Title

Integrating Population Structure into Metagenome-Wide Association Studies

Author

Goldman, Miriam

Publication Date

2024

Peer reviewed|Thesis/dissertation

Integrating Population Structure into Metagenome-Wide Association Studies

by
Miriam Goldman

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Signed by:
Katherine Pollard
9B0DDB91029D4F1...

Katherine Pollard
Chair

DocuSigned by:
Dara Torgerson
DAE0Signed by:
Sergio Baranzini
07D1896A49A94E7...

Dara Torgerson

Sergio Baranzini

Committee Members

Dedication and Acknowledgments

My journey toward a PhD has not been easy, and it would not have been possible without the many people supporting me. I would like to dedicate this thesis to my family, who have always supported my yearning for knowledge and for overcoming the challenges that have led me here and provided me with love, support, and guidance. I would like to thank my dad, who has always lent a listening and inquisitive ear, with a joke when I needed it. I would like to thank my mom for always providing food, safety, and the necessary resourcefulness to never give up on something, even when it feels too big and hard. My little siblings, Noah and Seth, have always provided me with the levity I needed. Lastly, my big sister, Hannah, who moved to SF halfway through my PhD, has always been there when I needed someone to tell me everything will be ok or just go for a nice long walk. I would not have been able to pursue this dream, much less succeed, and I am forever grateful to them all.

I would also like to thank Dr. Erika Camacho, my undergraduate research mentor at Arizona State University. Erika showed me there are many ways to explore the world, but math is the best one. She constantly encouraged me and pushed me to be the best mathematician I could be, and she is the reason I set out to get a PhD.

I also want to thank the community who supported me throughout my journey and made San Francisco my home. I would like to thank my friends, Calla, Laura, and Matt, from the BMI program, for always being there to understand my stresses and provide guidance. I want to thank my community at Magalit Jiu Jitsu for helping me have the appropriate distraction from work, which was also just learning something tough while exercising. I would like to thank my partner Sasha, who has always supported my goals, and Sasha's family, who have provided me with a home to live in and a garden to tend to as I completed the final years of my PhD. And I would like to thank my two adorable cats, Mochi and Coquito, who have sat in my lap as I read and write.

Next, I would like to thank my committee members, Dr. Sergio Baranzini and Dr. Dara Torgerson, who provided helpful feedback and ideas for my project. I would like to thank all of the members of the Pollard lab for their support and feedback through the years. I would like to especially thank Byron for providing a listening ear for my statistics problems, Abigail, who has given me good advice, and lastly, Chunyu, my partner in research, to whom I am incredibly in debt for helping me focus and listening to whatever I was interested in that week. I would also like to thank Francoise Chanut for editorial assistance throughout my PhD.

Most importantly, I would like to thank Dr. Katie Pollard, who has served as my advisor and mentor over the past years. I appreciate her constant patience and encouragement through a project that didn't always work how I wanted it to and through my struggles throughout the years. I can't overstate how much I could not have done this without her.

Contributions

The content of this dissertation is adapted from manuscripts both published and under review. For an expanded discussion and context for this thesis, see the following works:

Chunyu Zhao, Boris Dimitrov, **Miriam Goldman**, Stephen Nayfach, Katherine S Pollard, MIDAS2: Metagenomic Intra-species Diversity Analysis System, *Bioinformatics*, Volume 39, Issue 1, January 2023, btac713, https://doi.org/10.1093/bioinformatics/btac713

Chunyu Zhao., **Miriam Goldman**, Byron Smith, Katherine S Pollard, Genotyping microbial communities with MIDAS2: From metagenomic reads to allele tables. *Current Protocols*, vol. 2,12, December 2022, e604. doi: 10.1002/cpz1.604

Miriam Goldman, Chunyu Zhao, Katherine S. Pollard, Improved detection of microbiome-disease associations via population structure-aware generalized linear mixed effects models (microSLAM), *bioRxiv*, June 2024, doi:https://doi.org/10.1101/2024.06.27.600934

Chapter Two is adapted from my contributions to Zhao et al., 2023, Zhao et al., 2022, and Goldman et al., 2024. First author, Chunyu Zhao collected and ran MIDAS2 on the PREDICT dataset, which provides the basis for analysis in Chapter Two.

Chapter Three is adapted from my contributions Goldman et al., 2024. My co-author Chunyu Zhao performed the BlastKOALA analysis, collected the *Faecalibacterium prausnitzii* genomes, and aligned the operon to them. She also assisted with running MIDAS v3 on metagenomic data. This work provides the basis for analysis of Chapter Three.

Integrating Population Structure into Metagenome-Wide Association Studies

Miriam Goldman

Abstract

Diseases related to the human gut microbiome, such as inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), and colorectal cancer, have increased within developed nations. At the same time, increased access to metagenomics has allowed us to try to understand the microbiome. While many microbiome-disease association studies have been carried out, they primarily focus on genera or species that vary in abundance with disease status; this relatively coarse level of analysis limits our understanding of why microbes may act as disease markers and overlooks cases where disease risk is related to the presence or absence of specific strains with unique biological functions. My thesis shows that there are strong within-species phenotypic signatures across the gut microbiomes of different people and then introduces microSLAM, a statistical method that incorporates random effects to represent the population structure of the bacteria modeled in an association study. I applied microSLAM to a large set of gut metagenomes from IBD samples, where I discovered 49 species whose population structure correlates with IBD, meaning that people with the disease harbor distinct strains compared to healthy people. In addition, after controlling for population structure, I found 57 microbial genes that are significantly more common in healthy individuals and 26 that are more common in IBD patients, including a seven-gene operon in *Faecalibacterium prausnitzii* that is involved in the utilization of fructoselysine from the gut environment. In addition, I performed extensive simulations to understand the limitations and capabilities of microSLAM and found it was much more conservative and specific than the standard statistical approach, a

generalized linear model (GLM). These findings have highlighted the importance of considering within-species population genetic variation in microbiome studies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMR - antimicrobial resistance

ANI - average nucleotide identity

CD - Crohn's disease

FDR - false discovery rate

HGT - horizontal gene transfer

GalNAc - N-Acetylgalactosamine

GFR - glucoselysine/fructoselysine phosphotransferase

GI - gastrointestinal

GLM - generalized linear model

GLMM - generalized linear mixed model

GRM - genetic relatedness matrix

IBD - inflammatory bowel disease

IBS - irritable bowel syndrome

MAGs - metagenome-assembled genomes

MGEs - mobile genetic elements

MWAS - metagenome-wide association study

NCBI - National Center for Biotechnology Information

PC - principal component

PCG - preconditioned conjugate gradients

PCA - principal component analysis

PCoA - principal coordinates analysis

PREDICT - personalized responses to dietary composition trial

PTS - phosphotransferase

QP - quasi-phasable

Q-Q plot - quantile-quantile plot

SNP - single-nucleotide polymorphism

SNV - single-nucleotide variant

UC - ulcerative colitis

UHGG - unified human gastrointestinal genome

List of Symbols

ψ - genetic relatedness matrix

Φ - dispersion of the variance of fitted values

b - random variable for each sample

τ - estimated additive genetic variance or the population structure variance

β - gene-trait association parameter (log odds ratio in the case of a binary trait)

Chapter 1 Introduction

The human body is home to a diverse community of microorganisms, collectively known as the microbiota, comprising bacteria, fungi, and viruses. This microbial population varies greatly among individuals, influenced by factors such as diet [1–3], exercise [4,5], health [6–8], and host genetics[9,10]. Understanding the intricacies of the microbiome presents a unique opportunity for improving human health by addressing dysbiosis—an imbalance between the microbiome and its host—at its root. Harnessing this knowledge could potentially enhance drug efficacy [11–14], illuminate the impact of diet on our well-being [3,15,16], and offer significant health benefits [17,18]. In my thesis research, I aim to advance our comprehension of the human gut microbiome by investigating the genetic structure of the microbiome's bacterial population through metagenome-wide association studies. The following chapter will provide a concise overview of relevant background information.

1.1 The human gut microbiome

Microorganisms colonize the world around us and various sites on and in the human body. Those colonizing the human body are referred to as the human microbiota. The human gastrointestinal (GI) tract is one of the niches in which microbiota, including bacteria, archaea, viruses, and eukaryotes, make their homes. The GI tract is also one of the most significant interfaces between the host, environmental factors, and microbes in the human body. Every food humans eat and most medications humans take pass through the GI tract. The collection of microbiota colonizing the GI tract is termed the "gut microbiome" and has co-evolved with humans over thousands of years to form a complex, interdependent, and ultimately beneficial system [19,20]. From before childbirth, the species within the gut microbiome actively adjust to their specific habitats and hosts [21]. The bacteria within the human microbiome are constantly

evolving in response to host factors such as age, diet, lifestyle, hormonal changes, and host genetics. The microbiota has been shown to offer many benefits to the human body through a range of physiological functions, such as shaping the intestinal barrier [22], supplying energy [15], providing a protective effect against pathogens [23], and regulating host immunity [24] and drug metabolism [12]. However, the bacteria within the gut microbiome evolve quickly, which can lead to changes in these functions.

One of the ways gut bacteria evolve is through the capture and exchange of mobile genetic elements (MGEs). MGEs include plasmids, phages, integrative and mobilizable elements, transposons, and insertion sequence elements [25]. They rely on host cells and cellular machinery to multiply. MGEs alter the genome composition of bacterial species, changing functions and creating new lineages [26–29]. MGEs affect the health of the ecosystems that house bacteria by exchanging traits such as metabolism, virulence, symbiosis, and host specificity [30,31] via a process known as horizontal gene transfer (HGT). HGT is responsible for transferring Important functions such as antimicrobial resistance (AMR) [32] and is one of the forces contributing to gene losses and gains among microbiome species [26,33,34].

1.2 The Pangenome

Because of recent advances in genomic and metagenomic sequencing, we have started to understand the genetic structure of the human gut microbiome and its complex dynamics. Shotgun metagenomics, the non-targeted sequencing of fragments from multiple microbial genomes in a sample, allows for the profiling of taxonomic composition and functional potential of microbial communities [35], to recover whole genome sequences via metagenomic assembly [36], and to quantitatively profile the pangenome[35]. The pangenome is the full set of genes present across different versions (strains) of a given species. It includes both the genes present in all strains of a species (core genome) and the genes present only in some strains of a species

(accessory genome) [37]. Bioinformatic advancements such as MIDAS [38–40], StrainPhlAn [41], MetaSNV [42,43], panX [44], and Roary [45], also contribute to the ability to understand the pangenome. MIDAS first maps sequences from shotgun metagenomic data to universal single-copy gene families, allowing accurate identification of species with sufficient sequencing coverage for analysis of genetic variation. Then, to quantify the gene content of each species in each metagenome, MIDAS maps the reads to a pangenome database, the set of nonredundant genes across all sequenced genomes from each species. Last, to identify single nucleotide variants (SNVs) of individual species, MIDAS maps those to a genome database containing one representative genome sequence per species.

Mapping the pangenome has significantly enhanced our understanding of host-microbiome interactions and revealed a vast genetic diversity of bacterial species within and between human hosts [46]. Even when two individuals share the same microbial species, the cells within those populations are genetically and functionally very different [47,48]. Illustrations of this diversity include the identification of variable virulence and antibiotic resistance [49,50], of a set of pro-inflammatory genes from specific strains of *Ruminococcus gnavus* [51], of a *Faecalibacterium prausnitzii* GalNAc utilization pathway linked to cardiometabolic health [52], and of a strain of *Escherichia coli* with enhanced ability to live on the intestinal mucus that is associated with IBD [53]. These findings underscore the limitations of using species abundance alone to gain insight into host-microbiome interactions and the need for a more comprehensive approach. My thesis work aims to describe the microbiome at the strain level.

Strains within the microbiome can be defined as groups of genomes of a species that contain similar sets of genes. These sets of genes can provide diverse functions and be related to traits of the microbe or its host. Identifying and isolating these trait-associated strains can facilitate experimental investigations into host-microbiome interactions. Strains enriched in healthy hosts have been suggested as potential components of probiotics and therapies [17,54–56]. The structure of bacterial genomes is such that many genes are present or absent together

across strains, especially closely related strains. This means that genes from a strain could be promising biomarkers (e.g., for diagnosis or patient stratification). However, most are not good candidates for studies of causal mechanisms. On the other hand, we can also identify, from the pangenome, one or a small number of individual genes that predict a trait. Such predictions are most straightforward if genes are rapidly gained and lost (e.g., via mobile elements), so that their association with the trait is independent of evolutionary relationships amongst strains. These types of genes are promising candidates for discovering causal mechanisms through which microbes modify the health of their hosts and the host's response to treatments.

1.3 Population structure

From the first understanding of biological evolution, it has been noted that individuals of a species are not all the same. This is especially true of bacteria. Now that we have sequenced whole bacterial genomes in a sufficiently large number, we can see that some of these differences are related to genetic variations. This variation complicates our analysis of bacterial species, which had been simplified to assuming similarity within a species. To quantify similarities, we use population structure, which is the degree to which individuals within a species share a common evolutionary history [57]. Population structure is a biological reality in understanding how genetic variation relates to host phenotypic variation, and it is especially confounding in case-control designs. False positive associations are increased in cases where populations are subdivided, such as in a case-control design, because there is a nonrandom grouping of the cases and the controls and a non-random grouping of the population from each of the many the genetic alleles. Because we cannot assume that the populations we are testing are randomly sampled from a single distribution[58], we have an increase in false positives as the alleles are related to one another non-randomly, and the phenotype is related to those alleles in a cryptic fashion. Luckily, we can infer information about how microbiome samples are related to

their genetic variation. One approach, first proposed in 1978 in the context of human genetics, uses principal component analysis (PCA) of SNVs within a population to understand population structure [59,60,61]. The idea is to use PCA to quantify ancestry differences between cases and controls and integrate them into phenotypic modeling to adjust for confounding of genotype-phenotype associations. Since 2006, this idea of trying to decompose the genetic relatedness across many samples into a continuous variable and including the decomposition in models that test individual variants for their trait associations has been refined [62–65].

One of the methods that has seen success in tackling the problem of false positives in case-control studies is the generalized linear mixed model (GLMM). The GLMM is an extension of the generalized linear model (GLM) [66] that incorporates random effects. Random effects can be encoded as an unobserved vector that allows observations to be assumed to be conditionally independent. The means of the random effect depend on the linear predictor through a specified link function, and conditional variances are specified by a variance function, known prior weights, and a scale factor [67]. The use of random effects to model population structure based on the variance of the random effect calculated from a genetic relatedness matrix (GRM) was made computationally tractable for studies with a large number of samples by the EMMA and EMMAX software [63,68]. Later the GEMMA and SAIGE methods implemented more efficient algorithms and tactics for controlling the false positive rate in human studies [62,64].

Given the abundance of population structure within bacterial strains, the problems of false positive control in metagenomic association studies, and the robustness of GLMMs, the bulk of my dissertation focuses on developing a statistical model that can be used to perform microbial population structural aware metagenome-wide association studies (MWAS). This dissertation will describe the approach in detail and the methods used to assess the effectiveness of incorporating population structure into MWAS.

Chapter 2 Integration of population structure into metagenome-wide association studies through generalized linear mixed modeling

2.1 Why investigate population structure within the microbiome

Many previous studies have aimed to establish a connection between various microbiome features, primarily the relative abundance of genera or species, and differences in human health outcomes. However, most of these studies have not been able to translate their analytical findings into effective treatments or improvements in health for humans or model organisms [69]. One possible reason for this limitation is that large studies focusing on relative abundance assume that each individual harbors specific bacteria, and that these bacteria perform the same functions across different individuals. In reality,this assumption does not hold true. A particular bacterium might be beneficial in one environment, but harmful in another [70,71]. In addition, large relative abundance studies suffer from many false positives precisely because bacteria differ based on their environment. Suppose the gut environment of a healthy person is very different from that of a person with a disease. In that case, we expect the bacteria to be distinct, but this may not be directly related to the disease. Because of this confounding, it is hard to know when something new has been discovered or if it was simply a spurious difference. These spurious associations are time-consuming and difficult to investigate and validate, and when there are many false positives that limits the ability of scientists to focus our energy on the more likely functional associations.

But the limitations of relative abundance do not end there. There is evidence that people do not simply have one strain of a species; they can have a mix of strains at the same time. We can detect such mixtures based on the genes or SNVs present within their microbiome. One of the analyses I performed to investigate the amount of strain diversity within a large number of microbiome samples used the quasi-phasable (QP) species definition from Garud, Good et al. [72]

to quantify evidence for strain mixtures within metagenomes from the PREDICT study (NCBI accession: PRJEB39223) [73]. My goal was to determine whether each species in each metagenomic sample was QP, meaning that there is evidence of one strain as opposed to a strain mixture. The PREDICT study is a deeply sequenced gut metagenomics cohort of 1097 stool samples from people with a variety of diets and metabolic outcomes. From the PREDICT data, I determined that 44 species were present and deeply sequenced enough to be metagenotyped by MIDAS2. The SNV pipeline from MIDAS2 was used to generate the inputs for the QP model, namely synonymous sites (SNVs) in genes of the core genome of each given species. If the fraction of these sites with intermediate allele frequencies is high, this is taken as evidence for a strain mixture. I found that 65.9% (29/44) of the species analyzed had evidence that 30% or more of the samples contained more than one strain (**Figure 2.1**). This means that only having a single strain of a bacteria across many people is uncommon within most species seen in this metagenomic dataset.

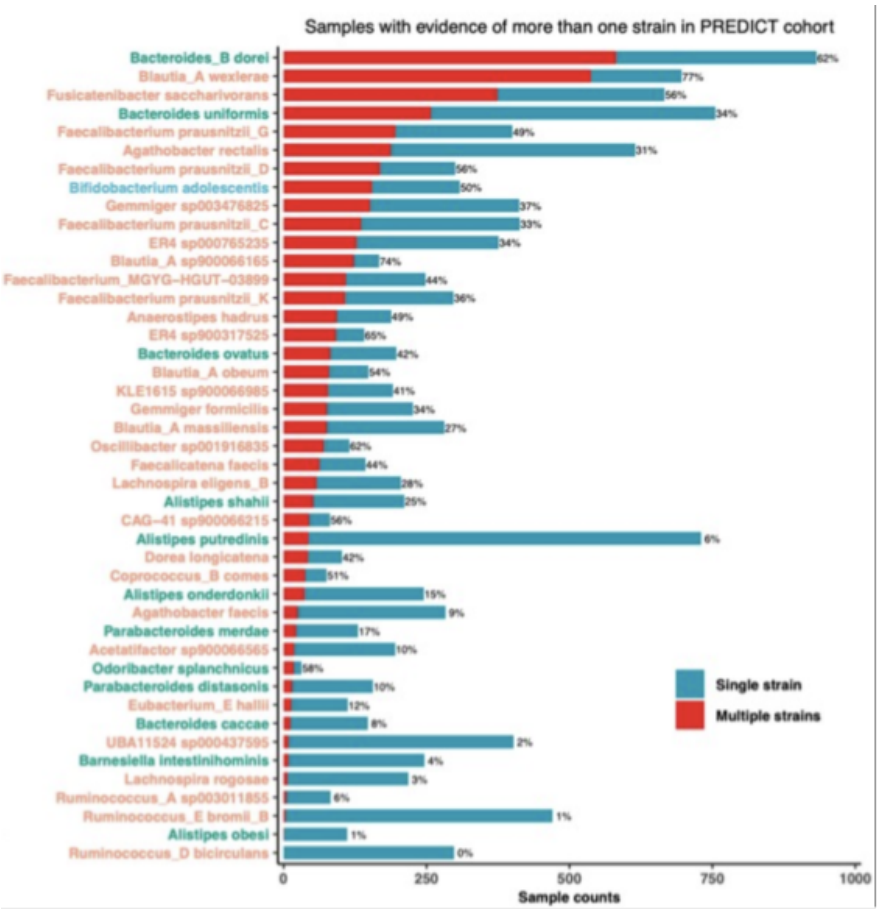

Figure 2.1 Strain mixtures within PREDICT cohort. Distribution of samples with evidence of a strain mixture versus one dominant strain for 44 species metagenotyped by MIDAS2 in 1097 samples from the PREDICT cohort (NCBI accession: PRJEB39223) using the quasi-phasable species model 2.3.1.

There was also great between-host diversity for the species determined to only have one strain present in most samples. I performed PCoA based on the pairwise Manhattan distance between samples computed from the population SNV minor allele frequency matrix of each single species. For one species, *Alistipes putredinis*, a species whose abundance has been linked to depression [7], there appear to be two or three distinct clusters of samples, each composed primarily of samples with a single strain. Another example of diversity within and across strains is *Barnesiella intestinihominis*, which has four distinct clusters of samples. This species has been associated with anti-cancer effects [74]. Given these known associations and my observations regarding the variety of the strains that exist for both of these species (**Figure**

2.2), I saw an opportunity to better understand if strain differences could contribute to the disease associations of these and other species.

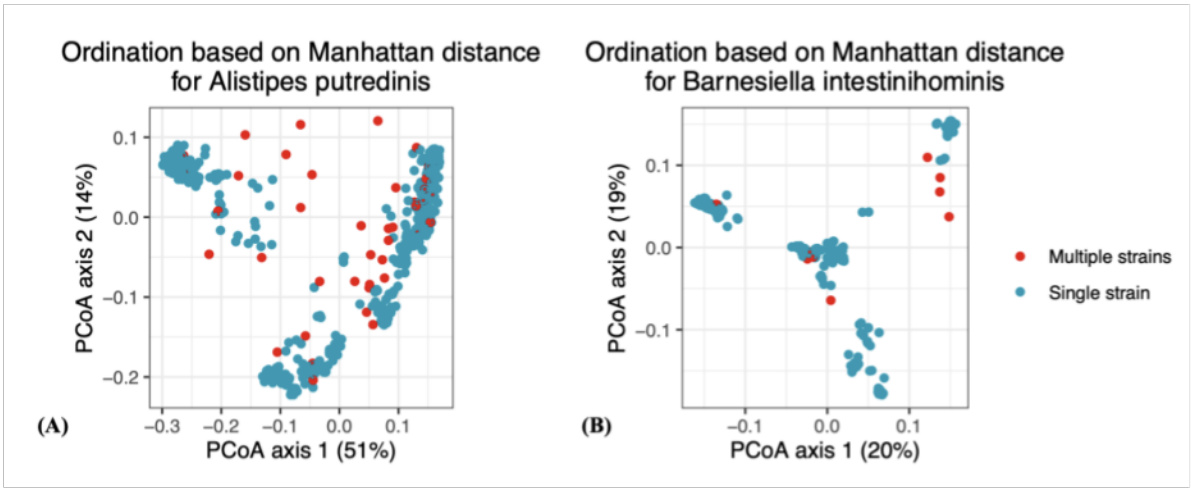

Figure 2.2 PCoA plot of two species with genetically distinct lineages across samples, yet a single dominant strain within most samples. PCoA was performed based on the pairwise Manhattan distance between samples computed from the population SNV minor allele frequency matrix of one single species. Each dot in the PCoA plot represents one sample, and the distance between a pair of dots represents the genetic similarity of that species in the two metagenomes. **A:** *Alistipes putredinis* - there appear to be two or three distinct clusters of samples, each composed primarily of samples with a single strain. **B:** *Barnesiella intestinihominis* - there are four distinct clusters of samples.

In addition to multiple distinct clusters of single strains, I also found evidence for a mixture of two single strains within samples. For example, in *Bacteroides B dorei (B. dorei)* 62% of the 932 PREDICT samples with *B. dorei* had evidence of more than one strain present. When I looked further into this I found that from the PcoA plot there were two genetically distinct clusters each consisting of one dominant strain. The samples that were not found to be QP, on the other hand, appeared to be between these two clusters, consistent with colonization by strains from both clusters (**Figure 2.3**). *F. prausnitzii_G,* present in 49% of the 401 PREDICT samples, also shows a similar pattern (**Figure 2.4**). These results suggest that there is, in fact, a variety of strains within the human gut microbiome and that the Manhattan distance is able to capture those strain distributions across people, and therefore could provide a useful estimator for the population structure of bacterial species. In addition, because the population structure is

continuous and related to a defined distance matrix, it made sense to build on approaches that have proven useful in human GWAS studies. One of the approaches from GWAS that has historically been very successful, a mixed modeling approach of incorporating a random variable based on the population structure, was a clear choice due to its speed, precision, and ability to adjust for false positives [75].

To conclude this section, there are significant challenges with understanding associations between the human gut microbiome and host traits, including disease phenotypes. We must go beyond using the relative audience of species alone. By incorporating population structure within the microbiome into metagenome-wide association studies, we have the potential to unlock new treatments and health improvements that could significantly impact future health outcomes. Building on the methods used in GWAS, I will next describe how I developed “microSLAM” (microbiome structure leveraged association models), an R package and a statistical model. This tool performs association tests that connect the presence/absence of genes within species to host traits, while accounting for population structure (i.e., strain genetic relatedness across hosts). MicroSLAM is implemented in three steps for each species. The first step estimates population structure across hosts. Step two calculates the association between population structure and the trait, enabling the detection of species for which a subset of related strains confer risk. To identify specific genes whose presence/absence across diverse strains is associated with the trait, step three models the trait as a function of gene occurrence plus random effects estimated from step two. The rest of this chapter will delve deeper into the details of how each of these steps are performed.

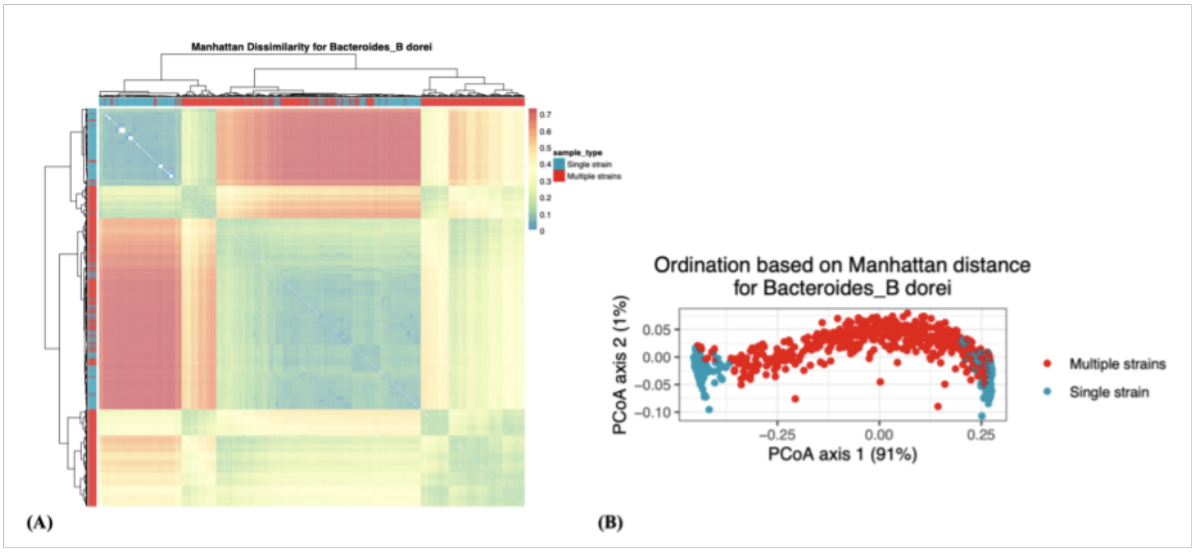

Figure 2.3 Many samples contain more than one strain of *Bacteroides_B dorei*. 62% of the 932 PREDICT samples with *B. dorei* have allele frequency spectra consistent with two or more distinct lineages colonizing the host. Samples with one dominant strain of *B. dorei* form two genetically distinct clusters, and the rest of the samples are intermediate between these, consistent with colonization by strains from both clusters. **A:** Heatmap of pairwise Manhattan distances between PREDICT samples. **B:** PCoA plot based on the Manhattan distances.

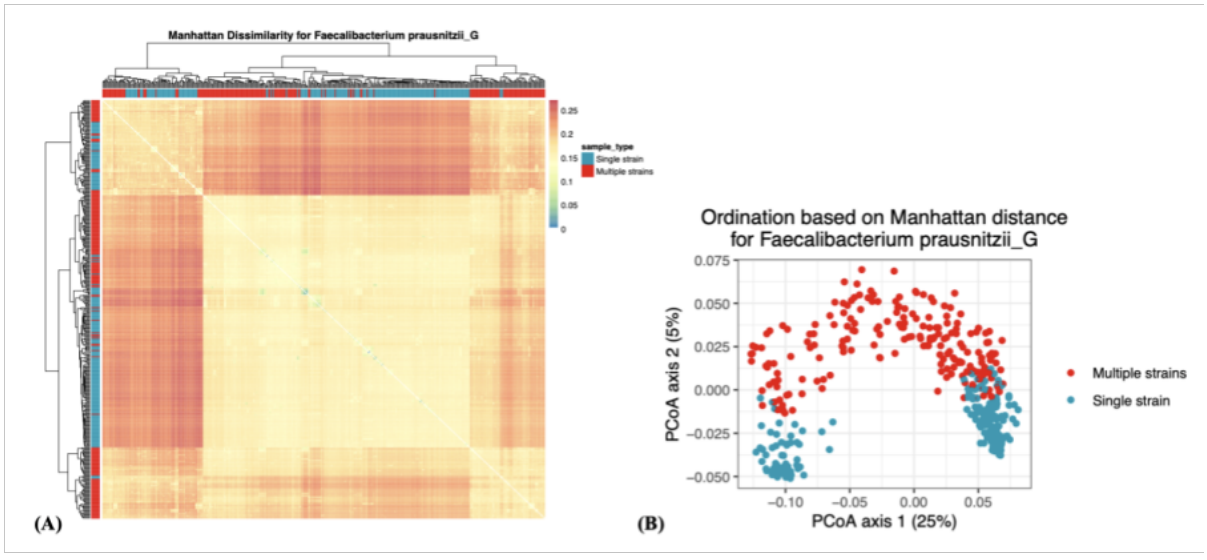

Figure 2.4 Many samples contain more than one strain of *Faecalibacterium prausnitzii_G*. 49% of the 401 PREDICT samples with *F. prausnitzii_G* have allele frequency spectra consistent with two or more distinct lineages colonizing the host. Samples with one dominant strain of *F. prauznitzii_G* form two genetically distinct clusters, and the rest of the samples are intermediate between these, consistent with colonization by strains from both clusters. A: Heatmap of pairwise Manhattan distances between PREDICT samples. **B:** PCoA plot based on the Manhattan distances.

2.2 Statistical model for strain-trait and gene-trait associations while accounting for population structure in metagenomes

In this section I present microSLAM, a 3-step modeling procedure for detecting within-species genetic variation associated with host biology. The inputs to microSLAM, for a given species, are a *p* x *n* binary matrix of gene family presence/absence values for *n* host samples and *p* gene families, a 1 x *n* vector of trait values for each sample (binary or quantitative), and optionally a *q* x *n* matrix *X* of data for *q* covariates. The outputs are a measure of population structure (τ) with a permutation p-value and, for each gene family, a coefficient (β) measuring the gene's association with the trait (e.g., log odds ratio for binary traits and logistic regression) with its local false discovery rate (localFDR) adjusted p-value [76]. Results from different species can be interpreted jointly to identify shared trends in trait-associations, such as enriched pathways.

MicroSLAM fits generalized linear mixed effects models that account for the genetic relatedness of strains of a given species across hosts. In Step 1, an *n* x *n* sample genetic relatedness matrix (GRM) is computed from the gene presence/absence matrix. To do so, we create an *n* x *n* Hamming distance matrix and then transform this into relatedness using 1-distance. The GRM is used in Step 2 to test if the species' population structure is associated with the trait, which would indicate that hosts with similar trait values tend to have similar strains. For example, for a case/control study, this step aims to detect species where a subset of related strains confers risk. We call Step 2 the τ test, because population structure is modeled using a parameter τ. In Step 3, random effects estimated from the GRM are used to adjust for population structure in a model that is used to test gene families for associations with the trait beyond simply being present in trait-associated strains. We call Step 3 the β test, because a parameter denoted β is used to quantify gene-trait associations.

In Step 2 (τ test), microSLAM fits a generalized linear mixed model. The trait $y_i: i = 1, \ldots, n$ is modeled as a function of any covariates *X* (with coefficients $\boldsymbol{\alpha}$) and random effects $b_i: i = 1, \ldots, n$ that are estimated from the GRM. The link function f() is the identity function for normalized quantitative traits (linear regression) or the logit function for binary traits (logistic regression):

$$E[f(y)] = \alpha X + b \qquad (1)$$

One key component of fitting Model 1 is estimating τ, the variance on the random effects, which depends on the association of the trait to the GRM. This is done iteratively using the average information restricted maximum likelihood (AI-REML) algorithm from the GMMAT [77] method. From this, we obtain a point estimate of τ, a point estimate of the random effects b_i, and a statistic, $T = b^2/N$, that measures how associated the species' population structure is with the trait. This T statistic is derived from [78] and computed using the linear setup from [79]. To assess the statistical significance of T, we randomly permute the trait values *B* times (e.g., *B*=1000), repeat model fitting, compute a T statistic for each permutation, and use these as an empirical null distribution to estimate a p-value based on how many of the permuted T statistics exceed the observed T statistic. Species with a significant T statistic have population structure that associates with the trait.

In Step 3 (β test), microSLAM fits a second model using the random effects (b) estimated in Step 2 and the presence/absence vector for each gene family, denoted *g* (with coefficients β):

$$E[f(y)] = \alpha X + \beta g + b \ (2)$$

Model 2 is fit separately for each gene family within each species. β measures the gene's association with the trait given the species' population structure and the covariates. Similar to the strategy used in SAIGE [64], we directly calculate the score statistic for each gene by fitting the covariate and population structure adjusted genotype vector to the phenotype.

Doing a direct computation given the random effect is an efficient strategy to reduce compute time; we only have to fit Model 1 once per species. Microbiome case/control studies are often unbalanced, for example, when a bacterial species is detected in many more controls than cases. To obtain accurate p-values in this scenario, we approximate the score statistics for testing the null hypothesis that β is zero using the Saddle Point Approximation (SPA) of the true distribution, as implemented in SAIGE.

To adjust the resulting p-values for multiple testing, we use localFDR [76], which accounts for the high correlation between gene families (i.e., when genes co-occur across strains) that invalidates methods such as Benjamini-Hochberg FDR [80] or Storey's q-value [81]. We transform SPA p-values into Z-values by dividing by two, multiplying times the sign of the estimated β coefficient, and converting the resulting numbers to quantiles. Then, localFDR uses maximum likelihood estimation to approximate the null Z-value distribution and identify Z statistics that deviate from this distribution. We implement this using the *locfdr* v1.1-8 package in R, fitting the null distribution to the Z-values between the 10th and 90th percentiles across all species.

MicroSLAM was developed for use with the outputs of the MIDAS software [38]. Throughout this project, I used different versions of the MIDAS software to enhance its functionality and performance. Metagenomic Intra-Species Diversity Analysis (MIDAS) is an integrated pipeline for profiling strain-level genomic variations in shotgun metagenomic data, originally developed by Stephan Nayfach in the Pollard lab. MIDAS2 [39] was developed by Chunyu Zhao and Boris Dimitrov, to work with more comprehensive MIDAS Reference Databases (MIDASDBs), and to run on large collections of samples in a fast and scalable manner. I worked on testing MIDAS2 installation, publishing MIDAS2 protocols, and the developing application user case to quantify evidence of a single dominant strain versus mixtures of multiple strains in each sample of MIDAS2. Most recently, the Pollard lab has released MIDAS v3 [40], which includes updates to its pangenome database and profiling pipeline which affect quantification and facilitate the interpretation of strain-specific gene content.

2.3 Methods

2.3.1 Quasiphasable species model

The QP model [72] was applied to the PREDICT study to determine whether each species in each metagenomic sample was quasi-phasable (QP) or not (i.e., one dominant strain versus colonization by multiple bacterial lineages). The model uses synonymous sites in genes of the core genome of a given species. If the fraction of these sites with intermediate allele frequencies was higher than .8, this was taken as evidence for a strain mixture.

I used similar sample and site filters for population single-nucleotide variants (SNV)s as in (Garud et al. 2019). Specifically, I filtered the per species *snps_info.tsv* and *snps_freqs.tsv* files from MIDAS2 as follows:

- Minimal per-sample median site depth of bi-allelic SNVs from protein-coding sequences ($\underline{D}$) is 20.
- Only include 4-fold degenerate synonymous sites
- Sample site depths must be between .3 *$\underline{D}$ and 3 * $\underline{D}$
- Minimal site depth is 10
- Minimal site prevalence is 5

This produces a filtered SNVs allele frequency file.

QP calculation. For each species, I estimated whether each sample is QP as follows:

1. For each non-intermediate site, I defined population major allele direction ($\underline{f_l}_{dir}$) as 1 if the majority of the allele frequencies are higher than 0.8 and 0 if not.

$$\bar{f}_{l_{dir}} = 1 \text{ if } (\sum_n^N f_n \geq .8 \geq \sum_n^N f_n \leq .2) \text{ else } \bar{f}_{l_{dir}} = 0$$

where n is the sample index, N is the number of relevant samples for the given site.

2. For each sample n , I computed the dominant haplotype of each non-intermediate site ($f_{n,l_{dir}}$) as 1 if the corresponding allele frequency is higher than 0.8 and 0 if not

$$f_{n,l_{dir}} = 1 \text{ if } f_{n,l_{dir}} \geq .8 \text{ else } f_{n,l_{dir}} = 0$$

where l is the site index.

3. For each non-intermediate site, the population major allele frequency ($\underline{f_l}$) is computed as:

$$\overline{f_l} = \frac{\sum_n^N f_{n,l_{dir}} == \overline{f}_{l_{dir}}}{N}$$

4. For each sample, I estimated N_D as the average genetic distance between sample *n* and the alleles present in the remainder of the samples ($\overline{f}_{l_{dir}}$) as:

$$N_D = \sum_l^L \overline{f_l} \text{ if } f_{n,l_{dir}} == \overline{f}_{l_{dir}} \text{ else } 1 - \overline{f_l}$$

where, l is site index, L is total number of non-intermediate sites; n is the sample index.

5. For each sample, I computed the number of intermediate alleles over all sites ($N_<$):

$$N_< = \sum_l^L 1 \text{ if } .2 < f_{n,l} < .8 \text{ else } 0$$

6. For each sample, if $\frac{N_<}{N_D} < 0.1$ then we say the sample is QP, because there is evidence of one dominant strain with sufficiently high coverage and sufficiently low rates of intermediate alleles.

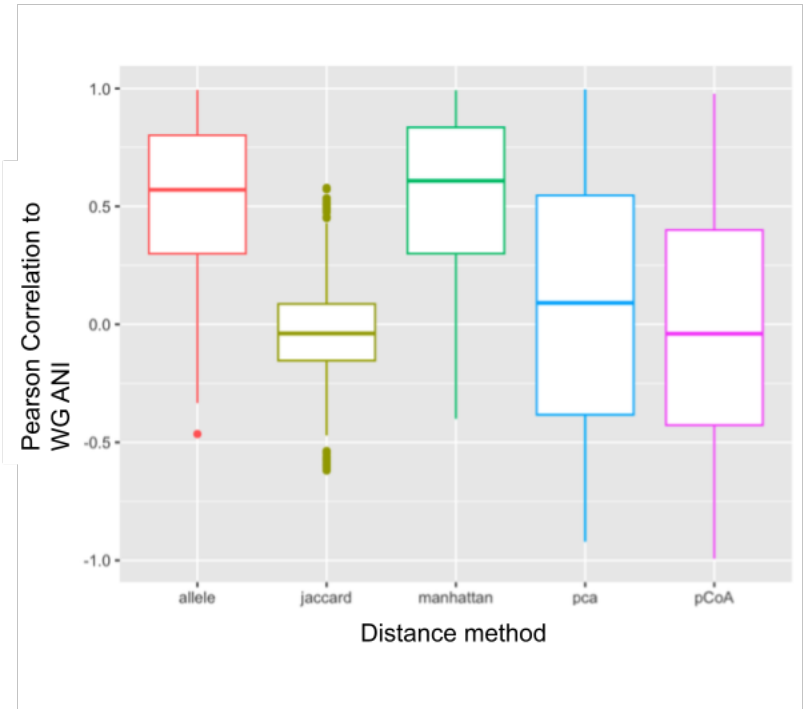

Figure 2.5 Comparison of different distance methods to whole genome ANI. Whole genome ANI was calculated from 100 relevant and abundant gut species from UHGG with 100 diverse genomes used to simulate the diversity between different strains within the human gut microbiome. This was done with different numbers of sites sampled between 1000 to 5000 and the Pearson correlation between the distance method and whole genome ANI (WG ANI) was calculated. Manhattan distance was the most similar to WG ANI.

2.3.2 PCoA and Manhattan distance using SNVs from microbiome

Manhattan distance was calculated based on the filtered SNV site-by-sample allele frequency matrix to evaluate the dissimilarity between samples. Principal Coordinate Analysis (PcoA) was calculated based on the Manhattan distance matrix using the ape package [82]. I decided to use Manhattan distance based on a small experiment of 100 relevant and abundant gut species. I downloaded 100 of their genomes from MGnify [83], filtering for those that were most different from each other, and I compared the whole genome ANI to different distance methods with drop outs and number of SNVs ranging from 1000 to 5000 in order to determine which distance method was most similar to whole genome ANI (**Figure 2.5**).

2.3.3 Genetic relatedness matrix

I used the pairwise Hamming distance based on the gene presence/absence matrix to create the genetic relatedness matrix (GRM) for a set of samples. For *N* genes where A_i and B_i are a binary presence or absence of gene *i* in sample A and in sample B, respectively, I defined genetic similarity as $1 - \sum_i^N \frac{|A_i - B_i|}{N}$. This is computed for all pairs of samples where a given species is present to create that species' GRM ψ.

If wanted, one could instead use a SNV presence/absence or frequency matrix instead of a gene presence/absence matrix to compute the GRM. For SNV distance, I recommend using the Manhattan distance. The resulting GRM based on the set of bi-allelic polymorphisms genotyped by MIDAS v3 or another tool would approximate the average nucleotide identity (ANI) between samples. I explored this approach and found that polymorphism based GRMs were generally very different from gene presence/absence based GRMs for the same species. In simulations, this led to higher false positive rates for the microSLAM β test (similar to GLM), presumably because there was trait-associated population structure in the gene presence/absence matrix for which this approach did not fully adjust. Further investigation into the selection of SNVs or distance metric could potentially make this strategy more effective.

2.3.4 Microbiome generalized linear mixed model for binary traits: additional details about microSLAM's modeling approach

In a case-control study with sample size N, we denote the status of the *i*th individual with $y_i = 1$ or 0, depending on whether it is a case or a control. Let the $1 \times (1 + p)$ vector $X_i$represent p covariates, plus an intercept term, and let G_i represent the presence or absence of a gene; this can also be replaced with the copy number of a gene, as estimated by MIDAS v3 or other tools. The logistic mixed model can be written as:

$$logit(\mu_i) = X_i\alpha + G_i\beta + b_i + \varepsilon_i$$

where $\mu_i = P(y_i = 1|X_i, G_i, b_i)$ is the probability that the i'th individual is a case given the covariates, gene presence/absence vector, and the random effect b_i that is estimated by microSLAM. The random effect b_i is modeled $N(0, \tau\,\psi)$ where ψ the N×N GRM described above, and τ is the estimated additive genetic variance. The $Var(y_i|\,b\,) = \phi\; var(\mu_i)$, in the case of a binary trait the random parameter Φ=1. The parameter α is a $1 \times (1 + p)$ coefficient vector of fixed effects and β is a coefficient representing the log odds ratio for the association between the gene's presence and the trait. For a quantitative trait, y_i is a real number and the model is a linear mixed model rather than logistic, so that β represents the expected change in the trait for the gene being present versus absent. Everything else is the same.

2.3.5 Estimating the coefficients and variance components

I employ the same restricted log-likelihood and average information matrix for estimating the coefficients and variance components as were employed in GMMAT [77] and Saige [64]. For more details on deriving these estimation procedures, refer to those studies and to Clayton and Breslow [67]. I also followed Saige's multi-step process to estimate the random effects and then use these in the logistic (or linear) model presented in the previous section. This helps us in two ways: 1) it reduces computational time significantly as random effects only have to be estimated one time for each species (not once for every gene in every species), and 2) avoiding refitting the random effect for every gene provides a more robust estimate.

Unlike Saige, I do not use PCG, randomized trace estimator, or a low-rank GRM. These are designed to reduce computation and memory costs within the context of human genomes with millions of genetic variants, but these are not major problems for us given the size of the datasets in this study. Also, our GRMs are naturally full-rank. These computational shortcuts could be implemented if needed.

2.3.6 Score testing for the GRM: $\boldsymbol{\tau}$ test modeling

I detail microSLAM's τ test, a new statistical procedure to inform the user whether the species' GRM is significantly related to the trait. This would indicate that a subset of related strains can predict the trait. I consider random effects $b_i \sim N(0, \tau\psi)$, as described above, then compare the models:

$$H_0: Y = X\alpha + \epsilon$$

$$H_a: Y = X\alpha + b + \epsilon$$

After the models have been fit (estimation converges), I have $\hat{\alpha}$; $\hat{b}$; $\hat{\phi}$; and $\hat{\tau}$. I also compute a working vector

$$\hat{Y} = X\,\hat{\alpha} + \hat{b} + \hat{\epsilon}, b \sim N(0, \hat{\tau}\,\psi);\ \hat{\epsilon} \sim N(0, W^{-1})$$

The test statistic for the τ test can be written as:

$T = \sum_{i=1}^{N} \widehat{b^2}_i / N$

This is the sample variance of the estimated random effects b_i. This statistic involves the sum of the squared random effect estimates. The null hypothesis is that T=0 (i.e., the random effects do not help to explain variation in the trait). To compute a p-value for T without making assumptions about its distribution, I used a permutation test.

2.3.7 Score testing for gene presence/absence: $\boldsymbol{\beta}$ test modeling

After I have fit the model described above for the τ test, I have estimates of the fixed effect coefficients $\hat{\alpha}$, the random effects $\hat{b}$, and the variance component parameters, $\hat{\phi}$; $\hat{\tau}$. Using these, I constructed a score test for each gene with the null hypothesis $H_0 : \beta = 0$. Suppose G is a $N \times 1$ genotype vector (where N is the number of samples). $\hat{\mu}$ are the probabilities of the samples having the trait (e.g., being cases) given the covariates X and the random effects $\hat{b}$: $P(Y = 1|X, \hat{b})$. Let $\widehat{W}$ be a diagonal vector with elements $\hat{\mu}\,(1 - \hat{\mu}\,)$ and $\tilde{G} = G -$

$X(X^t\widehat{W}X)^{-1}X^t\widehat{W}G$ is the covariate-adjusted genotype vector. With $\hat{\Sigma} = \widehat{W^{-1}} + \hat{\tau}\psi$ and $P = \widehat{\Sigma^{-1}} - \widehat{\Sigma^{-1}}X(X^t\widehat{\Sigma^{-1}}X)^{-1}X^t\widehat{\Sigma^{-1}}$ and a working vector $\hat{Y} = X\,\hat{\alpha} + \widehat{b_\iota} + g'(\hat{\mu})(y - \hat{\mu})$, the score test statistics, assuming $\hat{P}\tilde{G} = \hat{P}G$ is:

$$T = G^t(Y - \hat{\mu}) = G^t\hat{P}\hat{Y} = \tilde{G}^t\hat{P}Y = \widehat{\tilde{G}^t(Y} - \hat{\mu})$$

The variance of T is:

$$Var(T) = \tilde{G}W\tilde{G}$$

I estimated this directly for each gene G. As shown in [64] this is approximately equivalent to $\tilde{G}\,\hat{P}\tilde{G}$ but much faster to compute, plus the approximation is conservative.

The effect size $\hat{\beta}$ is the natural log of the odds ratio. We can estimate this using the variance component estimate under the null hypothesis.

$$\hat{\beta} = (\tilde{G}W\tilde{G})^{-1}(\tilde{G}\hat{P}\hat{Y}) = T/var(T)$$

The standard error of $\hat{\beta}$ is $SE(\hat{\beta}) = \left|\hat{\beta}/z\right|$ where z is the z-score corresponding to the p-value divided by 2.

Chapter 3 Improved Detection of Microbiome-Disease Associations Via Population Structure-Leveraged Association Model (microSLAM)

3.1 Abstract

Most microbiome-disease association studies focus on genera or species that vary in abundance with disease status, limiting our understanding of why these microbes act as disease markers and overlooking cases where disease risk is related to specific strains with unique biological functions. To bridge this knowledge gap, we developed “microSLAM” (microbiome Structure-Leveraged Association Model), an R package and a statistical model. This tool performs association tests that connect the presence/absence of genes within species to host traits, while accounting for population structure (i.e., strain genetic relatedness across hosts). Traits can be binary (such as case/control) or quantitative. MicroSLAM is fit in three steps for each species. The first step estimates population structure across hosts. Step two calculates the association between population structure and the trait, enabling detection of species for which a subset of related strains confer risk. To identify specific genes whose presence/absence across diverse strains is associated with the trait, step three models the trait as a function of gene occurrence plus random effects estimated from step two. Applying microSLAM to 710 gut metagenomes from inflammatory bowel disease (IBD) samples, we discovered 49 species whose population structure correlates with IBD. In addition, after controlling for population structure, we found 57 microbial genes that are significantly more common in healthy individuals and 26 that are more common in IBD patients, including a seven-gene operon in *Faecalibacterium prausnitzii* that is involved in utilization of fructoselysine from the gut environment. Overall, microSLAM detected IBD associations for 45 species that were not detected using relative abundance tests, and it identified specific strains and genes

underlying IBD associations for 13 other species. These findings highlight the importance of accounting for within-species genetic variation in microbiome studies.

3.2 Introduction

The human body is home to a complex community of microorganisms, known as the microbiome, which encodes millions of genes [84]. The species composition of the microbiome differs significantly between individuals and is associated with host genetics, diet, immune system, and several human diseases [9,85–87]. As microbiome species evolve, individual lineages lose and gain genes through horizontal gene transfer [26,33] and other processes that create structural variation [34,48,88]. The resulting pangenome can be quantified from shotgun metagenomics data [40,44,45], which has revealed immense genetic diversity between and within human hosts [46]. Even when two people harbor the same microbial species, the cells within those populations are likely to perform different functions [47,48]. For example, prior studies identified many cases of variable virulence and antibiotic resistance [49,50], a set of pro-inflammatory genes from specific strains of *Ruminococcus gnavus* [51], a *Faecalibacterium prausnitzii* GalNAc utilization pathway linked to with cardiometabolic health [52], and a strain of *Escherichia coli* with enhanced ability to live on the intestinal mucus that is associated with IBD [53]. These findings underscore the limitations of using species alone to gain insight into host-microbiome interactions.

We consider two ways to leverage within-species pangenomic diversity to discover associations between the microbiome and a trait of the host, such as disease. The first is designed for when a species has a strain or group of related strains that predicts the trait. Identifying and isolating trait-associated strains facilitates experimental investigations into host-microbiome interactions, and strains enriched in healthy hosts have been proposed as components of probiotics and therapies [17,54–56]. Due to the systematic structure of bacterial

genomes in which many genes have correlated presence/absence across strains–especially closely related strain–this approach will typically identify a large set of trait-associated genes. While any of these genes could be a good biomarker (e.g., for diagnosis or patient stratification), most of them are not good candidates for follow-up studies of causal mechanisms. Therefore, we also consider a second case in which one or a small number of individual genes predict the trait. Such associations are easiest to detect if the genes are rapidly gained and lost (e.g., mobile elements), so that they associate with the trait independently of evolutionary relationships amongst strains. Genes like these are promising candidates for discovering causal mechanisms through which microbes modify host health and treatment responses.

To identify trait-associated microbiome strains and genes, we developed a statistical model that can be used to perform a metagenome-wide association study (MWAS) for any continuous or binary host trait. Building on the work done on generalized linear mixed-effects models from human genetics [64,67,77,89], this modeling approach uses gene presence/absence data from cohorts with metagenomic sequencing to first estimate a between-sample genetic relatedness matrix for each microbiome species and associate this population structure with the host trait. Then, each gene in a species' pangenome is tested for its trait association after accounting for the relatedness of strains across hosts using random effects derived from the relatedness matrix. Our methodology is implemented in an open-source R package, called microbiome generalized linear mixed-effects model (microSLAM), which can be used with quantitative and binary traits (including unbalanced case/control studies), scales to thousands of samples, and has a controlled type one error rate. The two tests in microSLAM enable researchers to detect new associations and to refine associations discovered using relative abundance.

To investigate the utility of microSLAM, we analyzed a compendium of 710 publicly available metagenomes from IBD case/control studies. IBD is an inflammatory condition of the gastrointestinal tract characterized by its persistence [90]. IBD afflicts roughly 3 million Americans

[91], and its incidence has continued to increase in older adults in recent years [92]. The gut microbiome has long-standing links to IBD, including species abundance and gene associations [47,53,90,93–102]. Here, we combined MIDAS v3 pangenome profiling [40] with microSLAM to quantify associations between IBD and relative abundance, population structure, and gene presence/absence across 71 common members of the human gut microbiome. These analyses identified 49 species with IBD-associated population structure and 83 significant gene families, which we interpreted at the pathway level within and across microbiome species. Tests based on relative abundance would have missed these associations.

3.3 Results

3.3.1 MicroSLAM modeling approach

We present a new method, called microbiome population Structure Leveraged Association Model (microSLAM) using generalized linear mixed effects, that enables two complementary statistical tests of association between a host trait and within-species genetic variation (**Figure 3.1A**; Methods). The trait can be quantitative or binary (e.g., case/control). Both tests use the presence/absence of genes from a given species' pangenome across hosts, which can be quantified from metagenomic sequencing data using tools such as MIDAS v3, panX, and Roary [40,44,45]. The microSLAM method is implemented as an open-source R package at https://github.com/miriam-goldman/microSLAM.

The first test, the τ test, identifies species for which trait variation is associated with variation in overall gene content, as quantified by a random effect (b_i) for each host that is estimated using a GRM (1-Hamming distance of gene presence/absence) and quantifies the association between the sample's lineage and the trait. We refer to the output of the τ test as strain-level associations, because many gene families that are jointly present/absent across hosts are all equally likely to be associated with the trait. Identifying strain-trait associations is

important because this improves the precision of research and therapeutics development based on cultured strains beyond simply picking a random strain from a trait-associated species, which may in fact not be one of the strains driving the species-level association. The genes jointly defining trait-associated strains (i.e., those positively or negatively associated with b) also provide signatures that can be used for predictive modeling and potential diagnostics.

However, the many jointly evolving genes that define trait-associated strains will make it difficult to pinpoint causal mechanisms; some of them may play direct roles in the etiology of the trait or in enabling the strain to survive in hosts with that trait, while others are simply present in the same lineage. To address this, microSLAM implements a second test, called the β test, that identifies individual gene families that are significantly associated with the trait above and beyond what is expected given the GRM. This is accomplished by modeling the trait as a function of each gene family's presence/absence with a generalized linear mixed effects model that includes the random effect (b_i) for each sample (Methods). The resulting significant gene families may be recently and/or recurrently lost and gained (e.g., via mobile elements). To be detected, they must evolve somewhat independently of the gene families that distinguish strains and in patterns that strongly associate with the host trait. These are high-confidence candidates for studying causal mechanisms. Going beyond standard species relative abundance tests, microSLAM's two within-species tests are designed to enable (i) identification of specific strains and gene functions driving species-trait associations, and (ii) detection of novel trait associations not detectable at the species level.

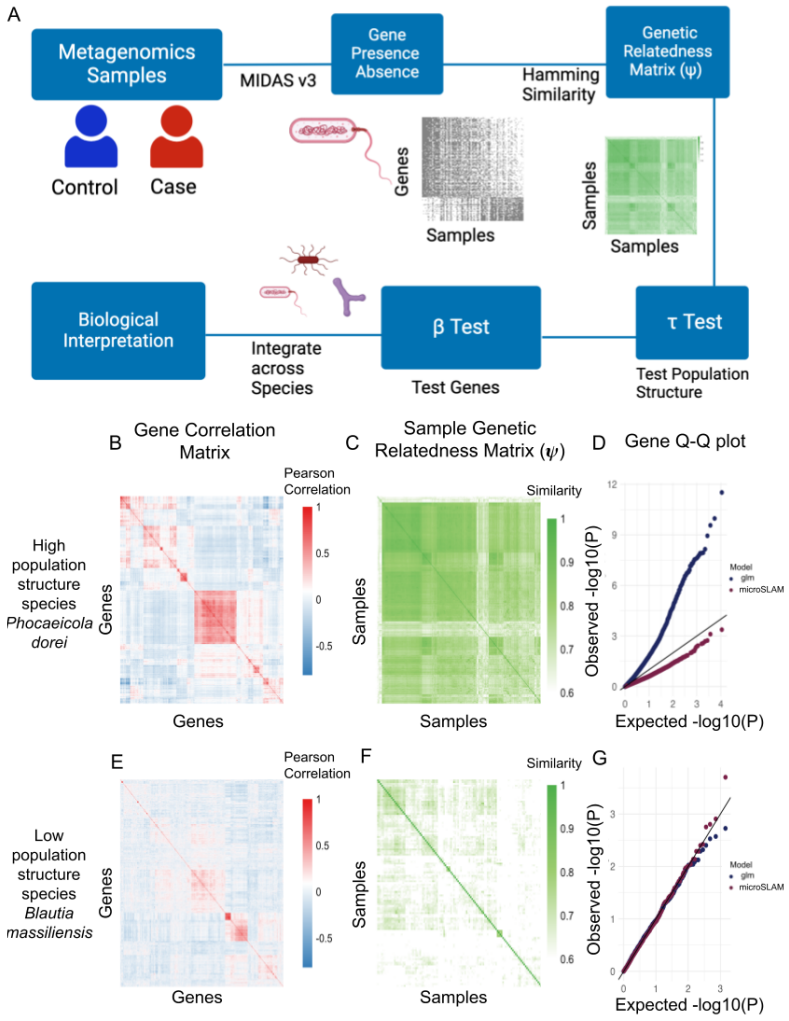

Figure 3.1 MicroSLAM motivation and approach. A) Flow chart of microSLAM modeling approach (diagram created in BioRender). **B – G)** Two bacterial species with different population structures. First row: *Phocaeicola dorei* (260 IBD cases, 218 controls); Second row: *Blautia massiliensis* (73 IBD cases, 44 controls). **B&E)** Heatmap of gene by gene correlation matrix based on gene presence/absence across IBD samples. Red: high positive correlation, Blue: high negative correlation. **C&F)** Heatmap of sample by sample genetic relatedness matrix (1 minus Hamming distance of gene presence/absence profiles). Dark green: high similarity, White: low similarity. **D&G)** Q-Q plot for p-values from tests of association between case/control status and presence/absence of individual genes in the pangenome. Tests are based on micoSLAM and standard logistic regression that does not adjust for population structure (glm). The diagonal line shows expected p-values under the null hypothesis of no association. Pangenome profiling for the metagenomes was done using MIDAS v3.

3.3.2 Population structure in inflammatory bowel disease gut microbiomes

We compiled 710 publicly available gut metagenomes from five inflammatory bowel disease (IBD) case/control studies and performed pangenome profiling of them using MIDAS v3. There were 71 species with sufficient sequencing coverage to analyze within-species genetic variation. After dropping gene families that are nearly always present or nearly always absent (Methods), we had an average of 2,254 gene families per species and a total of ~160 thousand across species. All species showed some population structure. In some species, such as *Phocaeicola dorei (P. dorei),* many gene families are co-evolving and show a high correlation in their presence/absence across hosts (**Figure 3.1B**). In turn, we see two distinct subgroups of strains in the GRM (**Figure 3.1C**). This high level of structure might be the result of selection pressures, drift, or a recent population expansion. When we perform MWAS for all *P. dorei* gene families using logistic regression (glm), we observe that most genes are significantly associated with IBD case/control status (**Figure 3.1D**). This inflation is similar to the well-known problem in human genetics in which ancestry-associated variants are all highly significant when genetic ancestry differs between cases and controls [103]. In contrast, the gene-level test in microSLAM does not show inflation, because our model adjusts for population structure when testing individual gene families for disease associations. We therefore hypothesize that inflation is a consequence of high population structure resulting from a high correlation between gene families. Supporting this, *Blautia massilensis* does not have many genes that are correlated (**Figure 3.1E**) and shows less structure in its GRM (**Figure 3.1F**). Accordingly, the glm p-values do not show inflation, and the microSLAM output is very similar to that of glm.

These results suggest that if we wish to identify individual gene families with unexpectedly high associations with a host trait given the species' GRM, the mixed modeling approach in microSLAM provides a way to adjust for population structure across hosts, just as

mixed models have enabled human geneticists to account for confounding from genetic ancestry. However, population structure is not necessarily a confounder in microbiome research, and it may also be of interest to identify trait-associated strains, defined by the presence and absence of many gene families relative to other strains. These genes would not be significant in the microSLAM β test, because they are highly correlated with population structure. For this reason, microSLAM also includes a strain-level test, the τ test.

Consider, for example, *Ruminococcus B gnavus (R. gnavus),* a species that has long been associated with IBD [51,104,105]. The *R. gnavus* GRM shows two distinct groups when hosts are sorted based on their b_i values (**Figure 3.2A**). One of these groups only contains healthy control individuals, while the other is split between IBD and controls. The b_i estimates better separate cases and controls than do the first two principal coordinates of the gene presence/absence matrix (**Figure 3.2B**). Not surprisingly, when we apply the microSLAM τ test to *R. gnavus*, we obtain a large and statistically significant measure of association (τ=4.67, permutation p-value=0.0001; **Figure 3.2C**), and the resulting model can classify IBD cases with high accuracy (ROC AUC = 0.987; **Figure 3.2D**). Now, if we look at the genes that are most highly correlated to the estimated b_i values of the samples, we identify 238 out of the ~1200 non-core, non-rare genes used in the analysis that are all nearly equally associated with IBD (**Figure 3.2E**). None of these genes are significantly associated with IBD after adjusting for population structure. Thus, we were able to identify several hundred highly correlated genes that form a predictive signature for the *R. gnavus* strains present in IBD patients versus controls. These observations illustrate the importance of including the τ test in microSLAM.

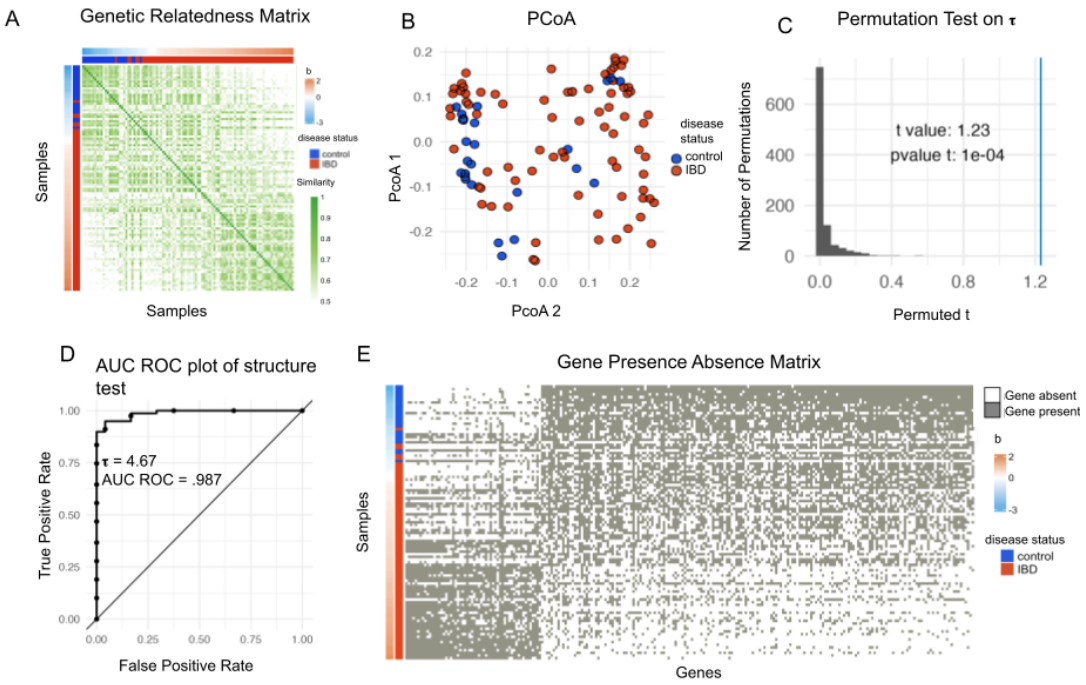

Figure 3.2 MicroSLAM detects both strain and gene associations. A) The GRM for *Ruminococcus B gnavus* with hosts sorted by their estimated *b* values and annotated by their disease status. **B)** PCoA from the *R. gnavus* gene presence/absence colored by host disease status (as in A). **C)** Histogram of permutation test statistics (t-values) from the τ test for *R. gnavus.* The line denotes the observed value of t. **D)** ROC plot for the microSLAM τ test model for *R. gnavus.* The statistic τ quantifies population structure. **E)** Gene presence/absence plot for a subset of genes associated with the random effect *b* for *R. gnavus.* Samples are ordered by *b* and annotated by their disease status.

3.3.3 MicroSLAM controls false positive rates and increases specificity in simulations

The examples in **Figure 3.1** and **Figure 3.2** suggest that microSLAM's τ test can detect strain-trait associations when species have a high degree of population structure and its β test may control false positive gene-disease associations better than a standard glm, albeit somewhat conservatively. But the ground truth is unknown in real data. Hence, we designed a series of simulations to assess the performance of both of microSLAM's tests. Our simulation strategy leveraged the IBD compendium in order to capture the range of patterns observed in real data while varying parameters such as effect size and sample size. For the β test, we

compared microSLAM to glm in order to evaluate the effects of adjusting for population structure via the random effects b_i.

First, we evaluated the τ test. To quantify type 1 error (i.e., false positive rate), we simulated gene presence/absence matrices with core and accessory genes, as well as a set of strain-specific genes, and for each host, a binary trait was simulated independently of the gene presence/absence matrix so that the GRM is not associated with the trait (τ test simulation 1; Methods). We investigated a sample size of 100 hosts, which is on the low end of what we observed for species in the IBD compendium, and repeated the simulation 1000 times, keeping track of how many iterations had a permutation p-value < 0.05. We observed a false positive rate of 0.054, which is very close to the expected value of 0.05. This indicates that the false positive rate of microSLAM's τ test is controlled (**Figure 3.3A)**.

To evaluate the power of the τ test, we modified the prior simulation so that the trait depends on the presence/absence of a particular strain (τ test simulation 2; Methods). We varied the strength of the strain-trait association (odds ratio) and explored sample sizes ranging from 60 to 250. As expected, power increases with the odds ratio and sample size (**Figure 3.3B**). MicroSLAM achieves ~80% power at an odds ratio of 1.5 with 250 samples, whereas an odds ratio greater than 2.0 is needed for similar power with only 60 samples. These results provide practical guidelines for the expected performance of the τ test.

Next, we investigated the type 1 error and power of microSLAM's β test compared to glm. We considered the case where gene presence/absence is not associated with a binary trait, but it is associated with population structure, and hence genes are correlated with each other. To do so, we simulated gene presence/absence using principal components of the observed GRM for each of the 71 species in the IBD compendium (β test simulation simulation 1; Methods). MicroSLAM controlled the false positive rate below 0.05 for all but two species where it is exactly 0.05 (*Dorea A longicatena, Roseburia sp900552665*). In contrast, the glm

without a random effect adjusting for population structure failed to do so for all but 9 species (*Faecalibacterium prausnitzii, Bifidobacterium adolescentis, Bariatricus comes, Blautia A faecis, Faecalibacillus intestinalis, Gemmiger qucibialis, Akkermansia muciniphila, Roseburia sp900552665, Acetatifactor sp900066565*), a failure rate of 87.3% (62/71). In addition, as the estimated τ increased, the false positive rate of the glm dramatically increased while the false positive rate for microSLAM decreased slightly (**Figure 3.3C**).

To explore if this conservative control of the false positive rate affects the power of microSLAM's β test, we performed simulations where 100 true positive genes are added to the previously stimulated genes, meaning that they have a presence/absence pattern that is associated with the simulated trait (β test simulation 2; Methods). We varied the strength of the association (odds ratio) and evaluated power at an empirical false positive rate of 0.05 (calculated using the non-trait-associated genes). These analyses show that microSLAM consistently has either the same or higher power than the glm at the same false positive rate (**Figure 3.3D**), with the difference between methods being most pronounced with a higher number of samples and a high degree of population structure (τ) (**Figure 3.6**). In order to understand exactly what types of genes lead to the higher false positive rate we simulated a data set completely *de novo* without using the IBD compendium (β test simulation 3; Supplement). This showed that the strain-associated genes tended to lead to an increase in false positives for the glm, while microSLAM was able to differentiate the true positives from genes linked to the strain but not directly associated with the trait (**Figure 3.7**). Thus microSLAM's β test has better specificity than does a standard glm.

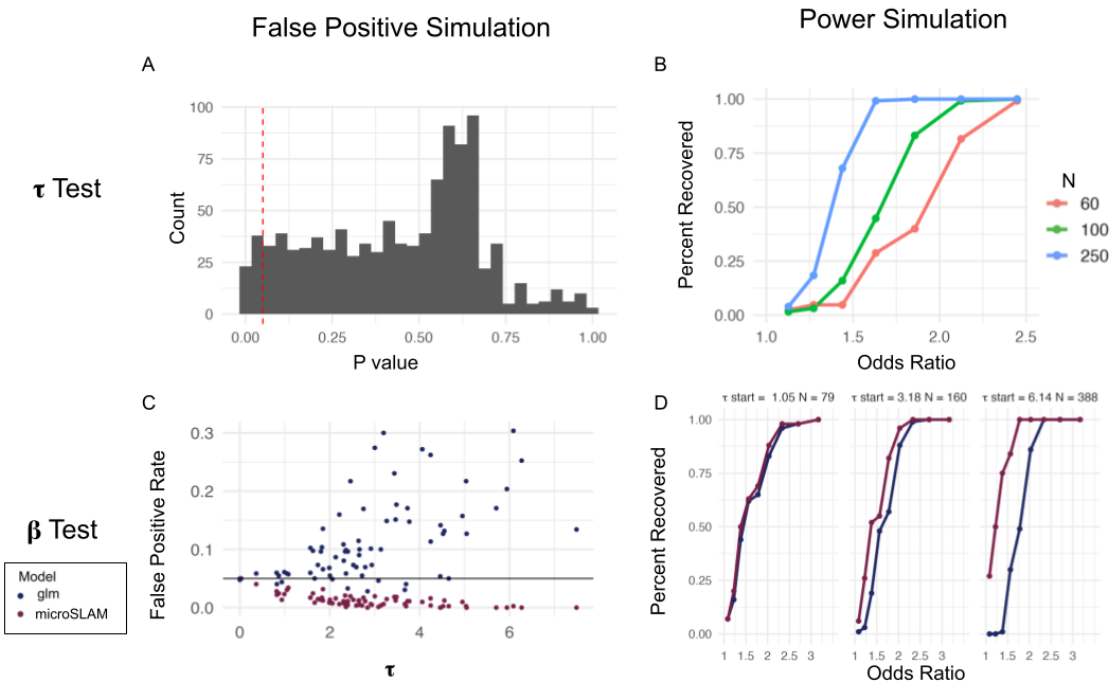

Figure 3.3 Simulations show that microSLAM improves power and false positive rates. **A**) The false positive rates of the τ test of microSLAM were estimated using simulations with varying GRMs but no trait associations. We simulated gene presence/absence and GRMs for the 1000 iterations (τ test simulation 1; Methods). A histogram of p-values for the τ tests shows that the percentage of tests with a p-value < 0.05 is 5.4%. **B)** Power of the τ test for simulations with a range of values for the odds ratio of the simulated y compared to the presence of the trait-associated strain (τ test simulation 2; Methods), repeated for different numbers of samples (N). **C)** False positive rates of the β tests for glm and microSLAM were estimated using simulations with varying levels of population structure (τ) but no trait associations. We simulated gene presence/absence using the GRMs for the 71 species in the IBD compendium (β test simulation 1; Methods). The false positive rate increases with τ for the glm and is generally above the targeted level (0.05; horizontal line), while it decreases and is generally below 0.05 for microSLAM. **D**) Power for 3 simulated species with different τ values and numbers of samples (N). For a subset of genes, presence/absence is simulated based on the trait using a range of odds ratios; other genes have presence probabilities that do not depend on the trait (β test simulation 2; Methods).

3.3.4 MicroSLAM reveals IBD associations across 71 gut microbiome species

We next sought to examine associations in our IBD compendium using microSLAM's population structure and gene tests. First, we performed the standard species-level analysis in which the relative abundance of each of the 71 gut species (quantified using kraken2 and bracken [106–108]; Methods) is tested for association with IBD case/control status using logistic regression, adjusting for host age. We also explored adjusting for study, but found that no species had significant study effects. Medications and other clinical covariates are important confounders but unfortunately were not provided in the publicly available datasets. We found that 13/71 species (18%) had significant relative abundance associations (localFDR<10%; **Figure 3.4A**).

To investigate strains potentially responsible for these relative abundance associations, and to explore the possibility that some species have strain associations without relative abundance associations, we next ran microSLAM using each species' gene presence/absence matrix and corresponding GRM (quantified using MIDASv3 [40]; Methods). We used microSLAM's τ test to identify species whose population structure is associated with IBD case/control status. At localFDR<10%, 49/71 species (69%) were significant, meaning that cases and controls tend to harbor distinct strains consistently across studies. Thirty-four of these species were not detected in the relative abundance test (**Figure 3.4A**). Twenty-seven of the species were only detected from the population structure test, of those 18/27 (67%) are from class *Clostridia*. These are well-powered and case/control matched studies coming from different geographical regions with different diets and lifestyles, as well as different DNA library preparation methods, so detecting strain-disease associations across studies suggests that these associations are truly linked to microbial population structure, rather than an unmeasured confounder (**Figure 3.8)**, though we cannot rule out confounding due to the limited amount of publicly available data

about the study subjects**.** As opposed to simply including a PC from the GRM to represent the structure of the population, the population structure component of the model pulls out from the GRM the cryptic relatedness that can be attributed to the phenotype given the covariates that are included. If we were to have more information about diet or exercise then the population structure component would be the portion of the GRM that can be attributed to the phenotype given the exercise of the host.

In addition to assessing the statistical significance of τ via a permutation test, we also report the area under the curve (AUC) for the receiver operating characteristic (ROC) from the τ test model. This shows how well the population structure component is able to separate the cases from the controls. The AUC is calculated within the same training data because the random effects *b,* which are per-host parameters (i.e., on per subject) generated from the GRM, are unknown for new hosts and hence the fitted model does not generalize beyond the training set. Overall the AUC from the τ tests was quite high; 54 species had AUC over 0.9. Class *Clostridia* tended to have the highest AUC values and a smaller variance in AUC values compared to *Bacteroidia* (**Figure 3.4B**). In addition to *R. gnavus* (**Figure 3.2**), species with significant τ tests included *Agathobacter rectalis* (previously found to be related to IBD under certain conditions [109]) and *Phocaeicola coprocola* (formally *Bacteroides coprocola*, which has been shown to have a relationship with ulcerative colitis [110]). In both of these species, there were no significant genes with the β test, but with the information from the τ test genes differentiating IBD-associated strains can be identified.

To investigate specific gene families associated with IBD case/control status, above and beyond the genes that define IBD-associated strains, we next applied microSLAM's β test. Across the 71 species, 83 genes from 27 species showed significant associations after adjusting for population structure (localFDR<20%, which is the threshold with optimal lift and somewhat more lenient than the 10% threshold used for the other two tests). Seven of these

species did not have significant relative abundance or population structure associations, underscoring the unique information captured by each of microSLAM's tests (**Figure 3.4A**).

Having analyzed IBD associations at the species, strain, and gene level, we integrated these results across the 71 species to look for phylogenetic trends (**Figure 3.4C**). Out of the 71 analyzed gut species, 13 (spread across the phyla *Firmicutes A*, *Firmicutes C*, *Bacteriodota*, and *Actinobacteria*) had no significant IBD associations, possibly due to a lower number of samples (N<100 for ten species). Of the 13 species with relative abundance associations, all were detected on one or both of microSLAM's tests (**Figure 3.4A**), suggesting that relative abundance differences are often accompanied by differences in gene content. Looking across the phylogenetic tree, Lactobacillus species tend to have the least IBD-associated population structure (low values of τ), although there is a subclade of two species with higher τ values. On the other hand, Oscillospirales tend to have high values of τ, and most species in this order do not have any significant genes. Finally, Bacteroidales stands out as the order with the most significant genes (60/83), consistent with species in this order having many mobile and accessory genes [111].

To further explore the functions of genes identified by microSLAM's β test (**Figure 3.4D**), we ran multiple gene annotation pipelines. As expected, most of the 83 significant genes had no functional annotation. For example, 39/83 are in the EggNOG COG category "function unknown". The remaining annotated genes were too few in number to perform well-powered enrichment analyses, but we did note several interesting trends (**Figure 3.4E**). The most common COG, encompassing 11 genes from 8 species, was "replication, recombination and repair". Four genes were annotated as transposases, 22 genes were from a family associated with plasmids (>10% of its members annotated as plasmids), and five genes were from a family associated with phages. With all of these annotations combined, in addition to the understanding these genes are significant beyond the overall population structure of their

species, we conclude that many of the significant genes identified by the β test are likely to be components of mobile elements.

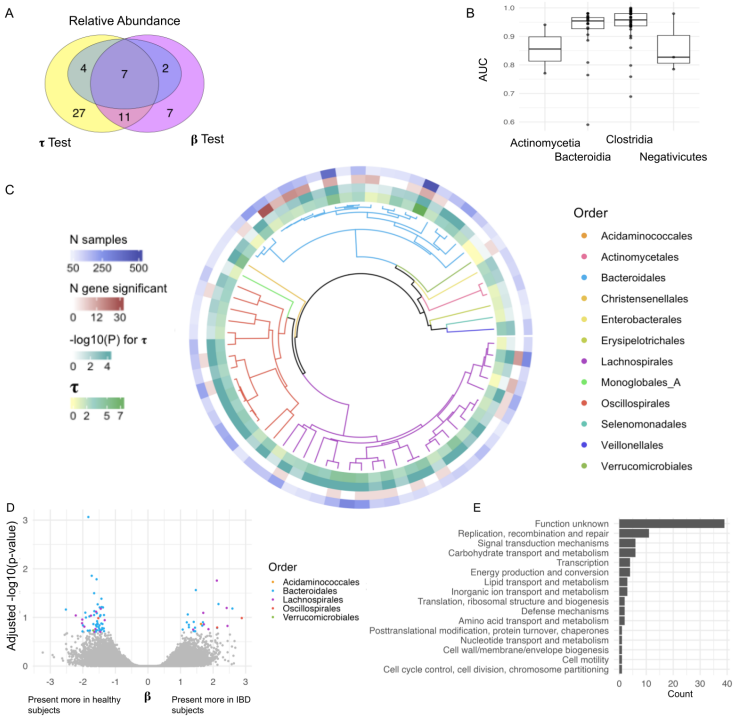

Figure 3.4 MicroSLAM identifies novel IBD associations. We analyzed all 71 species in our IBD compendium for three types of associations with case/control status: relative abundance (kraken2+braken: amount of the species predicts disease), population structure (microSLAM τ: strain predicts disease), and gene family (microSLAM β: gene presence/absence predicts disease). **A)** Venn diagram showing the number of species with significant IBD associations of each type. For genes, we counted the species if it had at least one significant gene family; species varied in the number of hits. All tests are localFDR adjusted for multiple testing. **B)** Boxplots showing the AUC ROC from τ test models for all 71 species, stratified by bacterial class. **C)** UHGG species tree for all 71 species, colored by order. The τ value, p-value for τ test, number of significant genes, and number of samples for each species are plotted in the outer rings. **D**) Volcano plot for β tests with significant genes (localFDR < 0.2) colored by bacterial order. **E**) Bar plot of COG categories for the 83 genes with significant β tests.

3.3.5 Seven-Gene GRF operon is a structural variant in *Faecalibacterium prausnitzii*

One significant gene identified by microSLAM was annotated as "subunit D of the fructoselysine/glucoselysine phosphotransferase (PTS) system" by BlastKOALA [112]. It was negatively associated with IBD case status in *Faecalibacterium prausnitzii D* (UHGG species id 102272), a species whose relative abundance is positively associated with IBD in our compendium. This hit intrigued us because *F. prausnitzii* is a well-studied bacteria with roles in short-chain fatty acid metabolism and inflammation [113,114]. Predicting a new molecular mechanism underlying this host-microbe interaction would enable future functional studies (e.g., in gnotobiotic mice) and potentially could be useful for developing diagnostics, dietary interventions, or other therapies.

To explore this gene family, we first compiled 85 high-quality and diverse *F. prausnitzii* genome assemblies from NCBI and clustered them into eight clades (**Figure 3.5A**; Methods). We observed that seven genes (plus occasionally an eighth gene) were consistently found together, with a conserved order and orientation across 53% of the NCBI *F. prausnitzii* genomes (49/85) (**Figure 3.9**). Annotations suggest that these genes encode a fructoselysine/glucoselysine PTS system operon. Having established that this operon is variably present across distantly related *F. prausnitzii* strains, we expanded our search to include all high-quality *F. prausnitzii* genomes available in the United Human Gut Microbiome Database (UHGG v2) [84]. This analysis indicated that the complete seven-gene operon is present in all nine *F. prausnitzii* clades, with between ~3% and ~24% of genomes per clade containing the operon (**Table 3.1**).

Table 3.1 A seven-gene operon present in all nine *F. prausnitzii* clades in UHGG v2.

Species ID	Species name	Annotated as plasmid	Not annotated as plasmid	Total genome counts	Proportion with operon present
100022	*F. prausnitzii C*	30	91	1258	0.096
100039	*F. prausnitzii H*	1	5	221	0.027
100195	*F. prausnitzii E*	6	4	244	0.041
101255	*F. prausnitzii F*	2	4	55	0.109
101300	*F. prausnitzii*	119	223	1446	0.237
102272	*F. prausnitzii D*	88	177	1280	0.207
102274	*F. prausnitzii I*	5	0	179	0.028
102545	*F. prausnitzii G*	124	127	2891	0.087
102619	*F. prausnitzii J*	25	32	236	0.242

Since genes can be syntenic without being functionally related, we conducted further analysis to determine the relationship between the seven-gene operon in *F. prausnitzii* and the well-characterized GFR operon in *Salmonella Typhimurium 14028s* [115]. We successfully mapped five genes from the *F. prausnitzii* operon to the corresponding *S. Typhimurium* operon (*gfrABCDF*) (**Figure 3.5B**). Notably, the *gfrE* gene, encoding a deglycase that cleaves glucoselysine 6-phosphate, is absent in *F. prausnitzii*. In addition, the operon in *F. prausnitzii* includes a gene without homology to any gene in the GFR *S. Typhimurium* operon. The regulatory genes, which are located at the start of the operons, also differ between the two species. We hypothesize the seven-gene *F. prausnitzii* operon identified in our analysis functions solely as a fructoselysine PTS system (**Figure 3.5C**).

Fructoselysine is a spontaneous product of Amadori rearrangements, a chemical reaction between amino acids, sugars, and heat that takes place in our food, and its presence in the human gut environment can promote the growth of bacteria capable of importing and using this carbohydrate as an energy source [116,117]. In *F. prausnitzii*, this is performed by a seven-gene operon encoding proteins that phosphorylate the substrate while transporting it across the bacterial cell membrane, making it available as a source of carbon. While only subunit D of this operon was significant in microSLAM's gene test after accounting for multiple comparisons, all

of the genes in the operon had unadjusted p-values less than 0.05 and flanking genes in the *F. prausnitzii* reference genome did not (**Figure 3.5D**). Variability in gene detection from shotgun metagenomics data is a likely source of the difference in significance for subunit D versus the other genes. MicroSLAM analysis for three other *F. prausnitzii* species in the IBD compendium did not yield any significant genes, primarily due to inadequate sample sizes, which restricts the statistical power of microSLAM's β test. Nonetheless, the gene presence/absence matrices for these species are consistent with the operon being variably present and depleted in IBD cases.

Altogether, these results suggest that the genes in this *F. prausnitzii* PTS operon are co-evolving in terms of their presence/absence across hosts, potentially independently of neighboring genes, and that the presence of this operon is more common in healthy hosts. Our data also suggest that the fructoselysine PTS system operon could be a mobile genetic element. Supporting this possibility, many NCBI and UHGG contigs carrying this operon are predicted by geNomad [118] to be plasmids (**Table 3.1**). We also observe sequences associated with mobile elements and horizontal gene transfer (HGT) in the genomic context surrounding the operon. These are computational predictions only, and no plasmids have been previously reported in *F. prausnitzii*. We therefore checked for the operon in the other 70 species in our microSLAM analysis, detecting it in strains from two other phyla: *Gemmiger qucibialis* (species_id 103937) and *Faecalibacterium sp90053945* (species_id 103899). It is also known to be present in *Escherichia coli*, *Bacillus subtilis* and *Agrobacterium tumefaciens Ti plasmid* [117]. While not conclusive, these data are consistent with HGT. Regardless of the mechanism of acquisition or loss, the variable presence/absence of the operon across *F. prausnitzii* strains indicates that certain lineages of this species can acquire and utilize fructoselysine, thereby enhancing their adaptability and competitiveness in the dynamic gut ecosystem relative to strains without the operon.

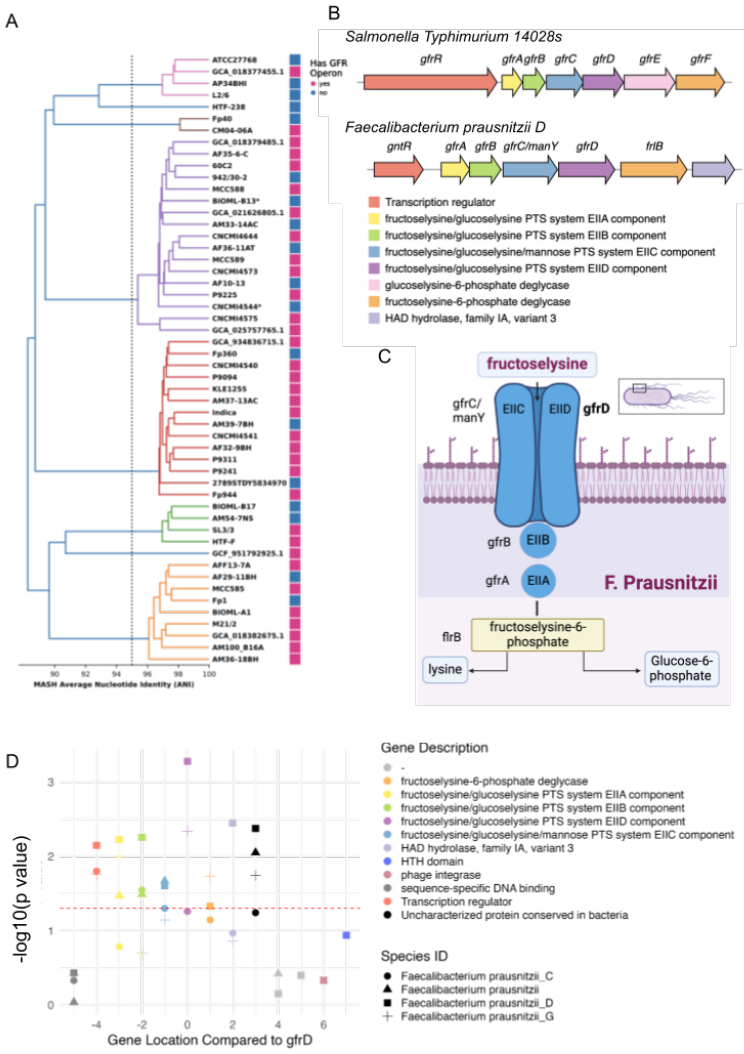

Figure 3.5 Investigation of *F. prausnitzii* fructoselysine PTS system operon. A) 52 representative genomes selected from NCBI and colored by the dRep secondary cluster (Selection of *Faecalibacterium prausnitzii* genomes**,** Methods). **B)** Comparison of *S. Typhimurium* operon to operon in *F. prausnitzii D*. **C)** Graphic of the *F. prausnitzii* fructoselysine PTS system operon and its products (made in BioRender). **D)** P-values for *F. prausnitzii* fructoselysine PTS system operon genes in microSLAM β tests across the four *F. prausnitzii* species defined by UHGG. The flanking genes are much less significant than the genes within the operon. Subunit D (most significant gene in microSLAM analysis) is located at 0, and all other indices are relative to this gene.

3.4 Discussion

In this paper we introduce microSLAM, a method that implements population structure-aware metagenome-wide association studies. Using a generalized linear mixed modeling approach, we are able to include information about the genetic relatedness of the microbiomes within diverse samples and to model the association of this population structure with host traits, while adjusting for other covariates. We focused on case/control study designs here, but microSLAM can also be applied to quantitative traits. In addition to testing if population structure itself (i.e., specific strains) are associated with a trait, microSLAM also includes a test aimed at identifying trait-associated genes that are evolving somewhat independently of strain lineages. Through realistic simulation studies, we demonstrated that microSLAM controls Type 1 error and has reasonable power in cohorts with more than one hundred samples. Compared to standard glm, microSLAM's gene-level β test controls false positives much more effectively, especially in species with notable population structure. When there is a significant population structure as well as a subset of genes that are more related to the phenotype than the strain signal, we showed that microSLAM increases specificity compared to glm. By providing microSLAM as an open-source R package, we provide a new tool for researchers to probe microbiome-host interactions with strain- and gene-level resolution.

In this study, we also put together a metagenomic compendium of IBD samples. Analyzing this data with microSLAM, we discover a wide variety of population structures within human gut metagenomes. We identified 49 species with a population structure related to IBD. In addition, after adjusting for population structure, 57 microbial genes are significantly enriched in healthy subjects, and 26 are enriched in IBD patients. From the genes enriched in healthy subjects, we identified a *F. prausnitzii* fructoselysine PTS system operon that is present in all clades of this species, but in only a minority of genomes within each clade, suggestive of being a mobile genetic element or other rapidly lost/gained structural variant. The presence/absence

of this operon may confer distinct metabolic advantages to different strains, including the ability of carriers to utilize fructoselysine as an energy source [116]. The potential impact on human health could be significant, given that *F. prausnitzii* is one of the most prominent butyrate producers in the human gut [16,93,119]. This might also lead to a greater resilience of the gut microbiota, offering enhanced protection against pathogenic bacteria and reducing risk of chronic disease. Therefore, future work aimed at understanding the mechanisms through which *F. prausnitzi*i acquires and disseminates this seven-gene operon is not only key to comprehending microbial ecology but also crucial for potential dietary or probiotic therapeutic interventions targeting the microbiome.

There are several limitations to the microSLAM method and implementation. First, we estimate the GRM using the same gene presence/absence data that we then used to test genes for their trait-associations. This approach has been used with mixed modeling with human genetic data, and has been shown to reduce power compared to estimating the GRM with an independent set of markers (e.g., variants on other chromosomes) [120]. We explored a similar approach by using single nucleotide polymorphisms (SNPs) in core genes for GRM and random effect estimation in the β test. But we found that for almost all species in our IBD compendium, the SNP data generated a GRM that was very different from the gene-based GRM, and hence SNPs were not good markers for estimating the population structure in gene presence/absence. Perhaps this approach could work with more investigation into how to pick SNPs for GRM estimation or with a different GRM distance metric.

Second, as is the case with any meta-analysis, we used samples collected from many studies, located in a variety of geographic locations. Publicly available metagenomics data rarely includes detailed information about potentially confounding variables, such as diets and medical care. Hence, these important covariates are not accounted for in our models. This means that there is a chance some of the significant τ tests were related to an unmeasured variable that happened to be associated with strain genetic differences. Since each study

included both cases and controls, and we did not observe any strong correlations between study and microSLAM's random effect estimates, we do not consider this a major problem. Nonetheless, with so little meta-data it is important to acknowledge that the strain-disease associations we detected across studies could be confounded by unmeasured variables (e.g., diet that selects for certain strains and alters IBD risk). In the future, when studies include more covariates, microSLAM can adjust for these just as we adjusted for age (the only consistently reported covariate) in this project. Beyond confounding, more complete metadata also would be helpful for understanding the capabilities of our method and for functionally interpreting our IBD findings.

Third, we do not find many individual significant genes within this study. This could be partially due to lack of power, especially for 55/71 species with ≤ 160 samples. A much larger dataset would increase our ability to find IBD associations for strains and genes. It would also enable separate modeling of associations for subtypes of IBD, which may have different microbiome signatures 8. Our simulations suggested that most species did not have sufficient power for separate Crohn's disease and ulcerative colitis models in our IBD compendium. We did investigate if any microSLAM discoveries were mostly driven by one subtype or the other, and we observed very few examples of this one significant gene cluster GUT_GENOME040547_00268 from species *Phascolarctobacterium faecium* was found to be positively associated with the disease but was only present in CD patients. In addition, there were three species with a significant τ where less than 1/3rd of the IBD patients are labeled as UC and there are less than 10 UC patients (*CAG-180 sp000432435, Ruminiclostridium E siraeum, and UBA11524 sp000437595*) meaning most of the signal is from CD alone in those species. While this finding could indicate that most associations are truly shared, it is more likely that we only had sufficient power to detect associations supported by both subtypes and that other subtype-specific associations remain to be discovered in the future with larger individual cohorts. As researchers move towards testing for strain and gene associations in studies with

hundreds or thousands of samples, microSLAM's improved specificity and controlled Type 1 error rate, as compared to glm, will be even more important.

Fourth, it is possible that some of our discoveries were driven by metagenomic reads from the wrong species or gene creating a false signal of gene presence (or absence). This cross-mapping is a frequent issue in read-mapping-based genomic analysis, especially for closely related species or highly conserved genes [121]. Hence, we recommend validating microSLAM's gene test results with complementary data. For example, in our investigation of the PTS operon, we confirmed the operon structure and variable presence across strains using high-quality genome assemblies. This, plus the fact that this operon is predominantly found within *F. prausnitzii* and not widely distributed in other species, substantially alleviates concerns about cross-mapping in this analysis.

Finally, many of the genes we identified were not annotated, leading to difficulty completing in-depth analyses of significant genes across species. For example, we lacked power to perform functional enrichment analyses despite seeing several consistent trends, such as mobile elements being discovered in multiple species. More gene annotations would help with this problem, but annotation alone is not enough to confirm function. We view microSLAM as an important first step for proposing candidate causal strains and genes that should be performed upstream of in vitro and in vivo experiments to test the hypothesized functions of the discovered strains and pathways. The ability of microSLAM to detect associations for species whose relative abundance is not correlated with host traits and to accurately disentangle associations of individual genes versus groups of strain-defining genes make it a useful new hypothesis-generating tool for microbiome research.

3.5 Methods

3.5.1 Compendium of IBD/healthy case/control metagenomic studies

We compiled a total of 2625 publicly available paired-end shotgun metagenomic samples, sourced from five studies related to either inflammatory bowel disease (IBD) or the Human Microbiome Project (HMP2) and having an average read count greater than 20 million (accession numbers: PRJNA400072, PRJNA398089, PRJEB15371, PRJEB5224 and PRJEB1220). A stringent sample selection process was implemented to ensure (1) all samples included comprehensive metadata, such as disease status, age, and antibiotic usage; and (2) only one sample was selected per subject, considering that multiple time points could have been sequenced from the same subject. Specifically, for multiple time point samples from HMP2, we adopted the same selection criterion used by (Lloyd-Price et al., 2019) – selecting week 20 or greater for all subjects, the maximum read count for healthy subjects and the time point with the highest dysbiosis score for IBD patients. The first time point was chosen for the MetaHit project (Almeida et al. 2021; Nielsen et al. 2014; Li et al. 2014). Ultimately, a total of 710 samples met these criteria and were included for downstream analysis.

3.5.2 Bioinformatics analysis

Preprocessing of the downloaded metagenomic sequencing libraries was performed using a QC pipeline that includes the following steps: (1) Adapter removal and quality trimming: adapter sequences were removed and low-quality reads were trimmed using Trimmomatic (v 0.39) [122]. (2) Human contamination removal: reads were aligned against the complete human reference genome (CHM13v2.0) [123] and a collection of 2250 genomes known to be contaminated by human sequences [124], using Bowtie2 (v 2.5.1) [125] to identify and remove human contamination. (3) Low-complexity read filtering: low-complexity reads were filtered out using BBduk (v 37.62) [126]. This step involved removing reads with an average entropy less than

0.5, with entropy k-mer length of 5 and a sliding window of 50 (parameters: entropy=0.5, entropy window=0.5, entropyk=4). Additionally, reads shorter than 50 base pairs (bp) post-filtering were removed (parameters: minlen=50). (4) Quality reporting: a quality report of the cleaned-up reads was generated with FastQC (v 0.12.1) [127]. After preprocessing, samples with read counts lower than 1 million were removed, resulting in 710 high-quality samples retained for further analysis.

3.5.3 Pangenome profiling using MIDAS v3

To determine which species within the 710 samples are both prevalent and sufficiently abundant for pangenome profiling, we implemented a two-step analysis using MIDAS v3 [40]. In the first step, we quickly scanned each sample to detect the presence of species by assessing the vertical coverage of 15 universal single-copy marker genes across 3956 distinguishable species in the UHGG v2 database [84]. In the second step, we adopted a whole-genome read alignment-based methodology to quantify the abundance of each species. This involved running MIDAS's single-nucleotide variant (SNV) pipeline for species that meet specific criteria: a median marker coverage of at least 2X and at least 50% of the marker genes uniquely covered. These steps ensure that our whole-genome based species abundance estimation analysis is restricted to species with substantial coverage across their genomes (horizontal coverage > 0.4, vertical coverage > 5). We further excluded sample-species pairs where the ratio of genome-wide vertical coverage to single-copy marker gene coverage exceeded 4, which helps us to eliminate potential false positives caused by cross-mapping of reads among closely related species and conserved gene families. This stringent criterion also improves computational efficiency [121]. After implementing the aforementioned filtering, 71 species that were present in more than 60 samples and met the abundance criteria were selected for subsequent pangenome profiling analysis. There were 619 samples with at least one species present.

To perform pangenome profiling, we utilized the Genes module from MIDAS v3 [40], which features careful curation of the pangenome database and comprehensive functional annotation. Specifically, a single Bowtie2 index was built for all 71 species, and QC-ed paired-end reads for each sample were aligned to this index. Our analysis included only genes covered by at least 4 reads (--read_depth 4). Genes with an estimated copy number greater than 0.4 were classified as present (--min_copy 0.4). This threshold was selected based on exploratory data analysis and simulations previously performed by our lab. The resulting sample-by-gene presence/absence binary matrix was then used for subsequent association analysis with microSLAM, excluding core genes (absent in less than 10 samples) and rare genes (present in less than 30 samples).

3.5.4 Generalized linear model

A standard generalized linear model (glm) [128] was fit for all genes for all species that were analyzed with microSLAM. This is a logistic regression model using *glm* in R with case/control status as the outcome, gene presence/absence as a predictor, and age as a covariate.

3.5.5 τ Test simulations

We performed simulations to assess the false positive rate and power of microSLAM's τ test. To assess the false positive rate (simulation 1), we set up a simulation where a binary trait was generated independently of GRMs. For each of 1000 iterations and n=100 samples, the trait *y* was simulated using a binomial distribution with a success probability of 0.5, and a covariate was simulated with a normal distribution centered at 45 with a standard deviation of 15, similar to the age distribution in our IBD compendium. Next, for each iteration, a gene presence/absence matrix was simulated with p=1000 genes. This included 400 "core" genes

simulated from a binomial distribution with a success probability of 0.8, 400 “accessory” genes simulated from a binomial distribution with a success probability of 0.2, and 200 genes simulated based on presence of a strain unrelated to the trait *y*, as follows. The strain’s presence/absence across samples was simulated using a binomial distribution with a success probability of 0.5, and then presence/absence for each of the 200 genes was set to absent if the strain was absent and simulated from a binomial distribution if the strain was present, where the success probabilities were chosen such that the average odds ratio of a given gene being present if the strain is present is 1.8. After the genes are simulated, the GRM is calculated and the population structure test is run with 100 permutations. The p-value is calculated for each iteration as the number of permutations with a more extreme T statistic than the observed T statistic. The false positive rate is calculated as the number of iterations with a p-value <0.05.

The power test (simulation 2) is carried out in a similar fashion, with two key changes. First, we explored a range of sample sizes (n=60, 100, 250) to assess the relationship between sample size and power. Second, we simulated the trait y based on presence/absence of the simulated strain. Specifically, we explored a range of effect sizes, quantified with an odds ratio $\theta/(1-\theta)$ ranging from 1.0 to 2.5. For a given odds ratio, we set $strain_{\delta} = (1 - strain) * (1 - \theta) + stain * \theta$ and then generated the trait y_{strain} using a binomial distribution with success probability equal to $strain_{\delta}$: $y_{strain} = rbinom(n, 1, strain_{\delta})$. This creates a stronger relationship between the trait and presence of the strain as the odds ratio increases. For each odds ratio and sample size, 125 iterations were run and power was calculated as the proportion of iterations with a significant τ test divided by 125.

3.5.6 **β** Test Simulations

We performed simulations to assess the false positive rate and power of microSLAM’s β test versus a standard glm. To assess false positives (simulation 1), we generated data in which

no genes were associated with the trait, so that all genes are false positives. We computed a p-value for each gene using each modeling approach and tracked the proportion of genes with p<0.05. In order to simulate real population structure, while introducing some random variation, we used the observed GRM for each of the 71 species in our metagenomic compendium to help generate simulated gene presence/absence matrices. Specifically, we first decomposed the observed GRM for each species into its first 10 principal components (PCs). We then standardized each PC by dividing each value by the PC's standard deviation $PC_{std} = PC/sd(PC)$ and computed the standard normal probability for each sample's loading on each standardized PC (one per sample *i* per PC dimension *j*): $p_{i,j} = pnorm(PC_{std})$. The probabilities $p_{i,j}$ retain relationships between samples across the 10 dimensions. For each of the 10 PCs, we simulated the presence/absence of 90 genes using a binomial distribution with a success probability equal to the sample's $p_{i,j}$ for that PC, for a total of 900 genes correlated with one dimension of the population structure. We also simulated 100 uncorrelated genes using a binomial distribution with a success probability chosen from a uniform distribution between 0.2 and 0.8. From the resulting 1000 x n gene presence/absence matrix, we simulated a binary trait (*y)* using the first two PCs (PC1 and PC2), as follows. We set *y* equal to one in a given sample if its loadings on PC1 and PC2 had opposite signs (either PC1>0 and PC2<0 or PC1<0 and PC2>0). This created a nonlinear relationship between *y* and 180 of the simulated genes (**Figure 3.11**, **Figure 3.6**).

For each simulated *y* and gene presence/absence matrix, we ran microSLAM to compute a GRM and estimate population structure (τ). These new τ's were different from the species' observed τ values and greatly varied across species, as in the observed data (**Figure 3.12**). After the GRM was calculated, and the τ test was run, both glm and microSLAM's β test were run to test for gene-trait associations. The false positive rate was determined by summing

the number of genes with p-values < 0.05 divided by the total number of genes, excluding genes simulated from PC1 or PC2 (p=820 genes).

To assess power for the β test (simulation 2), we start with the data from simulation 1 and set $y_\delta = (1-y)*(1-\theta)+y*\theta$ for an odds ratio of $\theta/(1-\theta)$. Then, we generated presence/absence for 100 additional genes from a binomial distribution with success probability y_δ: $G\_y = rbinom(n, 1, y_\delta)$. These genes G_y are positives (associated with y) and all other genes are negatives (independent of y). At $\theta = 0.5$ (i.e., an odds ratio of one), the generated genes *G_y* will not be associated with the trait. At $\theta = 0.55$, the average odds ratio will be 1.2. We investigated θ values between 0.52 and 0.78. We checked, and the odds ratios across simulations with the same θ value did not deviate more than 0.1 from the expected values.

We ran glm and microSLAM on each simulated dataset. As expected, the population structure test yielded estimated τ values that increase notably with the simulated odds ratio (i.e., as the association between y and the genes *G_y* increases). In order to assess power, we used the negative genes to establish a significance threshold for each modeling approach such that the empirical false positive rate would be no more than 0.05. Applying these thresholds to the positive genes *G_y*, we computed power as the proportion of positive genes detected. Power was compared between glm and microSLAM across odds ratios and species (each with different sample sizes and GRMs).

For the β test simulation 3, we sought to generate gene presence/absence matrices, trait values, and age values without using any real sample data. First, we simulated a trait from the binomial distribution with a success probability of 0.5. We assume that there are two strains, one that is correlated with the trait ($odds\ ratio = 2.22$) and one uncorrelated with it. For the correlated strain, we simulated 300 genes at a presence level of 0.5 and an odds ratio of 4.0 for the gene being present given a person has the strain. We simulated one-third of the samples having the uncorrelated strain and modeled this with 250 genes with low binomial presence

rates ($p = 0.3$) and an odds ratio of 4.0 for the gene being present given a person has the strain. Then we modeled 300 genes that were "core" across all samples; these were drawn from a binomial with a success probability of 0.8. We additionally simulated 150 genes that were non-strain associated "accessory" genes; these were drawn from a binomial success probability of 0.2. Last, we simulated at least one gene (G_y) that is even more highly correlated with the trait than is the correlated strain ($odds\ ratio = 2.44$). The more genes in G_y (we investigated 1, 2, or 3 genes) and the stronger the relationship between G_y and the trait, the higher the parameter τ will be. The resulting gene presence/absence matrices naturally have a range of different values of τ. Age was randomly generated with parameters similar to the IBD data: $ceiling(rnorm(N, mean = 45, sd = 15))$. We repeated Simulation 3 with the number of samples varying from 60 to 250.

3.5.7 Relative abundance test

We calculated the relative abundance of each species by downloading the UHGG v2 kraken database from MGnify [83] and running Kraken2 [106,108] with options *--paired --minimum-hit-groups 3* and then bracken [106,107] with options *-l S -t 1000.* We computed relative abundance as a given species' bracken coverage divided by the total coverage, and we removed species with less than 0.05% relative abundance. We then performed logistic regression, using the case/control label as the dependent variable (y) and relative abundance as the independent variable, with age as a covariate. The estimated log odds ratios and p-values from these logistic regression analyses were compared with outputs from the microSLAM τ test, after using localFDR to adjust p-values at a 0.1 level.

3.5.8 Identification of seven-gene GFR operon in *F. prausnitzii*

The gene *grfD* from *F. prausnitzii* (UHGG species id 102272) reported as significant by microSLAM's β test corresponds to the EIID component and is part of a putative GFR operon that encodes the Fructoselysine PTS system. A similar gene in *S. Typhimurium 14028s* has been identified as responsible for the utilization of fructoselysine [115,129]. To determine whether this EIID gene is part of a gene cluster that forms an operon - that is, genes that are sequentially arranged on the chromosome and co-regulated - we conducted the following analysis. First, we retrieved the neighboring genes upstream and downstream of this EIID gene in the UHGG reference genome for this *F. prausnitzii* species, considering up to five genes in each direction. We then used blastn [130] to identify homologous regions in three different sets of genomes: (1) 85 NCBI *F. prausnitzii* genomes, (2) all metagenome-assembled-genomes (MAGs) in the nine UHGG *F. prausnitzii* species clusters, and (3) all MAGs in the 71 species that were investigated in this study. These five genes from *F. prausnitzii* were also aligned to the corresponding *S. Typhimurium* operon (gfrABCDF) using Blastp [130]. All genomes were annotated using Prokka [131]. Genes were annotated with BlastKOALA [112] and eggNOG-mapper [132].

3.5.9 Selection of *Faecalibacterium prausnitzii* genomes

To avoid incorrectly assessing the seven-gene GFR operon as incomplete in an assembly simply due to fragmented contigs, we only selected high-quality *F. prausnitzii* genomes with assembly levels of scaffold, chromosome, or complete genome, and we specifically excluded atypical genomes. In total, we downloaded 105 genomes of *F. prausnitzii* from NCBI (using the taxon identifier 853). We assessed the genome quality using CheckM [133], and retained only genomes that met the following criteria: completeness >= 90, contamination <=5, and strain heterogeneity <= 10. After this filtering, 85 *F. prausnitzii* NCBI genomes were retained for the GFR operon screening analysis (**Figure 3.9**). Next, we used dRep [134] to perform

pairwise genome comparisons based on Average Nucleotide Identity (ANI). This dRep analysis involved first clustering all the genomes using the Mash heuristic for ANI [135] and subsequently using MUMMER [136] to compute ANI on sets of genomes that have at least 90% Mash ANI before performing a secondary clustering. As a result, 52 secondary clusters were formed at 98% MUMMER ANI (-comp 90 -con 5 -pa 0.95 -sa 0.98 -nc 0.8). Hierarchical clustering of the 52 genomes using average linkage was performed using the pairwise MASH similarity matrix (`scipy.cluster.hierarchy` package).

In addition to NCBI Genomes, we also collected all nine *F. prausnitzii* species clusters from UHGG v2 [84], using the same selection criteria to ensure assembly quality. We calculated the pairwise genome similarity between the resulting 52 *F. prausnitzii* genomes and the nine representative genomes from UHGG *F. prausnitzii* species using fastANI [137]. We also compared each NCBI *F. prausnitzii* genome to the nine representative genomes from UHGG, and each NCBI genome was assigned to the UHGG species cluster with the highest ANI (ANI >= 95%). If this similarity level was not reached, the NCBI genome remained unassigned. Eight *F. prausnitzii* species in UHGG were represented by the 52 NCBI *F. prausnitzii* genomes.

3.6 Supplement

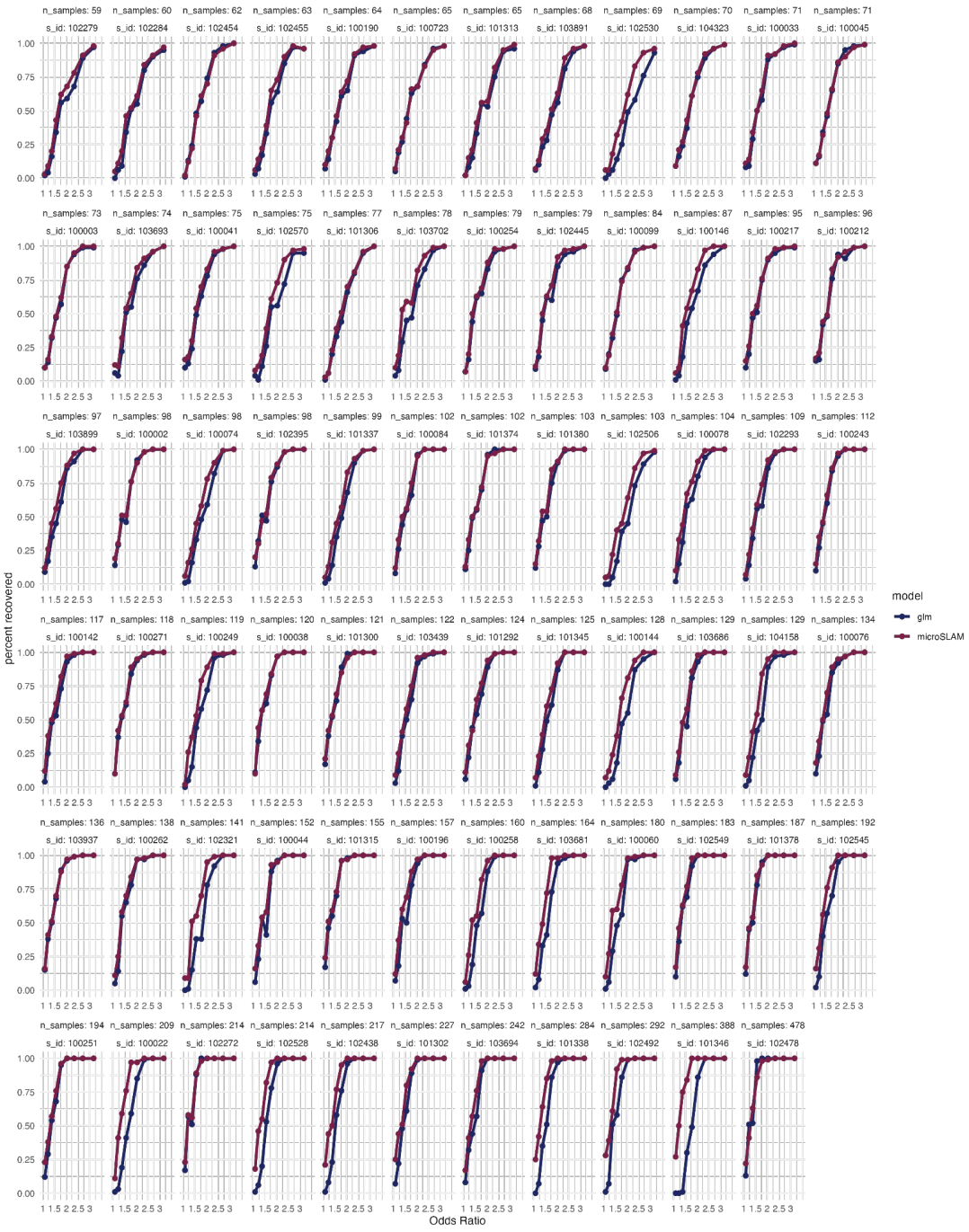

Figure 3.6 GLM and microSLAM β test power evaluations for 71 simulated species. These plots show estimated power of β tests using data from simulation 2 in which a gene presence/absence matrix and binary trait were simulated based on the observed GRMs from the 71 species in the IBD compendium using a range of different effect sizes (odds ratios, horizontal axes). There is one panel per GRM (labeled with species ID), (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) and panels are ordered from lowest to highest sample size. Power was computed as the proportion of positive genes discovered at an empirical localFDR of 0.05 for both microSLAM (*red*) and glm (*blue*). As the number of samples increases there tends to be a larger difference between the glm and the microSLAM models.

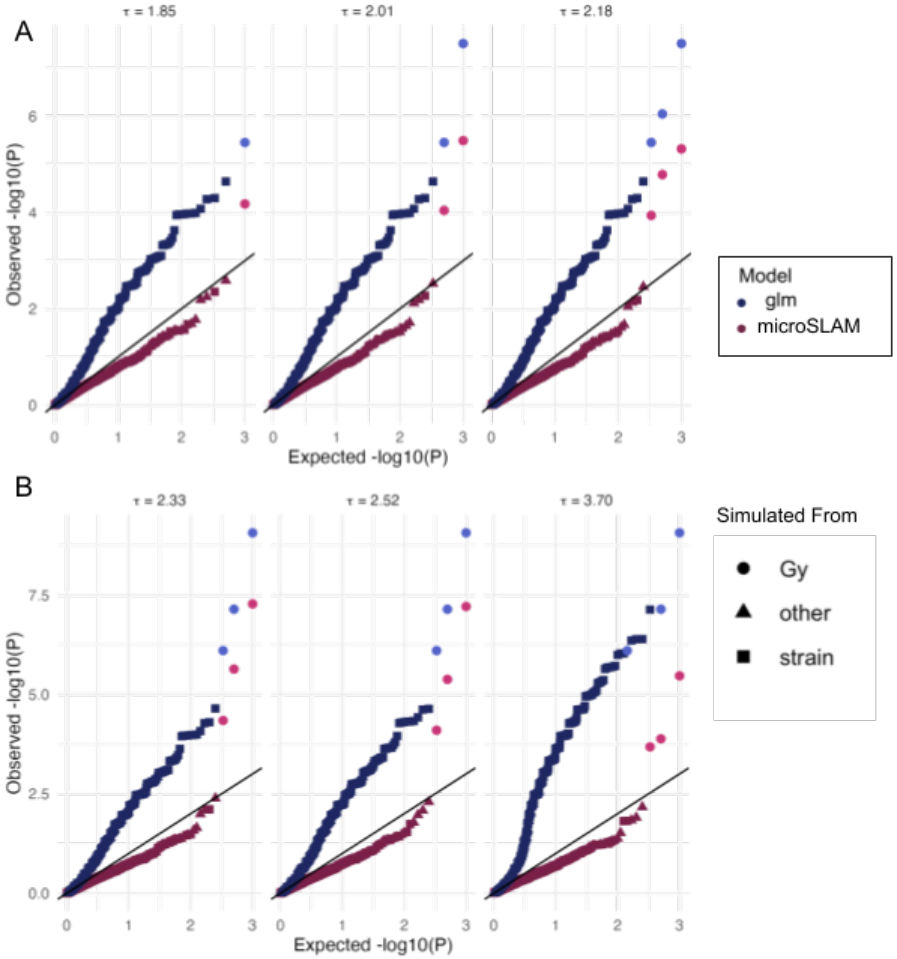

Figure 3.7 P-values from microSLAM's β test are somewhat conservative, while GLM's are inflated. Q-Q Plots of β Test results for a simulation with positive genes (G_y; *pink circles*: microSLAM, *light blue circles*: glm) plus negative genes that are linked to a strain (strain; squares) or randomly generated (other; triangles) (β test simulation 3, Supplemental Text)**. A)** Compared to glm (blue), microSLAM (red) better distinguishes the positive genes G_y from those simulated from the strain. The number of positive genes was one (*left*), two (*middle*), or three (*right*). The value of τ increases with each additional gene G_y. **B)** As the relationship between the strain and *y* is increased (left to right), the value of τ increases, and the rate of inflation increases for glm. Across different values of τ, microSLAM remains slightly conservative and continues to rank the positive genes G_y highest, indicating high specificity.

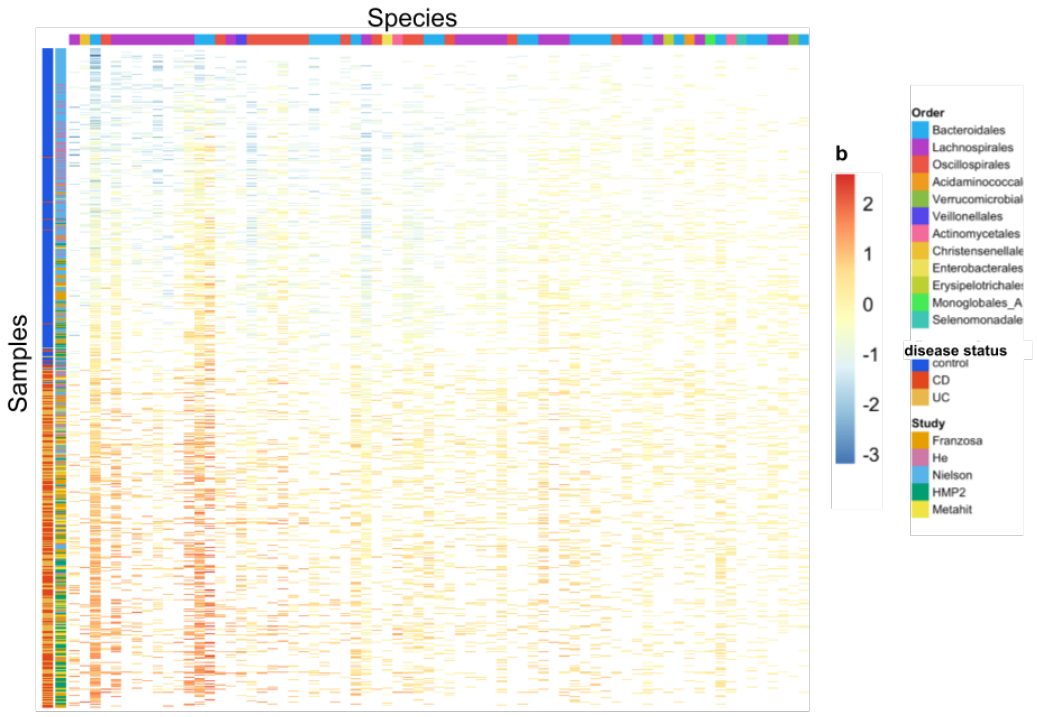

Figure 3.8 Random effect (b) values across studies and species. This heatmap shows microSLAM estimates of the random effect parameters (b) for each of 71 species across all samples where it was detected in the IBD compendium. The red-to-blue color scale denotes the association between strains and IBD (binary case/control status). Red: strains positively associated with IBD; Blue: strains negatively associated with IBD. The study and IBD subtype of each sample are shown on the left. IBD subtypes: Blue=control, red=Crohn's disease (CD), yellow=Ulcerative colitis (UC). CD and UC were combined as cases in the microSLAM modeling. Studies: Franzosa (NCBI BioProject PRJNA400072; orange), He (PRJNA398089; pink), Nielsen (PRJEB15371; blue), HMP2 (PRJEB5224; green), MetaHIT (PRJEB1220; yellow). Species are ordered by the standard deviation of *b* (left=highest standard deviation), where a higher standard deviation indicates greater strain diversity that is associated with case/control status. The samples in each column are ordered by lowest to highest average b value. A few studies (e.g., Nielsen) have more controls than others, but there is no systematic relationship between study and population structure. CD and UC tend to have similar distributions of *b* values (i.e., red and yellow are mixed on the left sidebar). While we cannot rule out confounders that were unmeasured in the publicly available data that we could access, these patterns suggest that our findings are not obviously biased by differences in study population (e.g., diet, medical care, geography, type of IBD) that could confound measured associations between case/control status and microbiome strains and genes.

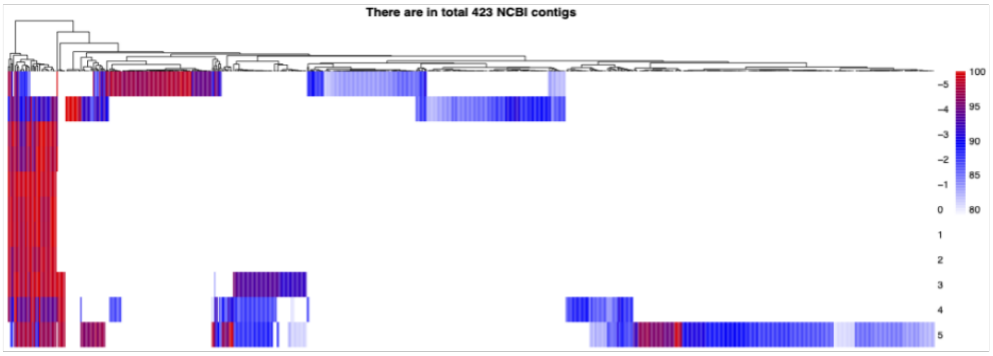

Figure 3.9 *F. prausnitzii PTS* operon evolves as a unit across diverse genomes. This operon comprises seven genes (occasionally eight genes) that were consistently present or absent together across 53% (49/85) of *F. prausnitzii* genomes from NCBI. The order and orientation of genes in the operon are conserved. This heatmap shows the genes (rows; position 0 is *gfrD,* which was significant after localFDR adjustment of microSLAM β test p-values). The other genes were significant before localFDR adjustment and are indexed relative to *gfrD* in the heatmap. Columns represent 423 contigs from 85 *F. prausnitzii* high-quality NCBI genomes. The color of the heatmap shows the blastn sequence similarity of the gene sequence in the contig compared to the sequence in the *F. prausnitzii* reference genome used in our microSLAM analysis (red=highest similarity, white=no significant match). The seven genes in the operon (middle rows of the heatmap) have high sequence similarity when they are present and are present together (red on left), whereas flanking genes are more variably present and have lower sequence identity (blue in top and bottom rows).

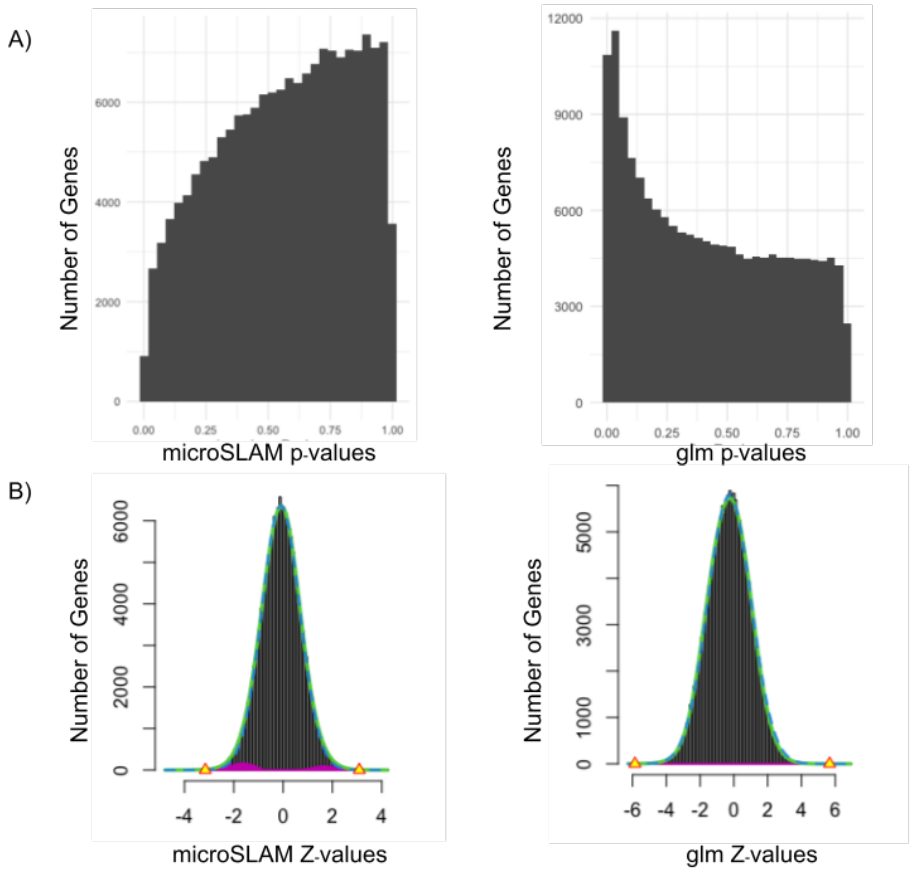

Figure 3.10 LocalFDR p-values and Z-values. A) Histogram of p-values for microSLAM's β test (left) and glm (right). **B)** Output from localFDR showing the distribution of the null z-values (green) versus the distribution of the z-values that do not follow the null (pink). Yellow triangles denote the z-value thresholds corresponding to a localFDR of 0.2.

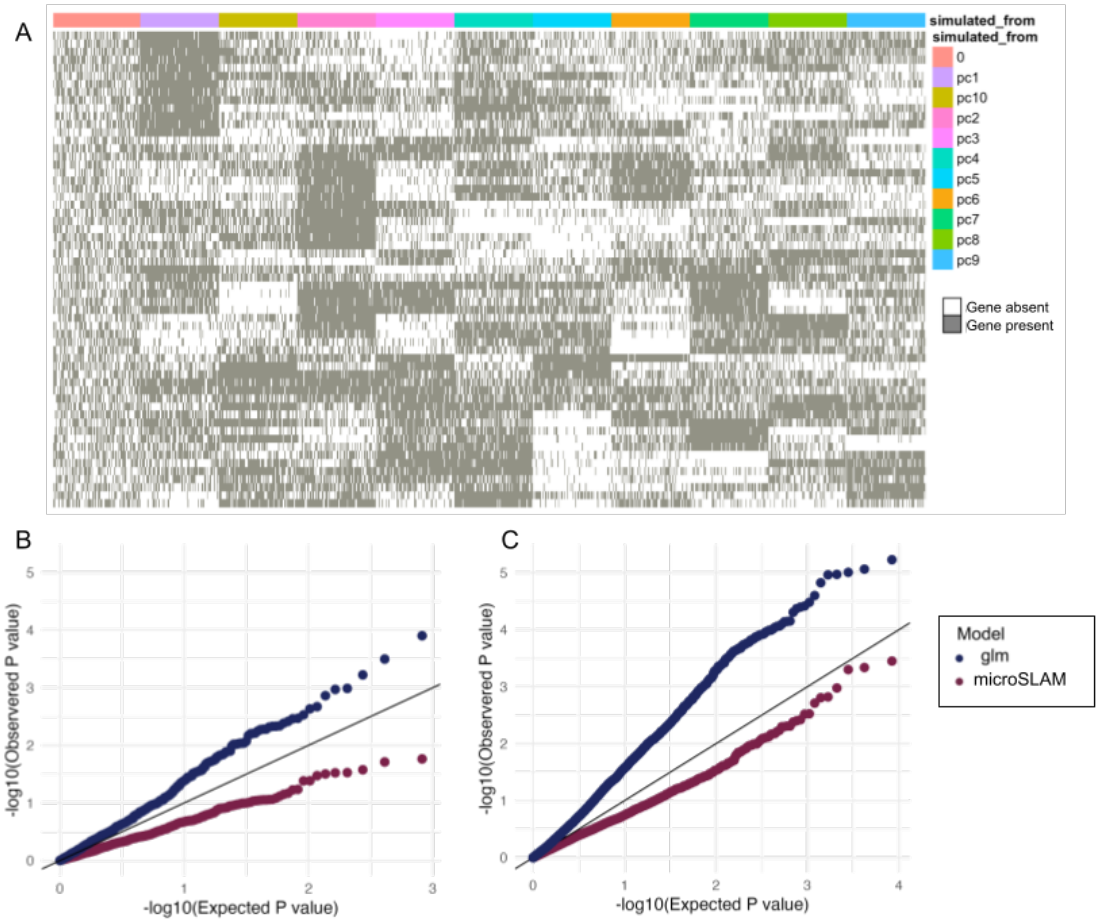

Figure 3.11 Example of data and results from microSLAM β test Simulation 1. **A)** Simulated gene presence/absence matrix based on the GRM of *Bacteroides thetaiotaomicron* plotted as a heatmap (grey=gene present in a given sample, white=gene absent). Genes are in columns and are labeled according to how they were simulated (0=random, pc1-10=using one of the first 10 principal components of the observed GRM for *B. thetaiotaomicron*. This presence/absence matrix has some population structure (estimated τ = 2.30), but no genes were simulated to be associated directly with the trait which is defined by the first two PCs. **B)** Q-Qplot of p-values from all genes not from PC1 or 2 from microSLAM's **β** test (red) and glm (blue) applied to the simulated gene presence/absence matrix in (A). There is a much higher error rate for the glm model. On the other hand, microSLAM is overly conservative (i.e., underpowered). **C)** Q-Q plot for microSLAM's **β** test (red) and glm (blue) applied to the observed *B. thetaiotaomicron* gene presence/absence matrix from the IBD compendium. The trends are very similar to those in the simulation.

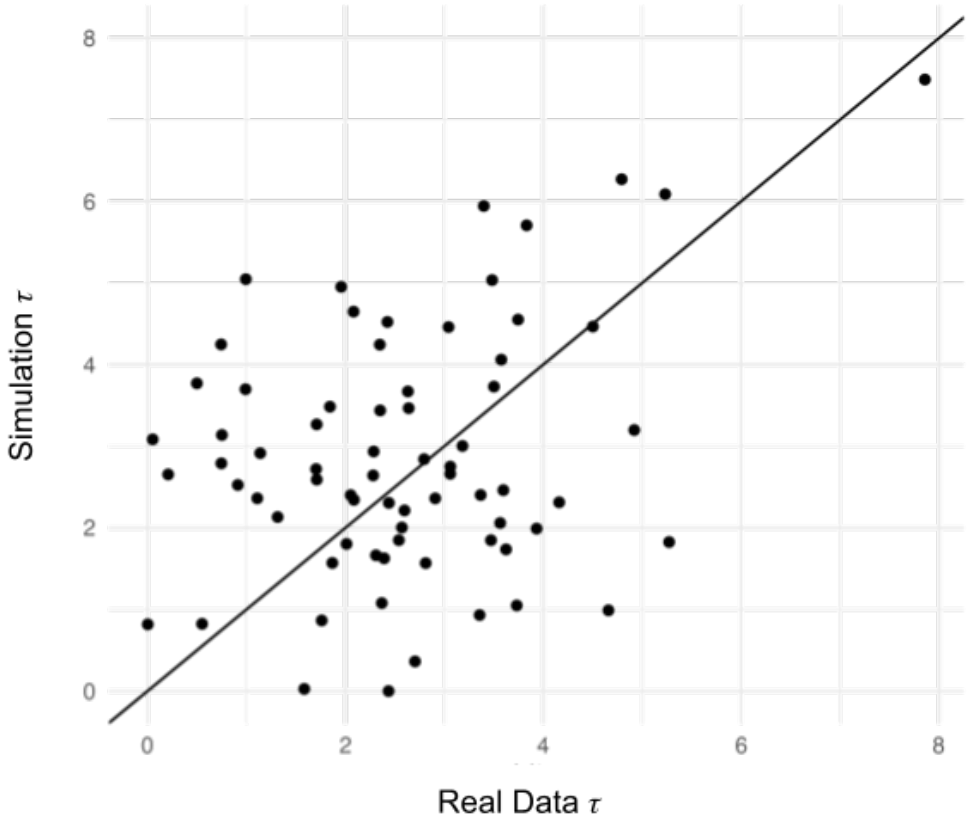

Figure 3.12 β Test simulation τ values versus observed τ values in IBD compendium. In the β Test Simulation 1 and 2 set up, we generated gene presence/absence matrices using the observed GRMs for the 71 species in the IBD compendium. Our objective was to generate simulated data that was similar to but not identical to the observed data (Methods). This scatter plot shows the τ values estimated by microSLAM on the simulated data (*y axis)* compared to the corresponding τ values estimated from the real data in the IBD compendium (*x axis*). The n τ from the simulation cover a similar range of values as those from the real data while not being highly correlated.

Chapter 4 Conclusions

4.1 Summary of conclusions

This dissertation introduces microSLAM, which implements population structure-aware metagenome-wide association studies. Using a generalized linear mixed modeling (GLMM) approach, I developed a method to include information about the genetic relatedness of the microbiomes within diverse samples and model the association of this population structure with host traits while adjusting for other covariates. I focused on case/control study designs here, but microSLAM can also be applied to quantitative traits. In addition to testing if population structure (i.e., specific strains) is associated with a trait, microSLAM also includes a test to identify trait-associated genes evolving somewhat independently of strain lineages. Realistic simulation studies demonstrated that microSLAM controls Type 1 error and has reasonable power in cohorts with more than one hundred samples. Compared to standard generalized linear modeling (GLM), microSLAM's gene-level β test reduces false positives much more effectively, especially in species with notable population structure. When there is a significant population structure and a subset of genes are more related to the phenotype than to the strain signal, I showed that microSLAM increases specificity compared to GLM. By making microSLAM an open-source R package, I am providing a new tool for researchers to probe microbiome-host interactions with strain- and gene-level resolution.

In this thesis, I also assembled a metagenomic compendium of IBD samples. Analyzing this data with microSLAM, I discovered various population structures within human gut metagenomes. I identified 49 species with a population structure related to IBD. In addition, after adjusting for population structure, I found that 57 microbial genes were significantly enriched in healthy subjects, and 26 were enriched in IBD patients. From the genes enriched in healthy subjects, I identified a fructoselysine PTS system operon in *F. prausnitzii* that is present

in all clades of this species but in only a minority of genomes within each clade, suggesting that this operon is a mobile genetic element or another rapidly lost/gained structural variant. The presence of this operon may confer distinct metabolic advantages, including the ability of carriers to utilize fructoselysine as an energy source [116]. The potential impact on human health could be significant, given that *F. prausnitzii* is one of the most prominent butyrate producers in the human gut [16,93,119]. This also leads to a greater resilience of the *F. prausnitzii*, offering enhanced protection against pathogenic bacteria and reducing the risk of chronic disease.

4.2 Limitations

There are several limitations to implementing my work and microSLAM, some of which I have discussed in Chapter 3. First, in the application portion I performed an analysis across datasets, which means I used samples collected from many studies located in various geographic locations. These publicly available metagenomics data do not include detailed information about sample variables, such as diets, medical care, and medication history. Hence, there are many significant covariates that can not be accounted for in our model. This means that there is a chance some of the significant τ tests were related to an unmeasured variable that happened to be associated with strain genetic differences. Beyond avoiding confounding factors, more complete metadata would also help us understand our method's capabilities and functionally interpret our IBD findings. A future implementation of microSLAM could add a random effect for an important known confounding factor such as the amount of fiber eaten by the subjects. Data such as this do exist, and hopefully, microSLAM will be run on a larger, more cohesive, and well-annotated dataset very soon.

Second, I estimated the GRM using the same gene presence/absence data I used to test genes for their trait associations. This approach has been used with mixed modeling with human genetic data. It has been shown to reduce power compared to estimating the GRM with

an independent set of markers (e.g., variants on other chromosomes) [120], which, in my case, affected the model's power. I explored the use of SNVs in core genes for GRM and random effect estimation in the β test. However, I found that the SNV data generated a GRM that was very different from the gene-based GRM for almost all species in our IBD compendium. Hence, SNPs were not good markers for estimating the population structure in gene presence/absence. Perhaps this approach could work with more investigation into how to pick SNPs for GRM estimation, including picking SNVs that are most related to the substrains of a species, choosing SNPs based on their location within the bacterial chromosome, or the average distance to a particular gene.

Another extension of microSLAM would be to test SNVs, rather than genes, for their associations with host traits. If SNVs are coded as presence/absence of the derived allele, then these tests could be done with the existing R code. It would also be relatively easy to extend the code to work with SNVs coded as 0, 1, or 2 copies of the derived allele. I chose to focus on gene presence/absence alleles, however, because their functional effects are easier to interpret and their smaller total number per species provides higher statistical power with currently available cohort sizes. Nonetheless, SNV-based MWAS with microSLAM is an exciting future direction to explore as cohort sizes increase and genome annotations improve. My analyses showing large differences in GRMs based on genes versus SNVs indicate that SNVs are likely to hold additional information beyond what we have studied with gene-based microSLAM analyses.

This limitation also led to one of the most significant innovations in the microSLAM model, the τ test. When running the version of the model that was separated into SNPs and genes, it became evident that in some very extreme cases, including a similarity matrix from data that had no relationship to the data later being modeled was, in fact, a considerable hindrance in the model's ability to control false positives accurately. The penalty for including non-related random effects meant that the model was worse than the GLM in those cases,

something I had seen mentioned in the human genetics literature [78]. This observation led to the innovative inclusion of the test on τ to see whether there was a good reason to include these variables in the model. This τ test has ended up being one of the significant findings from my project, as it demonstrated that there is a large variance in the amount of population structure related to the phenotype across different species and that this population structure can vary in different ways than the principal components alone. I hope this finding will play an important role in how people run MWAS in the future.

Finally, many of the genes I identified were unannotated, leading to difficulty completing in-depth analyses of significant genes across species. For example, I needed more power to perform functional enrichment analyses despite seeing several consistent trends, such as mobile elements being discovered in multiple species. More gene annotations would help with this problem, but more than annotation is needed to confirm function. MicroSLAM is an essential first step for proposing candidate causal strains and genes that should be performed upstream of in vitro and in vivo experiments to test the hypothesized functions of the discovered strains and pathways.

4.3 Future directions

Understanding how bacterial species acquire and disseminate functional genes is critical to comprehending microbial ecology and identifying dietary or probiotic therapeutic interventions targeting the microbiome. This work shows that microSLAM can detect associations for species whose relative abundance is not correlated with host traits and can accurately disentangle associations of individual genes versus groups of strain-defining genes. These capabilities make microSLAM a valuable new tool to generate hypotheses about the function of genes and operons that can be tested in in vitro and in vivo experiments. As previously stated, a much larger and better-annotated data set combined with microSLAM

would increase the power to detect associations and more specific associations. As researchers move towards testing for strain and gene associations in studies with hundreds or thousands of samples, microSLAM's improved specificity and control of Type 1 error, compared to GLM, will be even more critical, as in larger sets of samples, there is a natural increase in Type 1 error.

Incorporating the three-step mixed model for microSLAM leads to an efficient model, but microSLAM may need to be even more efficient with many more samples. In the future, a different backend for matrix computation, parallelization, or other techniques implemented in the much larger human GWAS studies could be added to the microSLAM algorithm. In addition, the population structure component could be used to not only separate strain effects but also to identify the genes that make up the strain-specific effects, which could be used as a diagnostic tool in many applications.

Distinguishing signals from noise in data will be a problem scientists continue to face forever. In metagenomics, there are many sources of noise and confounding. The goal of microSLAM is ultimately to help scientists take one source of understood confounding, population structure, and use that to distinguish the most evident effects of genes within a microbial population. I showed that including the structure of the population in metagenomic association studies increases our understanding of the discovered host trait associations and our ability to discover fascinating microbial ecology.

References

1. Laitinen, K. & Mokkala, K. Overall dietary quality relates to gut Microbiota diversity and abundance. *Int. J. Mol. Sci.* **20**, 1835 (2019).
2. Flint, H. J., Scott, K. P., Louis, P. & Duncan, S. H. The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 577–589 (2012).
3. Katsirma, Z., Dimidi, E., Rodriguez-Mateos, A. & Whelan, K. Fruits and their impact on the gut microbiota, gut motility and constipation. *Food Funct.* **12**, 8850–8866 (2021).
4. Monda, V. *et al.* Exercise modifies the gut Microbiota with positive health effects. *Oxid. Med. Cell. Longev.* **2017**, 3831972 (2017).
5. Mailing, L. J., Allen, J. M., Buford, T. W., Fields, C. J. & Woods, J. A. Exercise and the gut microbiome: A review of the evidence, potential mechanisms, and implications for human health. *Exerc. Sport Sci. Rev.* **47**, 75–85 (2019).
6. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 594–599 (2012).
7. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain Behav. Immun.* **48**, 186–194 (2015).
8. Pascal, V. *et al.* A microbial signature for Crohn's disease. *Gut* **66**, 813–822 (2017).
9. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
10. Rühlemann, M. C. *et al.* Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **53**, 147–155 (2021).
11. Zhao, Q., Chen, Y., Huang, W., Zhou, H. & Zhang, W. Drug-microbiota interactions: an emerging priority for precision medicine. *Signal Transduct. Target. Ther.* **8**, 386 (2023).

12. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R. & Goodman, A. L. Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* **570**, 462–467 (2019).
13. Javdan, B. *et al.* Personalized mapping of drug metabolism by the human gut microbiome. *Cell* **181**, 1661–1679.e22 (2020).
14. Tsunoda, S. M., Gonzales, C., Jarmusch, A. K., Momper, J. D. & Ma, J. D. Contribution of the gut microbiome to drug disposition, pharmacokinetic and pharmacodynamic variability. *Clin. Pharmacokinet.* **60**, 971–984 (2021).
15. den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J. Lipid Res.* **54**, 2325–2340 (2013).
16. Hodgkinson, K. *et al.* Butyrate's role in human health and the current progress towards its clinical application to treat gastrointestinal disease. *Clin. Nutr.* **42**, 61–75 (2023).
17. Martín, R. *et al.* Functional Characterization of Novel Faecalibacterium prausnitzii Strains Isolated from Healthy Volunteers: A Step Forward in the Use of F. prausnitzii as a Next-Generation Probiotic. *Front. Microbiol.* **8**, 1226 (2017).
18. Sanders, M. E., Merenstein, D. J., Reid, G., Gibson, G. R. & Rastall, R. A. Probiotics and prebiotics in intestinal health and disease: from biology to the clinic. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 605–616 (2019).
19. Neish, A. S. Microbes in gastrointestinal health and disease. *Gastroenterology* **136**, 65–80 (2009).
20. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307**, 1915–1920 (2005).
21. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509**, 357–360 (2014).
22. Natividad, J. M. M. & Verdu, E. F. Modulation of intestinal barrier by intestinal microbiota: pathological and therapeutic implications. *Pharmacol. Res.* **69**, 42–51 (2013).

23. Bäumler, A. J. & Sperandio, V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535**, 85–93 (2016).
24. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–544 (2016).
25. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
26. McCarthy, A. J. *et al.* Extensive horizontal gene transfer during Staphylococcus aureus co-colonization in vivo. *Genome Biol. Evol.* **6**, 2697–2708 (2014).
27. The, H. C., Thanh, D. P., Holt, K. E., Thomson, N. R. & Baker, S. The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nat. Rev. Microbiol.* **14**, 235–250 (2016).
28. Fitzgerald, S. F. *et al.* Genome structural variation in Escherichia coli O157:H7. *Microb. Genom.* **7**, 000682 (2021).
29. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
30. Dedrick, R. M. *et al.* The prophage and Plasmid mobilome as a likely driver of Mycobacterium abscessus diversity. *MBio* **12**, (2021).
31. Richardson, E. J. *et al.* Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nature Ecology & Evolution* **2**, 1468–1478 (2018).
32. von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 173 (2016).
33. McDougal, L. K. *et al.* Pulsed-field gel electrophoresis typing of oxacillin-resistant Staphylococcus aureus isolates from the United States: establishing a national database. *J. Clin. Microbiol.* **41**, 5113–5120 (2003).
34. Wang, D. *et al.* Characterization of gut microbial structural variations as determinants of human bile acid metabolism. *Cell Host Microbe* **29**, 1802–1814.e5 (2021).

35. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
36. Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic assembly: Overview, challenges and applications. *Yale J. Biol. Med.* **89**, 353–362 (2016).
37. *Genetics and Evolution of Infectious Diseases*. (Elsevier Science Publishing, Philadelphia, PA, 2017).
38. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
39. Zhao, C., Dimitrov, B., Goldman, M., Nayfach, S. & Pollard, K. S. MIDAS2: Metagenomic Intra-species Diversity Analysis System. *Bioinformatics* **39**, (2023).
40. Smith, B. J. *et al.* Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC. *bioRxiv* 2024.04.10.588779 (2024) doi:10.1101/2024.04.10.588779.
41. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
42. Costea, P. I. *et al.* metaSNV: A tool for metagenomic strain level analysis. *PLoS One* **12**, e0182392 (2017).
43. Van Rossum, T. *et al.* metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab789.
44. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
45. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
46. Song, H., Yoo, Y., Hwang, J., Na, Y.-C. & Kim, H. S. Faecalibacterium prausnitzii

subspecies-level dysbiosis in the human gut microbiome underlying atopic dermatitis. *J. Allergy Clin. Immunol.* **137**, 852–860 (2016).

47. Rossi, O. *et al.* Faecalibacterium prausnitzii Strain HTF-F and Its Extracellular Polymeric Matrix Attenuate Clinical Parameters in DSS-Induced Colitis. *PLoS One* **10**, e0123013 (2015).
48. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
49. Rowe-Magnus, D. A. *et al.* The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 652–657 (2001).
50. Gill, S. R. *et al.* Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. *J. Bacteriol.* **187**, 2426–2438 (2005).
51. Henke, M. T. *et al.* Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12672–12677 (2019).
52. Zhernakova, D. V. *et al.* Host genetic regulation of human gut microbial structural variation. *Nature* **625**, 813–821 (2024).
53. Fang, X. *et al.* Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Syst. Biol.* **12**, 66 (2018).
54. Hu, W. *et al.* Biodiversity and Physiological Characteristics of Novel Faecalibacterium prausnitzii Strains Isolated from Human Feces. *Microorganisms* **10**, (2022).
55. Yao, S., Zhao, Z., Wang, W. & Liu, X. Bifidobacterium Longum: Protection against Inflammatory Bowel Disease. *J Immunol Res* **2021**, 8030297 (2021).

56. Zhang, M. *et al.* Change in the Gut Microbiome and Immunity by Lacticaseibacillus rhamnosus Probio-M9. *Microbiol Spectr* **11**, e0360922 (2023).
57. Wares, J. P. Population Structure and Gene Flow. in *Encyclopedia of Evolutionary Biology* 327–331 (Elsevier, 2016). doi:10.1016/b978-0-12-800049-6.00035-4.
58. Li, C. C. Population subdivision with respect to multiple alleles. *Ann. Hum. Genet.* **33**, 23–29 (1969).
59. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans: These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic. *Science* **201**, 786–792 (1978).
60. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
61. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
62. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
63. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
64. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
65. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
66. *Generalized Linear Models*. vol. 2nd Edition (Chapman and Hall, London, 1989).
67. Breslow, N. E. & Clayton, D. G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* **88**, 9–25 (1993).
68. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

69. Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716–1725 (2018).
70. De Vadder, F. *et al.* Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metab.* **24**, 151–157 (2016).
71. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
72. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
73. Berry, S. E. *et al.* Human postprandial responses to food and potential for precision nutrition. *Nat. Med.* **26**, 964–973 (2020).
74. Daillère, R. *et al.* Enterococcus hirae and Barnesiella intestinihominis Facilitate Cyclophosphamide-Induced Therapeutic Immunomodulatory Effects. *Immunity* **45**, 931–943 (2016).
75. Gurinovich, A. *et al.* Evaluation of GENESIS, SAIGE, REGENIE and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data. *Front. Genet.* **13**, 897210 (2022).
76. Efron, B. Local False Discovery Rates. (2005) doi:10.1017/cbo9780511761362.006.
77. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
78. Lee, O. E. & Braun, T. M. Permutation tests for random effects in linear mixed models. *Biometrics* **68**, 486–493 (2012).
79. Zeng, P., Zhao, Y., Li, H., Wang, T. & Chen, F. Permutation-based variance component test in generalized linear mixed model with application to multilocus genetic association study. *BMC Med. Res. Methodol.* **15**, 37 (2015).
80. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful

approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

81. Storey, J. D. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).
82. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
83. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
84. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
85. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
86. Lopera-Maya, E. A. *et al.* Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat. Genet.* **54**, 143–151 (2022).
87. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
88. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
89. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
90. Loddo, I. & Romano, C. Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis. *Front. Immunol.* **6**, 551 (2015).
91. Dahlhamer, J. M., Zammitti, E. P., Ward, B. W., Wheaton, A. G. & Croft, J. B. Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years - United States, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 1166–1169 (2016).
92. Xu, F., Carlson, S. A., Liu, Y. & Greenlund, K. J. Prevalence of Inflammatory Bowel Disease Among Medicare Fee-For-Service Beneficiaries - United States, 2001-2018.

MMWR Morb. Mortal. Wkly. Rep. **70**, 698–701 (2021).

93. Machiels, K. *et al.* A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283 (2014).
94. Wu, X. *et al.* Biomarkers of Metabolomics in Inflammatory Bowel Disease and Damp-Heat Syndrome: A Preliminary Study. *Evid. Based. Complement. Alternat. Med.* **2022**, 3319646 (2022).
95. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
96. Carasso, S. *et al.* Inflammation and bacteriophages affect DNA inversion states and functionality of the gut microbiota. *Cell Host Microbe* (2024) doi:10.1016/j.chom.2024.02.003.
97. Zhang, Y. *et al.* Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature* **606**, 754–760 (2022).
98. Clooney, A. G. *et al.* Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut* **70**, 499–510 (2021).
99. Dalal, S. R. & Chang, E. B. The microbial basis of inflammatory bowel diseases. *J. Clin. Invest.* **124**, 4190–4196 (2014).
100. Sokol, H. *et al.* Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm. Bowel Dis.* **15**, 1183–1189 (2009).
101. Glassner, K. L., Abraham, B. P. & Quigley, E. M. M. The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* **145**, 16–27 (2020).
102. Vich Vila, A. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **10**, (2018).
103. Haldar, T. & Ghosh, S. Effect of population stratification on false positive rates of population-based association analyses of quantitative traits: Stratification effects on

population-based QTL analyses. *Ann. Hum. Genet.* **76**, 237–245 (2012).

104. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med.* **9**, 103 (2017).

105. Henke, M. T. *et al.* Capsular polysaccharide correlates with immune response to the human gut microbe Ruminococcus gnavus. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

106. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

107. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

108. Lu, J. *et al.* Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).

109. Lavelle, A. *et al.* Fecal microbiota and bile acids in IBD patients undergoing screening for colorectal cancer. *Gut Microbes* **14**, 2078620 (2022).

110. Nomura, K. *et al.* Bacteroidetes Species Are Correlated with Disease Activity in Ulcerative Colitis. *J. Clin. Med. Res.* **10**, (2021).

111. Coyne, M. J., Zitomersky, N. L., McGuire, A. M., Earl, A. M. & Comstock, L. E. Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio* **5**, e01305–14 (2014).

112. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

113. Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J. & Martinez-Medina, M. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *ISME J.* **11**, 841–852 (2017).

114. Quévrain, E. *et al.* Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease. *Gut* **65**, 415–425 (2016).

115. Miller, K. A., Phillips, R. S., Kilgore, P. B., Smith, G. L. & Hoover, T. R. A mannose family phosphotransferase system permease and associated enzymes are required for utilization of fructoselysine and glucoselysine in Salmonella enterica serovar Typhimurium. *J. Bacteriol.* **197**, 2831–2839 (2015).

116. Bui, T. P. N. *et al.* Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. *Nat. Commun.* **6**, 10062 (2015).

117. Deppe, V. M., Bongaerts, J., O'Connell, T., Maurer, K.-H. & Meinhardt, F. Enzymatic deglycation of Amadori products in bacteria: mechanisms, occurrence and physiological functions. *Appl. Microbiol. Biotechnol.* **90**, 399–406 (2011).

118. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01953-y.

119. Singh, V. *et al.* Butyrate producers, 'The Sentinel of Gut': Their intestinal significance with and beyond butyrate, and prospective use as microbial therapeutics. *Front. Microbiol.* **13**, 1103836 (2022).

120. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).

121. Zhao, C., Shi, Z. J. & Pollard, K. S. Pitfalls of genotyping microbial communities with rapidly growing genome collections. *Cell Syst.* **14**, 160–176.e3 (2023).

122. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

123. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

124. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**, 954–960 (2019).

125. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*

Methods **9**, 357–359 (2012).

126. Bushnell, B. BBDuk. *Jt Genome Inst. Available online: https://jgi. doe. gov/data-and-tools/bbtools/bb-tools-userguide/bbduk-guide/(accessed on 25 June 2020)* (2020).

127. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

128. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **135**, 370 (1972).

129. Wiame, E., Lamosa, P., Santos, H. & Van Schaftingen, E. Identification of glucoselysine-6-phosphate deglycase, an enzyme involved in the metabolism of the fructation product glucoselysine. *Biochem. J* **392**, 263–269 (2005).

130. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

131. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

132. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

133. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

134. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

135. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

136. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS*

Comput. Biol. **14**, e1005944 (2018).

137. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Miriam Goldman
5FB9AE3670E74E8...
Author Signature

8/7/2024
Date