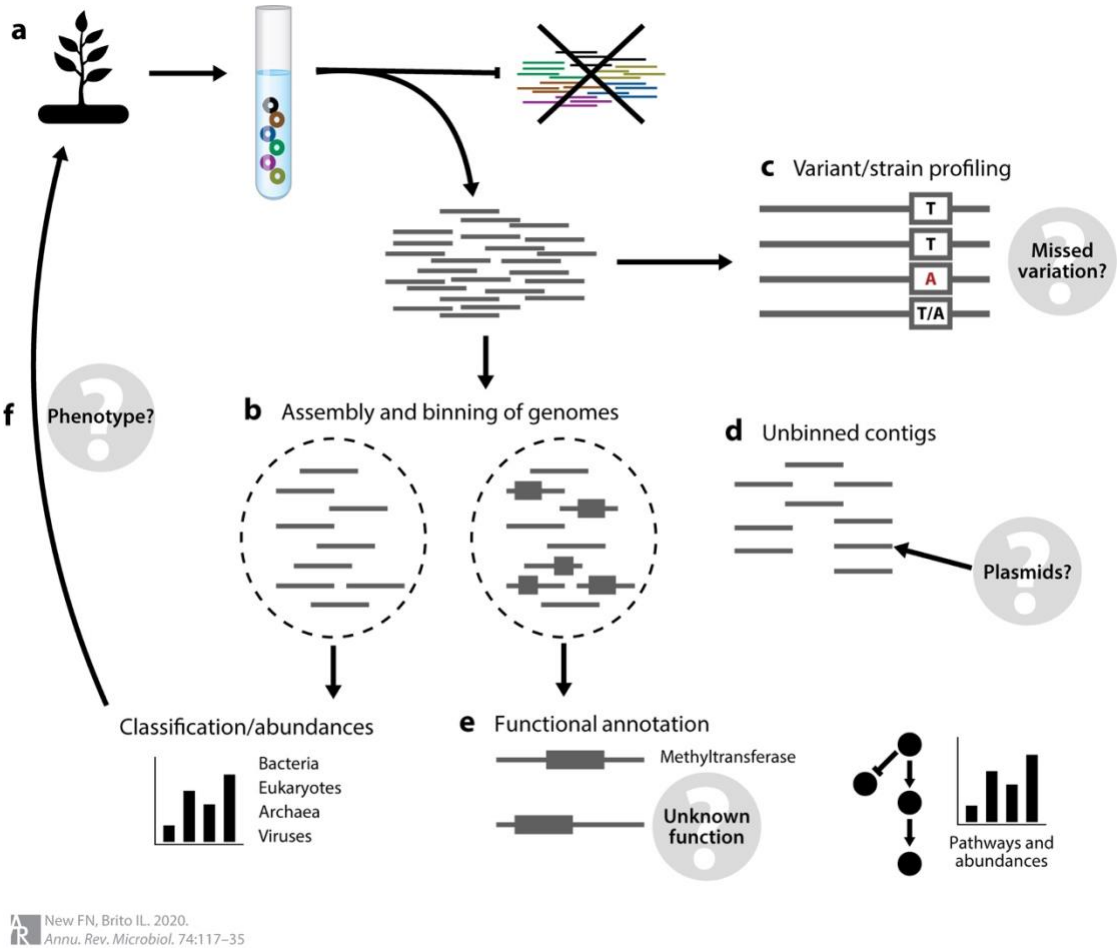


Figure 1.1. Metagenomic shotgun sequencing, the profiling of all DNA present within a microbiome sample, has many benefits and challenges. (*a*) DNA is extracted, made into libraries, and sequenced. The genomic and cellular context is lost during the process, and the output is reads that need to be organized into meaningful groups. (*b*) Assembly and binning are general steps for many analyses including taxonomic and functional profiling. However, many challenges and unknowns remain, here denoted by question marks: (*c*) Strain profiling methods have improved, but it is difficult to know what variation is missed; (*d*) unbinned contigs originating from plasmids, sequencing errors, or low-abundance (or low-sampled) organisms are indistinguishable; (*e*) many genes found in microbiome samples remain unannotated; and (*f*) it is difficult to link metagenomic data to host traits.

What has metagenomics taught us?

A trove of new (draft) genomes

Many of the earliest metagenomic studies used alignments to reference genomes to assess composition and function (62). Given a suitable reference catalog of genes, coding regions, or reference genomes, metagenomic data can be mapped to the reference using a typical alignment software. Yet, many environmental metagenomic samples, still to this day, lack appropriate representative reference genomes. A simple study where spores were selected from human gut microbiome samples revealed 45 novel candidate species (22), despite relatively deep study of the human gut microbiome. Advances in culturing of organisms are improving our reference libraries of previously unknown or unculturable species (65, 95), yet de novo assembly methods are still necessary to fill this knowledge gap. While single-genome assemblers were not appropriate for assembling metagenomes, because of the varying abundances of bacteria within a community, tools were designed to account for and leverage these distinctive metagenomic qualities to assemble draft genomes within complex, heterogeneous assemblies.

Methods to assemble metagenomes

Overlap or consensus assembly methods were initially applied for DNA sequence assembly. These methods use a greedy algorithm and were originally created for Sanger sequencing (4). Given that pairwise comparisons of every read must be made, they are computationally expensive and became less suitable for next generation sequencing (NGS). The algorithm underlying de Bruijn graph assemblers breaks the sequencing reads down into uniform k-mers of a specified size, k. The k-mers are used as nodes in the graph, and overlapping nodes are connected by an edge. The assembler then constructs sequences based on the compiled graph. These methods (4, 102a, 133a) reduced the computational memory requirements because they essentially compress the repetition inherent to NGS data, negating the need to perform pairwise read alignments. Despite the gains in performance, several challenges remain. As reads are broken down into k-mers, some genomic context is lost. Additionally, the choice of k-mer size and the choice of tools can significantly alter an assembly. One solution helpful for metagenomes has been to employ iterative de Bruijn graph assembly, which combines graphs from various k-mer sizes (78, 87). Currently, there is no single best practice. In the Critical Assessment of Metagenome Interpretation (CAMI) challenge (98) to assemble

metagenomes, the simulated data assembly had 39,140 contigs and was 1.97 Gbp long. However, other programs resulted in a range of results: MEGAHIT (66) resulted in the largest assembly, with 587,607 contigs, of 1.91 Gbp, whereas the smallest assembly, produced by Ray Meta (12), was 12.3 Mbp long and had just 13,847 contigs. We recommend the review paper by Ayling et al. (4) and the CAMI challenge (98) for comparisons of these methods.

A subsequent challenge is making sense of the thousands of relatively small contigs that can result from assembling heterogeneous microbiome data. Whereas contigs could be binned into genomes based on taxonomic markers, genomes are typically fragmented, incomplete, and contaminated. New algorithms to bin contigs have led to higher-fidelity genome assembly. Binning algorithms use several different metrics to group contigs: DNA composition, GC content, tetranucleotide frequency, depth of sequencing coverage, and abundance or co-abundance patterns across multiple samples (2, 54, 57, 128). An alternative method is to bin reads that are predicted to be derived from the same organism prior to assembly (29). There are also downstream tools that combine output from several binning tools (101, 108, 117) to refine and recombine contigs with the goal of identifying cleaner and more complete metagenome-assembled genomes (MAGs).

These techniques have been applied to large numbers of metagenomes to reconstruct draft genomes in a uniform manner. Nayfach et al. (77) reconstructed draft genomes from 3,810 publicly available human gut metagenomic samples, and Pasolli et al. (85) analyzed 9,428 metagenomes across microbiomes on different body sites. Both studies prioritized assemblies from international cohorts, including those in areas of the world in which fewer microbiome studies have been performed overall. The former study resulted in over 2,000 newly identified species, accounting for a 50% increase in the phylogenetic diversity of the human gut microbiome. Despite the creative use of recent tools to recover this vast diversity from individual gut samples, strain-level diversity and sequencing depth still pose a challenge to MAG assembly. The latter study reconstructed 154,723 MAGs, increasing the mappability of human metagenomic reads from around 67% to over 87% in gut microbiome samples, and from 65% to 82% in oral microbiome samples. These two studies highlight a particularly timely challenge to the field, the need to internationalize metagenomic sequencing efforts (89). Most of the unknown or uncharacterized species found in these studies were found in non-Western, low- and middle-income populations.

Metagenomic assembly quality

Following assembly, it is important to assess quality. Since there is often no ground truth in metagenomic sequencing for environmental samples, assembly quality is usually evaluated using summary statistics from single-genome assembly methods like size, contig N50, and maximum contig length (15). Completeness and contamination are the two main metrics that researchers rely on for assessing MAG quality. Completeness relies on the identification of marker gene sets and can miss strain heterogeneity. Likewise, contamination is derived from a set of single-copy marker genes and can be complicated if genes overlap contig gaps (38, 84, 102, 113). However, many of these tools rely on taxon-specific metrics that are better suited for those organisms that are well studied, and therefore they lack the same resolution in identification of marker genes for organisms that are more obscure (84, 127). Other methods address potential contamination by aligning the assemblies to many references with the ability to report chimeric contigs (75). A set of standards has been proposed, called the minimum information about a single metagenomic-assembled genome (MIMAG) (15), emphasizing manual curation and review. With more studies published obtaining single-cell genomes and/or cultured representatives, we urge a

systematic comparison of curated genomes with those obtained through metagenomic assembly.

Metagenomic genome assembly may still miss key aspects of the true underlying genomic variation. In order to detect low-abundance organisms, deeper sequencing is required. Shallow metagenomic sequencing, to as low as 500,000 reads, is a current alternative to amplicon sequencing in large cohort studies to gather species-level taxonomic and functional information on a large scale, at roughly the same cost as 16S rRNA sequencing (52). However, this method does not account for rare organisms and strains. Co-assembly of organisms present across genomes (66) or binning of reads from many samples prior to assembly (29), has a better chance of assembling low-abundance organisms. Novel methods that use co-barcoded sequencing reads derived from individual long DNA sequences to provide the origins of reads with which to construct scaffolds (9), as well as methods that combine high-fidelity short-read sequencing with long-read sequencing data (8), will undoubtedly aid metagenomic assembly. In fact, these technologies may replace traditional shotgun metagenomic sequencing one day, as their costs are reduced. Obtaining information on the genomic structure of organisms with multiple chromosomes or plasmids will still require additional innovations.

Better compositional data than taxonomic profiling alone

Metagenomic sequencing improves the resolution of bacterial community profiling compared to 16S rRNA profiling alone and has the added advantage of being kingdom-agnostic. Despite this benefit, this leads to one of the largest challenges of the metagenomics field, classifying and quantifying the species present in a metagenomic sample. Up-to-date comparisons and benchmarking of the available tools and databases are necessary in such a fast-paced field (132).

The diversity of host-associated viruses, fungi, and archaea

Along with the bacterial DNA present in microbiomes, metagenomic data sets often include viral, fungal, and host DNA. The eukaryotic component has often been ignored, as genome coverage generally compares poorly to the coverage of bacterial genomes and databases containing full genomes of eukaryotes found within microbiomes are limited. Nevertheless, efforts to assemble genomes from metagenomic data sets have revealed that diverse eukaryotes can inhabit the human infant gut (80) and environments such as geothermal geysers (124). Likewise, a measurable abundance of fungi inhabiting human skin has been detected (79).

Viruses with DNA-based genomes can also be found within metagenomic data, and occasionally in high abundance. Viral genomes, mainly from bacteriophage, are small compared to bacterial or eukaryotic genomes and can be well covered. However, because they are highly diverse and fewer of the genes in bacteriophage genomes can be assigned functions or even be found in reference databases (18), their roles remain elusive. There have been efforts to gain a better understanding of this microbiome component. Numerous computational tools have been developed to identify elements of phage genomes (43, 94), either integrated prophage or free-living phage. Deeper sequencing of viral particles isolated from samples reveals the active lytic phage within a community (17). Despite the challenges associated with analyzing phage within microbial communities, there have been several landmark observations, including a role for phage in inflammatory bowel disease pathogenesis (33). In the marine environment, metagenomic sequencing of a large number of samples, followed by co-occurrence analysis and abundance estimates of host and associated viral genomes, supports an ecological framework where lysogenic viruses dominate at high host density (31). Phage abundance data from longitudinal studies of the infant gut microbiome also resemble Lotka-Volterra dynamics (68). Assigning hosts remains a major challenge, although efforts continue to

improve in this sphere (76), where even the hosts of very large, ubiquitous bacteriophage present in the human gut remained elusive until recently (32, 36, 50).

Host DNA in animal and plant systems is also sequenced along with bacterial, viral, and eukaryotic symbiont DNA. Most often, host DNA is removed prior to analysis (93), and this step is often required in human metagenomic studies, where host DNA can be identifiable. In certain types of microbiomes, such as the oral and skin microbiomes, this results in removal of the vast majority of sequences, up to 90% (40). There are no commonplace uses for the host DNA. Increasing numbers of studies link human genotype data with microbiome composition and/or functions to perform combined genome-wide association studies (14, 91, 121), but we have yet to observe researchers utilizing human reads from metagenomic samples to this end.

Host-associated phenotypes

Metagenomic data analysis has shaped our understanding of the relationship between the microbiome and host phenotypes such as health outcomes, growth, and crop productivity. While there are myriad studies ranging from medicine, agriculture, and the environment, we highlight only a few notable

findings here from larger studies. The Environmental Determinants of Diabetes in the Young (TEDDY) study comprises almost 11,000 samples from 783 children beginning at 3 months of age until the onset of type 1 diabetes (T1D), or islet autoimmunity, with the goal of identifying compositional or functional aspects of the gut microbiome that may be predictive of the onset of T1D (119). It found that microbial factors associated with the onset of T1D were functionally similar but taxonomically diverse, that the gut microbiome matures faster with earlier cessation of breastfeeding, and that there are reproducible acquisitions of metabolic capabilities. Meta-analyses of smaller individual studies have also revealed important pathways to disease. Given that there have been over eight metagenomic microbiome studies on colorectal cancer on a geographically and culturally diverse set of populations, comparative metagenomic analysis can be used to find robust signals, such as an association with choline metabolism and pathways related to secondary bile acids (115, 126).

The Tara Oceans voyage was the largest metagenomic sequencing effort of marine environments to date, involving 243 size-fractionated samples that allowed for viral or prokaryotic enrichment (112). Assembly of these samples amounted to an excess of 117 million genes from over 35,000

species. Just a single drop (0.4 mL) of ocean water collected from the Sargasso Sea contained 6,236 genomes (average of 38% completeness) (82). Pairwise comparisons of the genomes found within these ocean water samples found that less than 0.1% of the genomes were from the same species, indicating the vast diversity of aquatic microbial communities. This example highlights the challenges inherent in metagenomic sequencing of microbial communities, where surveying an appropriate amount of diversity to answer a given biological question may be cost-prohibitive. Therefore, many aquatic studies on important systems like coral reefs rely on amplicon sequencing, where assessing community diversity at the species level may be more feasible (34, 49), especially in the case of time course or perturbation experiments where sampling error may complicate the interpretation of results.

Strain-level analyses

One of the major advantages of using metagenomic shotgun data is the ability to obtain strain-level data by resolving variations in single-nucleotide polymorphism (SNP) frequencies in microbial genomes across individuals harboring the same species. Strain-level differences can also be observed between individuals over time. There has not yet been a consensus among

researchers on the most appropriate method to use, although most programs use SNPs found in single-copy core genes, either retrieved from reference genomes (1, 41, 103, 116) or taken from the sample's MAGs (19). These genes are generally phylogenetically conserved; therefore, contamination and completion are easy to determine. Many of these genes encode proteins found within the ribosome, which are rarely horizontally transferred. Within a species, the mutation frequency in these genes is often no more than one SNP per read, thus complicating phasing methods that would link sets of SNPs into single genotypes.

The study of transmission of bacterial species between environments, between hosts and the environment, or between hosts typically has relied on full genome sequences. However, headway has been made in understanding the transmission of strains within complex communities by metagenomics studies. Simple examination of the dominant strain of each organism present in a community has revealed differences in the colonization of specific species after fecal microbiota transplantation (67). Improvements to this method have shown that often people are colonized by a consortium of strains within a single species from a donor through, rather than a single strain (103). Vertical transmission and colonization of strains from mother to newborn infant are observed by using SNP and strain-specific gene

content methods (41, 131). These patterns have been shown to persist even in adult family members. Adult twin pairs share higher frequencies of microbial SNPs in common strains than non-twin pairs; and shared SNPs and strain-specific flexible gene content are more commonly found for species in oral and gut microbiomes of family members (129). Interestingly, metagenomic data mining has also revealed evidence of transmission between spouses, showing the malleability of the adult human microbiome (19).

Functional assessments of metagenomes

The main advantage of metagenomic sequencing over amplicon sequencing is the ability to perform functional profiling of microbial communities. This normally entails aligning reads to either known or de novo–assembled genes to obtain gene abundances and infer functional abundances (by merging gene abundances by gene family or function), regardless of bacterial host. In other words, unlike taxonomic profiling, these methods do not rely on marker gene sets or even assembly in some cases. Many packages exist to streamline this process (44, 53). Caution needs to be taken when performing functional profiling because up to 50% of genes within host-associated and environmental microbiomes lack annotated functions (55). For example, the

earliest attempts at functionally profiling human gut microbiomes as part of the Human Microbiome Project, a large-scale effort to characterize the microbiomes of 300 individuals across several body sites with 16S rRNA and metagenomic whole-genome shotgun sequencing, led to the conclusion that functional profiles are conserved across body sites, despite vast differences in microbial composition (52a). This conclusion is largely based on the portion of genes that were capable of being functionally annotated at that time, which are largely conserved core genes present in all microbes. These conclusions have largely been revised by analyzing differences between the core and distinguishing functions between body sites, and by acknowledging the extent to which genes are annotated functionally (70).

Time courses are especially useful to examine relevant changes in the microbiome that occur alongside host physiology. Metagenomic sequencing of oral, vaginal, and gut microbiome samples of pregnant mothers reveals the dynamic nature of the microbiome during pregnancy. Aside from large intraindividual differences, gestational age of the fetus and health complications of the mother correlate with gene abundances of the mother's microbiome (47). Although the functional pathways of the various microbiota remained stable over the length of gestation, there were a few interesting examples of functions changing over time. For example, an

increase in fermenter activity in the guts of the subjects over gestational time seems to suggest that fermenters could be enriched during pregnancy.

For environmental samples, a similar approach can yield insight into the functional roles of specific microbial communities and the effects of anthropogenic change on these communities. For example, metagenomic sequencing has allowed us to see that sustained warming in grasslands leads to a shift in microbial metabolic processes such as organic matter decomposition (81). Antibiotic use also has a significant effect on environmental communities, in addition to host-associated communities. Built environment microbial communities, such as those found in urban sewage systems, could be a major route for the spread of antibiotic resistance genes. Within an urban Chinese sewage system, seasonal differences of 381 different antibiotic resistance genes were found using metagenomic data, and the majority of these genes were associated with known human gut commensal bacteria (110).

Interesting miscellanea

The microbiome has served as a unique platform for bioprospecting, and this has been aided by metagenomics approaches. In short, most analytical methods rely on examining metagenomic sequences for new enzymes that

share some homology or genetic architecture with known proteins or operons. Biosynthetic gene clusters (BGCs) can be identified by observing canonical operon structures harboring sequential enzymatic processes (11). Recent advances make direct use of metagenomics reads to find potentially interesting BGCs, such as those that produce type II polyketides, which comprise clinically important drugs such as doxorubicin and tetracycline (111). Similarly, novel CRISPR-Cas systems can be identified across diverse microbiomes, enabling new functionalities and the identification of novel PAM (protospacer adjacent motif) sites that differ from the canonical NGG sequence or with more compact genetic architecture (23). Bioprospecting for biofuel enzymes across environmental metagenomic data sets can also be used to prioritize which environmental samples to use for testing (28). These types of studies will often require cloning and expressing these genes exogenously to confirm function.

Experimental methods to determine gene functionality are well defined. One example is using metagenomic data to determine target gene(s), extracting the DNA sequence, and using expression vectors to transform the gene into bacteria for functional screening (16, 106). Although this method has had some success with larger gene operons, including identifying certain antiproliferative, anticancer, and antibiotic compounds (27), this approach is

generally more suited for those functions encoded by small operons of few genes. Samples from the environment such as soil microbiomes have been used to find new biologically and environmentally important phosphatases, as well as new domains encoding phosphatase activity, thus extending the classic categorization of known phosphatases (26).

Horizontal gene transfer

Horizontal gene transfer represents one of the major challenges to metagenomic assembly, yet metagenomics has been a useful tool in understanding this process. The flexible portion of a bacterium's genome allows the organism to rapidly adapt to changing environmental conditions by acquiring and incorporating novel functions, potentially altering its relationship with its host or providing a competitive edge against other organisms, and it is therefore of high importance to microbiome researchers. Although significant progress has been made using reference genomes (3, 104) or single-cell genomes (20, 64), current methods fall short on reliably assembling mobile genetic elements and assigning mobile genetic elements to a host genome. There is great variability in mobile genetic element structure. For example, integrated transposons and certain phage comprise inverted or direct repeats and can vary between hundreds to tens of

thousands of base pairs; plasmids and phage may contain large amounts of host genome. Recent evidence based on long-read metagenomic data reveals high mobility of transposable elements within a single organism, resulting in large heterogeneity within a single species in one microbiome (10). Several methods have been developed to try to apply alignment-based approaches to identify mobile genetic elements, either by examining the variation in reads aligning to reference genomes to identify flexible portions of genome assemblies (19, 35) or by aligning to reference genomes to identify those genomic regions that are not vertically conserved (107). Long-read metagenomic sequencing of microbiome samples will enable the capture of integrated mobile genetic elements and allow researchers to explore the heterogeneity of these elements within and across genomes. Additional tools, such as Hi-C sequencing, in which genomic DNA may be cross-linked and ligated with plasmid DNA (109, 130), may serve to enable better metagenomic assemblies that link plasmids with their hosts.

Replication rates of organisms

Relative rates of replication can be obtained using shotgun metagenomic data. Bacteria replicate their genomes bidirectionally from a singular origin of replication. Therefore, a replicating population of cells should have an

abundance of metagenomic reads that map near the origin, relative to the replication terminus. This works well in cultured bacteria, and there has been success using metagenomic data, most notably in identifying replication differences across inflammatory bowel disease (IBD) and type 2 diabetes cohorts (60). Although replication rates are most readily estimated when there are phylogenetically close reference genomes with well-known replication origins, these methods have also been applied to assembled metagenomes (21). In these cases, assembled contigs need to be binned into draft genomes and then ordered according to their relative coverage to determine the overall rate of replication. There are some inherent challenges with using such a technique on MAGs that arise from incorrect binning of contigs or scaffolds and the presence of promiscuous mobile genetic elements, which may skew coverage and overestimate replication rates. The success of this approach largely depends on the quality of the MAGs, and this method is much easier to perform in microbiomes for which there are good reference genomes.

What has metagenomics missed?

Analyzing metagenomic data requires careful consideration of the treatment of genome assemblies and abundances. Composite MAGs can lead to

inaccurate interpretations from inflated abundance or prevalence estimations, deflated diversity from ignoring or missing strain-level information, and reduced refinement in binning. Reporting quality metrics such as those proposed by the Genomics Standards Consortium (15) may lessen the burden on the end user to either determine the quality of publicly available MAGs or make incorrect assumptions. Metagenomic data sets almost exclusively rely on compositional quantification, further complicating analysis. We have only recently started reckoning with methods for assessing absolute microbial abundances and transferring this knowledge to metagenomic data (118). Aside from these technical issues, there are many aspects of microbial ecosystems that are missed when shotgun sequencing is performed alone on microbial communities.

Ecoevolutionary modeling

Metagenomic approaches have allowed us to obtain higher resolution than taxonomic profiling, and these approaches are poised to shed light on other aspects of microbial community assembly and evolutionary trajectories, yet the approaches are still fairly nascent in this regard. A remaining challenge is that information derived from individual genomes, which can be crucial for ecological or evolutionary inferences, can be lost. For example, population

variation in genetic architecture, mobile genetic elements, and SNP diversity may be difficult to ascertain. There have been several early efforts to draw evolutionary models from metagenomic data alone (46). From an examination of strain-level differences across gut microbiome samples, it appears that gene gains and losses are fairly common and can sweep to high frequencies relatively quickly, though strain replacement is the more dominant trend over longer periods of time. Differences in gene copy number variants within microbial genomes have proved to be informative about the function, and possibly the evolutionary trajectories, of specific organisms (48, 133). To some extent, it will take time for ecoevolutionary theory to develop, as we are still learning about the genomic structure in microbial communities in their natural environment. For example, a large number of small genes were recently uncovered in metagenomic data sets, many of whose functions are unknown (96).

Phenotype and comprehensive functional profiling

Phenotype is a complex trait, and metagenomic data alone are often insufficient in determining phenotypic traits. As an example, many of the metabolites in the human gut microbiome have strong associations with microbial species and pathways present in metagenomic data (120), but

predictions of metabolic output using metagenomic data alone can have high variance (73). First, phenotypic differences may be driven by large differences at the level of transcription. For example, *Verrucomicrobia* was identified as highly abundant in soil communities, leading to the assumption that it was vital for soil health and functioning. However, metatranscriptomics analysis revealed that *Verrucomicrobia* is metabolically inactive in the soil. As another example, metatranscriptomics of ruminant livestock showed that a mix of bacterial, archaeal, and eukaryotic species are active during plant degradation and methane production, which may be missed when focusing on either bacteria or eukaryotes alone (105). Second, gene-gene interactions across species may drive specific outputs of pathways, yet few of these in natural microbial communities are known. Examining co-occurrence networks may provide clues to codependent organisms (39, 45, 63), yet these methods have not been applied widely to metagenomic data sets. Modeling metabolic outputs using metagenomic data (71) is another important step toward this end, yet these models tend to be more accurate for less diverse microbiomes, such as those found in termites (61).

Despite these considerations, the predominant limitation in translating metagenomic data to phenotype is the overall proportion of genes we can

annotate. KEGG (56), COG (114), PFAM (42), TIGRFAM (51), MetaCyc (58), and other databases used to assign functions to assembled genes only capture roughly half of functions in a commonly assayed microbiome, such as the human gut (55); however, they capture a much smaller fraction in diverse, less-sampled microbiomes such as those from certain soil communities, less-studied animal microbiomes, and those from human populations living in low- and middle-income countries or remote areas (77, 85). Large-scale functional studies will be vital to improving functional databases, but these experiments are laborious, and curation of these databases is often done manually.

Alternative methods have been applied to microbial communities to gain additional functional insight. Stable isotope probing using isotopically labeled substrates can inform researchers about the specific bacteria utilizing the substrate (88, 125). The labeled substrate gets incorporated into DNA that can be separated and sequenced. One interesting example is the identification of new bile salt hydrolase genes in the gut microbiome using probes that label active enzymes that can be assayed with proteomic tools and metagenomic sequencing to identify those proteins (83). To determine which organisms are metabolically active in a sample, PMA (propidium monoazide) has been used to distinguish between live and dead cells. It

intercalates DNA, but only in those cells with compromised membranes. Light exposure causes covalent bonds to form with the DNA, resulting in fragmentation and rendering it unamenable to DNA sequencing. This method has been used widely for samples prior to 16S rRNA amplicon sequencing, but it was recently used to identify the live portion of a saliva microbiome sample that underwent metagenomic sequencing (74). Examples that probe specific functions will enhance our understanding of metagenomic communities above metagenomics alone.

Similarly, integrating metagenomic, metatranscriptomic, and metabolomic data can alternatively improve functional assessments of communities. There has been a sharp increase of methods that integrate omics data. The multi-omics approach is valuable in that it does not require bacterial culturing, which is an impediment to examining microbiome function. In phase 2 of the Human Microbiome Project, known as the Integrative Human Microbiome Project (iHMP), 1,785 individuals from three microbiome-associated condition cohorts were sampled: pregnancy and preterm birth, IBD, and type 2 diabetes. A wealth of data was collected, including gut microbiome metagenomes, metatranscriptomes, proteomes, metabolomes, and virome data (69). By integrating these data, the authors were able to associate functions and molecular dynamics to specific taxa of the gut

microbiome. Similarly, omics studies reveal that twin pairs share metabolic pathways on average almost twice as much as they share species (120), suggesting that in the search for therapeutic targets, genetic associations, or biomarkers, it may be more informative to study the functions of the gut microbiome rather than the organismal composition or species diversity.

In addition to omics performed on microbial communities, an increased number of studies are also incorporating measurements of the host. Zhou et al. (134) used a longitudinal multi-omics approach to study host-microbe dynamics in prediabetes. By integrating metagenomes, transcriptomes, metabolomes, cytokines, and proteomes from 106 individuals, they were able to detect molecular signatures in 1 person that preceded the onset of type 2 diabetes, which included the inflammation markers interleukin-1 receptor agonist and high-sensitivity C-reactive protein. More broadly, they were able to characterize thousands of host-microbe interactions that were distinctive between insulin-sensitive and insulin-resistant individuals. Techniques to integrate the diversity of data sources, each with their own benefits and limitations, are still under development (7).

Spatial analyses of the microbiome

Missing from many metagenomic analyses are temporal and spatial dynamics. Time course experiments are relatively expensive, but several large-scale temporal data sets are starting to emerge, such as a recent study of patients with IBD (90). Similar to 16S rRNA amplicon data sets, metagenomic time course data analysis requires not only careful managing of the compositionality but also autocorrelation. Techniques to obtain spatial data about microbiomes are emerging (92, 99, 100, 122, 123), by applying species-level probes or by sequencing proximate microbes captured in preserved microscale blocks or by performing laser dissection of fixed communities. These techniques capture species-level interactions and have not yet scaled to accommodate metagenomic sequencing approaches.

Innovation and future directions

Metagenomic analyses of microbiomes will be vastly altered in the coming decade by technological and accessibility improvements in DNA and RNA sequencing. As long-read sequencing becomes cheaper, it will negate the need for elaborate methods for genome assembly and phasing of SNPs. Since the quality of draft genomes assembled from short-read sequencing data may be highly variable, and few studies incorporate reference genomes

for comparison, long-read sequencing will go a long way to improve the quality of these metagenomic assembled genomes. The increased use of metagenomics will create new challenges in data storage and data reporting, especially as the types of data platforms (e.g., short-read, long-read, Hi-C, Tn-seq, functional screening) grow. The size of future data sets will necessitate solutions for data compression, high-speed search, and memory-efficient assembly methods, some of which are starting to become available (6, 29, 59). Standard protocols for data reporting and submission will be important, especially in terms of what information and metadata to provide.

If the expenses and error rates associated with long-read sequencing are reduced, many of the challenges associated with short-read sequencing will fade. Assemblies will be less fragmented, SNP phasing will be inferred based on co-occurrence on reads, and integrated mobile genetic elements will be associated with their flanking genomes. Nevertheless, this will not fully solve the problem of associating extrachromosomal elements with their host genomes. Alternatively, single-cell genome sequencing may provide the technological advance that surmounts some of these problems. Yet, even the largest of studies is several thousand cells, orders of magnitude below what is typically sampled in a shotgun metagenomic sample. Currently single-cell

sequencing technologies are limited by the cost, and the quality of the genomes is highly variable, resulting in a large amount of data loss.

Despite these projected improvements, methods to assign functions to the vast number of genes within the microbiome are still necessary to understand the mechanisms underlying microbiome-related phenotypic outcomes. It was recently discovered that a single-amino-acid-residue difference in the dopamine dehydroxylase (DahD) gene in *Eggerthella lenta* altered whether the pharmaceutical l-dopa remained active in microbiome samples from a cohort of patients with Parkinson disease (72). This level of detailed understanding of gene function will be required for the research field to go beyond characterization of microbiomes to understanding the mechanisms underlying an overall phenotype. Examination of modifications of DNA, such as methylation patterns obtained using single-molecule real-time (SMRT) sequencing (5), can reveal interesting patterns of plasmid mobility within natural microbial communities. Technological innovation will reveal interesting layers of organismal interactions, functional roles, evolutionary trajectories, and niche occupancy in microbial communities, which will lead to a better understanding of how to shape communities to achieve a prosperous outcome for health, agriculture, or environmental sustainability.

References

1. Albanese D, Donati C. 2017. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* 8(1):2260
2. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, et al. 2014. Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11(11):1144–46
3. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. 2019. A reverse ecology approach based on a biological definition of microbial populations. *Cell* 178(4):820–34.e14
4. Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* 21(2)584–94
5. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, et al. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* 36(1):61–69
6. Berger B, Peng J, Singh M. 2013. Computational solutions for omics data. *Nat. Rev. Genet.* 14(5):333–46
7. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, et al. 2016. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinform.* 17(Suppl. 2):15
8. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, et al. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 37(8):937–44
9. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, et al. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* 36(11):1067–80
10. Bishara A, Moss EL, Tkachenko E, Kang JB, Zlitni S, et al. 2018. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. bioRxiv 125211
11. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, et al. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* 47(W1):W81–87
12. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13(12):R122
13. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37(8):852–57

14. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, et al. 2016. The effect of host genetics on the gut microbiome. *Nat. Genet.* 48(11):1407–12
15. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35(8):725–31
16. Brady SF. 2007. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* 2(5):1297–305
17. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185(20):6220–23
18. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *PNAS* 99(22):14250–55
19. Brito IL, Gurry T, Zhao S, Huang K, Young SK, et al. 2019. Transmission of human-associated microbiota along family and social networks. *Nat. Microbiol.* 4(6):964–71
20. Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, et al. 2016. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535(7612):435–39
21. Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* 34(12):1256–63
22. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, et al. 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* 533(7604):543–46
23. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, et al. 2017. New CRISPR-Cas systems from uncultivated microbes. *Nature* 542(7640):237–41
24. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11(12):2639–43
25. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13(7):581–83
26. Castillo Villamizar GA, Nacke H, Boehning M, Herz K, Daniel R. 2019. Functional metagenomics reveals an overlooked diversity and novel

features of soil-derived bacterial phosphatases and phytases. *mBio* 10(1):e01966-18
27. Chang F-Y, Brady SF. 2013. Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *PNAS* 110(7):2478
28. Chaudhary N, Gupta A, Gupta S, Sharma VK. 2017. BioFuelDB: a database and prediction server of enzymes involved in biofuels production. *PeerJ*. 5:e3497
29. Cleary B, Brito IL, Huang K, Gevers D, Shea T, et al. 2015. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33(10):1053–60
31. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, et al. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* 8(1):15955
32. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, et al. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4(4):693–700
33. Duerkop BA, Kleiner M, Paez-Espino D, Zhu W, Bushnell B, et al. 2018. Murine colitis reveals a disease-associated bacteriophage community. *Nat. Microbiol.* 3(9):1023–31
34. Dunphy CM, Gouhier TC, Chu ND, Vollmer SV. 2019. Structure and stability of the coral microbiome in space and time. *Sci. Rep.* 9(1):6785
35. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* 27(1):140–53.e9
36. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, et al. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5(1):4498
37. Edgar RC. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv 074252
38. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, et al. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 3:e1319
39. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, et al. 2012. Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.* 8(7):e1002606
40. Ferretti P, Farina S, Cristofolini M, Girolomoni G, Tett A, Segata N. 2017. Experimental metagenomics and ribosomal profiling of the human skin microbiome. *Exp. Dermatol.* 26(3):211–19

41. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, et al. 2018. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* 24(1):133–145.e5
42. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42(D1):D222–30
43. Fouts DE. 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 34(20):5839–51
44. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15(11):962–68
45. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLOS Comput. Biol.* 8(9):e1002687
46. Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biol.* 17(1):e3000102
47. Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, et al. 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res.* 28(10):1467–80
48. Greenblum S, Carr R, Borenstein E. 2015. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160(4):583–94
49. Grottoli AG, Dalcin Martins P, Wilkins MJ, Johnston MD, Warner ME, et al. 2018. Coral physiology and microbiome dynamics under combined warming and ocean acidification. *PLOS ONE* 13(1):e0191156
50. Guerin E, Shkoporov A, Stockdale SR, Gonzalez-Tortuero E, Ross RP, Hill C. 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* 24:653–64
51. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31(1):371–73
52. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, et al. 2018. Evaluating the information content of shallow shotgun metagenomics. *mSystems* 3(6):e00069-18
52a. Hum. Microbiome Proj. Consort., Huttenhower C, Gevers D, Knight R, Abubucker S, et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–14
53. Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17(3):377–86

54. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*. 2:e603
55. Joice R, Yasuda K, Shafquat A, Morgan XC, Huttenhower C. 2014. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab*. 20(5):731–41
55a. Jump. Consort. Hum. Microbiome Proj. Data Gener. Work. Group. 2012. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLOS ONE* 7(6):e39315
56. 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(D1):D457–62
57. Kang DD, Li F, Kirton E, Thomas A, Egan R, et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359
58. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. 2002. The MetaCyc database. *Nucleic Acids Res*. 30(1):59–61
59. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 26(12):1721–29
60. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, et al. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* 349(6252):1101–6
61. Kundu P, Manna B, Majumder S, Ghosh A. 2019. Species-wide metabolic interaction network for understanding natural lignocellulose digestion in termite gut microbiota. *Sci. Rep.* 9(1):16329
62. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*. 14(4):169–81
63. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.* 11(5):e1004226
64. Labonté JM, Field EK, Lau M, Chivian D, Van Heerden E, et al. 2015. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front. Microbiol.* 6:349
65. Lagier J-C, Dubourg G, Million M, Cadoret F, Bilen M, et al. 2018. Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* 16(9):540–50

66. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10):1674–76
67. Li SS, Zhu A, Benes V, Costea PI, Hercog R, et al. 2016. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* 352(6285):586–89
68. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, et al. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* 21(10):1228–34
69. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569(7758):655–62
70. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550(7674):61–66
71. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, et al. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35(1):81–89
72. Maini Rekdal V, Bess EN, Bisanz JE, Turnbaugh PJ, Balskus EP. 2019. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science* 364(6445):eaau6323
73. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, et al. 2019. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10(1):3136
74. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* 6(1):42
75. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32(7):1088–90
76. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. *PLOS Genet.* 9(12):e1003987
77. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568(7753):505–10
78. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27(5):824–34
79. Oh J, Byrd AL, Deming C, Conlan S, NISC Comp. Seq. Program, et al. 2014. Biogeography and individuality shape function in the human skin metagenome. *Nature* 514(7520):59–64

80. Olm MR, West PT, Brooks B, Firek BA, Baker R, et al. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* 7(1):26
81. Orellana LH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT. 2018. Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. *Appl. Environ. Microbiol.* 84(2):e01646-17
82. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, et al. 2019. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 179(7):1623–35.e11
83. Parasar B, Zhou H, Xiao X, Shi Q, Brito IL, Chang PV. 2019. Chemoproteomic profiling of gut microbiota-associated bile salt hydrolase activity. *ACS Cent. Sci.* 5(5):acscentsci.9b00147
84. 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(7):1043–55
85. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, et al. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176(3):649–62.e20
87. 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(11):1420–28
88. Pepe-Ranney C, Campbell AN, Koechli CN, Berthrong S, Buckley DH. 2016. Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. *Front. Microbiol.* 7:703
89. Porras AM, Brito IL. 2019. The internationalization of human microbiome research. *Curr. Opin. Microbiol.* 50:50–55
90. Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, et al. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25(9):1442–52
91. Qin J, Li Y, Cai Z, Li S, Zhu J, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418):55–60
92. Riva A, Kuzyk O, Forsberg E, Siuzdak G, Pfann C, et al. 2019. A fiber-deprived diet disturbs the fine-scale spatial architecture of the murine colon microbiome. *Nat. Commun.* 10(1):4366

93. Rotmistrovsky K, Agarwala R. 2011. BMTagger : Best Match Tagger for removing human reads from metagenomics datasets, *Bioinformatics*, Unpublished.
94. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ*. 3:e985
95. Sarhan MS, Hamza MA, Youssef HH, Patz S, Becker M, et al. 2019. Culturomics of the plant prokaryotic microbiome and the dawn of plant-based culture media—a review. *J. Adv. Res.* 19:15–27
96. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, et al. 2019. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* 178(5):1245–59.e14
97. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75(23):7537–41
98. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, et al. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* 14(11):1063–71
99. Sheth RU, Li M, Jiang W, Sims PA, Leong KW, Wang HH. 2019. Spatial metagenomic characterization of microbial biogeography in the gut. *Nat. Biotechnol.* 37(8):877–83
100. Shi H, Zipfel W, Brito I, Vlaminck I De. 2019. Highly multiplexed spatial mapping of microbial communities. bioRxiv 678672
101. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3(7):836–43
102. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–12
102a. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19(6):1117–23
103. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, et al. 2018. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 23(2):229–40.e5
104. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–44

105. Söllinger A, Tveit AT, Poulsen M, Noel SJ, Bengtsson M, et al. 2018. Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. *mSystems* 3(4):e00038-18
106. Sommer MOA, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325(5944):1128–31
107. Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiomie* 7(1):36
108. Song W-Z, Thomas T. 2017. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* 33(12):1873–75
109. Stalder T, Press MO, Sullivan S, Liachko I, Top EM. 2019. Linking the resistome and plasmidome to the microbiome. *ISME J.* 13(10):2437–46
110. Su J-Q, An X-L, Li B, Chen Q-L, Gillings MR, et al. 2017. Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome* 5(1):84
111. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, et al. 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science* 366(6471):eaax9176
112. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, et al. 2015. Structure and function of the global ocean microbiome. *Science* 348(6237):1261359
113. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10(12):1196–99
114. 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(1):33–36
115. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, et al. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25(4):667–78
116. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27(4):626–38

117. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6(1):158
118. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, et al. 2017. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551(7681):507–11
119. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, et al. 2018. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* 562(7728):589–94
120. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, et al. 2019. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* 10(1):4505
121. Weissbrod O, Rothschild D, Barkan E, Segal E. 2018. Host genetics and microbiome associations through the lens of genome wide association studies. *Curr. Opin. Microbiol.* 44:9–19
122. Welch JLM, Hasegawa Y, McNulty NP, Gordon JI, Borisy GG. 2017. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *PNAS* 114(43):E9105–14
123. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. 2016. Biogeography of a human oral microbiome at the micron scale. *PNAS* 113(6):E791–800
124. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* 28(4):569–80
125. Wilhelm RC, Singh R, Eltis LD, Mohn WW. 2019. Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing. *ISME J.* 13(2):413–29
126. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, et al. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25(4):679–89
127. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20(1):257
128. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32(4):605–7
129. Xie H, Guo R, Zhong H, Feng Q, Lan Z, et al. 2016. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* 3(6):572–84.e3

130. Yaffe E, Relman DA. 2019. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat. Microbiol.* 5(2):343–53
131. Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, et al. 2018. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe* 24(1):146–54.e4
132. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking metagenomics tools for taxonomic classification. *Cell* 178(4):779–94
133. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, et al. 2019. Structural variation in the gut microbiome associates with host health. *Nature* 568(7750):43–48
133a. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–29
134. Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, et al. 2019. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* 569(7758):663–71

CHAPTER 2: Collective effects of human genomic variation on microbiome function

This part of the dissertation is adapted from New FN, Baer BR, Clark AG, Wells MT, & Brito IL. "Collective effects of human genomic variation on microbiome function". (2021). *In Review.*

Abstract

Studies of the impact of host genetics on gut microbiome composition have mainly focused on the impact of individual single nucleotide polymorphisms (SNPs) on gut microbiome composition, without considering their collective impact or the specific functions of the microbiome. To assess the aggregate role of human genetics on the gut microbiome composition and function, we apply sparse canonical correlation analysis (sCCA), a flexible, multivariate data integration method. A critical attribute of metagenome data is its sparsity, and here we propose application of a Tweedie distribution to accommodate this. We use the TwinsUK cohort to analyze the gut microbiomes and human variants of 250 individuals. Sparse CCA, or sCCA, identified SNPs in microbiome-associated metabolic traits (BMI, blood pressure) and microbiome-associated disorders (type 2 diabetes, some

neurological disorders) and certain cancers. Both common and rare microbial functions such as secretion system proteins or antibiotic resistance were found to be associated with host genetics. sCCA applied to microbial species abundances found known associations such as Bifidobacteria species, as well as novel associations. Despite our small sample size, our method is able to identify not only previously known associations, but novel ones as well. Overall, we present a new and flexible framework for examining host-microbiome genetic interactions, and we provide a new dimension to the current debate around the role of human genetics on the gut microbiome.

Introduction

Variation in gut microbiome composition underlies numerous human phenotypes and health outcomes, such as immune function, metabolic disorder, cancer and psychiatric traits. These variations have been attributed largely to environmental factors, including diet (Rothschild *et al.* 2018), antibiotic exposure (Francino 2016), and birth modality (Bokulich *et al.* 2016). However, twin and population-based studies have identified genetic associations with the overall composition of the gut microbiota (Blekhman *et al.* 2015; Davenport *et al.* 2015; Turpin *et al.* 2016; Bonder *et al.* 2016a; Igartua *et al.* 2017; Wang *et al.* 2018; Visconti *et al.* 2019; Hughes *et al.* 2020;

Kurilshikov *et al.* 2021), in addition to heritable taxa (Goodrich *et al.* 2014, 2016). Underpowered studies, population differences, as well as inconsistencies in experimental and computational methods, have led to a problem of replicability across microbiome genome-wide association studies (GWAS) and contribute to the doubt surrounding the role of genetics in shaping the gut microbiome (Costea *et al.* 2017; Rothschild *et al.* 2018).

Traditional methods to associate host genetics with microbiome traits involve comparing a single genotype to a single microbial species or pathway (one-vs-one) or a single genotype to a set of microbial species (one-vs-many association tests, Spearman's rank correlation) (Blekhman *et al.* 2015; Davenport *et al.* 2015; Goodrich *et al.* 2016; Turpin *et al.* 2016; Bonder *et al.* 2016a; Igartua *et al.* 2017; Wang *et al.* 2018; Hughes *et al.* 2020; Kurilshikov *et al.* 2021). As the number of species within a microbiome can be several orders of magnitude greater than the sample size, there is limited statistical power and a risk of overfitting. Until the cost and computational burden of microbiome profiling decreases, it will be difficult to reach the sample sizes of modern human GWAS, which now reach into the hundreds of thousands of participants even for single phenotypes (Tam *et al.* 2019). The sample size needed to perform reasonably efficient inference can be reduced by performing association tests between one genotype and the dissimilarity

between the study's microbiome samples, which reduces the dimension of multivariate phenotypes such as the gut microbiome (Goodrich *et al.* 2016; Rothschild *et al.* 2018). Although the microbiota function as a community, these gross metrics (e.g., beta diversity) obscure the sources and mechanisms of the host-microbe genetic relationship and their effect sizes tend to be small. Furthermore, like the observation that human GWAS studies focusing on single SNP-level associations do not account for the total observed heritability (Yang *et al.* 2010), microbiome GWAS studies may not account for vertical- and family-level heritability of microbiome components (Hildebrand *et al.*; Brito *et al.* 2019).

Since microbial genomes are not static due to recombination between closely related strains and horizontal gene transfer (Treangen and Rocha 2011; Shapiro 2016; Garud *et al.* 2019), species-level analyses can introduce avoidable variability in comparing results across studies. Metagenomic sequencing offers an opportunity to examine microbial genes, which offers a more consistent accounting of a microbiome's functional capacity across individuals and may capture more specific mechanisms underlying various phenotypes. Not surprisingly, GWA studies using metagenomes are few, as the cost of metagenomic data acquisition can be roughly 10-fold higher than 16S rRNA profiles. Furthermore, the number of genes within a metagenome

far outweighs the number of species, exacerbating the problem of dimensionality. Aggregating genes into functional units such as protein families, KEGG orthologs, or pathways reduces the dimensionality of the data, though these groupings can still outnumber sample sizes. Another challenge with metagenomic data is that they are right-skewed and are characterized by both zero-inflation and overdispersion. To appropriately model these data, we propose the use of a Tweedie distribution. Tweedie distributions are exponential dispersion families of distributions and are often used in generalized linear models (Jørgensen 1987). Depending on their parameterization, they can have mass at zero along with non-negative continuous support. Tweedie distributions can also describe the mean to variance power law relationship that is present in several types of data including metagenomic abundances. These two characteristics have made it very useful in fields as diverse as ecology and natural language processing (Kendal and Jørgensen 2011; Foster and Bravington 2013; Warton and Hui 2017; Baer *et al.* 2018). In ecology the distribution is commonly characterized through Taylor's law.

To address the challenge of identifying associations between human genetics and the composition and function of the gut microbiome, we apply a flexible, unsupervised, multivariate data integration method, known as

canonical correlation analysis (CCA) (Hotelling 1936). This method bypasses the need for multiple hypothesis testing and leverages the combination of many small effect sizes. CCA is the earliest multi-table analysis method, first used in 1936 to relate multidimensional variables (Hotelling 1936). CCA creates low-dimensional representations of features, similarly to principal component analysis (PCA). However, unlike PCA, it allows for comparisons across multiple measurements, where the low dimensional representations of each set of features, or canonical components, represent the maximum correlation between the two sets of linear combinations. This method works well when the number of samples exceeds the number of measured features, which is not the case with modern genomics data. As an example, for a set of individuals, it is possible to correlate their RNA-seq gene expression data with their DNA copy number or SNP data. In this case, the number of genes with expression data and the number of polymorphisms will most likely exceed the number of individuals that can feasibly be sampled. Penalized CCA methods have been developed to overcome this issue (Witten and Tibshirani 2009; Witten *et al.* 2009; Rodosthenous *et al.* 2020). In penalized CCA, also known as sparse CCA or sCCA, sparsity is induced in the features through penalization, for example using an l_1 penalty (lasso) (Tibshirani 1996). This method is flexible in two ways: 1) it can be applied to

any number of measurements sampled from the same individuals or units, and 2) it can handle different penalization schemes, and 3) it does not require summarization of tables as preprocessing, unlike the beta diversity example. Here, we apply this method to human genotypes and corresponding metagenomic features from a set of 125 twin pairs that are part of the TwinsUK project (Xie *et al.* 2016). Using sCCA and Tweedie distributions, we model the relationship between human SNPs and microbial species or gene family profiles, using appropriate penalties for each.

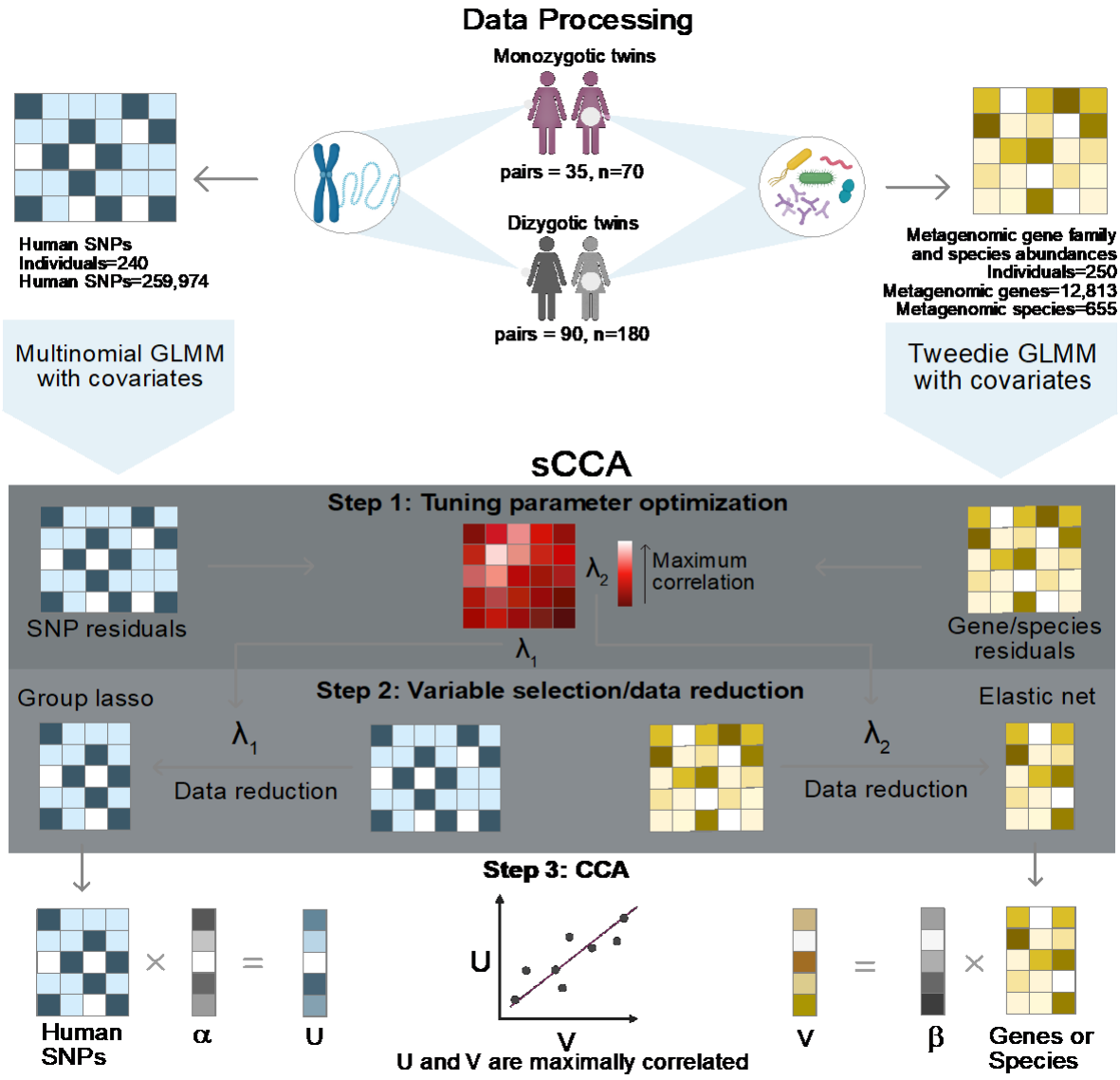


Figure 2.1. Overview of the SCCA method applied to the TwinsUK cohort. Microbiome gene abundances or species abundances estimated from whole shotgun metagenomic sequencing, and host genotype data were inputted into generalized linear mixed models to extract residuals for downstream analysis. Step 1 involves the tuning parameter optimization for sparse CCA. The pair of regularization parameters that represent the maximum correlation in the data are chosen for Step 2: Variable selection. In Step 2, we apply elastic net to the sets of genes or species abundances and group lasso to the human genotype data to reduce the size of the datasets. In Step 3, CCA is applied to the reduced data to find the maximally correlated linear combinations of the data.

Results

Study cohort and metagenomic processing

The TwinsUK project includes corresponding human genotypes and microbiome metagenomic shotgun sequences of 250 individuals. This dataset consists of female twin pairs from the United Kingdom, of which 35 are monozygotic and 92 are dizygotic (Moayyeri *et al.* 2013; Xie *et al.* 2016). A subset of 240 individuals were previously genotyped (Goodrich *et al.* 2016). The SNP data were filtered to remove missing data, rare alleles (present in fewer than 10% of individuals), and loci violating HWE or in high linkage disequilibrium (>80%) with another. After standard quality filtering, metagenomic sequences were assembled into contigs. We generated a custom microbial gene catalog from the metagenomic assemblies, totaling 5,025,174 microbial genes. These were combined into 12,813 annotated gene families present in at least 10% of the samples. We also estimated species abundances from the metagenomes and found that 655 species are found in at least 10% of the samples.

Metagenomic shotgun sequencing data are compositional, which arises due to sequencing DNA from each sample that is equal to the library size, and proportional to the community size within each sample, and therefore results in relative abundance data of community members rather

than absolute abundances. Such composition-based data imposes strong constraints on the correlations in relative abundances, and subsequent analysis must be done with caution. Mishandling of compositionality can make most results uninterpretable (Gloor *et al.* 2017). Microbiome 16S rRNA sequencing analyses have used rarefaction or sub-sampling techniques to mitigate the effects of compositionality, but this can result in the loss of data and can result in false positives (McMurdie and Holmes 2014). Normalization methods for read counts assigned to a genomic feature, such as those used in RNA sequencing analyses, account for the number of reads sequenced per sample and other technical variability (Mortazavi *et al.* 2008), however, normalized abundances are still relative values (Zhao *et al.* 2020). Rather, by modifying a commonly used normalization method, RPKM (reads per kilobase sequenced per million) (Mortazavi *et al.* 2008), we include a term for the geometric mean of each sample's abundances, thereby accounting for the compositionality of the data as well as the library size of each sample(Aitchison 1982). Our modified RPKM approach is a hybrid between Aitchison's center log ratio transformation (CLR) (Aitchison 1982) and the standard RPKM approach. Rather than divide the number of mapped reads by the number of reads sequenced, we divide by the geometric mean of the sample and adjust for the length of the gene and the magnitude

of the library in one step.

Tweedie distribution for metagenomic abundance data

Normalized abundances from metagenomic sequencing are also right-skewed, overdispersed, and tend to be zero-inflated. The negative binomial and Poisson distributions are typically used to model sequencing data including RNA-seq and shotgun metagenomic sequencing (Love *et al.* 2014; Calle 2019). However, the mean and variance of the Poisson distribution are the same. The negative binomial has a more flexible dispersion parameter than the Poisson, but it's still not flexible enough: when comparing the log variance and the log mean, its intercept is always at *0* like the Poisson distribution (Figure 2.2A, B). To account for zero-inflation and overdispersion within the normalized metagenomic counts, we introduce the Tweedie distribution to model the taxa and gene abundances from metagenomic shotgun sequencing. This contrasts with some genomic methods which use zero-inflated Poisson and negative binomial distributions, which add an additional parameter and make them no longer exponential dispersion families. Metagenomic abundance data have a large mass at *0* and a long, positive tail, like the probability distribution for certain Tweedie distributions. The shape of a Tweedie distribution is determined by the shape parameter, p. When $1 < p < 2$, the distribution is continuous for Y

> 0, with a positive mass at $Y=0$. When $p > 2$, the distributions are continuous for $Y > 0$, without the mass at zero, and are no longer appropriate for zero-inflated data. The variance of the response variable is related to the mean through the Tweedie power and dispersion parameters, p and ϕ, where $Var(Y) = \phi\mu^p$. We show that when $1 < p < 2$, the Tweedie distribution is flexible enough to capture the mean-to-variance power relationship in the metagenomic taxa and gene abundances (Figure 2.2A, B) (Jørgensen 1987). This Tweedie distribution captures the relationship better than either the negative binomial or Poisson, but it also accounts for the number of zeros present in the data. The expected number of zeros, given the count data, roughly matches the observed number of zeros specified by the Tweedie relationship, modeled as $P(Y=0) = \exp(\frac{-\mu^{2-p}}{\phi(2-p)})$, where ϕ is the dispersion parameter and p is the power parameter. For fixed ϕ and p, the probability of a gene count within a sample being zero decreases as the mean increases linearly (Figure 2.2C, D).

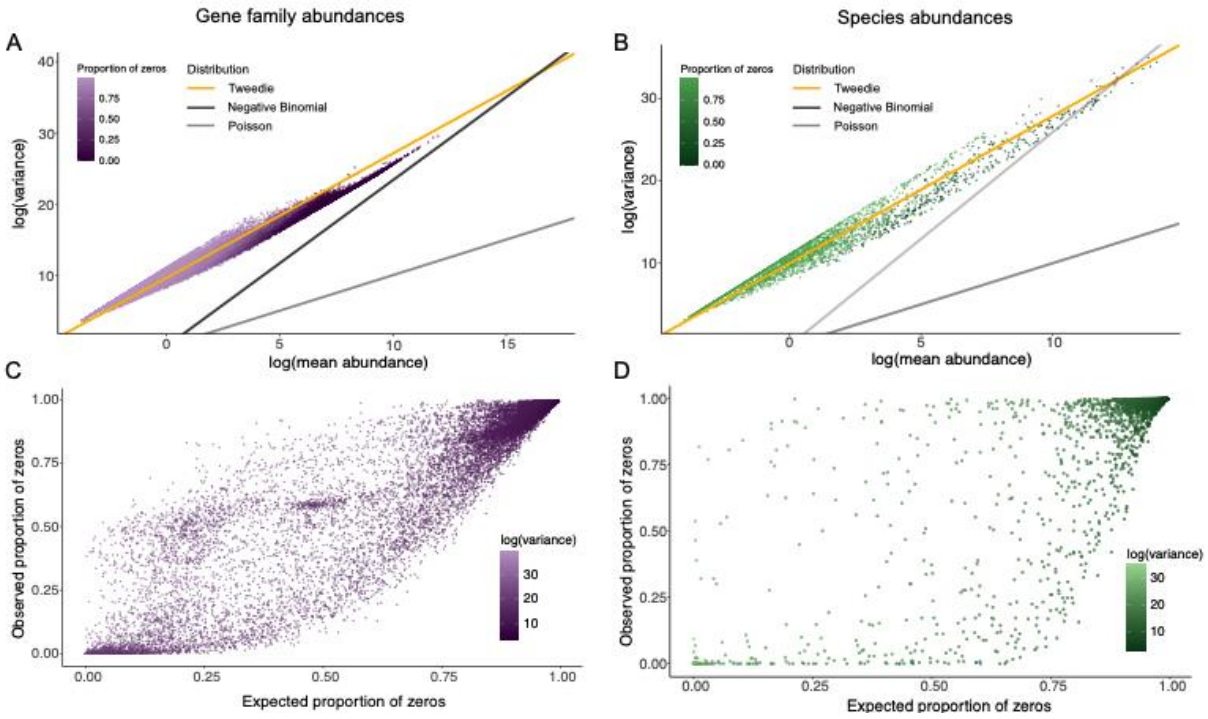


Figure 2.2. Tweedie distribution captures the mean-to-variance power structure of metagenomic abundance data. A-B) the relationship between the log mean and log variance of the microbiome gene family abundances (A) and microbiome species abundances (B). The yellow line represents the Tweedie distribution that best captures the mean to variance relationship, while the Poisson and negative binomial do not. C-D) show the expected proportion of zeros per microbial gene family or microbiome species plotted against the observed proportion of zeros for the gene family abundances (C) and species abundances (D). The expected proportion is calculated from the Tweedie distribution where a linear relationship is expected.

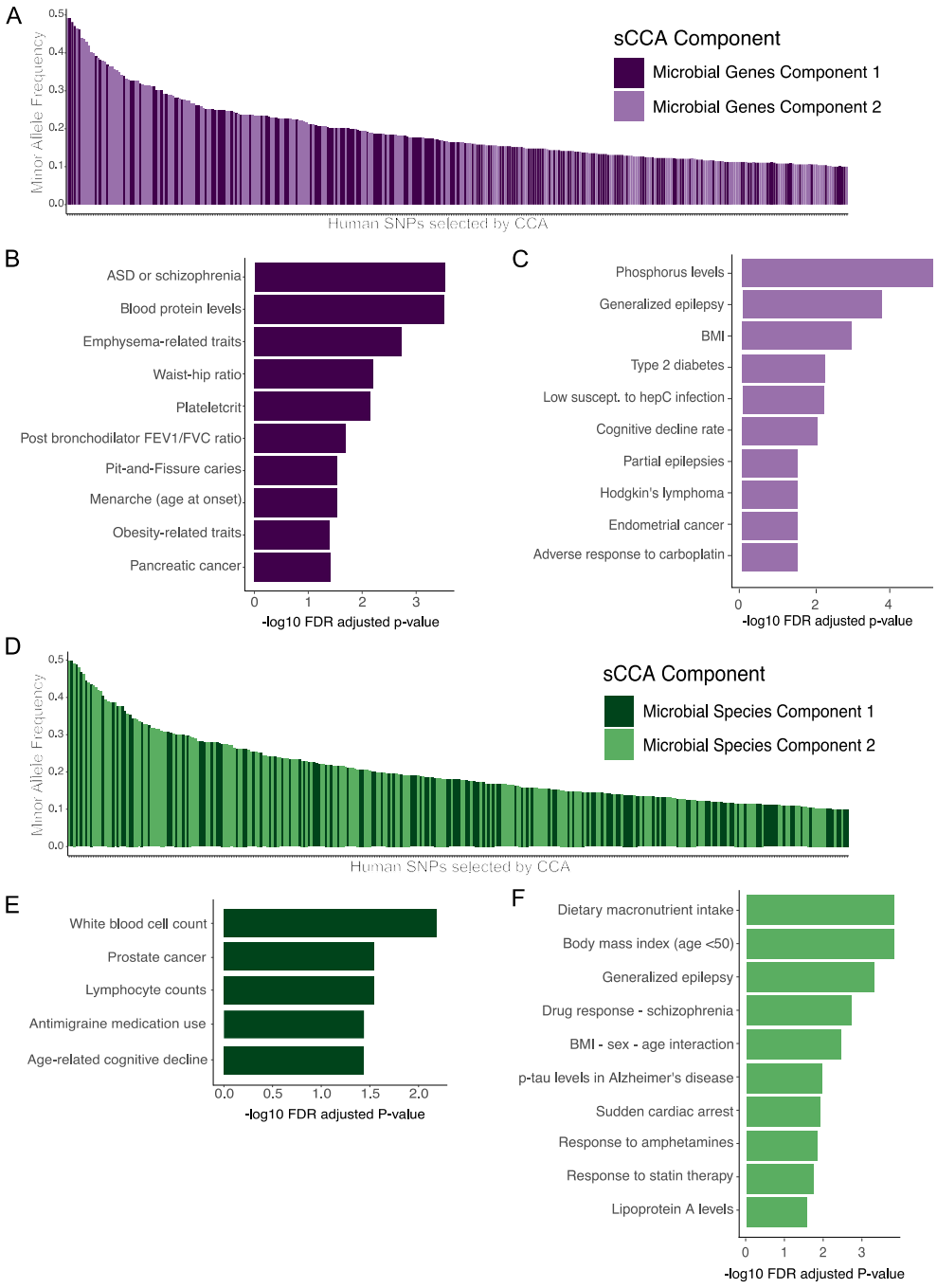


Figure 2.3. Human SNPs selected by sCCA as correlated with the gut microbiome from each analysis. A-C) results from the correlation test of host genetics with the microbial gene family abundances of the gut microbiome, and D-F) results from the test of host genetics with microbiome species abundances. A and D) show the prevalence of the

selected SNPs in the TwinsUK population, colored by which CCA component they were selected in. B, C, E, and F) show the annotation of the human SNPs and their associated human genes that are associated with the gut microbiome genes and gut microbiota, respectively.

Sparse CCA identifies novel associations of host genetics with the gut microbiome

We use a sparse variant of CCA to identify associations between human genetics and features of the gut microbiome (Figure 2.1). First, we identify appropriate penalization methods for each data type. To the human SNPs, which we have dummy-coded for additivity and dominance, we apply a group lasso (Yuan and Lin 2006), and to the microbial species and gene family abundances we apply an elastic net (Zou and Hastie 2005). We apply a general linear mixed model to the SNPs and the microbial gene family and species abundances to control for the effects of sample shipment number, age at sampling, zygosity and family, and BMI. We include a random effect for the twin's zygosity (monozygotic or dizygotic) to remove the effect of a person's correlation with their own twin. The residuals are used as inputs to the sCCA pipeline. The first step in sCCA is tuning parameter optimization for the group lasso and elastic net. Next, using these optimized parameters for variable selection, sCCA is applied (Figure 2.1). The output is the first canonical component, comprising a list of SNPs and metagenomic gene

families or species, each with a corresponding weight. Features with non-zero weights are associated with each other across tables. To obtain additional orthogonal components from sCCA, we deflate the original residuals with the output from the first component and run sCCA again with the deflated data (Witten *et al.* 2009).

To determine whether the same human SNPs associate with both species and bacterial gene families, we performed sCCA four times: one set of analyses on the species abundances and one set on the gene family abundances. For each of these two sets, we extracted two canonical components, for a total of four lists of human SNPs associated with either species or gene abundances from the gut microbiome.

The first component of the sCCA test with microbial gene families identified 161 human SNPs and the second identified 171 human SNPs collectively associated with microbial gene family abundances. The analysis of the species abundances identified a similar number: 141 and 181 human SNPs that are associated with species abundances within the first and second component, respectively. We find limited overlap between previous microbiome GWAS studies and our results at the level of SNPs. Twelve SNPs were found in common with previous studies: (rs9327097 [located within TNFAIP8], rs11578436, rs193466 [RARS1], rs60701 [DAP],

rs6947185 [COL26A1], rs7638704, and rs1882926 [PDC-AS1]) from Kurilshikov et al. (Kurilshikov *et al.* 2021) and rs906351 from Davenport et al. (Davenport *et al.* 2015). This is expected given the differences in methodologies.

We annotated the selected SNPs by their gene-disease associations and functions (GO terms) and performed enrichment analyses in FUMA GWAS (Watanabe *et al.* 2017) (Figure. 2.3, for SNP enrichment results, see Supplemental Table S2). We see similarities between the many significantly enriched traits (Fisher's exact test, FDR adjusted p-value) of our bacterial species and bacterial pathway sCCA analyses, which is also reflected in a small number of shared SNPs between the two analyses (53 SNPs overlapped in the first component of our bacterial species and bacterial genes sCCA analyses; and 12 overlapped in the second components of both analyses.) In both analyses, we see overlap of metabolic traits and diseases (obesity, BMI, blood pressure, type 2 diabetes), neurological diseases (Alzheimer's disease, Schizophrenia, epilepsy, age-related cognitive decline), and cancer (pancreatic cancer, endometrial cancer, adverse response to chemotherapy, prostate cancer, Hodgkin's lymphoma). Many of these traits reflect known links between the microbiome and metabolic disorders (Yang

et al. 2021), cancers (Sepich-Poore *et al.* 2021), and psychiatric disorders (Bastiaanssen *et al.* 2019; Szeligowski *et al.* 2020).

Microbial gene family CCA results

We hypothesized that human genetics may have associations that span species and reflect specific functional gene families. From our two analyses on microbial gene family abundances, we find that 168 and 171 microbial gene families are associated with human genetics. Any two components of sCCA extract orthogonal information from the data. Here, this is reflected prominently in the abundances of the gene families between components 1 and 2. Component 1 selected more prevalent genes and is enriched for common enzymes (Fisher's exact test, FDR p-value) (Figure 2.4A). Conversely, component 2 selects rare genes (generally present in fewer than 50% of the population) and is enriched for rarer functions such as antimicrobial resistance (Figure 2.4C). The first component's results include functions with enrichment for secretion system and enzymes (Fisher's exact test, FDR $p<0.1$) (Figure 2.4B). The second component is enriched for enzymes, two-component system, peptidoglycan biosynthesis and degradation proteins, DNA repair and recombination proteins, and antimicrobial resistance (Fisher's exact test, FDR $p<0.1$) (Figure 2.4C). To further investigate the KEGG BRITE module known as "enzymes", we

performed an enrichment test of the subcategories of enzymes within the results. Within component 1, the enriched enzymes are translocases and lyases, and translocases are enriched in component 2 (Figure 2.4B, C).

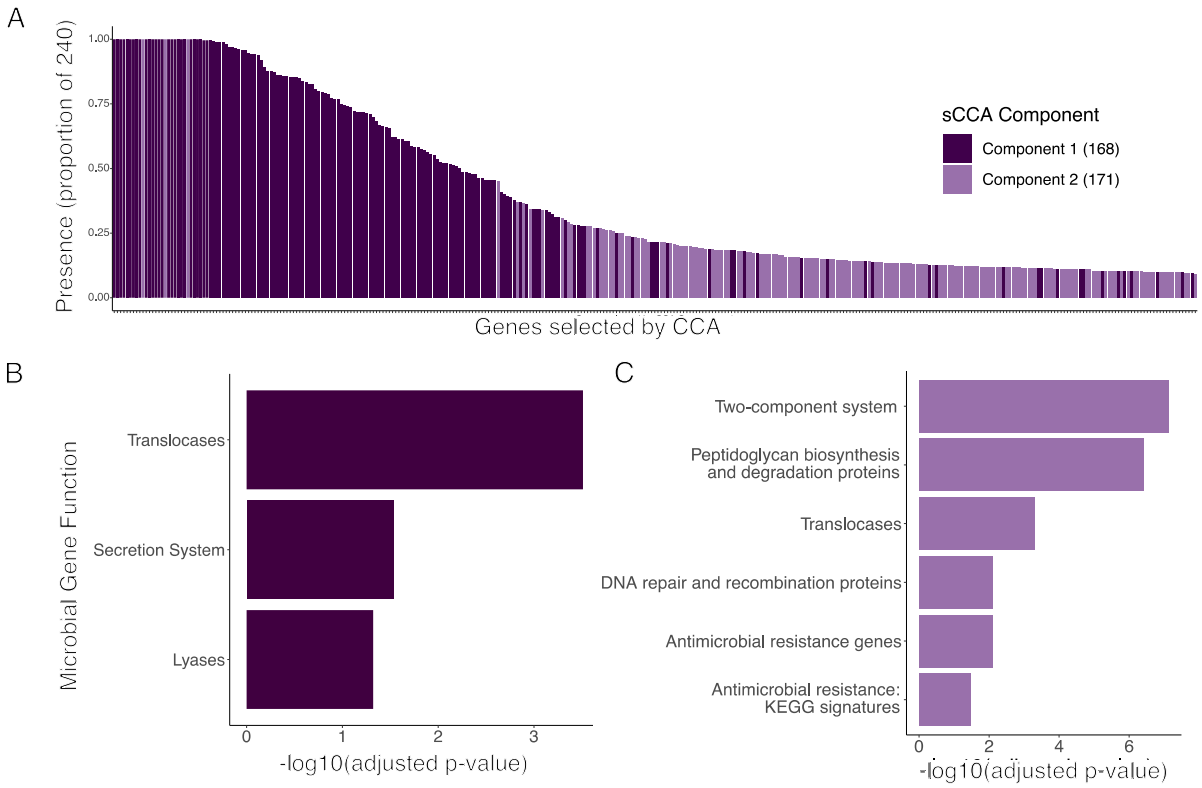
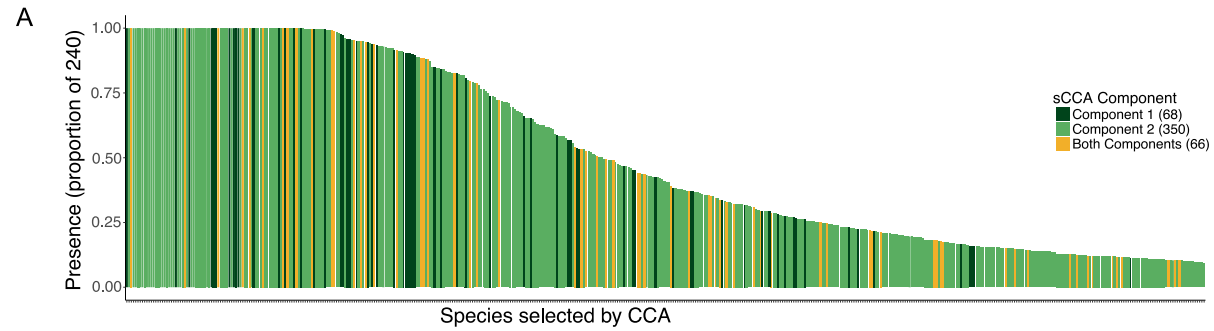


Figure 2.4. Microbial gene results. A) the prevalence of the microbial genes in the TwinsUK cohort colored by CCA component. B-C) the statistically significant functional enrichment of the microbial genes, Fisher's exact test, FDR corrected -$\log_{10}$ p-values.

Microbial species CCA results

When we apply this method to investigate microbial taxa abundances, we find 134 microbial taxa associated in the first component, and 417 microbial taxa associated in the second component from sCCA. Sixty-six of the detected species representatives overlapped between the two components. Most of the species reflect organisms that are well-established with the commensal gut microbiome, although our analysis did detect associations involving potential pathogens such as *Phocaeicola vulgatus* (Figure 2.5A-D). Similar to previous studies, we find that *Faecalibacterium prauznitsii* is associated with host genetics (Wang *et al.* 2016; Bonder *et al.* 2016b), but we also find phage known to infect this species including Taranis, Toutatis, Lugh, Epona, and Oengus are associated with host genetics. We also identify crAssphage, uncultured crAssphage, and crAss001. The phageome is known to be unique to individuals and consistent over time with some evidence for a core gut phageome (Townsend *et al.* 2021). Several other species overlap with previous microbiome GWAS including the genera Streptococcus the genera *Barnesiella, Campylobacteracea, Lactococcus,* and *Akkermansia* (Davenport *et al.* 2015), *Bacteroides xylanisolvens* (Bonder *et al.* 2016b), the genera *Enterobacteriaceae* and *Bacillus* (Wang *et al.* 2016), the genera *Blautia* and *Lachnospira* (Bonder *et al.* 2016a), and the genus *Bifidobacterium* (Goodrich *et*

al. 2014, 2016; Blekhman *et al.* 2015; Wang *et al.* 2016; Bonder *et al.* 2016b; Kurilshikov *et al.* 2021).



B

Both components

Species	Prevalence	Mean Abundance
Bacteroides xylanisolvens	1.0000	2.314×10^5
Roseburia hominis	1.0000	1.024×10^5
Dysosmobacter welbionis	1.0000	9.444×10^4
Intestinimonas butyriciproducens	1.0000	2.972×10^4
Faecalitalea cylindroides	1.0000	2.732×10^4
Alistipes indistinctus	1.0000	1.912×10^4
Oscillibacter sp. NSJ-62	1.0000	1.759×10^4
Ruminococcus champanellensis	1.0000	1.088×10^4
Streptococcus thermophilus	1.0000	5.771×10^3
Butyrivibrio proteoclasticus	0.9958	2.953×10^2

C

Component 1

Species	Prevalence	Mean Abundance
Bifidobacterium adolescentis	1.0000	1.573×10^5
Collinsella aerofaciens	1.0000	1.085×10^5
Coprococcus comes	1.0000	1.021×10^5
Faecalibacillus intestinalis	1.0000	3.459×10^4
Bacteroides sp. HF-5287	1.0000	3.135×10^4
Phocaeicola salanitronis	1.0000	9.092×10^3
Clostridium scindens	1.0000	8.299×10^3
Streptococcus pasteurianus	1.0000	6.162×10^3
Lacrimispora saccharolytica	1.0000	2.224×10^3
Enterococcus faecium	1.0000	2.086×10^3

D

Component 2

Species	Prevalence	Mean Abundance
Bacteroides uniformis	1.0000	1.154×10^6
Phocaeicola vulgatus	1.0000	9.019×10^5
Faecalibacterium prausnitzii	1.0000	4.527×10^5
Bacteroides ovatus	1.0000	3.790×10^5
Bacteroides cellulosilyticus	1.0000	3.181×10^5
Parabacteroides distasonis	1.0000	2.642×10^5
Bacteroides fragilis	1.0000	2.531×10^5
Bacteroides caccae	1.0000	2.496×10^5
Bacteroides thetaiotaomicron	1.0000	2.323×10^5
Anaerostipes hadrus	1.0000	1.636×10^5

Figure 2.5. The prevalence of microbial species selected by CCA as being associated with host genetics. A) Prevalence of the species out of 240 individuals. Data from components 1 and 2 are shown, with 66 species selected by both components. B-D) The top 10 most prevalent species sorted by mean abundance selected by components (B), only component 1 (C), or only component 2 (D).

Discussion

We apply both the Tweedie distribution, for modeling gene and species abundances in metagenomic data, and sparse CCA to the challenge of identifying correlations between overall host genetics and the composition of the gut microbiome or its composite functions. Common species or gene families will tend to have a higher mean across samples, while also having larger variation, and a lower probability of being *0* (which here can mean either missing in the sample or not measured) in any one sample. Conversely, a rare gene function or low-abundance bacterium, may have a low mean, high variance, and a high probability of being *0* in many samples. We show that Tweedie distributions, with a power parameter, p, between 1 and 2 best captures this relationship within the data (Figure 2.2). While the Tweedie distribution is a superior fit for both metagenomic gene family and species abundance data, it has not yet been widely applied to genomics.

Our choice of penalties for sCCA have known features that affect the interpretation of results: 1) group lasso, which we applied to the human variants, enables us to select SNPs as a whole (Yuan and Lin 2006) and 2) elastic net, which we applied to the metagenomic features, will tend to select a larger set of correlated features within a sample if genes or species are correlated with each other (Zou and Hastie 2005). When considering

metagenomic gene family abundances, we can intuitively imagine that once one gene from a pathway is selected, the model will select the whole group of correlated genes from the same pathway. This behavior of elastic net variable selection makes it ideal for analyzing gene abundances. sCCA with elastic net searches for the overall effect of many smaller, combined effects, which while appropriate for genes that can be grouped into functional pathways, may also capture biologically relevant relationships between microbiota and their host. For example, co-occurrence of bacterial species in the gut microbiomes may capture how multiple species can work together to perform a function such as metabolic cross feeding that results in the production of a biologically important metabolite.

While sCCA identifies a small number of known gene-microbiome relationships, it expands the number of broadly associated human SNPs with microbiome composition and function. Our results show that human genes known to be associated with human diseases such as type 2 diabetes and schizophrenia are associated with components of the gut microbiome and are significantly enriched in our data. These results may provide new routes of study for identifying links between the gut microbiome and disease risk. We identify many microbial species that are known components of a healthy gut microbiome and expand upon the list of species that may be heritable.

We also identify abundant crAssphage, which are the most abundant bacteriophage in the human gut, as being associated with human genetics. CrAssphage are a gut-associated bacteriophage which is broadly associated with primates (Siranosian *et al.* 2020). In addition, while most members of the gut virome are unique between twins and family members, crAssphage genomes tend to be nearly identical between mothers and infants, indicating a heritable component (Siranosian *et al.* 2020).

A drawback of our penalized CCA method is that by using regularization methods in the first part of the method, we cannot readily interpret the results as we would for a traditional CCA. For example, the coefficients for the metagenomic species and gene abundances cannot be ranked since the standard errors are unknown and the shrinkage from the penalization may have had a different effect across features. That is, properly controlling for variable selection makes inference on the sCCA coefficients more difficult than it is classically. Additionally, the results have a specific interpretation that the associated SNPs and metagenomic abundances share a common latent factor (Bach and Jordan 2005). In other words, there is no implication of directionality between the SNPs and metagenomic gene abundances.

Given that this method requires paired metagenomic (shotgun

metagenomic sequencing or 16S rRNA sequencing) and host genome data, we are still limited by what is publicly available. Our analyses are performed on a small, healthy cohort of women from the United Kingdom. We do not expect large differences in gut composition/function within a mostly healthy cohort, so effect sizes are small. A larger sample from more diverse geographic and cultural backgrounds would enhance our ability to identify novel associations between the human gut microbiome and human genetics. Despite the usual limitations of a small sample size, we were still able to uncover previously known associations as well as novel ones. It would be interesting to perform our method on a large disease cohort where strong differences between disease and control microbiomes are expected and may provide more context for disease-associated mechanisms. Additionally, there are many known SNPs involved in microbiome-related disorders and larger microbiome differences between disease and control samples would make it more likely to find additional SNPs that correlate with these overall shifts.

Methods

Subject details and data

The study involved metagenomic sequencing from a subset of individuals from the TwinsUK Project at King's College London as reported previously

(Moayyeri *et al.* 2013; Xie *et al.* 2016). Deidentified metagenomic data were processed to remove adapter sequence, low quality, duplicate, and human reads (Rotmistrovsky and Agarwala 2011; Schmieder and Edwards 2011; Bolger *et al.* 2014). High quality reads were assembled into contigs using MetaSpades v3.13.1 (Nurk *et al.* 2017). All work involving human subjects was approved by the Cornell University IRB (Protocol ID 1108002388) and all methods were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants.

Gene Catalog construction

Prodigal v2.6.3 was used to predict proteins on the assembled contigs. Duplicate proteins were removed using vsearch v2.15.0 and sequence redundancy was further removed by CD-hit v4.8.1 (Li and Godzik 2006) (90% identity) leading to a catalog of 12,563,449 nonredundant proteins. DNA sequences for the clustered proteins were used to form the final gene catalog for alignments.

Metagenomic gene family abundance calculation

Metagenomic reads were aligned to the gene catalog using BWA mem v0.7.17 (-a -bwtsw -t 4) (Li 2013). The output was filtered to retain primary

alignments and reads aligning with at least 90% identity. Next, genes were removed if less than 80% of the gene was covered. Abundances were then normalized using a modified RPKM approach according to the gene length and the geometric mean of abundances per sample to account for the compositionality of the data. Genes were annotated using KEGG orthologs (downloaded 2016) and abundances were aggregated by KO into functional gene families leading to a final abundance table with 12,645 microbial gene family abundances.

Genotype Data

Human genotyping data are available for 240 individuals from an Illumina HumanHap300 Bead Chip or Illumina HumanHap610 Quad Chip and imputed using IMPUTE version 2 (Howie *et al.* 2009; Moayyeri *et al.* 2013). One individual from each monozygotic twin pair was genotyped and the data were duplicated for the full pair. SNP data were filtered to remove missing data, rare alleles, and loci violating HWE or in high LD (>80%) with one another (Plink v1.9 --geno 0 ---maf 0.10 –indep-pairwise 50 10 0.8 –hwe 0.001) (Chang *et al.* 2015). Variants are coded for additive and dominance effects such that the additive component is {1, 0, -1} for the number of alleles, and the dominance effect is coded as {-1, 1, -1} such that these

effects are orthogonal to the additive effects.

Microbial taxa abundances

First, metagenomic reads were annotated to the species level using Kraken2 (v2.1.0, --confidence 0.1) (Wood *et al.* 2019). Then species abundances were estimated using Bracken (v2.0, -r 100, -l S) (Lu *et al.* 2017). Species abundances were normalized by the geometric mean per sample and species present in at least 90% of the population were retained. A total of 655 species were analyzed.

Statistical analyses

Prior to association testing, a Tweedie generalized linear mixed model was performed in R using the 'lme4' and 'statmod' packages to extract residuals from the gene, species, and SNP data. The models include fixed terms for the individual's age at sampling, their BMI, the shipment number of their sample, and a random effect for their twin status (either monozygotic or dizygotic) to remove the effect of a person's correlation with their own twin otherwise known as relatedness. The use of the twins' zygosity and genetic differences between them is a topic for another paper. Further, we included ten principal components from the human SNPs in the models to remove

any effects of ancestry from the data, where the principal components were fit after removing all co-twins (i.e., only one individual from each twin pair is used to perform PCA).

Sparse canonical correlation analysis, sCCA

We use a multivariate method to associate human variants and microbial gene abundances called sparse canonical correlation analysis, or sCCA. Like principal components analysis, CCA projects observations into lower dimensional space, but has the added benefit of comparing across tables. CCA aims to identify the linear combination of two tables that maximizes the correlation between the two. Traditional CCA requires low-dimensional data where the sample size is larger than the number of features, which is not the case for modern genomic datasets. In order to apply CCA to high-dimensional data, we apply a penalized variant of CCA, or sCCA, that first performs variable selection to reduce the number of features to be fit. For the genes and species input data, we apply elastic net regularization, which combines the l_1 and l_2 penalties from lasso and ridge regression methods, that is:

$$\underset{\alpha,\beta:\, ||X\beta||_2=||Y\alpha||_2=1}{\operatorname{argmin}} ||X\beta - Y\alpha||_2^2 + \lambda_1 pen_1(\beta) + \lambda_2 pen_2(\alpha)$$

We apply group lasso to the variants in order to select the entire variant,

however variations of this method could use different variant coding or different regularization methods. The first stage of this method is to perform tuning parameter optimization to select the l_1 and l_2 penalty via cross validation. Twelve pairs of penalty parameters are tested via cross validation. Once the penalties are selected, we perform sparse CCA with the gene or species residuals, the variant residuals, and the selected penalties. The output is a list of all inputs (genes or species, and variants) with coefficients that are either zero or nonzero. Features with a nonzero coefficient was selected by sparse CCA as having the maximum correlation between the two input tables. We fit the sCCA through block coordinate descent by iterating regression on X and on Y.

Statistical analyses and association tests were performed in R. Code adapted from glmnet, gglasso, and lme4, to perform penalized CCA with elastic net and group lasso. The new reduced set of gene family abundances and human variants are then fit by CCA for each component requested.

sCCA results analyses

BiomaRt (Durinck *et al.* 2009) was used to annotate the human SNPs using HG19. FUMA GWAS was used to annotate the selected SNPs by their gene-disease associations and functions (GO terms) and performed

enrichment analyses to identify GWAS associations for human SNPs and genes (Watanabe *et al.* 2017).

Using the KEGG API, pathway information was linked to the microbial gene family abundances selected by CCA. Fisher's exact test with false discovery rate correction performed in R for pathway enrichment test.

References

Aitchison J., 1982 The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological) 44: 139–177.

Bach F. R., and M. I. Jordan, 2005 *A Probabilistic Interpretation of Canonical Correlation Analysis*.

Baer B. R., S. Seto, and M. T. Wells, 2018 Exponential Family Word Embeddings: An Iterative Approach for Learning Word Vectors, in *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, Montreal, Canada.

Bastiaanssen T. F. S., C. S. M. Cowan, M. J. Claesson, T. G. Dinan, and J. F. Cryan, 2019 Making Sense of … the Microbiome in Psychiatry. The international journal of neuropsychopharmacology 22: 37–52. https://doi.org/10.1093/ijnp/pyy067

Blekhman R., J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski, *et al.*, 2015 Host genetic variation impacts microbiome composition across human body sites. Genome biology 16: 191. https://doi.org/10.1186/s13059-015-0759-1

Bokulich N. A., J. Chung, T. Battaglia, N. Henderson, M. Jay, *et al.*, 2016 Antibiotics, birth mode, and diet shape microbiome maturation during early life. Science translational medicine 8: 343ra82. https://doi.org/10.1126/SCITRANSLMED.AAD7121

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonder M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, *et al.*, 2016a The effect of host genetics on the gut microbiome. Nature Genetics 48: 1407–1412. https://doi.org/10.1038/ng.3663

Bonder M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, *et al.*,

2016b The effect of host genetics on the gut microbiome. Nature Genetics 48: 1407–1412. https://doi.org/10.1038/ng.3663
Brito I. L., T. Gurry, S. Zhao, K. Huang, S. K. Young, *et al.*, 2019 Transmission of human-associated microbiota along family and social networks. Nature Microbiology 1. https://doi.org/10.1038/s41564-019-0409-6
Calle M. L., 2019 Statistical Analysis of Metagenomics Data. Genomics & informatics 17: e6. https://doi.org/10.5808/GI.2019.17.1.e6
Chang C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4: 7. https://doi.org/10.1186/s13742-015-0047-8
Costea P. I., G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, *et al.*, 2017 Towards standards for human fecal sample processing in metagenomic studies. Nature biotechnology 35: 1069–1076. https://doi.org/10.1038/nbt.3960
Davenport E. R., D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober, *et al.*, 2015 Genome-Wide Association Studies of the Human Gut Microbiota, (B. A. White, Ed.). PLOS ONE 10: e0140301. https://doi.org/10.1371/journal.pone.0140301
Durinck S., P. T. Spellman, E. Birney, and W. Huber, 2009 Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols 4: 1184–1191. https://doi.org/10.1038/nprot.2009.97
Foster S. D., and M. V Bravington, 2013 A Poisson-Gamma model for analysis of ecological non-negative continuous data. Environ Ecol Stat 20: 533–552. https://doi.org/10.1007/s10651-012-0233-0
Francino M. P., 2016 Antibiotics and the Human Gut Microbiome: Dysbioses and Accumulation of Resistances. Frontiers in Microbiology 6: 1543. https://doi.org/10.3389/fmicb.2015.01543
Garud N. R., B. H. Good, O. Hallatschek, and K. S. Pollard, 2019 Evolutionary dynamics of bacteria in the gut microbiome within and across hosts, (I. Gordo, Ed.). PLOS Biology 17: e3000102. https://doi.org/10.1371/journal.pbio.3000102
Gloor G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, 2017 Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology 8: 2224. https://doi.org/10.3389/fmicb.2017.02224
Goodrich J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, *et al.*, 2014 Human genetics shape the gut microbiome. Cell 159: 789. https://doi.org/10.1016/J.CELL.2014.09.053
Goodrich J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, *et al.*, 2016 Genetic determinants of the gut microbiome in UK Twins HHS Public

Access. Cell Host Microbe 19: 731–743. https://doi.org/10.1016/j.chom.2016.04.017
Hildebrand F., T. I. Gossmann, C. Frioux, E. Özkurt, P. N. Myers, *et al.*, Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host & Microbe 0. https://doi.org/10.1016/J.CHOM.2021.05.008
Hotelling H., 1936 Relations Between Two Sets of Variates. Biometrika 28: 321. https://doi.org/10.2307/2333955
Howie B. N., P. Donnelly, and J. Marchini, 2009 A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies, (N. J. Schork, Ed.). PLoS Genetics 5: e1000529. https://doi.org/10.1371/journal.pgen.1000529
Hughes D. A., R. Bacigalupe, J. Wang, M. C. Rühlemann, R. Y. Tito, *et al.*, 2020 Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. Nature microbiology 5: 1079–1087. https://doi.org/10.1038/s41564-020-0743-8
Igartua C., E. R. Davenport, Y. Gilad, D. L. Nicolae, J. Pinto, *et al.*, 2017 Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. Microbiome 5: 16. https://doi.org/10.1186/s40168-016-0227-5
Jørgensen B., 1987 Exponential Dispersion Models. Journal of the Royal Statistical Society. Series B (Methodological) 49: 127–145.
Kendal W. S., and B. Jørgensen, 2011 Tweedie convergence: A mathematical basis for Taylor's power law, 1/f noise, and multifractality. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 84. https://doi.org/10.1103/PhysRevE.84.066120
Kurilshikov A., C. Medina-Gomez, R. Bacigalupe, D. Radjabzadeh, J. Wang, *et al.*, 2021 Large-scale association analyses identify host factors influencing human gut microbiome composition. Nature Genetics 53: 156–165. https://doi.org/10.1038/s41588-020-00763-1
Li W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. https://doi.org/10.1093/bioinformatics/btl158
Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
Love M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15: 550. https://doi.org/10.1186/s13059-014-0550-8
Lu J., F. P. Breitwieser, P. Thielen, and S. L. Salzberg, 2017 Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science 3: e104. https://doi.org/10.7717/peerj-cs.104

McMurdie P. J., and S. Holmes, 2014 Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible, (A. C. McHardy, Ed.). PLoS Computational Biology 10: e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Moayyeri A., C. J. Hammond, A. M. Valdes, and T. D. Spector, 2013 Cohort Profile: TwinsUK and Healthy Ageing Twin Study. International Journal of Epidemiology 42: 76. https://doi.org/10.1093/IJE/DYR207

Mortazavi A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5: 621–8. https://doi.org/10.1038/nmeth.1226

Nurk S., D. Meleshko, A. Korobeynikov, and P. A. Pevzner, 2017 metaSPAdes: a new versatile metagenomic assembler. Genome research 27: 824–834. https://doi.org/10.1101/gr.213959.116

Rodosthenous T., V. Shahrezaei, and M. Evangelou, 2020 Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics 36: 4616. https://doi.org/10.1093/BIOINFORMATICS/BTAA530

Rothschild D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, *et al.*, 2018 Environment dominates over host genetics in shaping human gut microbiota. Nature 555: 210–215. https://doi.org/10.1038/nature25973

Rotmistrovsky K., and R. Agarwala, 2011 *BMTagger : Best Match Tagger for removing human reads from metagenomics datasets BMTagger screening*.

Schmieder R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864. https://doi.org/10.1093/bioinformatics/btr026

Sepich-Poore G. D., L. Zitvogel, R. Straussman, J. Hasty, J. A. Wargo, *et al.*, 2021 The microbiome and human cancer. Science 371.

Shapiro B. J., 2016 How clonal are bacteria over time? Current Opinion in Microbiology 31: 116–123. https://doi.org/10.1016/J.MIB.2016.03.013

Siranosian B. A., F. B. Tamburini, G. Sherlock, and A. S. Bhatt, 2020 Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nature Communications 11: 280. https://doi.org/10.1038/s41467-019-14103-3

Szeligowski T., A. L. Yun, B. R. Lennox, and P. W. J. Burnet, 2020 The Gut Microbiome and Schizophrenia: The Current State of the Field and Clinical Applications. Frontiers in Psychiatry 11: 156. https://doi.org/10.3389/fpsyt.2020.00156

Tam V., N. Patel, M. Turcotte, Y. Bossé, G. Paré, *et al.*, 2019 Benefits and limitations of genome-wide association studies. Nature Reviews Genetics 20: 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tibshirani R., 1996 Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58: 267–288.
Townsend E. M., L. Kelly, G. Muscatt, J. D. Box, N. Hargraves, *et al.*, 2021 The Human Gut Phageome: Origins and Roles in the Human Gut Microbiome. Frontiers in Cellular and Infection Microbiology 11: 498. https://doi.org/10.3389/fcimb.2021.643214
Treangen T. J., and E. P. C. Rocha, 2011 Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genetics 7. https://doi.org/10.1371/journal.pgen.1001284
Turpin W., O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans, *et al.*, 2016 Association of host genome with intestinal microbial composition in a large healthy cohort. Nature genetics 48: 1413–1417. https://doi.org/10.1038/ng.3693
Visconti A., C. I. Le Roy, F. Rosa, N. Rossi, T. C. Martin, *et al.*, 2019 Interplay between the human gut microbiome and host metabolism. Nature Communications 10: 4505. https://doi.org/10.1038/s41467-019-12476-z
Wang J., L. B. Thingholm, J. Skiecevičienė, P. Rausch, M. Kummen, *et al.*, 2016 Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nature genetics 48: 1396–1406. https://doi.org/10.1038/ng.3695
Wang J., A. Kurilshikov, D. Radjabzadeh, W. Turpin, K. Croitoru, *et al.*, 2018 Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. Microbiome 6: 101. https://doi.org/10.1186/s40168-018-0479-3
Warton D. I., and F. K. C. Hui, 2017 The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017). Methods in Ecology and Evolution 8: 1408–1414.
Watanabe K., E. Taskesen, A. van Bochoven, and D. Posthuma, 2017 Functional mapping and annotation of genetic associations with FUMA. Nature Communications 8: 1826. https://doi.org/10.1038/s41467-017-01261-5
Witten D. M., R. Tibshirani, and T. Hastie, 2009 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10: 515–534. https://doi.org/10.1093/biostatistics/kxp008
Witten D. M., and R. J. Tibshirani, 2009 Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8. https://doi.org/10.2202/1544-6115.1470
Wood D. E., J. Lu, and B. Langmead, 2019 Improved metagenomic analysis with Kraken 2. Genome Biology 20: 257. https://doi.org/10.1186/s13059-019-

1891-0
Xie H., R. Guo, H. Zhong, Q. Feng, Z. Lan, *et al.*, 2016 Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Systems 3: 572-584.e3. https://doi.org/10.1016/J.CELS.2016.10.004
Yang J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565–569. https://doi.org/10.1038/ng.608
Yang G., J. Wei, P. Liu, Q. Zhang, Y. Tian, *et al.*, 2021 Role of the gut microbiota in type 2 diabetes and related diseases. Metabolism: clinical and experimental 117: 154712. https://doi.org/10.1016/j.metabol.2021.154712
Yuan M., and Y. Lin, 2006 Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 68: 49–67.
Zhao S., Z. Ye, and R. Stanton, 2020 Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA 26: 903–909. https://doi.org/10.1261/RNA.074922.120
Zou H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

CHAPTER 3: Improving accessibility to latent strain analysis for metagenomics

Introduction

Resolving strain-level resolution from metagenomic shotgun sequencing

The human gut microbiome is an important factor in human health and the immune system. There is a lot of inter-personal diversity of the gut microbiome in terms of function and composition as seen by many population-level analyses, yet we are only beginning to understand the functional consequences of such diversity. Metagenomic analyses at the species level, such as those using 16S rRNA sequencing, allow us to see species compositional changes that may correlate with health or immune status. Metagenomic shotgun sequencing analyses reveal microbial functions, species, and even strain-level associations with human health. Understanding the strain-level diversity between people can unlock much more information than species alone. The Human Gut Microbiome project found that species specific to certain human body sites showed strain-specific clustering and are stable over time (Lloyd-Price *et al.* 2017).

Culture-independent methods for identifying strains from

environmental samples are becoming more available for use with metagenomic shotgun sequencing. Many of these methods rely on sequence variation across genomes or single-copy core genes from metagenomic assemblies (Ferretti *et al.* 2017; Truong *et al.* 2017; Albanese and Donati 2017; Smillie *et al.* 2018). It can be difficult to generate enough data for many species within a single microbiome to accurately and precisely resolve strain-level genomes. Latent Strain Analysis, or LSA, is a pre-assembly method for binning metagenomic reads into bins that represent species or strains, by using pooled variation across metagenomic samples, thus allowing the user to assemble genomes of species and strains from metagenomes (Cleary *et al.* 2015). By pooling together metagenomic reads across many samples, LSA identifies and bins together strings of nucleotides known as k-mers that covary together across samples. It uses a hyperplane hashing function and streaming singular value decomposition in order to find covariance relations between k-mers in a scalable and memory-efficient way. The pre-assembled bins from LSA can be assembled using standard genome assembly software such as Velvet (Zerbino 2010) and annotation methods like AMPHORA2 (Wu and Scott 2012) or BLAST (Altschul *et al.* 1990) to identify bacterial strains from metagenomes.

In this chapter, I will demonstrate the utility and value of LSA using a

real application to metagenomic data to find low-abundance microbial species within the gut microbiome. At the end, I will describe the steps for performing LSA on metagenomic data in the interest of increasing accessibility to the tool.

Identifying gut-associated microbial genes that are associated with anxiety symptoms

Anxiety disorders affect approximately 18% of Americans, and about 1 in 13 people worldwide. While most people suffering from these disorders do not seek or receive treatment, of those who do, only about half of patients respond positively to treatment ("Facts & Statistics | Anxiety and Depression Association of America, ADAA"). Because of the vast burden anxiety places on people worldwide, it is important that we understand the underlying pathophysiology of anxiety in order to develop better treatments.

There is growing evidence for bi-directional communication between the human gut microbiome and the host's central nervous system via the gut-brain axis (Dinan and Cryan 2015, 2017). Bacteria of the gut may influence the central nervous system through several mechanisms. Bacterial species can regulate the production of neurotransmitters, produce or upregulate the production of proteins or metabolites involved in

neurological processes and gut hormones, and modulate the host's immune system via cytokine induction (Dinan and Cryan 2017; Foster *et al.* 2017). Inflammation that starts in the gut could also contribute to inflammation that reaches the brain. A prolonged stress response in the body may lead to increased permeability of the intestines, leading to circulating microbes or microbial metabolites. It is notable that gastrointestinal symptoms are often seen together with anxiety and other psychiatric disorders (Simpson *et al.* 2020).

A systematic review of the available evidence linking the gut microbiome and anxiety revealed that, in general, neither species diversity between people nor species richness within people seem to explain anxiety symptoms, however the abundances of specific bacteria do correlate with anxiety symptoms or appear differentially abundant in people with anxiety compared with healthy controls (Simpson *et al.* 2021). Many studies report that patients with anxiety have higher abundances of certain species known to be associated with gastrointestinal inflammation such as *Enterobacterales*, *Eggerthella*, and *Desuflovibrito*. An interesting finding is that *Bifidobacterium* was observed in higher abundances in patients with anxiety. Some strains within this group are anti-inflammatory and used as probiotics, but there is evidence that other strains could contribute to inflammation (Simpson *et al.*

2021). Therefore, studying strains of gut microbial species could shed more light on the pathophysiology of the microbiome-CNS connection.

Twins offer a unique opportunity to study traits in the context where one twin has a condition and the other does not, also known as discordance. For identical twins, by studying a discordant trait (where one twin sibling is affected and the other is not) you can rule out the effect of genetics on the trait. In other words, since identical twins share genetics, any difference between the two siblings must be due to the environment. In this chapter, I leveraged data from the TwinsUK cohort (Xie *et al.* 2016) in collaboration with the Clark Lab at Cornell to study microbiome features that correlate with anxiety in adult twins. Dr. Xu Wang, a former post-doc in the Clark Lab, performed a differential gene abundance analysis in twins discordant for anxiety and identified genes that were abundant in the twins without anxiety and missing in the twins with anxiety symptoms. The bacterial sources of the genes were not identifiable using standard methods such as BLAST. Here, I used latent strain analysis (LSA) to create metagenomic assembled genomes (MAGs) that I then used to identify the bacterial sources of these genes.

Methods

Subject details and data

This work uses the metagenomic sequencing from a subset of individuals from the TwinsUK Project at King's College London as reported previously (Moayyeri *et al.* 2013; Xie *et al.* 2016). Deidentified metagenomic data were processed to remove adapter sequence, low quality, duplicate, and human reads (Rotmistrovsky and Agarwala 2011; Schmieder and Edwards 2011; Bolger *et al.* 2014). All work involving human subjects was approved by the Cornell University IRB (Protocol ID 1108002388) and all methods were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants.

Anxiety was measured using an anxiety scale questionnaire administered by a healthcare professional. Anxiety scores are calculated from a survey of 40 questions. Anxiety scores, or AS, are grouped into three classes: AS1: low (0-17), AS2: moderate (18-23), AS3: high (> 24). Twin pairs with discordant anxiety phenotypes (AS 1 vs 3, or an AS score difference greater than 10 to increase the sample size) were chosen from the larger cohort resulting in 32 individuals being analyzed. These 32 samples represent 11 monozygotic and 9 dizygotic twin pairs.

Target gene identification

Paired-end reads were mapped to the Integrated Gene Catalog, or IGC (Li *et al.* 2014). Genes significantly enriched in between high and low anxiety individuals were identified using EdgeR (Robinson *et al.* 2010). Differential gene abundance analysis identified 187 highly significant genes (FDR p-value < 0.05) that were enriched in low anxiety individuals. Genes were searched using NCBI BLAST for taxonomic identificatin (Altschul *et al.* 1990; NCBI Resource Coordinators 2016).

Metagenomic processing

Metagenomic reads were first sorted into species- and strain-specific bins using LSA (Cleary *et al.* 2015). This process generated 1,406 metagenomic bins containing raw reads that represent putative species. Each bin was assembled using Velvet (v1.2.10) (Zerbino 2010) and contigs greater than 500 base pairs were retained for further analyses.

Taxonomic annotation of metagenome assembled genomes

To identify the bacterial origin of each MAG, I used a marker gene-based approach to annotate the MAGs taxonomically. AMPHORA2 (Wu and

Scott 2012) contains a set of 31 conserved, single copy marker genes. The first step is to predict whether each of these 31 core genes is present within MAG. For MAGs with low contamination, in other words, ones that represent a single species, the expectation is that the MAG will contain one copy of each of the AMPHORA2 core genes. Once marker genes have been identified for each MAG, the genes are mapped to the BLAST nr database to identify the species. MAGs that contain more than one copy of any marker genes are considered to be more contaminated, and it's important to check the identity of each marker gene copy. MAGs that contain only copy of each gene, or few copies of them, have little contamination and may represent a genome of a single species.

Linking target genes to species using metagenome assembled genomes

To identify which MAGs the anxiety-associated genes could be coming from, I mapped the genes of interest to each MAG using BLAST (v2.3.0) (Altschul *et al.* 1990). Alignments were filtered to keep the longest and highest quality matches.

Results

Gene enrichment in twins discordant for anxiety

Reads were mapped to the IGC and differential abundance tests were performed in EdgeR (Robinson *et al.* 2010) to look for differences between the anxious and non-anxious twins. There were 187 genes that are highly significantly enriched in individuals that report no or low anxiety symptoms (Figure 3.1). All of these genes are missing in the ten twins with anxiety, and six out of nine of the non-anxious twins have these genes in high abundance. The abundances of 175 of these genes are tightly correlated indicating they may share a common taxon. None of these genes had a taxonomic annotation in the IGC database. When input to NCBI BLAST, the results were inconclusive because the genes did not map to any taxa.

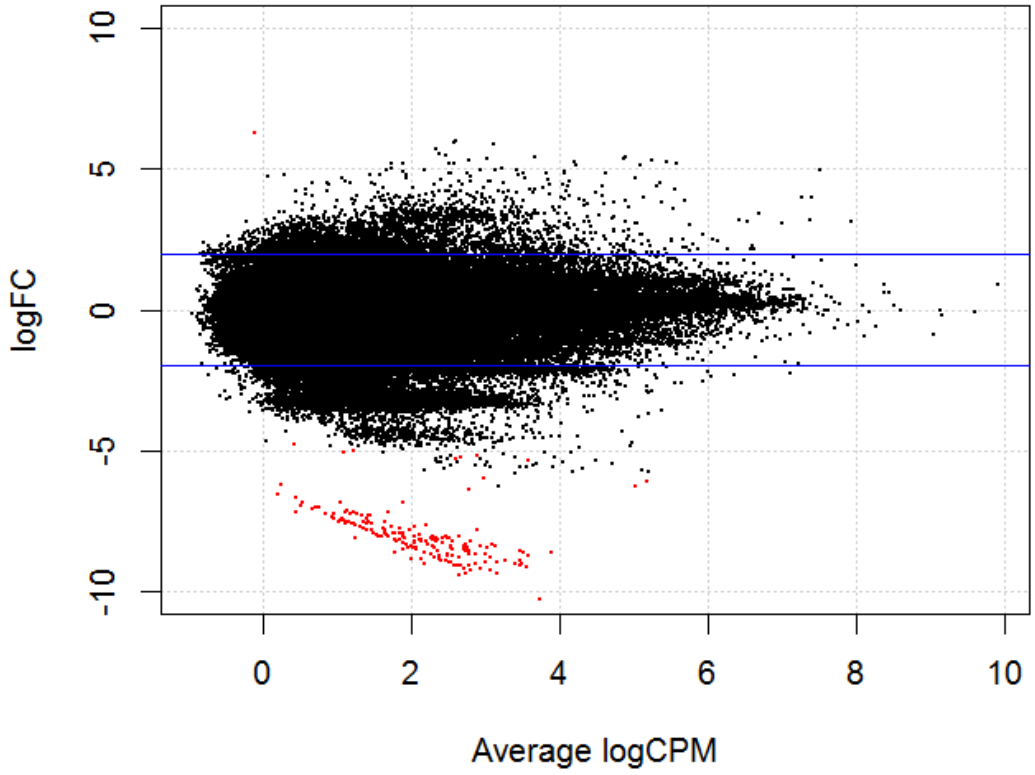


Figure 3.1. Differential (microbial) gene abundance analysis. In total, 160k genes had enough coverage in 9 or more individuals for analysis. Here, 187 genes are highly, significantly enriched with high fold change between anxious and non-anxious twins.

Latent strain analysis

To identify the source of these genes enriched in low anxiety individuals, I needed to first identify the species within the metagenomic samples. I performed pre-assembly binning of the metagenomic reads using LSA in order to create more accurate and more clean metagenome-assembled genomes. Essentially, LSA groups reads into bins according to covariance structures within the data based on k-mer abundance across samples. This process resulted in 1,406 bins. By tuning LSA, it is possible to influence the resolution of binning to create more or fewer bins based on prior expectations. With deep sequencing, it is possible to get many more bins and recover rare species or strains. Each bin was assembled using Velvet and contigs greater than 500 base pairs were retained (v1.2.10) (Zerbino 2010). To identify bins that contained any of the genes found to be enriched in non-anxiety samples, I mapped the enriched genes to all of the MAGs and filtered for the best and longest alignments. The genes enriched in the non-anxious samples map to 26 of the metagenome-assembled genomes. Bin 46 had the most of any MAG with 116/175 of the genes, or about 66%. The next top bin has 64 of the genes, or about 36%.

Azospirillum species are the source of most of the genes enriched in unaffected twins

Next, to identify the taxonomic identity of the MAGs, I used the software AMPHORA2 which uses a marker gene approach. AMPHORA2 looks for a specific set of 31 single-copy, conserved bacterial or archaeal sequences, and outputs a protein and nucleotide sequence file for each marker gene. Once the marker genes were identified for each bin, the genes were mapped to the BLAST nucleotide database to identify their genomic origin. The marker genes from bin 46 all map to the genus *Azospirillum* with a high percent identity and no other matches. *Azospirillum* is a Gram-negative, microaerophilic bacterial genus from the family *Rhodospirillaceae* and is known to promote plant growth. Interestingly, species within this genus are found in some commercialized probiotics for plants.

The next few bins with the most matches to the enriched genes, bins 107, 303, and 505, all had marker genes that map to *Azospirillum*, in addition to other taxa. The marker genes from bin 107 mapped to *Azospirillum sp.*, *Ruminococcus sp.*, and *Clostridium sp*. Marker genes from bin 303 were annotated as *Clostridium sp.* and *Azospirillum sp*. And bin 505 was annotated as *Eubacterium sp.*, *Ruminoccocus sp., Azospirillum sp.*, and *Clostridium sp*. With LSA, it is possible that that some bins are more contaminated than others, and

these results indicate that some reads may be erroneously binning together with different species. However, bin 46, with the highest number of enriched genes of interest, only maps to *Azospirillum* (Figure 3.2A).

These four bins, 46, 107, 303, and 505, that all map to *Azospirillum* at least partially, all contain some of the enriched microbial genes. It is possible that they are all capturing different parts of the same genome, (or genomes of different *Azospirillum* species) that got split into multiple bins. There is some overlap in the enriched genes that map to each of these bins (Figure 3.2). Bin 46 has overlap with each of the four bins, and nine of the genes enriched in unaffected twins are present in bins 46, 107, and 303, but no genes overlap all four bins. Between these four bins, 155/175, or 88.5%, of the enriched genes are represented (Figure 3.2B). Each of the four bins contains genes that bin 46 does not, indicating that these other bins could be capturing more genes of interest and were just not binned with the same *Azospirillum* genome.

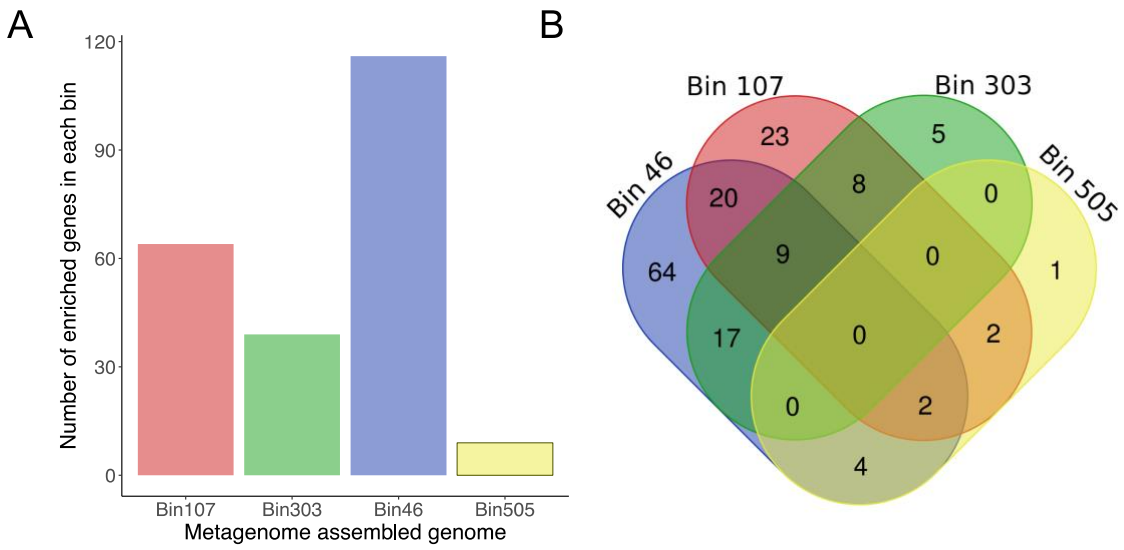


Figure 3.2. Venn diagram of the overlap for the enriched genes across the four bins fully or partially annotated as *Azospirillum*. A) The number of genes enriched in the unaffected twin found within each bin that was annotated as *Azospirillum*, with bin 46 containing the most. B) The overlap of these genes between the four bins. Bin 46 contains 64 genes not seen in any other bin. Bins 46, 107, and 303 overlap by 9 genes. Together, these four bins contain 155 unique genes from the set of enrichment results.

Both of these analyses using MAGs are based on the presence of genes within contigs. Only two contigs for bin 46 had an amphora gene and an anxiety gene. More occurrences of this would be good evidence that indeed *Azospirillum* is the source of these genes, but it does not negate the fact that both the genes of interest and the taxonomic marker genes are found within the same MAG. The species could be rare in the gut, and were only identifiable using meticulous methods for pre-assembly binning. Many contigs have multiple amphora genes on them for partition 46, which gives more weight to the identification of the taxon.

Discussion

Latent strain analysis identified the bacterial taxon that expresses the genes found to be enriched in low anxiety individuals, and also missing in patients with anxiety, in this study of discordant twin pairs. We found that 187 genes were significantly enriched in these unaffected twins, whereas these genes were missing in their sibling who experienced anxiety. Further, 175 of those genes were tightly correlated with one another, indicating a common genomic source. Using standard annotation methods such as BLAST, we were unable to track the source of these enriched genes. However, by

performing pre-assembly binning I was able to fine tune the resolution of metagenomic assembly to uncover many species in fine enough detail to be able to identify their origin.

LSA benefits from having many samples yet also works well with a relatively small sample size (n=32). These samples are sequenced deeply, and LSA can use all the information to identify patterns of covariance in the data to bin reads into MAGs. A larger dataset of anxiety patients and controls may allow us to identify the exact strain of *Azospirillum* contributing these genes to the gut microbiome. Some of the results indicate that other additional bins may include data from *Azospirillum* that binned ambiguously with other genera. By looking at all bins that map to *Azospirillum*, I was able to link 155/175 of the enriched genes to this genus. Future work on refining these bins could generate cleaner and more complete genomes, which may also lead to identifying the species or even strain of *Azospirillum* present in these individuals.

Certain *Azospirillum* species, such as *A. brasilense*, are used in probiotics for plant growth (Menendez and Garcia-Fraile 2017). This may represent the first time *Azospirillum* has been identified as a potential human probiotic species. Since this is a known plant probiotic, it is also possible that this organism is present in these individuals microbiomes because of food they

consume. Without more dietary information, it would be hard to say, but it could be the case that the non-anxious twins ate more plant-based diets than their anxious sisters.

Future work is needed to understand if *Azospirillum* species do act or have the potential to act as probiotic species in the human gut. *In vitro* experimentation could be used to test for certain probiotic properties such as survival of host stress, colonization of a host, antimicrobial properties, or immunomodulation. However, in order to show that a probiotic makes it to the appropriate body site (the gastrointestinal tract, in this case), *in vivo* work is required once strains of interest have been identified using *in vitro* methods (Papadimitriou *et al.* 2015).

In addition to asking whether *Azospirillum* could act as a probiotic in the human gut microbiome, my results indicate that *Azospirillum* could potentially have a protective effect against anxiety-related symptoms and disorders. Mouse models of anxiety could be used to test the effects of fecal microbiota transplants amended with *Azospirillum* species on the host's anxiety related symptoms (Steimer 2011).

Performing Latent Strain Analysis

In this section of the chapter, I will outline the steps to run latent strain analysis on a typical Unix-based server with a sun grid engine scheduler like the ones found in the Brito Lab. I have created an online version of this information at https://fnew.github.io/posts/2019/05/blog_post_lsa/. The source code required extensive modification as many features were hardcoded for specific computing environments.

LSA is run in 14 sequential steps that must be monitored. All scripts are maintained on the in-house server (/workdir/scripts/latent_strain_analysis/). LSA comes prepackaged with a script that will automatically create submission scripts for each step. I have modified this script to work with the Sun Grid Engine (SGE) that is installed in-house. The first steps of the process initialize the environment for running LSA and prepare your metagenomic reads. This means setting up the directory, copying scripts over, and establishing paths. Here, you should run 'python LSFScripts/setupDirs.py -I /input/reads/ -n 9'. This will create several directors for you including one for original reads, scripts, and more. Original, raw metagenomic reads should be in a folder called "/original_reads/". Next, the reads need to be split into even chunks of 1,000,000 fastq records each (or 4,000,000 lines). Once the fastq files are

split into chunks, the next step is to generate the hash function.

The next phase of LSA involves creating the metagenomic kmer abundance matrix which is done with a hash function. A hash function is a function that takes inputs of any size and maps them to a table or matrix with fixed sizing. LSA uses a sequence-sensitive hash function to create the k-mer abundance matrix out of metagenomic sequencing reads. Reads are split into 33 base pair long k-mer (by default) and n hyerplanes are randomly drawn to create 2^n bins, then k-mers that fall within the same k-dimensional space are assigned to a bin and thus a column of the k-mer matrix, where rows represent samples and columns are k-mers. Because the hash function takes sequence into account, nearby columns will be closely related k-mers.

To begin, you can run the script "CreateHash_Job.q" which calls the python script "create_hash.py". Within the python script, there are two important arguments, "-k" and "-s", for the k-mer length and the hash size. The defaults are k-mer length of 33 and hash size of 31.

Next, LSA hashes all the split reads using "HashReads_ArrayJob.q". The job array is set by the number of chunks created in the read splitting step. For example, if I have 100 metagenomic samples that were split into 1,000 chunks, the array number here is '1000'. At this stage, if some jobs fail, the process can continue, but those chunks will be missing downstream.

This is the longest step of LSA and can take up to 48 hours to complete read hashing on ~1,500 chunks. The output will go into a folder called "/hashed_reads/" within the LSA directory.

Once read hashing is complete, the next step is to tabulate the k-mer counts of the reads. LSA counts the hashed k-mers in $1/5^{th}$ of each sample using the script "MergeHash_ArrayJob.q". It is very important that within the "/original_reads/" folder, there is only one ".fastq" file per sample, as this step uses that information to set the job array size. After this point, it is wise to remove the original data files from the "/original_reads/" folder, and only keep the split reads. Note that I have modified the python code at this step to be able to run in-house, and the original LSA code will not work (hash_counting.py and merge_hashq_files.py). The output should include five files per sample. The array job size is the number of samples multiplied by 5. For example, if you have 100 samples, the array size is 500 and the script header will include: "#$ -t 1-500". The next step is to combine the five files for each sample using the script "CombineFractions_ArrayJob.q". All of these jobs must finish successfully. The number of array job tasks here is equal to the number of samples, which is 100 in this example.

Now, using the hashed reads and k-mer counts, the next step is to create the abundance matrix using global k-mer conditioning in the

"GlobalWeights_Job.q" script. This step requires the use of anaconda libraries and can use up to (or more than) 70GB of RAM. This step must succeed to continue. Once this step is complete, rows of the abundance matrix are written to separate files and local sample conditioning is computed using the script "KmerCorpus_ArrayJob.q". If this job failes, you may see a "core" output file. This step requires around 60GB per sample of RAM, but runs fairly quickly. It will create one file per sample that looks like "…/hashed_reads/{FILE_NAME}count.hash.conditioned". All these jobs must finish successfully.

The next phase of LSA is calculating the streaming SVD, clustering the k-mers, and partitioning the metagenomic reads into bins for assembly. The first step in this phase, calculating the streaming SVD, required some modification to run. The manual states that this step will need 60GB of distributed RAM and scales in time with the number of samples: for example, it will run for 7 hours with 64 small samples, or 6 hours for 12 large samples. This step requires the addition of several lines of code to run:

```
$ export
PYRO_SERIALIZERS_ACCEPTED=serpent,json,marshal,pickl
e
$ export PYRO_SERIALIZER=pickle
$ export PATH=/programs/Anaconda2/bin:$PATH
$ export
LD_LIBRARY_PATH=/programs/Anadonda2/lib:$LD_LIBRARY_
PATH
```

Further, this script must be run in a screen in bash: "`bash KmerLSI_Job.q`".

Next, there are four steps to run to cluster the k-mers. First, "KmerClusterIndex_Job.q" creates the clustering index which ultimately determines the resolution of read partitioning. This is the step that often requires the most tuning. After this job is run, there will be a file in the newly created directory called "…/cluster_vectors/numClusters.txt". Checking this file will let you know the number of partitions or bins to expect at the end of LSA. If the number of bins is quite different than your expectations, rerun this script with a new value for the argument "-t". The authors suggest values of 0.5-0.65 for large scale data in the order of terabytes, 0.6-0.8 for medium size data in the order or 100 gigabytes, or greater than 0.75 for small scale data around 10 gigabytes in size. However, these metrics are not clearly distinguishable and do not work well on real data. When running LSA on a medium size dataset, about 500GB, "-t 0.6" results in 19 partitions, whereas "-t 0.8" results in 1,406 partitions. This step will also create the file "/cluster_vectors/cluster_index.py".

Next, LSA will cluster blocks of k-mers using "KmerClusterParts_ArrayJob.q". The job array size here, or number of tasks, is 2 ** hash size / 10^6 + 1. By default, the hash size is set to 31 and

therefore the number of tasks will be 2,148 by default. This step creates numbered directories in the folder "/cluster_vectors/" for the number of partitions that was set in the previous step. During this step it is important to monitor the jobs as they tend to silently fail due to memory constraints. Thoroughly check the logs for any mention of memory issues. If jobs unknowingly fail here, it will not become apparent until the end of LSA, at which point a lot of scripts will need to be rerun. If this happens, the scripts at the end will let you know that the files are corrupt.

The clustered blocks of k-mers will next be merged using "KmerClusterMerge_ArrayJob.q". This step creates files in "/cluster_vectors/" with the extension ".npy" and deletes the directories created in the previous step. The number of array jobs is equal to the number of clusters or partitions, which comes from "/cluster_vectors/numClusters.txt".

The final step before read partitioning is to arrange the k-mer clusters on the disk using "KmerClusterCols_Job.q". This step creates many files in "/cluster_vectors/": cluster_cols.npy, cluster_probs.npy, cluster_vals.npy, kmer_cluster_sizes.npy, and *cluster.npy for each sample.

The final phase of LSA is to partition the original reads according to the k-mer clusters created using the k-mer abundance matrix. First, reads are

partitioned using "ReadPartitions_ArrayJob.q". Create a temporary directory for read partitioning, such as "partitions_tmp" in the LSA folder. This step will create numbered directories for each chunk (from the initial step of splitting your fastq reads) in the temporary folder, and will eventually write files into numbered directories within /cluster_vectors/. The number of tasks here equals the number of chunks created in the beginning when the fastq reads were split, not the number of clusters or partitions from the k-mer clustering stage. Check for any failed jobs and resubmit. Finally, the partitions are merged from the /cluster_vectors/ directory into their final partitions within /read_partitions/ using "MergeIntermediatePartitions_ArrayJob.q". The number of tasks in this step is the number of partitions, or the number of clusters found in "…/cluster_vectors/numClusters.txt". If any of these jobs fail, resubmit them.

This is the end of the LSA pipeline. From here, the typical steps include assembly of the bins (partitions), quality assessment of the bins, and taxonomic annotation using a marker gene approach.

References

Albanese D., and C. Donati, 2017 Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nature Communications 8: 2260. https://doi.org/10.1038/s41467-017-02209-5

Altschul S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. Journal of molecular biology 215: 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cleary B., I. L. Brito, K. Huang, D. Gevers, T. Shea, *et al.*, 2015 Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nature biotechnology 33: 1053–60. https://doi.org/10.1038/nbt.3329

Dinan T. G., and J. F. Cryan, 2015 The impact of gut microbiota on brain and behaviour: implications for psychiatry. Current opinion in clinical nutrition and metabolic care 18: 552–8. https://doi.org/10.1097/MCO.0000000000000221

Dinan T. G., and J. F. Cryan, 2017 Brain-Gut-Microbiota Axis and Mental Health. Psychosomatic medicine 79: 920–926. https://doi.org/10.1097/PSY.0000000000000519

Facts & Statistics | Anxiety and Depression Association of America, ADAA,

Ferretti P., S. Farina, M. Cristofolini, G. Girolomoni, A. Tett, *et al.*, 2017 Experimental metagenomics and ribosomal profiling of the human skin microbiome. Experimental Dermatology 26: 211–219. https://doi.org/10.1111/exd.13210

Foster J. A., L. Rinaman, and J. F. Cryan, 2017 Stress & the gut-brain axis: Regulation by the microbiome. Neurobiology of stress 7: 124–136. https://doi.org/10.1016/j.ynstr.2017.03.001

Li J., H. Jia, X. Cai, H. Zhong, Q. Feng, *et al.*, 2014 An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 2014 32:8 32: 834–841. https://doi.org/10.1038/nbt.2942

Lloyd-Price J., A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, *et al.*, 2017 Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550: 61–66. https://doi.org/10.1038/nature23889

Menendez E., and P. Garcia-Fraile, 2017 Plant probiotic bacteria: solutions to feed the world. AIMS microbiology 3: 502–524. https://doi.org/10.3934/microbiol.2017.3.502

Moayyeri A., C. J. Hammond, A. M. Valdes, and T. D. Spector, 2013 Cohort

Profile: TwinsUK and Healthy Ageing Twin Study. International Journal of Epidemiology 42: 76. https://doi.org/10.1093/IJE/DYR207
NCBI Resource Coordinators N. R., 2016 Database resources of the National Center for Biotechnology Information. Nucleic acids research 44: D7-19. https://doi.org/10.1093/nar/gkv1290
Papadimitriou K., G. Zoumpopoulou, B. Foligné, V. Alexandraki, M. Kazou, *et al.*, 2015 Discovering probiotic microorganisms: in vitro, in vivo, genetic and omics approaches. Frontiers in microbiology 6: 58. https://doi.org/10.3389/fmicb.2015.00058
Robinson M. D., D. J. McCarthy, and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26: 139–40. https://doi.org/10.1093/bioinformatics/btp616
Rotmistrovsky K., and R. Agarwala, 2011 *BMTagger : Best Match Tagger for removing human reads from metagenomics datasets BMTagger screening*.
Schmieder R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864. https://doi.org/10.1093/bioinformatics/btr026
Simpson C. A., O. S. Schwartz, and J. G. Simmons, 2020 The human gut microbiota and depression: widely reviewed, yet poorly understood. Journal of affective disorders 274: 73–75. https://doi.org/10.1016/j.jad.2020.05.115
Simpson C. A., C. Diaz-Arteche, D. Eliby, O. S. Schwartz, J. G. Simmons, *et al.*, 2021 The gut microbiota in anxiety and depression – A systematic review. Clinical Psychology Review 83: 101943. https://doi.org/10.1016/J.CPR.2020.101943
Smillie C. S., J. Sauk, D. Gevers, J. Friedman, J. Sung, *et al.*, 2018 Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell host & microbe 23: 229-240.e5. https://doi.org/10.1016/j.chom.2018.01.003
Steimer T., 2011 Animal models of anxiety disorders in rats and mice: some conceptual issues. Dialogues in clinical neuroscience 13: 495–506. https://doi.org/10.31887/DCNS.2011.13.4/TSTEIMER
Truong D. T., A. Tett, E. Pasolli, C. Huttenhower, and N. Segata, 2017 Microbial strain-level population structure and genetic diversity from metagenomes. Genome research 27: 626–638. https://doi.org/10.1101/gr.216242.116
Wu M., and A. J. Scott, 2012 Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28: 1033–1034. https://doi.org/10.1093/bioinformatics/bts079
Xie H., R. Guo, H. Zhong, Q. Feng, Z. Lan, *et al.*, 2016 Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut

Microbiome. Cell Systems 3: 572-584.e3. https://doi.org/10.1016/J.CELS.2016.10.004
Zerbino D. R., 2010 Using the Velvet de novo assembler for short-read sequencing technologies. Current protocols in bioinformatics Chapter 11: Unit 11.5. https://doi.org/10.1002/0471250953.bi1105s31

CHAPTER 4: The predictive power of the gut microbiome for host health

Introduction

Observational studies of the gut microbiome have uncovered many links to health for the initiation and maintenance of metabolic and gastrointestinal (GI) diseases. For example, differences in community stratification have been observed between lean and obese individuals using taxonomic composition information from 16S rRNA sequencing (Turnbaugh *et al.* 2009). Using shotgun metagenomic sequencing, functional differences in the gut microbiomes of healthy people and those with type 2 diabetes have been observed (Wang *et al.* 2012; Qin *et al.* 2012). For GI-related conditions such as irritable bowel disease, IBD, there is a lot of evidence that the gut microbiome is integral in its onset and progression (Glassner *et al.* 2020). In fact, transferring proinflammatory bacteria from diseased mice to healthy mice induces inflammation in the gut, and specific species of microbes have been found to be involved in driving or suppressing inflammation (Jostins *et al.* 2012; Ananthakrishnan *et al.* 2017; Glassner *et al.* 2020).

Beyond these metabolic disorders, community differences have been observed in the gut microbiota between healthy and affected individuals seen

within many other conditions such as autoimmunity (Xu *et al.* 2019), psychiatric (Simpson *et al.* 2021), and cardiac diseases (Witkowski *et al.* 2020). By studying the associations of the gut microbiome with more diseases in more populations, we can generate more targets for hypothesis testing. Beyond differences at the community level, it is important to identify functional differences between healthy and disease gut microbiomes to get closer to understanding disease mechanisms. A deeper understanding of disease mechanisms as related to the gut microbiome could lead to the development of diagnostics of potential interventions. In order to move from observational studies of community or functional differences to identifying causal links for more conditions, we need to first identify potential targets of study.

Although sequencing costs have come down, large-scale studies of the microbiome are still limited by patient recruitment for many diseases. The Human Microbiome Project (HMP) has over 2,000 samples from stool collected and sequenced (Consortium *et al.* 2012). The second iteration of the HMP, iHMP, includes data relating to type 1 diabetes, pregnancy, and IBD (Lloyd-Price *et al.* 2017). While it can be difficult to recruit large cohorts for every disease that might be important to study, some consortiums are taking a different approach. The TwinsUK Registry is the largest cohort of

adult twins in the United Kingdom (Moayyeri *et al.* 2013). Started in 1992, it has over 14,000 adult twins enrolled. Twin data is useful for distinguishing the relative roles that environment and genetics play in shaping phenotypic variation. Based on the assumption that monozygotic twins share 100% of their genomes, any phenotypic differences between them should be due to environmental factors. These participants are deeply phenotyped and genotyped, with whole genome, transcriptomic, epigenetic, metabolomic, and metagenomic data available (16S rRNA gene sequencing available for most of the participants, and shotgun metagenomic sequencing available for 250 adults at the time of this work) (Moayyeri *et al.* 2013). By using this large cohort of individuals, it is possible to study different subsets according to diseases for specific analysis, as well as studying the gut microbiomes of generally healthy people. In this chapter, I will be using metagenomic and phenotype data for a subset of 250 adult women from the TwinsUK registry. These individuals were sampled randomly from the TwinsUK participants and are representative of the overall cohort. Additionally, through our collaboration with the TwinsUK program, I have access to about 400 phenotypes that span many different biometrics and diseases. The phenotypes that I have access to broadly fit into several categories including, but not limited to, anxiety, immune system, cardiac, lifestyle/physical,

metabolic and autoimmunity. As mentioned, I am especially interested in studying the anxiety, cardiac, and autoimmune traits within this population, as there is less known about these conditions compared with GI-related or metabolic diseases.

Commonly used methods for studying features of the microbiome, such as the abundances of microbial genes, rely on linear models for normally distributed data or nonparametric ranking tests such as Kruskal-Wallis or Spearman's rank correlation tests (Xia and Sun 2017). Methods that assume data are normally distributed are likely to be flawed because metagenomic data is often overdispersed, zero-inflated, and structured with gene-gene interactions which violate the independence assumption required by many models. The Kruskal-Wallis test (Kruskal and Wallis 1952), essentially a non-parametric one-way analysis of variance (ANOVA) are useful when comparing more than two groups (for example when a trait is not binary). Either of these tests are useful for studies comparing taxonomic diversity within or between samples, or across body sites. The Wilcoxon rank-sum test (Wilcoxon 1945), essentially a non-parametric two-sample t-test, can be used for comparing two groups of continuous variables (Xia and Sun 2017). Rank-based analyses should be used with caution because it is

likely the case in metagenomics that these methods could miss weakly associated features that in aggregate account for a lot of variation.

Ensemble machine learning approaches such as random forest (RF), are increasingly applied to genomic and metagenomic data for prediction, classification, and variable selection (Breiman 2001). RF is effective at handling datasets where the number of predictors vastly outnumbers the number of samples (the "large p, small n" problem), and the aggregation of decision trees allows for the handling of correlations and interactions among metagenomic gene abundances. RF was found to perform the best for microarray analysis as well as other high-dimensional data (Breiman 2001; Lee *et al.* 2005). In a review of machine learning approaches applied to the study of the gut microbiome, RF was found to perform the best across five benchmark datasets, while support vector machines performed the poorest (Knights *et al.* 2011). The elastic net method performed worse than RF with higher error rates, but the authors noted that it can be useful as a preprocessing step for other methods such as RF.

In this chapter, I will present my work on associating microbial functions (these will be referred to as gene families throughout this chapter) from the human gut microbiome with various human traits and diseases that

are less often studied in the context of the gut microbiome, in order to identify microbial traits that may be related to host health. To begin, I will use classical, non-parametric statistical methods to find correlations of the microbial gene abundances with host traits. I will then apply random forest for feature selection using backward elimination to identify the microbial genes whose abundances are the most predictive of certain host traits.

Methods

Subject Details and Data

This study used metagenomic sequencing from a subset of individuals from the TwinsUK Project at King's College London as reported previously (Moayyeri *et al.* 2013; Xie *et al.* 2016). Deidentified metagenomic data were processed to remove adapter sequence, low quality, duplicate, and human reads (Rotmistrovsky and Agarwala 2011; Schmieder and Edwards 2011; Bolger *et al.* 2014). All work involving human subjects was approved by the Cornell University IRB (Protocol ID 1108002388) and all methods were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants.

This subset of the TwinsUK metagenomic data consists of 250 individuals representing 90 pairs of dizygotic (fraternal) twins (n=180) and 35 pairs of monozygotic (identical) twins (n=70). Trait data exist for a range

of these individuals, but it is highly dependent on survey completion. For many traits, there is vast missing data, and for others the majority of responses are negative. In total, there are 214 traits and phenotypes with data for at least 10 individuals (130 categorical traits, 84 continuous traits) which fall into broad categories of cardiac, rheumatic or autoimmune, and metabolic traits (Figure 4.1). When filtering for rare genes (genes present in at least 10/250 individuals are retained), I did not take into account the balance of affected and unaffected individuals. Many of these traits are disaggregated survey questions that represent a composite diagnosis. For example, to diagnose anxiety in these individuals, they were given a diagnostic questionnaire with 40 questions. The raw phenotype data has these questions listed individually, but it is also possible to score the results of this questionnaire to calculate an overall anxiety score (see Chapter 3). For some traits, aggregating will be necessary to overcome issues of missing data or class imbalance.

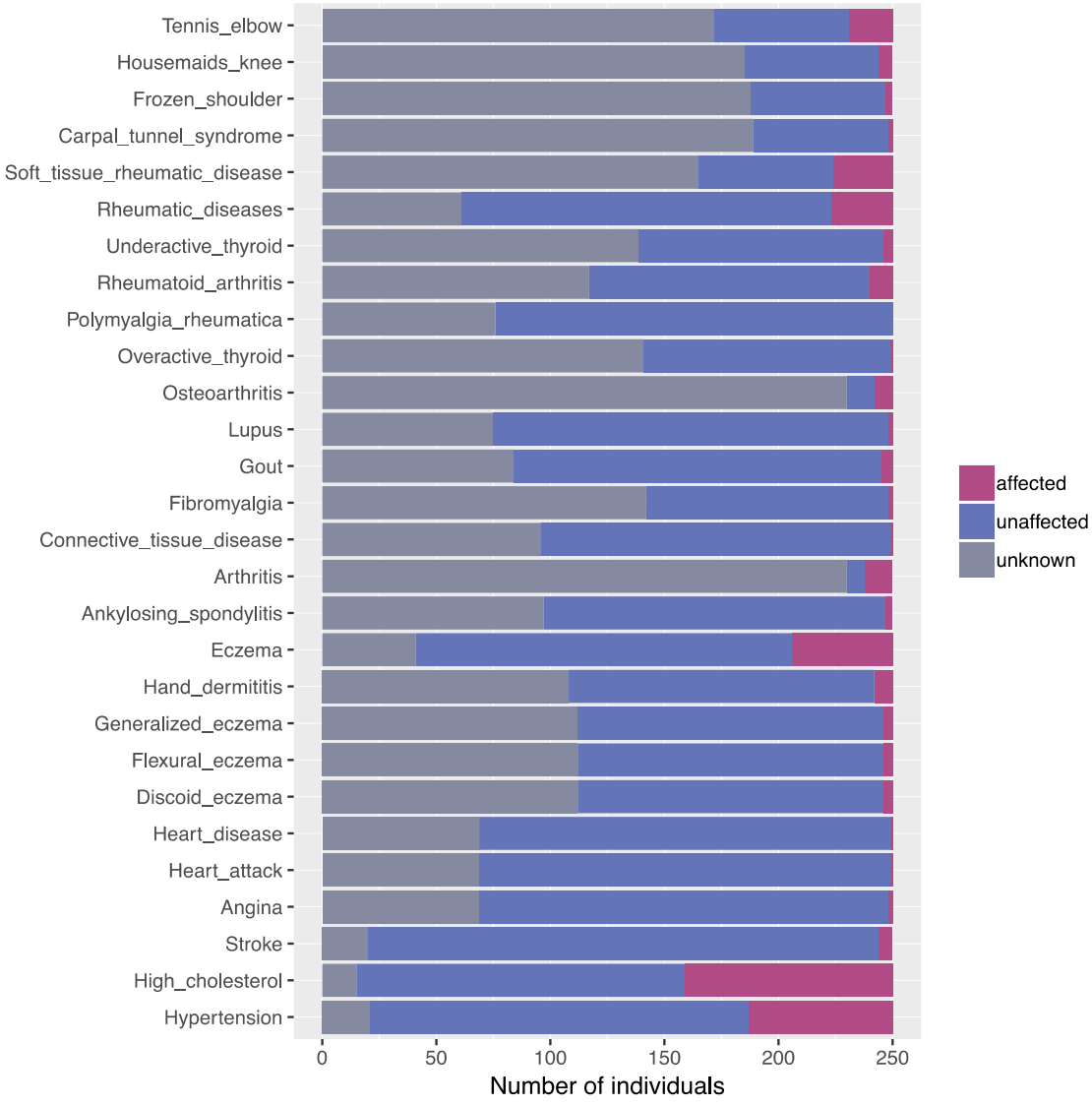


Figure 4.1. A subset of the phenotypes and traits available in the 250 TwinsUK dataset. Most of the individuals in this population are either unaffected or their status is unknown for these phenotypes.

Metagenomic data analysis

High quality metagenomic reads were aligned to the Integrated Gene Catalog, IGC (Li *et al.* 2014), (~9.8 million genes of the human gut microbiome), and only primary alignments were retained. Reads that aligned with less than 90% mapping identity were removed. Genes with at least 80% of the gene body length covered by aligned reads were retained. Gene abundances were normalized for the length of the gene and number of reads sequenced per sample, also known as reads per kilobase per million (RPKM) units (Mortazavi *et al.* 2008). The normalized gene abundances were then aggregated by function using annotations from KEGG (Kanehisa and Goto 2000), eggNOG (Huerta-Cepas *et al.* 2019), and Pfam (Finn *et al.* 2014). These aggregated functional groups will be referred to as gene families. Finally, only gene families that were present in at least 10/250, or 4%, of the individuals were retained in this study. The result is a gene family catalog consisting of 23,176 gene family abundances.

Statistical analyses

Multidimensional scaling analysis (Cox and Cox 2008) was performed on a distance matrix (Brays-Curtis distance (Bray and Curtis 1957)) generated from the microbial gene abundances using the 'vegan' package in R.

Wilcoxon rank sum tests performed on clustered continuous traits in R. Continuous traits were clustered into two groups, for example BMI was clustered into high BMI or low BMI with a threshold set at 30. Spearman's correlation analysis was performed in R. For categorical traits, many of which contain more than 2 categories, I performed Kruskal-Wallis tests in R. A Bonferroni correction for multiple testing was made to all p-values from these analyses.

Random forest for feature selection

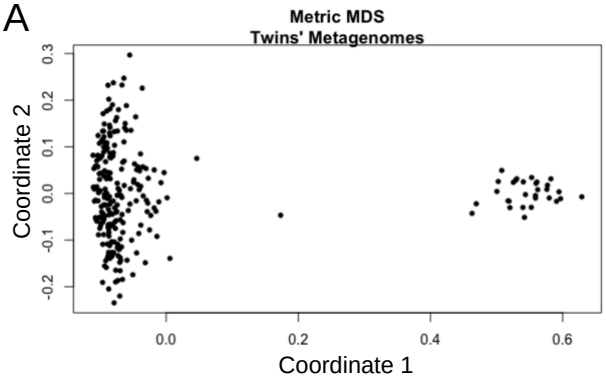
I constructed a random forest classifier to determine the microbial functions that best predict host traits using the 'randomforest' package in R. Feature selection was performed using the 'varSelRF' package in R. I constructed a random forest of decision trees, which ranked gene families by variable importance and iterated through 701 times for a total of 701 random forests. At each iteration, the bottom 10% of least informative gene abundances are dropped. The set of genes that minimizes the out-of-bag error is selected as being the most important for determining host status.

Results

Technical and other confounding variables in the data

To identify confounding variables in the data, I performed a nonparametric multidimensional scaling analysis (Cox and Cox 2008), or MDS, on a Bray-Curtis dissimilarity (Bray and Curtis 1957) matrix calculated from the metagenomic profiles of each person. There is a lot of variation along the first MDS coordinate, creating essentially two groups of data points (Figure 4.2A, C). To understand what could be causing this structure in the data, I performed a series of Chi-squared tests of all the known metadata and phenotype data for the dataset (including health/disease information, BMI, age, shipping information, and more) (Figure 4.2B). Several phenotypes were found to be significant after a Bonferroni correction (p-value < 0.05) including disaggregated anxiety symptoms, birth modality, BMI, and some technical variables like when samples were collected, shipped, or received. The most significant variable was the sample shipment number ($p = 5.364 \times 10^{-8}$) (Figure 4.2B). One important thing to note is that samples for each twin pair always shipped together (Figure 4.2C).

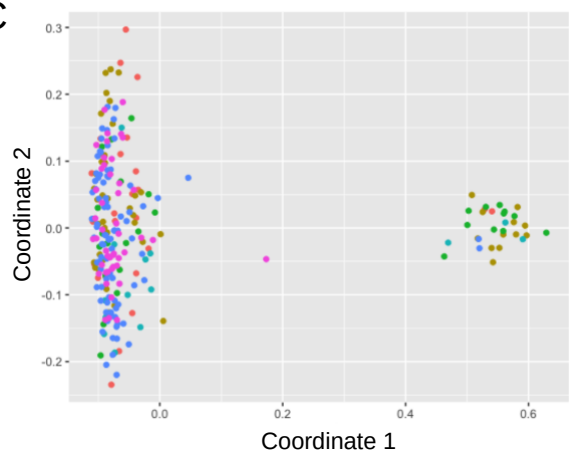
A



B

Categorical variable	p-value
Anxiety phenotype1	0.03007
Anxiety phenotype2	0.01371
Anxiety phenotype3	0.03976
Birth modality	0.01396
Anxiety phenotype4	0.00709
Anxiety phenotype5	0.04249
Anxiety phenotype6	0.03939
Chronic skin rash	0.03894
Gout	0.01386
Continuous variable	**p-value**
BMI	0.03381
CD8+ VB2 count	0.00705
IL6 count	0.02966
Technical variable	**p-value**
Sample shipment number	5.3641e-8
Sample date received	0.04265
Sample collection date	0.02134

C



Shipment number	Number of individuals*
1	24
2	50
3	28
4	14
5	88
6	46

*Twins always shipped with their sister

Figure 4.2. Dimensionality reduction of the twins' metagenomic functional data reveals structure from technical variables. A) A metric multidimensional scaling (MDS) plot showing the first two coordinates from the MDS analysis of the metagenomic functional abundances. The first MDS coordinate shows the highest level of variation across the samples. B) Significant ($p < 0.05$) results of a series of Chi-square tests to look for traits or variables that might explain the separation of data along the first MDS coordinate in A. Sample shipment number has the smallest p-value indicating a relationship with the structure seen in the MDS plot. C) The same MDS plot now colored by sample shipment number. The figure legend for C also includes a count of the number of individuals in each shipment. The twin pairs were always shipped together.

Many microbial functions correlate with many host traits and diseases

I next wanted to test whether any microbial gene family abundances correlate with any of the host phenotypes or diseases. A Wilcoxon rank sum test was performed on the continuous traits clustered into two categories (k=2) as this worked for every trait in my preliminary pass. A Kruskal-Wallis test was performed on the categorical traits for comparisons between more than two categories. Following these tests, a Bonferroni correction was applied to account for multiple testing, which resulted in 92 traits with significant results (p-value < 0.05).

An initial round of testing using the full dataset (no filtering on the presence of gene families across samples) resulted in many significant results for genes that were only present in a handful of subjects. Using the filtered dataset (each gene family must be present in at least ten individuals) results in far fewer significant results than using the entire dataset. These results indicate a bias for sample size for each gene family when performing these analyses and justify potentially stricter filtering.

The traits with the highest number of significant correlations are anxiety-related symptoms (disaggregated questions from an anxiety diagnosis questionnaire), arterial thickness, asthma, congenital heart disease, the use of

medicine for treating diabetes, and various rheumatic diseases (Figure 4.3). Age and BMI are also significantly correlated with several microbial gene abundances (Figure 4.3).

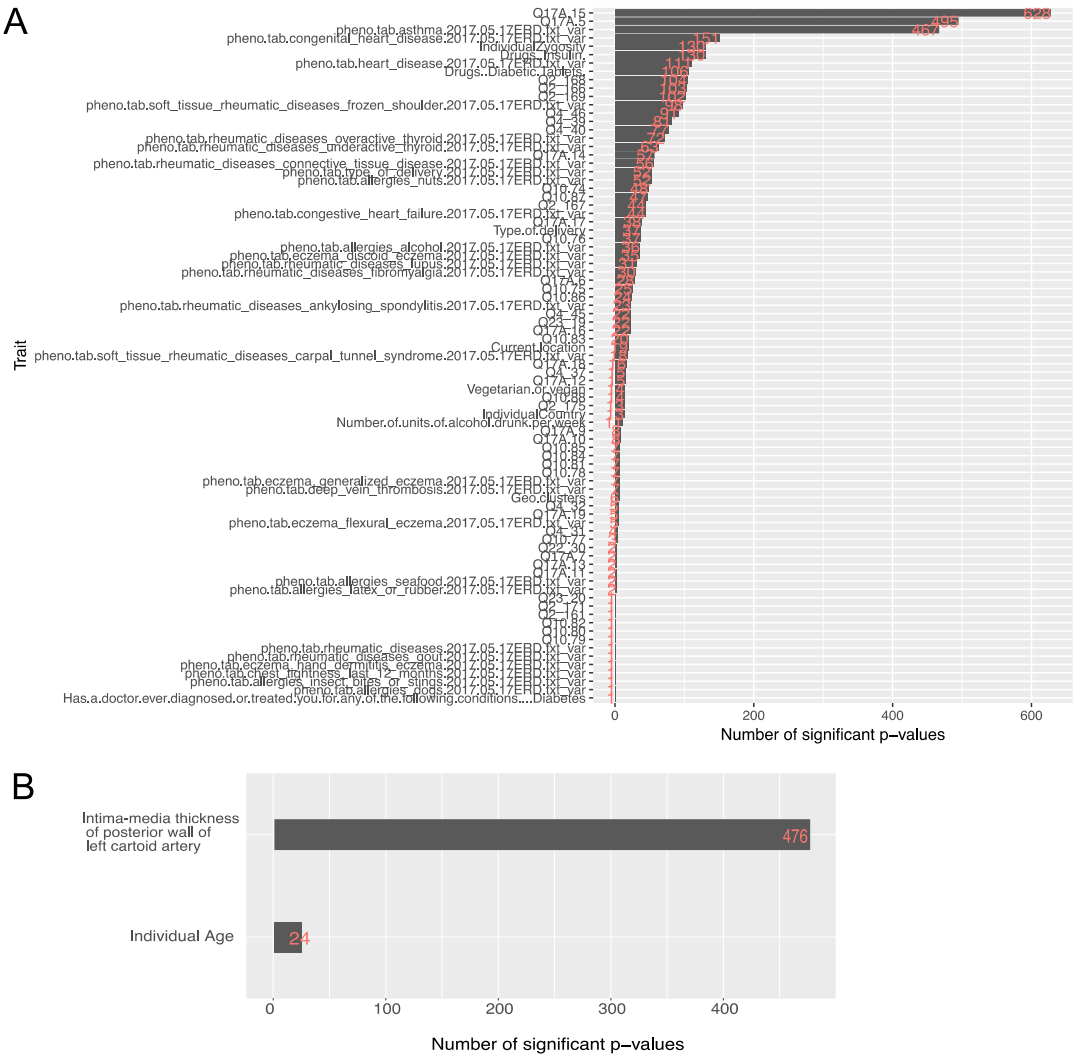


Figure 4.3. Number of microbial gene families significantly associated with various human diseases and traits. A) For these categorical traits, a Kruskal-Wallis test was performed for each microbial gene family to look for associations with each human disease or trait. Of these traits tested, 90 have at least one significant association with microbial gene family abundances of the human gut microbiome (Bonferroni corrected p-value < 0.05). B) Continuous traits were tested against the microbial gene family abundances using a Wilcoxon rank-sum test. After a Bonferroni correction, only two traits were significantly correlated with any abundances.

Many traits fall within certain health-related categories such as rheumatic diseases, allergies, anxiety, cardiac conditions, diabetes, and personal lifestyle. Rheumatic or autoimmune diseases represented here include lupus and fibromyalgia, which share 38% of their significantly associated microbial gene families. The functions of these gene families (from KEGG, eggNOG, or pfam) range from transport systems, membrane fusion proteins, bacterial motility proteins, ion channels, chaperones and folding catalysts, membrane trafficking proteins, and lipopolysaccharide and peptidoglycan biosynthesis proteins. High cholesterol is another trait that has high coverage for individuals in this dataset. There are 86 microbial functional groups significantly associated with high cholesterol. These functions include autophagy, transferases, metabolism, and secretion system proteins. As these gene families were aggregated by functional group prior to analysis, it is not possible to look at the individual microbial genes associated with these diseases.

Microbial functions can be used to distinguish disease status in adult microbiomes

Given that many microbial gene families were indeed significantly correlated with host traits or diseases, I next tested if any of these microbial

gene families could be used to predict those host traits. The results from the non-parametric statistical tests made it clear that only certain phenotypes would be appropriate for studying within this population. Specifically, only phenotypes with high coverage across individuals in this cohort. With this in mind, I chose to analyze five specific traits: soft tissue rheumatic disease, rheumatic disease (other than soft tissue), hypertension, high cholesterol, and eczema (Figures 4.1 and 4.4).

To test my hypothesis, I constructed a random forest of binary decision trees which ranks features (the microbial gene family abundances) by their variable importance in terms of splitting the data between case and control within the forest for each trait. At each iteration, the bottom 10% of variables were dropped, this is known as backward elimination variable selection. I repeated this step for a total of 701 random forests constructed. The most important variables for distinguishing case from control data are those present when the out-of-bag error is at a minimum (Figure 4.4A).

Similar to the rank sum tests, the results of the random forest classifier for variable selection are sensitive to sample size and the number of cases and controls. Except for soft tissue rheumatic disease, there is a linear relationship between the number of cases for a trait and the number of observed associations with metagenomic gene family abundances (Figure

4.4B). Therefore, I may be missing interesting associations for the more unbalanced traits.

The microbial gene families associated with these five traits fit into eight KEGG pathways: metabolism, microbial metabolism in diverse environments, secondary metabolite synthesis, antibiotic synthesis, carbon metabolism, amino acid synthesis, ABC transporters, and quorum sensing (Figure 4.4C). The fact that all five of these traits share the same KEGG pathways could be indicative of biases in the data or within the KEGG database itself as some of these, such as metabolism, are very broad categories. However, I do not see a lot of overlap of the individual gene families between the traits. Of the 2,189 gene families that were found to be predictive for these five traits, only 149 of them were found within two traits (6.8%), just 2 gene families were found in 3 of the traits (0.09%), and no genes were found within all five traits.

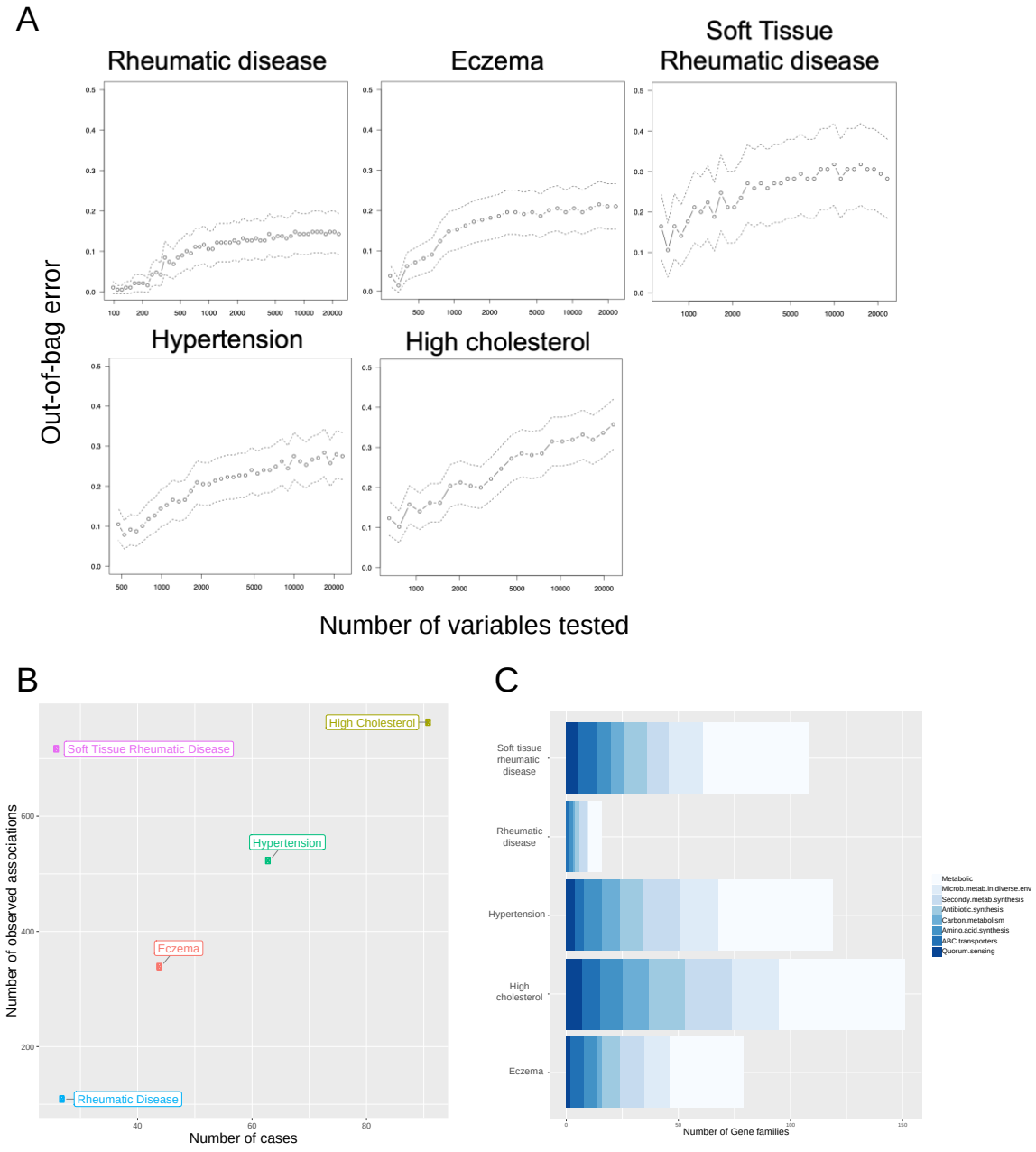


Figure 4.4. Random forest for variable selection and classification. Five traits had sufficient coverage in the dataset to be analyzed using random forest for variable selection: eczema, rheumatic disease, soft tissue rheumatic disease, hypertension, and high cholesterol. A) The out-of-bag error at each iteration of random forest for backward elimination variable selection. The optimal set of variables for distinguishing cases and controls minimizes the out-of-bag error. B) The number of microbial genes used to predict each trait plotted against the number of cases of each trait in the dataset. There is an

almost linear relationship between the number of cases and the number of positive associations. C) The number of gene families found to be associated with each trait, separated by functional category.

Discussion

Using a subset of the TwinsUK data, and a combination of nonparametric statistical and machine learning methods, I was able to identify interesting links between the functions of the gut microbiome and host traits or diseases. To begin, I identified confounding variables within the data using an MDS analysis and enrichment tests. Sample shipment number was most significantly correlated with the variation seen in the MDS plot, but other traits such as various anxiety symptoms, BMI, and birthing modality were also significantly correlated. These results are useful for further analyses of the TwinsUK dataset, as it will be important to control for these variables.

Using nonparametric statistics, I was able to identify many microbial functions that are significantly associated with host traits and diseases. While I was able to identify many significantly associated microbial gene families, this method only allows for univariate testing, and I might be missing out on aggregate pathway-level information. Another complication is the effect of sample sizing on these methods. When I analyzed the full, unfiltered dataset, I found many more significant results, however these results were often found for very rare genes which were present in fewer than 10 individuals.

When I filtered out these rare genes, the number of significant results decreased, but there were still 92 traits and diseases with significant associations after a Bonferroni correction.

These microbial genes that I analyzed were aggregated into functional categories, or gene families, so it is not possible to know the exact microbial genes that may be associated with these host traits. This may limit their usefulness for identifying biomarkers or potential therapeutic targets. However, it is possible to look at pathways from KEGG and eggNOG, or protein families from pfam. Two traits in particular are associated with interesting microbial pathways: autoimmune diseases and high cholesterol. Lupus and fibromyalgia shared 38% of their significantly associated microbial functions. These pathways and proteins include transport proteins, membrane trafficking proteins, ion channels, and chaperones. High cholesterol was associated with bacterial proteins and pathways such as autophagy, transferases, and secretion system proteins.

Given that certain microbial functions are associated with host traits or diseases, I wanted to go a step further and ask if these, or any other, microbial functions could be used to predict a host trait or disease. Features that can be used to consistently and accurately predict a trait could be powerful biomarkers and targets for further study. To test this theory, I

applied random forest for feature selection using backward elimination to identify microbial gene abundances that are predictive for host status for five diseases: eczema, high cholesterol, hypertension, rheumatic disease, and soft tissue rheumatic disease. I chose these traits on the basis of their coverage in the dataset. Each of these five host traits correlated with microbial genes from similar functional categories (KEGG pathways): metabolism, microbial metabolism in diverse environments, secondary metabolite synthesis, antibiotic synthesis, carbon metabolism, amino acid synthesis, ACB transporters, and quorum sensing. Despite this overlap in KEGG pathways, the individual gene families found to be associated with each trait themselves have little overlap. Because the data were aggregated into functional gene families prior to analysis, it is not possible to study the individual genes that may be predictive of these host traits.

This method worked well for these traits as they contained relatively balanced data compared with the other traits in the dataset, in terms of cases and controls, which is a concern when running random forest. Methods exist for handling unbalanced data such as sampling schemes, class weighting or imputation for missing data (Chen *et al.* 1999; Xu-Ying Liu *et al.* 2009; Hong and Lynn 2020). Missing data imputation can be done with random forest (for continuous or categorical data), but caution should be taken especially

when data are highly skewed or there are interactions between variables (Hong and Lynn 2020).

Random forest with class weighting can be used to mitigate the effects of unbalanced data (Chen *et al.* 1999). Essentially, during a random forest analysis, as decision trees are constructed with a subset of the available features (microbial gene families in this case), it is possible to change the weight for each class when calculating the impurity score at a split point in the decision tree. The impurity score is a measure of how heterogeneous the groups are at each split in the training dataset. In other words, was the feature used to make a split good or bad at distinguishing the cases from the controls. Methods for weighted random forest for unbalanced data basically add a bias so that mixed groups that skew towards the minority class are favored, which leads to false positives for the majority class (Hong and Lynn 2020).

Another way to handle unbalanced data with random forest is to explicitly resample from the training data to create bootstrap samples with more balanced data distributions. This method, known as balanced random forest, under-samples the majority class and oversamples the minority class (Chen *et al.* 1999; Xu-Ying Liu *et al.* 2009). When data are extremely unbalanced, for example, case=5, control=100, the bootstrap samples end

up having the same few case samples over and over again, creating dependence between the decision trees. Overall, each method should be used with caution, but could be beneficial for future work on this dataset.

Another alternative approach for this work would be to use a variable selection method such as elastic net (Zou and Hastie 2005) or lasso (Tibshirani 1996) as a preprocessing step on the microbiome gene families prior to performing random forest. This combination has been shown to improve model accuracy (Knights *et al.* 2011).

Data availability and cleaning were complications for this project. Very few traits had wide or balanced coverage across the subjects and many of the traits were duplicated with conflicting responses. With a larger subset of the TwinsUK project, there might have been more coverage of traits and diseases, but it is not likely given that these individuals were not recruited for any specific conditions. I wanted to include an analysis of traits that distinguished discordant twins within the traits or diseases, but I lacked the data to perform statistically sound analyses. For example, within all rheumatic diseases (aggregated across diseases in this category), there are twelve pairs of twins discordant for a rheumatic disease, with just four monozygotic twin pairs. All traits that had any discordant pairs (eight traits), were similarly limited and unbalanced between zygosities (Table 4.1). It is

possible to aggregate traits or find other ways to score phenotypes such that we could create subsets of the data for analysis. For example, in Chapter 3, I presented an analysis of anxiety, where we were able to create a subset of the TwinsUK data for twins that were either discordant for anxiety or had very different anxiety scores (greater than a 10 point difference on the anxiety disorder diagnostic tool).

Table 4.1. A subset of the TwinsUK phenotypes that are discordant within twin pairs and the number of those pairs that are monozygotic. Few traits with discordant traits are represented by many identical twins.

Trait	Number of discordant twin pairs	Number of monozygotic pairs
High cholesterol	33	9
Hypertension	23	5
Stroke	5	1
High blood pressure	15	1
High blood pressure (in pregnancy)	22	4
Eczema	22	5
Rheumatoid arthritis	7	4
Rheumatic diseases	12	4

Future work, especially with an extended set of TwinsUK data, should identify or develop methods for incorporating the familial structure into the analysis. This could be done with a mixed effects random forest that incorporates a random effect for the twins' families and a fixed effect for

zygosity. Twin data allows us to distinguish genetic from environmental effects and could be useful in this situation when looking for microbial genes or functions that may play an environmental role in host disease. This may be complicated by the fact that some microbial species and even their functions are associated with host genetics, so separating the genetic effect from their environmental effect could be complicated.

References

Ananthakrishnan A. N., C. Luo, V. Yajnik, H. Khalili, J. J. Garber, *et al.*, 2017 Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases. Cell Host & Microbe 21: 603-610.e3. https://doi.org/10.1016/J.CHOM.2017.04.010

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bray J. R., and J. T. Curtis, 1957 An ordination of upland forest communities of southern Wisconsin. ECOLOGICAL MONOGRAPHS 27: 325–349.

Breiman L., 2001 Random Forests. Machine Learning 45: 5–32. https://doi.org/10.1023/A:1010933404324

Chen C., A. Liaw, and L. Breiman, 1999 *Using Random Forest to Learn Imbalanced Data*.

Consortium T. H. M. P., C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, *et al.*, 2012 Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214. https://doi.org/10.1038/nature11234

Cox M. A. A., and T. F. Cox, 2008 Multidimensional Scaling, pp. 315–347 in *Handbook of Data Visualization*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Glassner K. L., B. P. Abraham, and E. M. M. Quigley, 2020 The microbiome and inflammatory bowel disease. The Journal of allergy and clinical immunology 145: 16–27. https://doi.org/10.1016/j.jaci.2019.11.003

Hong S., and H. S. Lynn, 2020 Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction.

BMC Medical Research Methodology 20: 199. https://doi.org/10.1186/s12874-020-01080-1

Huerta-Cepas J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, *et al.*, 2019 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47: D309–D314. https://doi.org/10.1093/nar/gky1085

Jostins L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, *et al.*, 2012 Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119–124. https://doi.org/10.1038/nature11582

Kanehisa M., and S. Goto, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30. https://doi.org/10.1093/nar/28.1.27

Knights D., E. K. Costello, and R. Knight, 2011 Supervised classification of human microbiota. FEMS microbiology reviews 35: 343–59. https://doi.org/10.1111/j.1574-6976.2010.00251.x

Kruskal W. H., and W. A. Wallis, 1952 Use of ranks in one-criterion variance analysis. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 47: 583–621.

Lee J. W., J. B. Lee, M. Park, and S. H. Song, 2005 An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48: 869–885. https://doi.org/10.1016/J.CSDA.2004.03.017

Li J., H. Jia, X. Cai, H. Zhong, Q. Feng, *et al.*, 2014 An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology 2014 32:8 32: 834–841. https://doi.org/10.1038/nbt.2942

Lloyd-Price J., A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, *et al.*, 2017 Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550: 61–66. https://doi.org/10.1038/nature23889

Moayyeri A., C. J. Hammond, A. M. Valdes, and T. D. Spector, 2013 Cohort Profile: TwinsUK and Healthy Ageing Twin Study. International Journal of Epidemiology 42: 76. https://doi.org/10.1093/IJE/DYR207

Mortazavi A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5: 621–628. https://doi.org/10.1038/nmeth.1226

Qin J., Y. Li, Z. Cai, S. Li, J. Zhu, *et al.*, 2012 A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490: 55–60. https://doi.org/10.1038/nature11450

Rotmistrovsky K., and R. Agarwala, 2011 *BMTagger : Best Match Tagger for removing*

human reads from metagenomics datasets BMTagger screening.
Schmieder R., and R. Edwards, 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864. https://doi.org/10.1093/bioinformatics/btr026
Simpson C. A., C. Diaz-Arteche, D. Eliby, O. S. Schwartz, J. G. Simmons, *et al.*, 2021 The gut microbiota in anxiety and depression – A systematic review. Clinical Psychology Review 83: 101943. https://doi.org/10.1016/J.CPR.2020.101943
Tibshirani R., 1996 Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58: 267–288.
Turnbaugh P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, *et al.*, 2009 A core gut microbiome in obese and lean twins. Nature 457: 480–484. https://doi.org/10.1038/nature07540
Wang J., J. Qin, Y. Li, Z. Cai, S. Li, *et al.*, 2012 A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490: 55–60. https://doi.org/10.1038/nature11450
Wilcoxon F., 1945 Individual comparisons by ranking methods. Biom. Bull. 1: 80–83.
Witkowski M., T. L. Weeks, and S. L. Hazen, 2020 Gut Microbiota and Cardiovascular Disease. Circulation Research 127: 553–570. https://doi.org/10.1161/CIRCRESAHA.120.316242
Xia Y., and J. Sun, 2017 Hypothesis Testing and Statistical Analysis of Microbiome. Genes & diseases 4: 138–148. https://doi.org/10.1016/j.gendis.2017.06.001
Xie H., R. Guo, H. Zhong, Q. Feng, Z. Lan, *et al.*, 2016 Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Systems 3: 572-584.e3. https://doi.org/10.1016/J.CELS.2016.10.004
Xu H., M. Liu, J. Cao, X. Li, D. Fan, *et al.*, 2019 The Dynamic Interplay between the Gut Microbiota and Autoimmune Diseases. Journal of immunology research 2019: 7546047. https://doi.org/10.1155/2019/7546047
Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, 2009 Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39: 539–550. https://doi.org/10.1109/TSMCB.2008.2007853
Zou H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

CHAPTER 5: Discussion and Conclusions

In this thesis, I have presented my work on statistical and computational analyses of metagenomic data for the study of the human gut microbiome. Shotgun metagenomic sequencing allows us to analyze environmental communities such as the gut microbiome in a wide variety of ways. In Chapter 1, I presented a wide-ranging review of the current literature and future directions for the field of metagenomics (New and Brito 2020). We highlighted the benefits of using metagenomic sequencing such as greater resolution of species, strains, and functional content of environmental samples such as the gut microbiome. With advances in experimental and computational methodologies, it will be possible to glean even more information from metagenomes in the future. In Chapters 2-4 I presented my work analyzing the metagenomes of participants from the TwinsUK project. The TwinsUK project represents a large collaboration with Drs. Andrew Clark, Ruth Ley, Timothy Spector, Ilana Brito, and their respective teams. This collaboration allowed me to have access to genotype and phenotype information for the individuals. When I began my work within this collaboration, I initially pursued the question of whether features of the gut microbiome, such as genes or pathways of the microbial genomes, can

be used to predict host diseases or phenotypes. I eventually focused on one phenotype in particular, anxiety, to understand the underlying associations of this psychiatric disease with the functions and species of the gut microbiome. I then switched gears to studying the link between host genetics and the gut microbiome's functions. Throughout my work, I have focused on selecting the most appropriate statistical and computational methods.

When I began working with the TwinsUK data, I had access to the metagenomes and a list of phenotypes for 250 adult female twins. Dr. Emily Davenport from the Clark Lab was instrumental in acquiring and cleaning the phenotype data. Even still, additional data cleaning was the first step for my work with this population cohort. I assessed the coverage of traits, the amount of missing data, and the overall scale of the phenotypes available to me from this cohort. I wanted to understand the data and so to start, I looked for confounding variables within the dataset. This preliminary work with the dataset proved useful for my later work when I would need to control for these covariates in my analyses. The main variables that I control for in this dataset are sample shipment number, BMI, and age.

With access to metagenomic and phenotype data, my first goal was to identify any microbial genes or functions that associated with host traits. Although I had access to over 400 individual phenotypes for these twins, the

majority of traits would not be usable because of missing data or extreme imbalance in the cases and controls. I removed all traits that were missing in 240/250 individuals, which resulted in 214 traits. These traits roughly fell into the categories of cardiac conditions or diseases, rheumatic or autoimmune, metabolic, or immune traits. One option to increase the number of traits I could use would be to aggregate the traits into categories. For example, I could combine all autoimmune diseases into a category for autoimmunity to increase the coverage.

This question was difficult in terms of selecting methods for a couple of reasons. The number of bacterial genes present in a typical microbiome is often in the millions. Even with filtering out rare genes, it is difficult to decrease the size of this type of data without some form of aggregation. I chose to aggregate the gene abundances by functional categories or protein families which I refer to as "gene families" throughout this dissertation (KEGG, eggNOG, and pfam). While this was useful in terms of the dimensionality of the data and computation, it made interpreting the results sometimes difficult because it is impossible to identify the exact proteins involved.

To address my questions, I applied both classic statistical and machine learning methods. I explored non-parametric statistics for associating

microbial gene family abundances with host traits, given that these abundances are highly skewed and zero-inflated. I then applied random forest for feature selection to identify the microbial gene families which could best distinguish cases from controls. This method worked for a few of the traits, but unbalanced data and missing data were challenges.

Future work on this question should aim to incorporate the twin relationship into any model. For example, if two identical twins are discordant for a trait such as hypertension, then any microbiome differences between identical twins should be weighted higher than differences between fraternal twins. This is because any differences between identical twins cannot be due to genetics, whereas with fraternal twins it is harder to distinguish the influence of the environment from genetics. Advances in mixed effects random forest could be useful for this strategy. Similarly, by using twin data with paired host genotype information, it would be interesting to incorporate host genetics into this analysis. For instance, one could ask if the microbial functions associated with these diseases are heritable and whether any quantitative trait loci for the gut microbiome overlap with those for the specific disease, if there are any known.

From this collaboration stemmed another project involving the TwinsUK dataset. Dr. Xu Wang, a former post-doc in the Clark Lab, had

been similarly pursuing this question of linking microbial genes to host traits and was studying anxiety. He subset the TwinsUK data into two datasets in terms of the anxiety phenotype: one where the twin pairs were discordant for anxiety (i.e., one twin was diagnosed with anxiety and the other was not), and another dataset where the twin pairs had a large difference (more than 10 point score difference) in their anxiety diagnostic scores. With these datasets, Dr. Wang performed a differential gene abundance analysis between the anxious and non-anxious (or low anxious) twins. He found that 187 genes were differentially abundant in the twins. Of those genes, 175 were tightly correlated with one another, but he was unable to figure out where these genes originated from by using typical methods to identify species from gene sequences. I had been interested in strain-level metagenomic analyses, so I had the idea to apply latent strain analysis (LSA) to this challenge (Cleary *et al.* 2015). In Chapter 3, I used LSA to identify a low-abundance microbiome-associated species that is inversely correlated with anxiety symptoms in twins. LSA is a pre-assembly binning software tool that is particularly adept at identifying low-abundance bacterial organisms from metagenomes by pooling covariance data across samples. In the interest of increasing accessibility for this tool, I also included a step-by-step, detailed overview of how to run LSA.

By performing pre-assembly binning with LSA of the metagenomic reads, I was able to more accurately and cleanly generate metagenome-assembled genomes, or MAGs from the samples' metagenomes. These MAGs could then be taxonomically annotated using readily available software that uses conserved, single-copy marker genes. Once I had taxonomically annotated MAGs from the twins' samples, I could map the genes of interest from the differential gene enrichment analysis to the MAGs and identify their source. I was surprised to see that one MAG in particular contained so many of the genes, and it happened to belong to the genus *Azospirillum*. Species within this genus are known to have probiotic effects for plants, so it is possible this species entered the person's body on food they ate. However, future work to fine-tune the LSA bins and identify the species of *Azospirillum* present in these individuals' guts is needed. Experimental work could be done to test for probiotic effects *in vivo*, and mouse studies using models for anxiety could be used to test for the protective effects of microbiome-associated *Azospirillum*.

Overall, my work linking host phenotypes and diseases to features of the gut microbiome was limited by data in terms of the number of samples and the phenotypes available. A larger subset of the TwinsUK dataset may be beneficial in this regard, but given that the cohort was not formed in

order to study specific diseases, this problem of low coverage for phenotypes may be unavoidable. Access to more metagenomic datasets from larger and more diverse populations from around the world, particularly those with accompanying phenotype data, would be useful for future analyses.

From there, I switched gears away from the traits of the individuals in the TwinsUK cohort to studying their genetics in relation to their gut microbiomes. The literature for metagenome-wide association studies of the gut microbiome is fairly limited, however there are several published metagenome-wide association studies. Twin and population-based studies have identified genetic associations with the overall composition of the gut microbiota (Blekhman *et al.* 2015; Davenport *et al.* 2015; Turpin *et al.* 2016; Bonder *et al.* 2016; Igartua *et al.* 2017; Wang *et al.* 2018; Visconti *et al.* 2019; Hughes *et al.* 2020; Kurilshikov *et al.* 2021), in addition to heritable taxa (Goodrich *et al.* 2014, 2016). A couple of these papers looked at functions of the gut microbiome, but did so in terms of gross dissimilarity metrics (Rothschild *et al.* 2018) or very coarse aggregation such as GO terms (Bonder *et al.* 2016).

In Chapter 2, I presented a number of methodological advancements for the analysis of metagenomic sequencing data. I proposed using a

multivariate data integration methods known as canonical correlation analysis (Hotelling 1936), or CCA, for multi-omic analysis such as associating host genetics with microbial features of the gut microbiome. By using a multivariate method, we are no longer plagued by issues from multiple hypothesis testing, as we are simply fitting one model. This method also allows us to identify the composite effect of the human genome on the overall gut microbiome, which follows from the assumption that any effects of genetics on the composition or function of the gut microbiome will be individually weak. To handle the high dimensionality of the data (human genotyping and microbiome gene families or species abundances), we applied a sparse variant of CCA (Witten and Tibshirani 2009; Rodosthenous *et al.* 2020), or sCCA, using methods for variable selection or penalization: elastic net (Zou and Hastie 2005) and group lasso (Yuan and Lin 2006). CCA and sCCA, rely on the normal distribution and are not typically used for highly skewed data such as metagenomic abundances, which are zero-inflated, overdispersed, and compositional. For the latter constraint, I modified a typical normalization scheme for the gene abundances, to take into account the geometric mean of the abundances, according to best practices for compositional data (Aitchison 1982). To address the former two issues, in collaboration with Dr. Martin Wells and Dr. Benjamin Baer,

we identified an appropriate probability distribution that can capture the structure of the data. We applied a member of the Tweedie distributional family for modeling metagenomic abundance data (Jørgensen 1987). By estimating the appropriate parameters we were able to select a Tweedie distribution that fit the metagenomic abundances better than other common zero-inflation or abundance data distributions could. With these two improvements, I performed a multivariate metagenome-wide association study to identify human variants that are correlated with microbiome-associated gene families or species abundances.

From this analysis, I identified many novel associations between human variants and the gut microbiome, with few human variants that overlapped with previous metagenome-wide association studies. This is likely due to the major differences in methodologies. Either way, my work expands the list of human loci known to be associated with the composition or function of the gut microbiome. We did however find many overlapping microbial species that were known to be correlated with host genetics, and many microbial functions that were previously unknown to be associated with host genetics.

Although sCCA with Tweedie was an appropriate model to use for this analysis, there are a number of improvements that I can imagine to this

work. For instance, this process takes multiple steps, where a generalized linear mixed model using a Tweedie distribution is first used to extract residuals from the metagenomic abundance data. Then, tuning parameter optimization takes place, and finally sCCA is run. It would be ideal to incorporate a Tweedie distribution directly into the model. Additionally, sCCA is a non-directional model. This makes interpretation of the data difficult. Another downside to this method when working with twin data is that it does not take the twins family structure into account. And finally, the ideal method would handle the compositionality of the data natively.

A related but alternative method that could handle some of these issues is known as reduced rank regression (RRR) analysis which estimates regression coefficients while simultaneously performing dimensionality reduction on the dependent variable and is often used in redundancy analysis (Anderson 1951). RRR is very similar to CCA except for one major difference, while both algorithms maximize the correlation between linear combinations of the X and Y, RRR seeks to identify the linear combinations of maximum correlation that explain the maximum variance in Y. In other words, RRR selects a subset of SNPs (by imposing a maximum rank on the SNP table, a form of variable selection or regularization) that are most associated with the gene abundances, and then performs a regression

analysis to estimate the regression coefficients on the remaining SNPs. This is a directional method that would allow for direct interpretation of the results. An ideal model would use a multivariate Tweedie distribution to handle the twin pairs, but this is not possible as there is currently no closed form of a multivariate Tweedie distribution and methods to approximate one are difficult to implement.

Future work with the TwinsUK dataset should use the genetic relationship of twins to full advantage. In this work, we treated the dependence within monozygotic twins and within dizygotic twins as a nuisance parameter that was roughly estimated. However, future work can leverage techniques in the twin studies literature to isolate narrow sense heritability (Visscher *et al.* 2008; Yang *et al.* 2010; Speed *et al.* 2012; Tenesa and Haley 2013). This would allow direct estimation of the heritability of the entire gut microbiome, in contrast to approximate approaches which appear widely in the literature and are based on dominant principal components summaries.

Overall, my work would benefit from larger and more diverse (geographically and ethnically) population-level data. Much of the work done to study gut microbiomes has been performed on individuals in the West or the global North. There are known differences in microbiome community

structure across the globe based on geography, diet, and cultural habits that we are missing by not studying more people (Porras and Brito 2019). Large metagenomic studies that have paired host genomes (fully sequenced genomes or SNP array data), are difficult to generate and often difficult to access. Increasing access to these types of published data, within the ethical bounds of privacy laws and guidelines, would be a welcome improvement to the field of metagenome-wide association studies.

An important aspect of my work has been identifying and applying appropriate methods for analyzing metagenomic sequencing data. When I first began working with metagenomic abundance data, I noticed many similarities to gene expression data. This was a useful observation, as many methods are established in the RNA-sequencing field that can be ported over for use in metagenomic analysis. One of the biggest challenges to working with sequencing abundance data of any type is the inherent compositionality of the data. In other words, the data reflect relative abundances within a sample, constrained by an arbitrary amount (generally the library size). The field is working on methods for handling relative abundances as well as obtaining absolute abundances experimentally. Experimental methods for absolute abundances are generally more straightforward for fewer species of interest using 16S rRNA sequencing in

combination with polymerase chain reaction (PCR) methods (Barlow *et al.* 2020), but it may be a while before these methods can be applied to entire metagenomes.

References

Aitchison J., 1982 The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological) 44: 139–177.

Anderson T. W., 1951 Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann. Math. Statist. 22: 327–351.

Barlow J. T., S. R. Bogatyrev, and R. F. Ismagilov, 2020 A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. Nature Communications 11: 2590. https://doi.org/10.1038/s41467-020-16224-6

Blekhman R., J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski, *et al.*, 2015 Host genetic variation impacts microbiome composition across human body sites. Genome biology 16: 191. https://doi.org/10.1186/s13059-015-0759-1

Bonder M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, *et al.*, 2016 The effect of host genetics on the gut microbiome. Nature Genetics 48: 1407–1412. https://doi.org/10.1038/ng.3663

Cleary B., I. L. Brito, K. Huang, D. Gevers, T. Shea, *et al.*, 2015 Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nature biotechnology 33: 1053–60. https://doi.org/10.1038/nbt.3329

Davenport E. R., D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober, *et al.*, 2015 Genome-Wide Association Studies of the Human Gut Microbiota, (B. A. White, Ed.). PLOS ONE 10: e0140301. https://doi.org/10.1371/journal.pone.0140301

Goodrich J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, *et al.*, 2014 Human genetics shape the gut microbiome. Cell 159: 789. https://doi.org/10.1016/J.CELL.2014.09.053

Goodrich J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, *et al.*, 2016 Genetic determinants of the gut microbiome in UK Twins HHS Public Access. Cell Host Microbe 19: 731–743. https://doi.org/10.1016/j.chom.2016.04.017

Hotelling H., 1936 Relations Between Two Sets of Variates. Biometrika 28: 321.

https://doi.org/10.2307/2333955
Hughes D. A., R. Bacigalupe, J. Wang, M. C. Rühlemann, R. Y. Tito, *et al.*, 2020 Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. Nature microbiology 5: 1079–1087. https://doi.org/10.1038/s41564-020-0743-8
Igartua C., E. R. Davenport, Y. Gilad, D. L. Nicolae, J. Pinto, *et al.*, 2017 Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. Microbiome 5: 16. https://doi.org/10.1186/s40168-016-0227-5
Jørgensen B., 1987 Exponential Dispersion Models. Journal of the Royal Statistical Society. Series B (Methodological) 49: 127–145.
Kurilshikov A., C. Medina-Gomez, R. Bacigalupe, D. Radjabzadeh, J. Wang, *et al.*, 2021 Large-scale association analyses identify host factors influencing human gut microbiome composition. Nature Genetics 53: 156–165. https://doi.org/10.1038/s41588-020-00763-1
New F. N., and I. L. Brito, 2020 *What Is Metagenomics Teaching Us, and What Is Missed?*
Porras A. M., and I. L. Brito, 2019 The internationalization of human microbiome research. Current Opinion in Microbiology 50: 50–55. https://doi.org/10.1016/J.MIB.2019.09.012
Rodosthenous T., V. Shahrezaei, and M. Evangelou, 2020 Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics 36: 4616. https://doi.org/10.1093/BIOINFORMATICS/BTAA530
Rothschild D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, *et al.*, 2018 Environment dominates over host genetics in shaping human gut microbiota. Nature 555: 210–215. https://doi.org/10.1038/nature25973
Speed D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved Heritability Estimation from Genome-wide SNPs. American Journal of Human Genetics 91: 1011. https://doi.org/10.1016/J.AJHG.2012.10.010
Tenesa A., and C. S. Haley, 2013 The heritability of human disease: estimation, uses and abuses. Nature Reviews Genetics 14: 139–149. https://doi.org/10.1038/nrg3377
Turpin W., O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans, *et al.*, 2016 Association of host genome with intestinal microbial composition in a large healthy cohort. Nature genetics 48: 1413–1417. https://doi.org/10.1038/ng.3693
Visconti A., C. I. Le Roy, F. Rosa, N. Rossi, T. C. Martin, *et al.*, 2019 Interplay between the human gut microbiome and host metabolism. Nature Communications 10: 4505. https://doi.org/10.1038/s41467-019-12476-z

Visscher P. M., W. G. Hill, and N. R. Wray, 2008 Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics 9: 255–266. https://doi.org/10.1038/nrg2322

Wang J., A. Kurilshikov, D. Radjabzadeh, W. Turpin, K. Croitoru, *et al.*, 2018 Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. Microbiome 6: 101. https://doi.org/10.1186/s40168-018-0479-3

Witten D. M., and R. J. Tibshirani, 2009 Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology 8. https://doi.org/10.2202/1544-6115.1470

Yang J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565–569. https://doi.org/10.1038/ng.608

Yuan M., and Y. Lin, 2006 Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 68: 49–67.

Zou H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x